UNISTAT Statistical Package

For Windows™

USER'S GUIDE

Version 6.5

October 2013

http://www.unistat.com

ADDITIVE
SOFT- & HARDWARE FÜR TECHNIK & WISSENSCHAFT

Vertrieb durch:
ADDITIVE Soft- und Hardware für Technik und Wissenschaft GmbH
Max-Planck-Straße 22b • 61381 Friedrichsdorf
http://www.additive-net.de/unistat • eShop: http://eshop.additive-net.de
Verkauf: +49-6172-5905-30 • unistat@additive-net.de

Published by UNISTAT Limited.

UNISTAT House, 4 Shirland Mews, London W9 3DY, England.

http://www.unistat.com

Section titles of procedures which are not included in Light Edition are shown in dark blue font.

Table of Contents

USER'S GUIDE 1

Chapter 1 Introduction 13

1.1. New Procedures and Features 15

1.2. Using UNISTAT User's Guide 16
1.2.1. Conventions and Notations..16
1.2.2. Reproducing the Examples..17

1.3. Modes of Running UNISTAT 18
1.3.1. Stand-Alone Mode..18
1.3.2. Excel Add-In Mode..21
1.3.3. Background Mode..22

Chapter 2 Common Features 23

2.1. Procedure Dialogues..25
2.1.1. Variable Selection Dialogue..26
2.1.2. Categorical Data Analysis 32
2.1.3. Multiple Dependent Variables 35
2.1.4. Creating Interaction, Dummy and Lag/Lead Variables 36
2.1.5. Output Options Dialogue..41

2.2. Output Medium..42
2.2.0. Output Medium Toolbar 43
2.2.1. UNISTAT Output Window..45
2.2.2. Output to Word 48
2.2.3. Output to Excel..49
2.2.4. Output to Web Browser 51

2.3. Graphics Editor..53
2.3.1. Graphics Toolbar..54
2.3.2. On-Screen Editing 55
2.3.3. File Menu..62
2.3.4. Edit Menu..65
2.3.5. View Menu..86
2.3.6. Export Menu..89

2.4. Tools..92
2.4.1. Options..92

2.4.2. Macros 109
2.4.3. Log File 113
2.4.4. Example Files 114
2.4.5. Help 114

Chapter 3 Data Processor 115

3.0. Overview 117
3.0.1. Screen Layout 117
3.0.2. Data Types 120

3.1. File Menu 128
3.1.0. File Formats 128
3.1.1. New 142
3.1.2. Open 142
3.1.3. Merge 144
3.1.4. Save 144
3.1.5. Save As 144
3.1.6. ODBC and SQL 145
3.1.7. Print 147
3.1.8. Goto 148
3.1.9. Auto Load 149
3.1.10. Exit 149

3.2. Edit Menu 150
3.2.1. Undo 150
3.2.2. Redo 150
3.2.3. Cut 150
3.2.4. Copy 151
3.2.5. Paste 151
3.2.6. Clear 152
3.2.7. Delete 152
3.2.8. Insert 153
3.2.9. Drag-Drop Action 155
3.2.10. Column Labels 155
3.2.11. Row Labels 157
3.2.12. Long String Table 160

3.3. Data Menu 162
3.3.1. Information 162
3.3.2. Transpose Matrix 163
3.3.3. Column Sort 164
3.3.4. Key Sort 164

3.3.5. Matrix Sort 165
3.3.6. Recode Column 166
3.3.7. Aggregate 167
3.3.8. Range Statistics 168
3.3.9. Stack Columns 168
3.3.10. Unstack Columns 170
3.3.11. Format Columns 171
3.3.12. Select Row 172

3.4. Formula 174
3.4.0. Overview 174
3.4.1. Quick Formula 177
3.4.2. Formula Editor 178
3.4.3. Compute Matrix 196
3.4.4. Calculate 196

Chapter 4 Data and Function Plots 197

4.0. Overview 199

4.1. 2D Plots 200
4.1.1. X-Y Plots 200
4.1.2. Polar Plot 220
4.1.3. Spectral Diagram 222
4.1.4. Fan Grid Plot 223

4.2. 3D Plots 224
4.2.1. X-Y-Z Scatter Plot 224
4.2.2. X-Y-Z Grid Plot 232
4.2.3. Spin Plot 235

4.3. Charts 238
4.3.1. Pie Chart 238
4.3.2. Bar Chart 240
4.3.3. Area Chart 245
4.3.4. Ribbon Chart 247
4.3.5. 3D Bar Chart 248
4.3.6. High-Low-Close Chart 249

4.4. Plot of Functions 251
4.4.1. Plot of 2D Functions 252
4.4.2. Plot of 3D Functions 253
4.4.3. Plot and Roots of Polynomials 254
4.4.4. Plot of Distribution Functions 256

Chapter 5 Descriptive Statistics and Distributions.........261

5.0. Overview..........263
5.0.1. Multisample Data Types..........263
5.0.2. One-Sample Data Types..........265
5.0.3. Paired Data Types..........267

5.1. Descriptive Statistics..........268
5.1.1. Summary Statistics..........268
5.1.2. Confidence Intervals..........275
5.1.3. Quantiles (Percentiles)..........279
5.1.4. Sample Statistics..........285
5.1.5. Frequency Distributions..........293
5.1.6. Stem and Leaf Plot..........296
5.1.7. Sequence Diagram..........297
5.1.8. Scatter Diagram..........298

5.2. Distribution Functions..........299
5.2.1. Cumulative Probability..........301
5.2.2. Critical Value..........303
5.2.3. Probabilities and Critical Values..........304
5.2.4. Random Numbers..........309
5.2.5. Expected Frequencies..........311

5.3. Descriptive Plots..........314
5.3.1. Box-Whisker, Dot and Bar Plots..........314
5.3.2. Normal Probability Plot..........324
5.3.3. Histogram..........326
5.3.4. 3D Histogram..........335
5.3.5. Bland-Altman Plot..........338
5.3.6. Ladder Plot..........344

Chapter 6 Statistical Tests, Correlations and Tables.... 347

6.0. Overview..........349
6.0.1. One Sample Tests..........349
6.0.2. Two Sample Tests..........350
6.0.3. Tests with Paired Data..........353
6.0.4. Multisample Tests..........354
6.0.5. Tests with Matrix Data..........355
6.0.6. Tests with Binary Data..........355
6.0.7. 2 x 2 Tables..........357
6.0.8. R x C Tables..........360

6.1. Parametric Tests 362
6.1.1. t- and F-Tests 362
6.1.2. Equivalence Test for Means 375
6.1.3. Parametric Tests Matrix 377
6.1.4. Hotelling's T-Squared Test 379

6.2. Correlations 381
6.2.1. Correlation Coefficients 382
6.2.2. Pearson-Spearman-Kendall Correlations Matrix 390
6.2.3. Partial Correlation Matrix 392
6.2.4. Intraclass Correlation Coefficients 394

6.3. Goodness of Fit Tests 397
6.3.1. Chi-Square Tests 397
6.3.2. Kolmogorov-Smirnov Tests 401
6.3.3. Normality Tests 405

6.4. Nonparametric Tests with One or Two Samples 410
6.4.1. Unpaired Samples 410
6.4.2. Paired Samples 421
6.4.3. Binomial Proportion 430
6.4.4. Unpaired Proportions 443
6.4.5. Paired Proportions 449

6.5. Multisample Nonparametric Tests 461
6.5.1. Kruskal-Wallis One-Way ANOVA 461
6.5.2. Jonckheere's Trend 470
6.5.3. Multisample Median Test 471
6.5.4. Friedman Two-Way ANOVA 474
6.5.5. Quade Two-Way ANOVA 480
6.5.6. Kendall's Concordance Coefficient 483
6.5.7. Page's L Trend 485
6.5.8. Cochran's Q 486
6.5.9. Kappa Test for Inter-Category Variation 488
6.5.10. Kappa Test for Inter-Observer Variation 489

6.6. Tables 491
6.6.1. Contingency Table 491
6.6.2. Cross-Tabulation 496
6.6.3. Break-Down 520

6.7. Sample Size and Power Estimation 522
6.7.1. One Sample 523
6.7.2. Two Samples 529

6.7.3. Variance 535
6.7.4. Correlation 538
6.7.5. Two Correlations 540
6.7.6. Two Proportions 544
6.7.7. ANOVA 548
6.7.8. Phi Distribution 553

6.8. Meta Analysis 555
6.8.1. Data Preparation 556
6.8.2. Input Data Types 558
6.8.3. Summary of Effect Sizes 570
6.8.4. Output Options 576
6.8.5. Examples 583

Chapter 7 Regression and Analysis of Variance 593

7.0. Overview 595
7.0.1. Matrix Data 595
7.0.2. Categorical Data 597

7.1. Matrix Statistics 598
7.1.1. Matrix Statistics Variable Selection 598
7.1.2. Matrix Statistics Output Options 600
7.1.3. Matrix Statistics Examples 601

7.2. Regression Analysis 605
7.2.1. Linear Regression 607
7.2.2. Polynomial Regression 626
7.2.3. Stepwise Regression 630
7.2.4. Nonlinear Regression 636
7.2.5. Logit / Probit / Gompit 663
7.2.6. Logistic Regression 684
7.2.7. Multinomial Regression 709
7.2.8. Poisson Regression 718
7.2.9. Box-Cox Regression 730

7.3. Analysis of Variance and General Linear Model 739
7.3.0. Overview 739
7.3.1. Analysis of Variance 759
7.3.2. General Linear Model 766

7.4. Tests for ANOVA 782
7.4.1. Table of Means 782
7.4.2. Homogeneity of Variance Tests 784

7.4.3. Multiple Comparisons 792
7.4.4. Regression with Replicates 812
7.4.5. Heterogeneity of Regression 815

Chapter 8 Multivariate Analysis 823

8.0. Overview....825

8.1. Cluster Analysis 828
8.1.1. Hierarchical Cluster Analysis....829
8.1.2. K-th Neighbour Cluster Analysis 839
8.1.3. K-Means Cluster Analysis....840

8.2. Discriminant Analysis 844
8.2.1. Multiple Discriminant Analysis....846
8.2.2. K-th Neighbour Discriminant Analysis....859

8.3. Multidimensional Scaling....862
8.3.1. Classical Multidimensional Scaling....863
8.3.2. Ordinal Multidimensional Scaling 865

8.4. Principal Components Analysis 869

8.5. Factor Analysis....873
8.5.1. Factoring Methods....873
8.5.2. Unrotated Output Options....875
8.5.3. Rotations....877
8.5.4. Rotated Output Options....878
8.5.5. Factor Analysis Example 879

8.6. Canonical Correlations 884

8.7. Reliability Analysis....888

8.8. Multivariate Plots....893
8.8.1. Matrix Plot....893
8.8.2. Rectangular Plot 894
8.8.3. Icon Plots 895

Chapter 9 Time Series Analysis....901

9.1. Box-Jenkins ARIMA....903
9.1.0. Overview 903
9.1.1. Differencing Input Options 903
9.1.2. Differencing Output Options 905
9.1.3. Model Fitting 908

9.1.4. Model Output Options 911
9.1.5. Forecasting 913

9.2. Forecasting and Smoothing.... 916
9.2.0. Exponential Weights Moving Average....917
9.2.1. Brown's Exponential....918
9.2.2. Holt's Linear 920
9.2.3. Winters' Additive Seasonal....923
9.2.4. Winters' Multiplicative Seasonal....926
9.2.5. Dixon-Grubbs-Neumann Outlier Tests....930

9.3. Quality Control 932
9.3.0. Overview 932
9.3.1. Variable Control Charts 936
9.3.2. Attribute Control Charts....957
9.3.3. Pareto Chart....968
9.3.4. Hotelling's T-Squared Analysis....971
9.3.5. Weibull Analysis 975
9.3.6. Process Capability Analysis 983
9.3.7. Data Transformation.... 1005
9.3.8. Gauge R&R Analysis.... 1017

9.4. Survival Analysis 1028
9.4.0. Survival Variable Selection 1028
9.4.1. Life Table 1031
9.4.2. Kaplan-Meier Analysis 1043
9.4.3. Survival Comparison Statistics.... 1054
9.4.4. Cox Regression.... 1060

9.5. Fourier Analysis 1079
9.5.1. Fourier Transform 1080
9.5.2. Inverse Fourier Transform.... 1083

Chapter 10 Analysis of Bioassays 1085

10.0. Overview....1087

10.1. Parallel Line Method....1087
10.1.1. Data Preparation.... 1088
10.1.2. Parallel Line Variable Selection.... 1089
10.1.3. Parallel Line Output Options.... 1091
10.1.4. Parallel Line Examples.... 1104

10.2. Slope Ratio Method....1128
10.2.1. Slope Ratio Variable Selection.... 1128

10.2.2. Slope Ratio Output Options 1129
10.2.3. Slope Ratio Examples 1137

10.3. Quantal Response Method 1145
10.3.1. Quantal Response Variable Selection 1145
10.3.2. Quantal Response Output Options 1147
10.3.3. Quantal Response Examples 1152

10.4. Combination of Assays 1159
10.4.1. Variable Selection 1159
10.4.2. Output Options 1160
10.4.3. Example 1164

Appendix 1167

A.1. Continuous Distributions 1169
A.1.1. Normal 1169
A.1.2. Student's t 1169
A.1.3. Chi-Square 1169
A.1.4. F 1170
A.1.5. Beta 1170
A.1.6. Gamma 1170
A.1.7. Uniform 1171
A.1.8. Triangular 1171
A.1.9. Lognormal 1171
A.1.10. Exponential 1171
A.1.11. Erlang 1172
A.1.12. Weibull 1172

A.2. Discrete Distributions 1173
A.2.1. Poisson 1173
A.2.2. Bernoulli 1173
A.2.3. Binomial 1173
A.2.4. Negative Binomial 1174
A.2.5. Geometric 1174
A.2.6. Hypergeometric 1175
A.2.7. Discrete Uniform 1175

References 1177

Index 1187

UNISTAT Statistical Package

Chapter 1
Introduction

1.1. New Procedures and Features

Since the release of Version 6.0, the following changes have been made in UNISTAT Statistical Package:

New procedures

5.3.5. Bland-Altman Plot
6.4.5.6. Statistics for Diagnostic Tests
6.8. Meta Analysis
7.2.6.4.4. ROC Analysis

Improvements in existing procedures

3.0.2.5. Date-Time Data
5.3.2. Normal Probability Plot
5.3.3. Histogram
5.3.4. 3D Histogram
5.3.6. Ladder Plot
6.2.1.3. Kendall's Rank Correlation
6.2.1.4. Point Biserial Correlation
6.2.3. Partial Correlation Matrix
7.2.1.2. Linear Regression Output Options
7.2.2. Polynomial Regression
7.2.4.4. Nonlinear Regression Output Options predictions
7.2.5. Logit / Probit / Gompit loglog link function, predictions, marginal effects, average effects
7.2.6. Logistic Regression ROC, AUC analysis
7.3.2.1. GLM Variable Selection orthogonal polynomial contrasts
7.4.3. Multiple Comparisons output options
8.1.1.4. Hierarchical Cluster Output Options
10.1.3. Parallel Line Output Options
10.1.3.2. Homogeneity of Variance Tests
10.3.2.3. Effective Dose (or Lethal Dose) Spearman-Karber method
10.4.2.3. Combined Potency USP according to *US Pharmacopoeia* (2009)

1.2. Using UNISTAT User's Guide

UNISTAT Users Guide is supplied as a single PDF document. Context-sensitive help is available at all times, providing instant information on the currently selected procedure. Where relevant, the algorithm used in the procedure is given and examples are provided. The data for these examples are supplied in example files so that results can be easily reproduced by the user. Section titles of procedures which are not included in Light Edition are shown in dark blue font.

1.2.1. Conventions and Notations

All references to computer keys are enclosed within less than (<) and greater than (>) signs. For instance <2> refers to the number key 2, <f2> to the function key 2, <Enter> to the enter (or return) key. If two keys are to be pressed simultaneously, a plus (+) sign will separate the keystrokes. For instance <Shift> + <Tab> means pressing the shift and tab keys simultaneously. Square brackets ([) and (]) are used to enclose button labels in dialogues. For instance, [Variable] refers to a button on the Variable Selection Dialogue (clicking on which selects the highlighted variable for analysis). The underlined character is the hot-key for this button. Optional parameters that can be appended to some functions are also enclosed in square brackets. The hyperlinks are in blue colour.

Menu selections are represented in bold sans serif font. Multi-level menu selections are separated by an arrow. For example, **Graph** → 2D Plots → X-Y Plots means clicking on **Graph**, selecting 2D Plots from the pull-down menu and X-Y Plots from the daughter menu.

References to the sections of this guide, and thus references to all graphics and statistics procedures are in sans serif font, such as X-Y Plots. References to prompts, messages, dialogues and all other aspects of the user interface are in narrow sans serif font, such as Syntax Error. UNISTAT's own spreadsheet functions are represented in bold gothic font such as **HCoSec()**. References to variable names are in generic serif *italic* font and the file names are in upper case.

In order to avoid unnecessary repetitions, information common to more than one section is given under a higher level section. For instance, to find out how to perform a Mann-Whitney U Test, one can start from section 6.4.1.1. Mann-Whitney U Test. However, it will help to have a look at section 6.4.1. Unpaired Samples, 6.4. Nonparametric Tests with One or Two Samples and 6.0. Overview as well. Usually, the types of data used in procedures are explained at the beginning of each chapter.

1.2.2. Reproducing the Examples

UNISTAT's help system and User's Guide contain a large number of examples many of which have been published in commonly used books. Data sets used in the examples can be found in the UNISTAT example data files.

The data files can be accessed from Tools → Example Files. Each of these files contains data for more than one example, but in general the data for procedures under the same section are grouped within the same file. For instance, the file PARTESTS contains the data for Parametric Tests and REGRESS contains the data for Linear Regression examples.

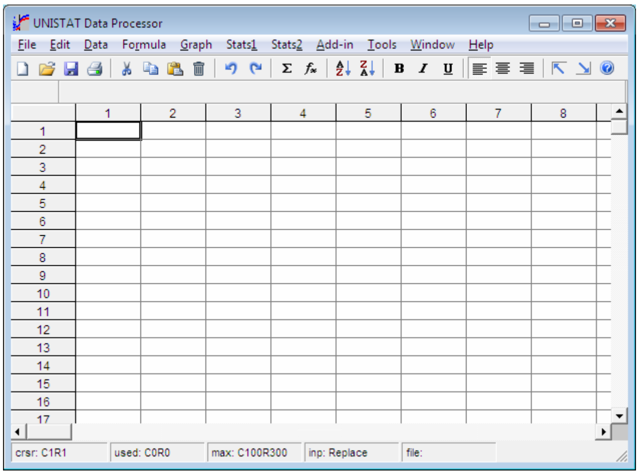
Examples will usually make a reference to the book or article in which they were first published, followed by the name of the file containing the example data. Where a suitable published example cannot be found, an example will be provided using the generic DEMODATA file. The user will then be told which columns are to be assigned to which tasks. When all the instructions are followed, it should be possible to precisely reproduce the results given in this manual. These may then be compared with the results in the original published source. In most cases, UNISTAT output would include many more statistics than were given in the original source. Also, UNISTAT's results will generally be more accurate, as the examples in books often display the intermediate results leading to rounding-off errors.

We believe that by including published examples in this User's Guide we not only provide a valuable tool to learn UNISTAT by solving real-life problems, but also provide the user with the possibility of verifying and validating the accuracy of UNISTAT's results by comparing them against respected academic sources.

1.3. Modes of Running UNISTAT

1.3.1. Stand-Alone Mode

Start UNISTAT from the *Unistat 6.5* icon. Loading and initialising the program will take a few moments. When this is completed, the UNISTAT spreadsheet (Data Processor) will be displayed.



crsr: C1R1 | used: C0R0 | max: C100R300 | inp: Replace | file:

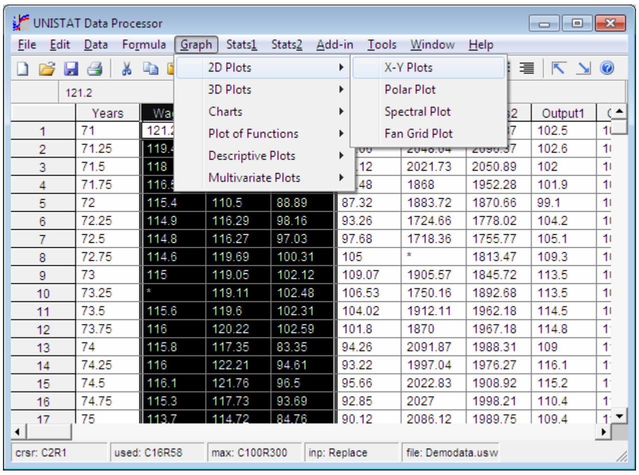
The top line of this window displays pull-down menu options for the Data Processor and is called the Menu Bar. The panel just under the Menu Bar contains a number of buttons and is called the Toolbar. The next line is used for text input and various prompts and will be referred to as the Input Panel. Note that in the above screen shot the Input Panel is currently inactive and so appears as simply a grey area. When it becomes active, a white text box with a flashing cursor will appear. At the very bottom of the window, information is displayed on certain Data Processor parameters. This area is called the Status Bar.

Before doing anything else, the new user of UNISTAT may wish to browse through the pull-down menu items and see what is available. Try moving the mouse over buttons on the toolbar. If the mouse pointer is held still over a button for a few seconds a short description of the button’s action will be displayed in a Tool Tip box.

At this stage there is no data in the spreadsheet to analyse. So, the available operations which can be performed are limited (such as Plot of Functions, Cumulative Probability and Sample Size and Power Estimation).

To enter data from keyboard, just type a number. This is done by typing directly into the current cell. Either press <Enter> or an arrow key or just click on another cell to complete the entry. To enter a column or row label, double-click on the label. A text editor will be placed on the Input Panel. Enter or edit the label. When finished, press <Enter> or click [OK].

As an exercise, you may like to open one of the example data files supplied with UNISTAT. To do this select Tools → Example Files. A standard Windows Open dialogue will appear allowing the selection of a data file. Then select DEMODATA. Because the Data Processor already contains some data which you may have just typed in the previous exercises, the program may ask you whether you wish to save the existing data to a file first, or to clear it. Selecting [No] will clear the existing data and the screen will be redrawn displaying the contents of the file DEMODATA.

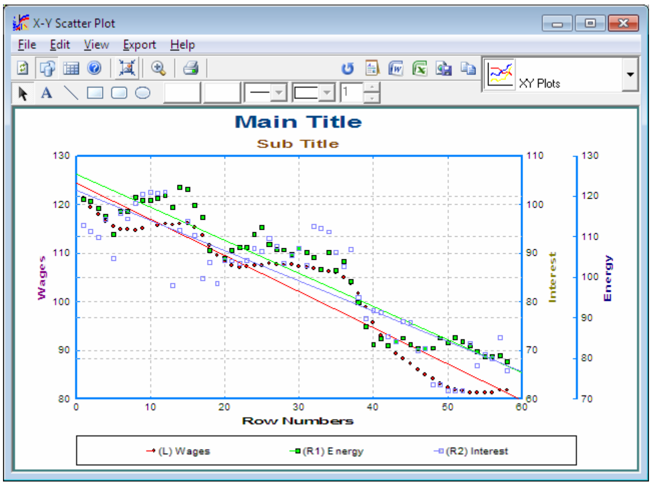


121.2

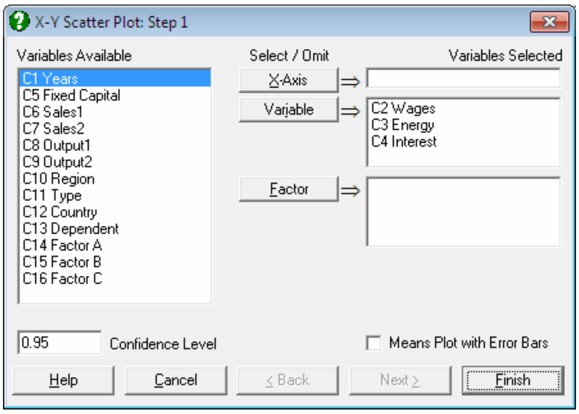
	Years	Wa					2	Output1	(
1	71	121.2					7	102.5	1
2	71.25	119.			[illegible]	[illegible]	[illegible]	102.6	1
3	71.5	118			12	2021.73	2050.89	102	1
4	71.75	116.			48	1868	1952.28	101.9	1
5	72	115.4	110.5	88.89	87.32	1883.72	1870.66	99.1	1
6	72.25	114.9	116.29	98.16	93.26	1724.66	1778.02	104.2	1
7	72.5	114.8	116.27	97.03	97.68	1718.36	1755.77	105.1	1
8	72.75	114.6	119.69	100.31	105	*	1813.47	109.3	1
9	73	115	119.05	102.12	109.07	1905.57	1845.72	113.5	1
10	73.25	*	119.11	102.48	106.53	1750.16	1892.68	113.5	1
11	73.5	115.6	119.6	102.31	104.02	1912.11	1962.18	114.5	1
12	73.75	116	120.22	102.59	101.8	1870	1967.18	114.8	1
13	74	115.8	117.35	83.35	94.26	2091.87	1988.31	109	1
14	74.25	116	122.21	94.61	93.22	1997.04	1976.27	116.1	1
15	74.5	116.1	121.76	96.5	95.66	2022.83	1908.92	115.2	1
16	74.75	115.3	117.73	93.69	92.85	2027	1998.21	110.4	1
17	75	113.7	114.72	84.76	90.12	2086.12	1989.75	109.4	1

crsr: C2R1 | used: C16R58 | max: C100R300 | inp: Replace | file: Demodata.usw

In order to draw a multiple line plot of the data columns *Wages*, *Energy* and *Interest*, move the mouse pointer to the label of *Wages* (where *Wages* appears on a grey background), press the left mouse button, drag the mouse pointer to the label of *Interest* and release the button. All three columns will be highlighted. Then select Graph → 2D Plots → X-Y Plots. A line graph of *Wages*, *Energy* and *Interest* will be plotted against the index (row numbers).



Highlighting spreadsheet columns first and then selecting a procedure will execute this procedure immediately with the default options. Suppose, for instance, that you wish to plot *Wages*, *Energy* and *Interest* against *Years* on the X-axis. To do this you will need to switch the highlighting off first (by simply clicking on any data cell) and then select Graph → 2D Plots → X-Y Plots. In this case, a dialogue will pop up allowing you to select specific columns for specific tasks. This is called a Variable Selection Dialogue and displays a list of all available variables on the left.


X-Y Scatter Plot: Step 1

Variables Available
C1 Years
C5 Fixed Capital
C6 Sales1
C7 Sales2
C8 Output1
C9 Output2
C10 Region
C11 Type
C12 Country
C13 Dependent
C14 Factor A
C15 Factor B
C16 Factor C

Variables Selected
C2 Wages
C3 Energy
C4 Interest

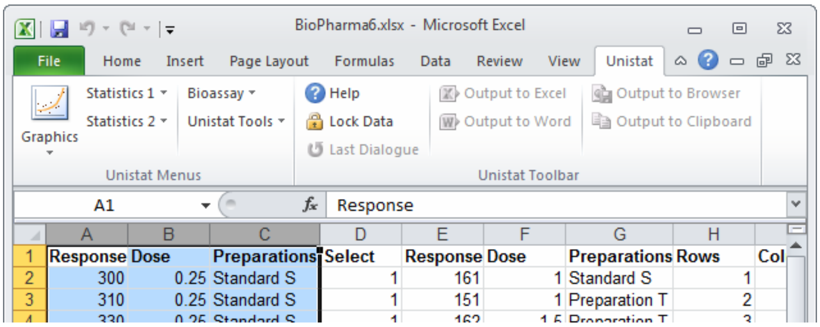
0.95 Confidence Level

□ Means Plot with Error Bars

Highlight the first column *Years* by clicking on it and then select it as the X-axis variable by clicking on the [X Axis] button or pressing <Alt> + <X>. The *Years* variable will be transferred to the Variables Selected list on the right. Finally, to display the graph simply click on [Finish]. If a variable is selected by mistake, or you wish to de-select a variable for some other reason, you can highlight the variable in the Variables Selected list and click on its corresponding [Select / Omit] button.

1.3.2. Excel Add-In Mode

Start Excel using the *Unistat 6.5 for Excel* icon. This will start Excel with a new top level menu option Unistat which contains the new UNISTAT menu items Graphics, Statistics 1, Statistics 2 and Unistat Tools and a new UNISTAT toolbar. Bioassay is an optional module.



	A	B	C	D	E	F	G	H	
1	Response	Dose	Preparations	Select	Response	Dose	Preparations	Rows	Col
2	300	0.25	Standard S	1	161	1	Standard S	1	
3	310	0.25	Standard S	1	151	1	Preparation T	2	
4	330	0.25	Standard S	1	162	1.5	Preparation T	3	

When the data in Excel is ready for analysis, highlight the block of cells you wish to analyse. In order to transfer the highlighted block to UNISTAT properly, it should conform to a few rules.

1) A column in the block should contain either numeric or String Data, but not a mixture of both (with the exception of Column Labels).
2) Row 1 of the block may contain Column Labels or data.
3) Column 1 of the block may contain Row Labels or data.

In most cases, the program will detect automatically whether the first row of the block contains Column Labels or data. If there is an ambiguity, a dialogue will pop up and ask for clarification.

Statistics: Suppose you wish to construct a table for the columns in the highlighted block, displaying basic statistical information. Select Statistics 1 → Descriptive Statistics → Summary Statistics to display UNISTAT's Summary Statistics Variable Selection Dialogue. A list on the left (the Variables Available list) will show the columns in the highlighted block as *C1, C2,…,* and their labels, if any. Highlight the columns to be analysed and then click on the [Variable] button. The selected columns will be transferred to the Variables Selected list on the right. Then click on [Next] to display the Output Options Dialogue. Here you can select the statistics you want to display in the table. Finally, click [Finish] to send the output to a new worksheet in Excel.

Graphics: From UNISTAT menus select Graph → 2D Plots → X-Y Plots. If you did not disturb the highlighted data block in Excel data sheet, you will see the variables selected for the previous procedure still present in the Variables Selected list on the right. Clicking on [Finish], the graph will be displayed in UNISTAT Graphics Editor, where you can edit and customise it. When finished, click on the Excel button to send the graph to Excel. The top-left corner of the graphics object will be placed at the active cell of the active Excel sheet.

For further information see 2.2.0. Output Medium Toolbar and 2.2.3. Output to Excel. The Unistat Tools menu option provides access to UNISTAT setup options, macros, log file and example files (see 2.4. Tools).

1.3.3. Background Mode

It is possible to use UNISTAT from another application as a statistics and scientific graphics engine. By adding a few lines to the application (which can be in any language, including C++, VB, VBA) it is possible to start UNISTAT in the background mode, pass the data and instructions and receive results without any part of UNISTAT appearing on the screen.

A Developer's Pack available from UNISTAT Ltd. contains a full description of all the calls and actions the developer needs to know.

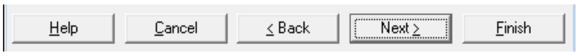
UNISTAT Statistical Package

Chapter 2
Common Features

2.1. Procedure Dialogues

All graphics and statistics procedures of UNISTAT are accessed via standard pull down menus. When a procedure is selected, one or more dialogues will be opened, requesting further input. Each procedure has its own specific sequence of dialogues. However, the most commonly used ones are the Variable Selection Dialogue and Output Options Dialogue.

All selections made in Procedure Dialogues can be saved in UNISTAT macro files. These macros can then be replayed to run the same procedure with the same settings (see 2.4.2. Macros).



The five buttons provided at the bottom of Procedure Dialogues have the following tasks:

Help: Displays context sensitive help on the current procedure. The keyboard equivalent is <Alt> + <H>.

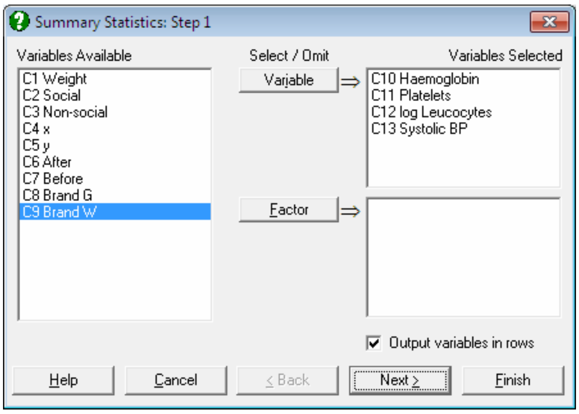
Cancel: Closes the dialogue and returns the control to the spreadsheet. The keyboard equivalent is <Alt> + <C>.

≤ Back: Drops the control back one level in the hierarchy of Procedure Dialogues. It is disabled at the first dialogue. The keyboard equivalent is <Alt> + <<>or <Escape>.

Next ≥: Displays the next dialogue. If there are no further dialogues, it is disabled. The keyboard equivalent is <Alt> + <>> or <Enter>.

Finish: Accepts all forthcoming dialogues in the procedure dialogue hierarchy with default options and executes the procedure. The keyboard equivalent is <Alt> + <F>.

2.1.1. Variable Selection Dialogue



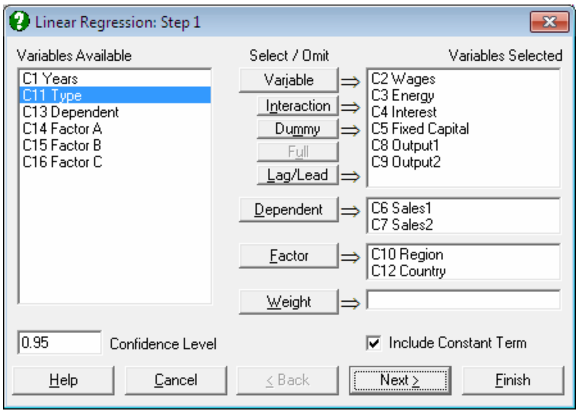
Summary Statistics: Step 1

Variables Available: C1 Weight, C2 Social, C3 Non-social, C4 x, C5 y, C6 After, C7 Before, C8 Brand G, C9 Brand W

Variables Selected (Variable): C10 Haemoglobin, C11 Platelets, C12 log Leucocytes, C13 Systolic BP

☑ Output variables in rows

When a procedure is selected from the pull-down menu, a Variable Selection Dialogue will be opened first (with the exception of a few procedures that do not require data, such as Plot of Functions, Critical Value, Cumulative Probability, Sample Size and Power Estimation). In some dialogues, there will be a section on top, offering various options for the type of data to be analysed. Under this, there will be a list box on the left displaying all available variables. There will also be one or more buttons (task buttons) displayed at the centre and a list box on the right corresponding to each button. These are used to assign specific tasks to selected variables.



Linear Regression: Step 1

Variables Available: C1 Years, C11 Type, C13 Dependent, C14 Factor A, C15 Factor B, C16 Factor C

Variable: C2 Wages, C3 Energy, C4 Interest, C5 Fixed Capital, C8 Output1, C9 Output2

Dependent: C6 Sales1, C7 Sales2

Factor: C10 Region, C12 Country

Weight:

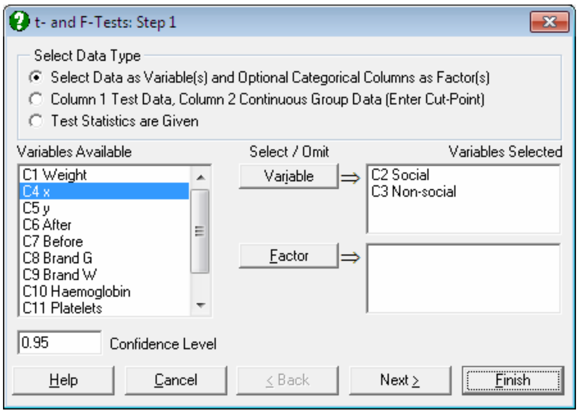
0.95 Confidence Level

☑ Include Constant Term

For instance, in X-Y Plots procedure, the X-Axis variable and Y-Axis variables will have two separate task buttons. The type and number of task buttons and their corresponding list boxes are specific to each procedure. They will also differ within a procedure according to the type of data used. It is possible to highlight multiple items in any list and send them to any other list either by clicking on task buttons or by drag-dropping using the right-mouse button. In some procedures, Variable Selection Dialogues may contain other controls like check or text boxes.

Not all selections are compulsory in a Variable Selection Dialogue. It is possible, for instance, to plot an X-Y line diagram without selecting an X-Axis variable, or to display a summary statistics table without having to select a categorical data (factor) variable. Compulsory variables are marked bold and [Next] and [Finish] buttons are not enabled unless all compulsory variables are selected.

2.1.1.1. Data Type Selection



t- and F-Tests: Step 1

Select Data Type
- ● Select Data as Variable(s) and Optional Categorical Columns as Factor(s)
- ○ Column 1 Test Data, Column 2 Continuous Group Data (Enter Cut-Point)
- ○ Test Statistics are Given

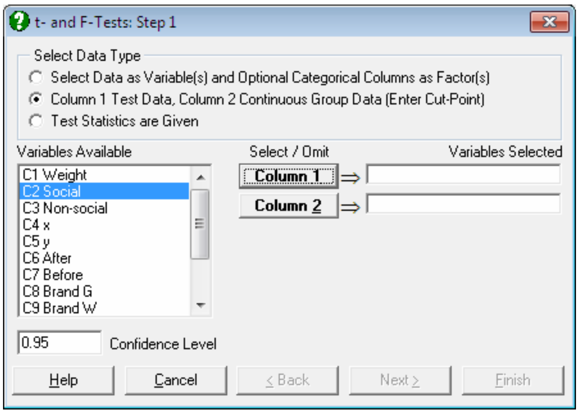
Variables Available: C1 Weight, C4 x, C5 y, C6 After, C7 Before, C8 Brand G, C9 Brand W, C10 Haemoglobin, C11 Platelets

Variables Selected (Variable): C2 Social, C3 Non-social

Factor:

0.95 Confidence Level

Variable Selection Dialogues of some procedures will have a section on top, displaying options for the type of data to be analysed. For instance, in t- and F-Tests procedure, it is possible to run a test between two columns in the spreadsheet by selecting them as [Variable]s. It is also possible to select one or more optional categorical variables (factors) to run the test between the subgroups defined by categories. There will also be an option to run the tests between the selected variables, but only for the rows defined by some categories. All this is possible while the first data option is selected.



t- and F-Tests: Step 1

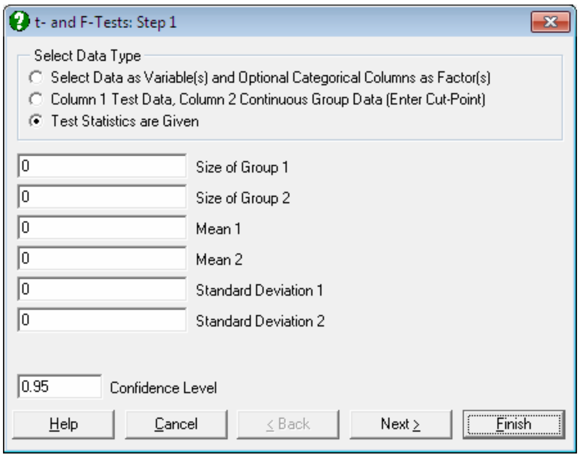
Select Data Type

○ Select Data as Variable(s) and Optional Categorical Columns as Factor(s)
● Column 1 Test Data, Column 2 Continuous Group Data (Enter Cut-Point)
○ Test Statistics are Given

Variables Available: C1 Weight, C2 Social, C3 Non-social, C4 x, C5 y, C6 After, C7 Before, C8 Brand G, C9 Brand W

0.95 Confidence Level

When the second data option is selected, the dialogue will be updated to display a different set of task buttons. In this case, you can select two data columns by clicking on [Column 1] and [Column 2]. Next, the program will ask for a cut-point for the second column, which will divide it into two groups smaller than and greater than or equal to the cut-point. The test will then be performed between these two groups using the values in data in Column 1.



t- and F-Tests: Step 1

Select Data Type

○ Select Data as Variable(s) and Optional Categorical Columns as Factor(s)
○ Column 1 Test Data, Column 2 Continuous Group Data (Enter Cut-Point)
● Test Statistics are Given

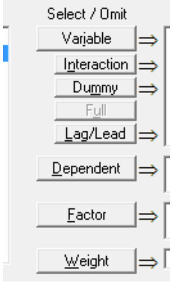
0 Size of Group 1
0 Size of Group 2
0 Mean 1
0 Mean 2
0 Standard Deviation 1
0 Standard Deviation 2

0.95 Confidence Level

When the third data option is selected, you can perform t- and F-Tests without selecting any data columns. If you know all the parameter values required, you will be able to run a t-test or F-test without having the data set itself.

This t-test example demonstrates only one of the many data type options available in UNISTAT. For other types of options see sections 5.0. Overview for descriptive statistics and 6.0. Overview for statistical tests.

2.1.1.2. Variable Selection by Task Buttons



On entry, all available data columns will be listed on the left (the Variables Available list) and the lists on the right (the Variables Selected lists) will be empty. In these lists, the numeric variables are referred to as *C1*, *C2*, etc., followed by their Column Labels, if any. String variables are distinguished from numeric variables in that the letter *C* in their column reference is replaced by *S* or *L*. Similarly, date variables will be represented by the letter *D* and time variables by *T*. For the details of these different types of data see 3.0.2. Data Types.

When one or more items are highlighted on the Variables Available list, an arrow will appear on the right hand side of each task button pointing to the right. When you click on a task button, the highlighted variables on the Variables Available list will be moved to the list that is immediately to the right of this particular button. Variables selected for one task are removed from the Variables Available list and therefore they cannot be selected for another task simultaneously. Exceptions to this are the regression and GLM Variable Selection Dialogues where variables can be selected for more than one purpose.

In order to deselect a variable from the analysis, click on this variable on the right Variables Selected list. A left-pointing arrow will be located to the left of the task button corresponding to this list. Clicking on the button will deselect the variable and add it to the Variables Available list on the left. Multiple highlighting and drag-and-drop will work for all list boxes. There are two types of selection buttons:

1) Buttons that allow any number of items (like [Variable], [Factor]).

2) Buttons that allow a limited number of items (like [X axis], [Column 1]). If a procedure allows only one variable to be assigned a certain task, then the button for this task will be disabled for further selections. To assign the task to a different variable, the selected variable must first be deselected.

Buttons used for selecting columns are specific to procedures. However, it is possible here to give a brief description of the most commonly used buttons.

[Variable]: Selects any number of variables for analysis. Examples: Y-axis variables in X-Y Plots, independent variables in Regression Analysis, test variables in statistical tests, etc. The order of selection is significant, that is, the analysis will be carried out on the selected variables in the order they appear in the Variables Selected list.

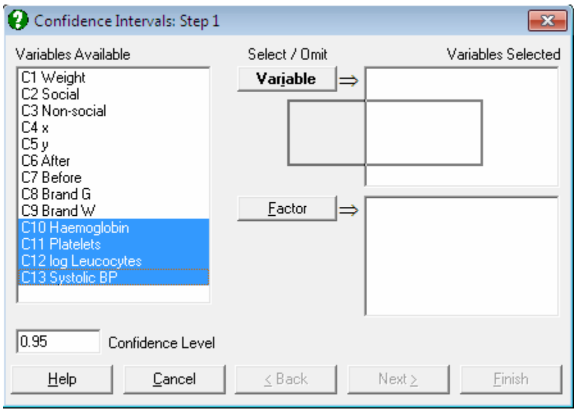
[Factor]: Selects categorical variables that define subgroups of one or more continuous variables. The order of selection is significant. The number of factors selected is usually unlimited but in some procedures it may be limited to one (as in Survival Analysis).

[Dependent]: Selects columns containing continuous data, usually for use in Regression Analysis and Analysis of Variance procedures.

[X-axis], [Y-axis], [Z-axis]: Select columns for X, Y or Z axis of a graph.

[Weight]: Selects a column as weights in the analysis of other columns.

2.1.1.3. Variable Selection by Drag-Drop

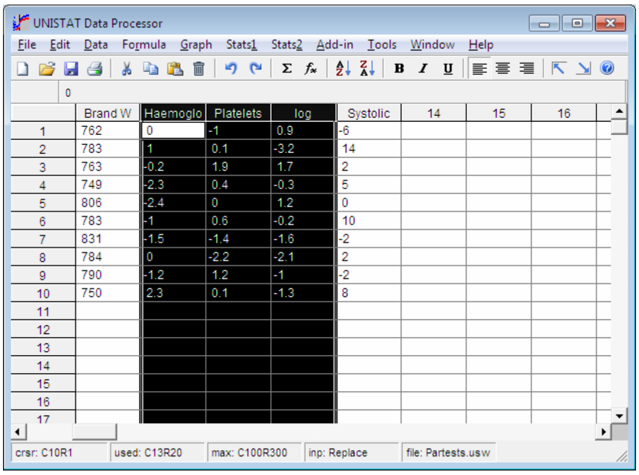


C1 Weight
C2 Social
C3 Non-social
C4 x
C5 y
C6 After
C7 Before
C8 Brand G
C9 Brand W
C10 Haemoglobin
C11 Platelets
C12 log Leucocytes
C13 Systolic BP

0.95 Confidence Level

It is also possible to assign tasks to selected (highlighted) variables by pressing down the right mouse button, dragging them on a Variables Selected list and dropping. Conversely, the highlighted variables can be deselected by drag-dropping them from a Variables Selected list to the Variables Available list. Drag-drop also works between different Variables Selected lists.

2.1.1.4. Variable Selection from Data Processor


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help
B I U
0

	Brand W	Haemoglo	Platelets	log	Systolic	14	15	16
1	762	0	-1	0.9	-6			
2	783	1	0.1	-3.2	14			
3	763	-0.2	1.9	1.7	2			
4	749	-2.3	0.4	-0.3	5			
5	806	-2.4	0	1.2	0			
6	783	-1	0.6	-0.2	10			
7	831	-1.5	-1.4	-1.6	-2			
8	784	0	-2.2	-2.1	2			
9	790	-1.2	1.2	-1	-2			
10	750	2.3	0.1	-1.3	8			
11								
12								
13								
14								
15								
16								
17								

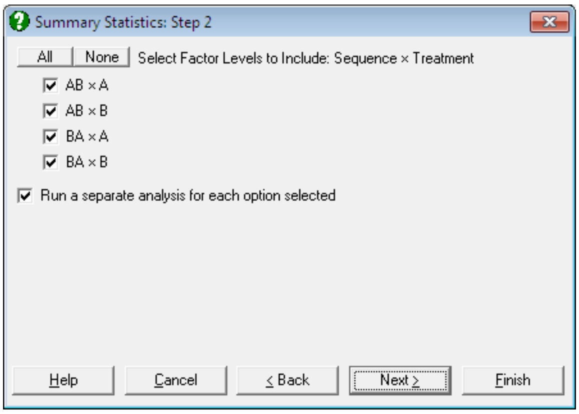
crsr: C10R1 | used: C13R20 | max: C100R300 | inp: Replace | file: Partests.usw

In Stand-Alone Mode, another method of selecting variables for analysis is to highlight them in the Data Processor. Non contiguous blocks of columns can be selected by holding down the <Ctrl> key whilst clicking on the column label. When a procedure is selected, the program will automatically issue a [Finish] command and proceed to perform the procedure with default selections. Normally, the user will obtain the output without seeing any dialogues.

Columns of data highlighted in the Data Processor are assigned the non-specific [Variable] task. This will be sufficient to generate an output with default settings in many procedures. If the selected procedure requires further compulsory variable assignments (e.g. a dependent variable for Regression Analysis), then the program will issue a warning and display the relevant Variable Selection Dialogue.

It is also possible to select a block of cells (instead of entire columns) to run a procedure on the selected range only. If a block of cells is highlighted and, say Summary Statistics is selected, the program will automatically generate a Select Row variable and run the procedure on the selected block only. All procedures will run on the selected cases as long as the Select Row column remains in effect.

2.1.2. Categorical Data Analysis



Summary Statistics: Step 2

Select Factor Levels to Include: Sequence × Treatment

☑ AB × A
☑ AB × B
☑ BA × A
☑ BA × B

☑ Run a separate analysis for each option selected

Many UNISTAT procedures offer the possibility to perform analyses on or between subgroups of data columns, as defined by one or more factor columns (i.e. categorical variables). Usually, selection of a factor variable is optional. If at least one factor column is selected, then a further dialogue will pop up, displaying a check list of all levels (i.e. the distinct values) of the factor. If two or more factor columns are selected, combinations of levels in selected factors will be listed.

There will also be a check box Run a separate analysis for each option selected, which is used to determine whether the variables or the factors will take priority. When this box is checked, the test will be performed on or between variables for each level combination checked. In other words, factors will be in the outer loop. When the Run a separate analysis for each option selected box is unchecked, then the program will perform the procedure on or between all selected levels (or combinations of levels) of factor columns, for each variable. In other words, variables will be in the outer loop. Note that this check box may have slightly different semantics in different groups of procedures. For an example demonstrating this point see 5.1.1. Summary Statistics.

When the list of levels (or level combinations) is displayed, all entries will be checked by default. You can uncheck all by clicking on the None button on top. However, if you wish to have all boxes displayed unchecked on entry, enter the following line in *Unistat65.ini* file under the [Options] section:

```
CheckAllFirst=0
```

By default, the Run a separate analysis for each option selected box is displayed checked. Enter the following line in *Unistat65.ini* file under the [Options] section in order to display this box unchecked on entry:

```
RunSeparate=0
```

When the program determines the levels of a factor, by default, it sorts them alphabetically in increasing order. If you wish to have factor levels displayed unsorted, i.e. in the order of their occurrence in each factor column, then enter the following line in *Unistat65.ini* under the [Options] section:

```
SortFactorLevels=0
```

The way the categorical analysis option works is slightly different in each of the following three groups of procedures.

Multisample Data

In this group of procedures, where [Variable] and [Factor] lists are displayed together, it is optional to select a factor variable. If no factor variables are selected, then the test is performed on the selected variables with no categorisation. If at least one factor variable is selected, then it is possible to run the analysis on subgroups defined by the combination of factor levels. Procedures in this category are as follows:

Graph → 2D Plots →
X-Y Plots
Polar Plot
Graph → Charts →
Pie Chart
Bar Chart
Area Chart
Ribbon Chart
3D Bar Chart
Graph → Descriptive Plots →
Box-Whisker, Dot and Bar Plots
Normal Probability Plot
Histogram
Statistics 1 → Descriptive Statistics →
Summary Statistics
Confidence Intervals
Quantiles (Percentiles)
Statistics 1 → Parametric Tests →
Parametric Tests Matrix

Statistics 1 → Goodness of Fit Tests →
Normality Tests
Statistics 1 → Nonparametric Tests (Multisample) →
Kruskal-Wallis One-Way ANOVA
Multisample Median Test

Two Independent Samples Data Option

Tests for two unrelated samples allow selecting an unlimited number of factors as in the multisample data option above. These procedures are:

Statistics 1 → Parametric Tests →
t- and F-Tests
Equivalence Test for Means
Parametric Tests Matrix
Statistics 1 → Goodness of Fit Tests →
Kolmogorov-Smirnov Tests
Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Samples →
Mann-Whitney U Test
Hodges-Lehmann Estimator (Unpaired)
Wald-Wolfowitz Runs Test
Moses Extreme Reaction Test
Two Sample Median Test

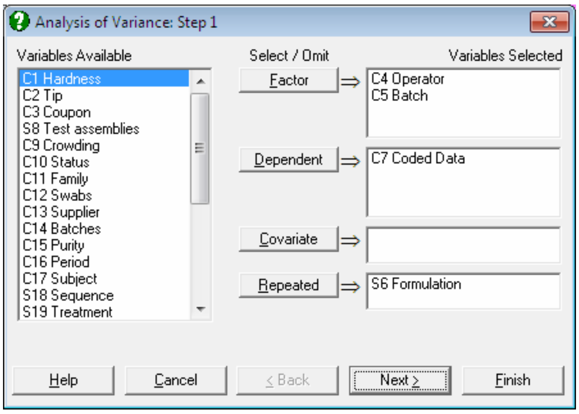
However, unlike the multisample data option, these procedures support two additional data types (see 6.0.2. Two Sample Tests).

Matrix Data

In procedures where data is supplied in matrix format, it is possible to select one or more factors in order to define the groups of cases to be included in the analysis. It is then possible to run a separate analysis on each subsample in one go, or to run a combined analysis on all selected subsamples. The procedures where this feature is available are:

Statistics 1 → Matrix Statistics
Statistics 1 → Regression Analysis →
Linear Regression
Polynomial Regression
Stepwise Regression
Logit / Probit / Gompit
Logistic Regression
Multinomial Regression
Poisson Regression

2.1.3. Multiple Dependent Variables



Variables Available

C1 Hardness
C2 Tip
C3 Coupon
S8 Test assemblies
C9 Crowding
C10 Status
C11 Family
C12 Swabs
C13 Supplier
C14 Batches
C15 Purity
C16 Period
C17 Subject
S18 Sequence
S19 Treatment

Variables Selected

Factor ⇒ C4 Operator
C5 Batch

Dependent ⇒ C7 Coded Data

Covariate ⇒

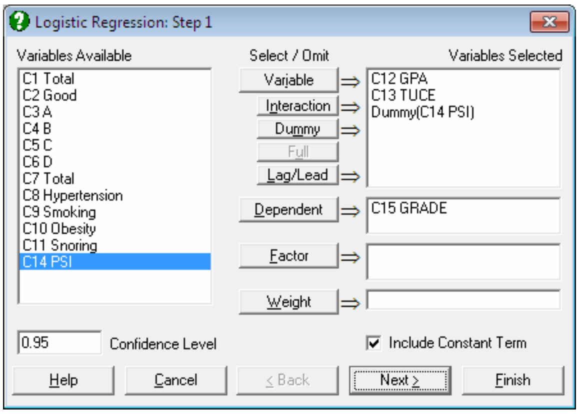
Repeated ⇒ S6 Formulation

UNISTAT allows selection of more than one dependent variable in the following procedures:

Statistics 1 → ANOVA and GLM →
Analysis of Variance
General Linear Model
Statistics 1 → Regression Analysis →
Linear Regression
Polynomial Regression
Stepwise Regression
Logistic Regression
Multinomial Regression
Poisson Regression

In these procedures, when more than one dependent variable is selected, the analysis will be repeated as many times as the number of dependent variables, each time changing only the dependent variable and keeping other selections unchanged. Missing values are handled independently for each run.

2.1.4. Creating Interaction, Dummy and Lag/Lead Variables



Variables Available
C1 Total
C2 Good
C3 A
C4 B
C5 C
C6 D
C7 Total
C8 Hypertension
C9 Smoking
C10 Obesity
C11 Snoring
C14 PSI

Variables Selected
C12 GPA
C13 TUCE
Dummy(C14 PSI)

Dependent: C15 GRADE

0.95 Confidence Level

☑ Include Constant Term

In most regression models, it is possible to add new terms to the model using transformations of existing variables, thus eliminating the need to create them as data columns in the spreadsheet beforehand. In these procedures, interaction, dummy and lag/lead terms can be specified during the variable selection phase. The program will then create these terms internally in its temporary memory. This feature is available in the following procedures:

Statistics 1 →
Matrix Statistics
Statistics 1 → Regression Analysis →
Linear Regression
Polynomial Regression
Stepwise Regression
Logit / Probit / Gompit
Logistic Regression
Multinomial Regression
Poisson Regression
Statistics 2 → Survival Analysis →
Cox Regression

Although it is not a regression procedure as such, this feature is also included in the Matrix Statistics procedure to provide you with information on the terms of the models selected in other regression procedures. A particularly useful feature here is the option to send the entire final raw data (**X**) matrix of the regression model to the Output Medium, so that you can see the actual values of the interaction, dummy and lag/lead terms generated by the program internally. In Stand-Alone Mode, this output can then be sent to the Data Processor and used in other procedures if necessary (see 7.1. Matrix Statistics).

In the Variable Selection Dialogue of the above procedures, up to four smaller buttons [Interaction], [Dummy], [Full] and [Lag/Lead] will appear just under the [Variable] button. The Cox Regression procedure does not have a [Lag/Lead] button, as it is irrelevant for this procedure.

The behaviour of these buttons differs from other standard task buttons in that:

1) They can be used to select items from both the Variables Selected list, as well as the Variables Available list.
2) When they are used on a selection, the source (selected) items are not omitted from the list.
3) To de-select the new variables created by these buttons from the Variables Selected list, the [Variable] button should be used (not the button with which they were created).

The functionality of these buttons are as follows:

Interaction: This button becomes activated when one or more items are selected from one of the Variables Available or Variables Selected lists. If only one variable is highlighted, then a new term will be added to the Variables Selected list, which is the product of this variable by itself, e.g. *C2 Wages* × *C2 Wages*. If two variables are highlighted, then the new term will be the product of these two variables, e.g. *C10 Region* × *C11 Type*. Maximum three-way interactions are allowed. Interactions of string, date, time, dummy or lag/lead variables are not allowed. In order to create interaction terms for dummy variables, create interactions first, and then create dummy variables for them.

A special cross sign is used between the variables in an interaction term. If there is a problem with this character on a non-English operating system, you can enter and edit the following line in *Unistat65.ini* file under the [Options] group to display any other character (say *):

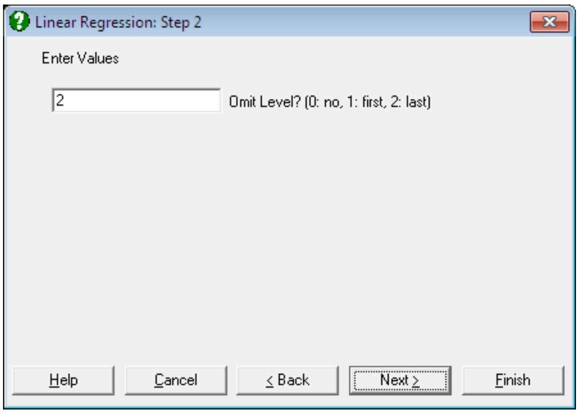
```
InteractionCross=×
```

This sign also appears in ANOVA and GLM output with interaction terms.

Dummy: This button is used to create *n* new (dummy) variables for a factor column containing *n* levels. Each dummy variable corresponds to a level of the factor column. A case in a dummy column will have the value of 1 if the factor contains the corresponding level in the same row, and 0 otherwise.

One or more categorical variables can be selected from one of the Variables Available or independent variables lists. A new entry will be created in the independent variables list in the form of Dummy(*C10 Region*). If the selection contains interactions, dummies will be created in the form of Dummy(*C10 Region* × *C11 Type*).

Sometimes, it may be important to select the interaction terms in a specific order which may be different to their default order in the Variables Available list. In this case, you can select the columns for the independent variables list by clicking on the [Variable] button in the desired order before selecting them for interactions or dummies.


Linear Regression: Step 2

Enter Values

2 Omit Level? (0: no, 1: first, 2: last)

Once one or more dummy variables have been included in the model, the next dialogue will ask whether you wish to include all dummy variables corresponding to all levels of a factor (or an interaction term of factors), or omit the first level or the last level. This dialogue may have other fields in some procedures. The purpose of this exercise is to provide you with a facility to remove the linear dependencies created when all levels of factors are included in the model. For example, suppose the first option was selected and a dummy variable created for each level of a factor. This results in an over-parameterised model, since the full set of dummy variables for any factor will always add up to the unity vector. If a model is run in this configuration, the regression algorithm will detect and omit the dummy variables that cause the collinearity as and when they occur. In regression models where a constant term is included, UNISTAT will naturally omit the last dummy variable for each factor, and each factor with i levels will end up contributing (i - 1) degrees of freedom to the model (if no other dependencies exist in data). The interaction terms will also include dummy variables for all possible combinations of their individual factor levels. Again, if no other dependencies exist, they will only contribute to the model with (i - 1)(j - 1) degrees of freedom in the case of a two-way interaction and (i - 1)(j - 1)(k - 1) in the case of a three-way interaction. However, if a constant term is not included in the model, then all levels of the first factor will be included. You are advised to consider these issues before running a model with dummy variables and omit either the first or the last level in order not to end up with unexpected results.

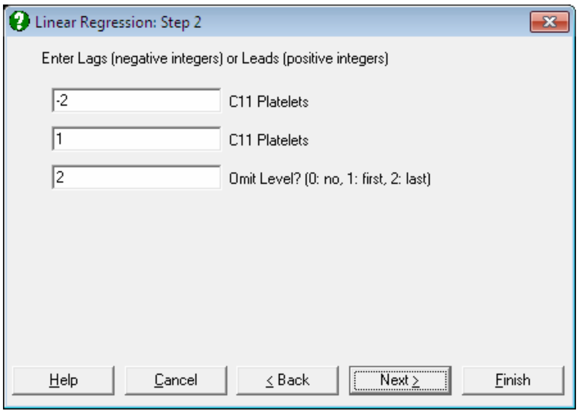
If you wish to omit levels other than the first or last, or include dummy variables with values other than 0 or 1 (i.e. in order to apply contrasts), it is advised that you construct dummy variables as data columns first. In Stand-Alone Mode, you can do this automatically using the Data Processor's **Dummy()** function (see 3.4.2.5. Statistical Functions). Dummy variables created in this way should be included in the analysis by selecting them as [Variable]s.

Full: This button becomes active when two or more items are selected from one of the independent variables or Variables Available lists. Dummy variables for each column selected, as well as dummies for all possible interaction terms (up to 3-way) will be created automatically. The number of new independent variables added to the model is determined by the number of distinct values (levels) of the selected columns. For each interaction term, this number is equal to the product of the number of levels in all variables in the term.

For example, suppose two columns are highlighted, *C10 Region* and *C11 Type*. Then the following three terms will be created: Dummy(*C10 Region*), Dummy(*C11 Type*), Dummy(*C10 Region*C11 Type*). If four columns are highlighted, *C1, C2, C3, C4,* then 14 new dummy terms will be created in the following order: Dummy(*C1*), Dummy(*C2*), Dummy(*C3*), Dummy(*C4*), Dummy(*C1*C2*), Dummy(*C1*C3*), Dummy(*C1*C4*), Dummy(*C2*C3*), Dummy(*C2*C4*), Dummy(*C3*C4*), Dummy(*C1*C2*C3*), Dummy(*C1*C2*C4*), Dummy(*C1*C3*C4*), Dummy(*C2*C3*C4*). Also suppose that *C1, C2, C3, C4* contain 2, 3, 4, 5 levels respectively. Then the total number of variables added to the model will be 239.

WARNING! *Ensure that the columns selected as Dummy are categorical variables containing a limited number of distinct values.*

Lag/Lead: This button is used to create new variables by shifting the rows of an existing variable up or down. If you highlight some columns and click on the [Lag/Lead] button, these will be transferred to the Variables Selected list as Lag(*C1 Label1*;0), Lag(*C2* Label2;0), etc. In this case, after clicking [Next] a further dialogue will ask for the size of the lags (or leads) for each [Lag/Lead] variable selected.


Linear Regression: Step 2

Enter Lags (negative integers) or Leads (positive integers)

-2 C11 Platelets

1 C11 Platelets

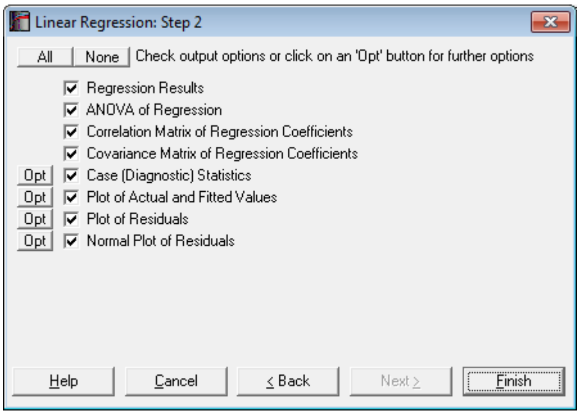
2 Omit Level? (0: no, 1: first, 2: last)

You can select the same variable an unlimited number of times, in order to include it in the model with different sizes of lags/leads. Enter a negative integer to define a selection as a lag and a positive integer for a lead. Leaving an entry as zero means that the selected variable will be included in the model without modification. When, for instance, -2 is entered for a lag, the program will create a new variable internally, starting from the third case of the original column. Therefore this variable will have two observations missing at the end. On the other hand, when 2 is entered for a lead variable, its first observation will correspond to the third row of the data matrix, the first two cases will be defined as missing and the last two cases will be omitted.

WARNING! *Selection of lag/lead variables results in a loss of degrees of freedom. You must ensure that there is a sufficient number of cases (rows) in the data matrix when lags and / or leads are selected.*

WARNING! *The interpretation of lag/lead variables may not be clear when they are selected along with factors for categorical analysis (see* 2.1.2. Categorical Data Analysis*) or with the Data Processor's* Data → Select Row *function. The use of lag/lead variables is not prevented in such cases you should ensure that their effect is unambiguous.*

2.1.5. Output Options Dialogue



Check output options or click on an 'Opt' button for further options

☑ Regression Results
☑ ANOVA of Regression
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
☑ Case (Diagnostic) Statistics
☑ Plot of Actual and Fitted Values
☑ Plot of Residuals
☑ Normal Plot of Residuals

The Output Options Dialogue is displayed after an analysis is run in procedures with potentially long output, such as Linear Regression, General Linear Model, Multiple Comparisons, etc. This dialogue lets you choose only those output options you are interested in.

If an output option requires further user input (i.e. if it has further dialogues and windows to display) then an [Opt] button will be placed to the left of its check box. When you click [Finish] without clicking on an [Opt] button first, the program will dump this output option (alongside other options checked) with its default values. If you want to change the default values, you can click on the [Opt] button to display the further dialogues for this particular output option. Then you can either obtain this particular output option on its own by clicking [Next] or [Finish], or click [Back] to display the Output Options Dialogue again and output all selected options together.

UNISTAT stores selections made in the Output Options Dialogue. User-selected output preferences will persist across UNISTAT sessions.

2.2. Output Medium

One of the most popular aspects of UNISTAT Statistical Package is its ability to create output in Word and Excel, using the powerful automation features of these applications.

When UNISTAT is run in Stand-Alone Mode, by default, all output is sent to a WordPad-like window, which is an integral part of UNISTAT Statistical Package. In Excel Add-In Mode, the output is sent to a new worksheet within Excel by default. After this, it is possible to send the same output to a number of other applications, without having to run the procedure again. The Output Medium Toolbar allows you to send the output to Word, Excel, web browser or to the Windows system clipboard.

It is also possible to set any one of these output media as default, in which case, the output is created directly in the new Default Output Medium.

UNISTAT does not simply send the old style *line printer* output to these applications. Instead, it re-formats its output fully utilising the specific formatting capabilities of each application. When the Word button is clicked, UNISTAT will format its tables directly within Word, in the form of Word tables. Likewise, when the Excel button is clicked, output tables will be formatted in the form of Excel tables directly within Excel. When the browser button is clicked, UNISTAT creates an HTML file and formats the output as HTML tables.

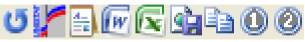
In this way, it is possible to use Output Window as the primary preview window and send only the final results to Word (or Excel or web browser) for inclusion in a final report.

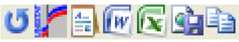
In Stand-Alone Mode, UNISTAT can also send its output tables to its own spreadsheet (the Data Processor) for further analysis.

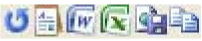
2.2.0. Output Medium Toolbar

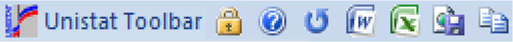
The Output Medium Toolbar appears on Data Processor, Output Window and Graphics Editor window after performing a procedure. In Excel Add-In Mode, the UNISTAT toolbar is visible at all times, but the Output Medium buttons will be functional only after a procedure has been performed. All toolbars have similar Output Medium buttons, but there may be additional window-specific buttons and not all output buttons may appear depending on which application is installed on the system. The Word, Excel and web browser buttons will appear only if their respective applications have been installed previously.

When all options are available, the Output Medium Toolbar will look like this on different windows:

Data Processor: 

Output Window: 

Graphics Editor: 

Excel add-in: 

The toolbar buttons have the following tasks:

Last Procedure Dialogue: The last dialogue in the hierarchy of Procedure Dialogues will be displayed allowing you to change the output options without having to re-run the entire procedure.

Send Matrix to Data Processor (Stand-Alone Mode **only):** This button is available only when the output from a procedure is in table format. When it is clicked, the table will be copied to Data Processor cells starting from the first blank column available, so that it can be used as input for further analysis. Even if numbers in the output are formatted to display a limited number of digits, they will be copied to the spreadsheet cells with the full 15 digits of precision. Abbreviated labels will be generated for each column added to the spreadsheet.

Output to Output Window (Stand-Alone Mode **only):** This is UNISTAT's own WordPad-like Output Window. The text output is in old style *line printer* format and graphics are in enhanced metafile format.

Output to Word: Output is sent to the current document in Word, at the current cursor position, as formatted Word tables and enhanced metafile format pictures. See 2.2.2. Output to Word below.

Output to Excel: Output is sent to a new Excel worksheet in the form of formatted Excel tables and enhanced metafile format pictures. See 2.2.3. Output to Excel below.

Output to Web Browser: Output is sent to the default web browser in the form of HTML tables and PNG format bitmap images. See 2.2.4. Output to Web Browser below.

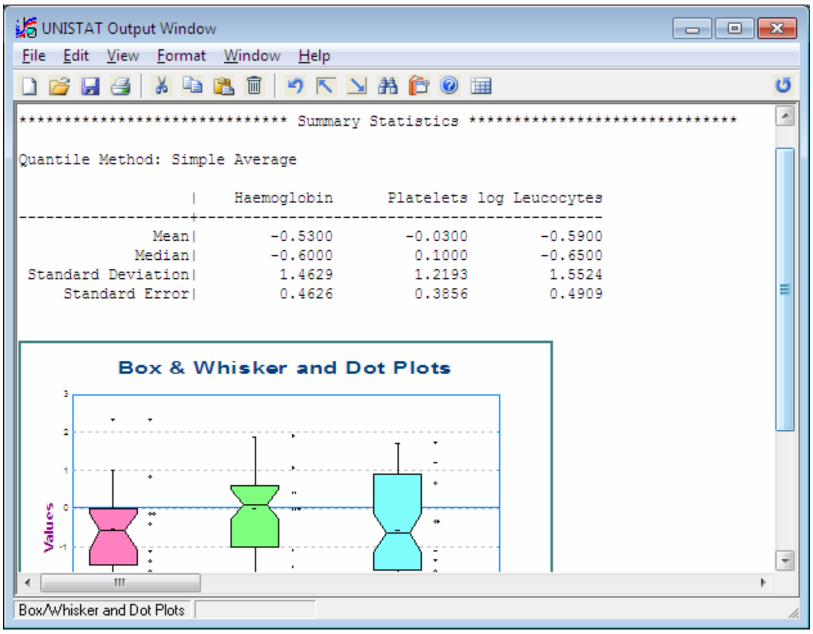
Output to the Clipboard: Output is copied to the clipboard using a fixed width font (as in Output Window). Graphics are in enhanced metafile format.

Macro Shortcut Buttons (Stand-Alone Mode, **Data Processor only):** The macro file *Documents*\Unistat65\UsrBtn1.usm is run and output is sent to the Default Output Medium. See 2.4.2.4. Macro Shortcut Buttons.

Lock UNISTAT Data (Excel Add-In Mode **only):** The highlighted block of data in the active Excel worksheet is locked. While this button is depressed, highlighting other blocks will not change the data to be analysed by UNISTAT. To change the analysis range click on this button again and then highlight a new block of data.

UNISTAT Help (Excel Add-In Mode **only):** This button launches the UNISTAT help system.

2.2.1. UNISTAT Output Window


UNISTAT Output Window
File Edit View Format Window Help
Box & Whisker and Dot Plots
3
2
1
0
-1
Values
Box/Whisker and Dot Plots

```
****************************** Summary Statistics ******************************

Quantile Method: Simple Average

                   |    Haemoglobin      Platelets log Leucocytes
-------------------+---------------------------------------------
               Mean|        -0.5300        -0.0300        -0.5900
             Median|        -0.6000         0.1000        -0.6500
 Standard Deviation|         1.4629         1.2193         1.5524
     Standard Error|         0.4626         0.3856         0.4909
```

In Stand-Alone Mode, by default, all output is sent to UNISTAT's own Output Window. All text output is created in old style *line printer* format (where grid lines are composed of minus signs and grid line intersections are plus signs) and is displayed with a fixed-width font. The graphic output is sent to this window in the form of an enhanced metafile object.

The Output Window offers all the functionality of WordPad, including cut-and-paste editing, undo, print, change fonts, etc. Its contents can be saved in one of Rich Text (.RTF) or text (.TXT) formats. The Rich Text format will save all formatting information as well as graphics objects, while the text format saves the unformatted text only.

Although an integral part of UNISTAT Statistical Package, the Output Window works quasi-independently, so that it can be made active for editing the output at any time during a UNISTAT session. The contents of this window are not saved to a file automatically by the program. The user should take care to save the contents in a file in case a copy is to be kept on disk.

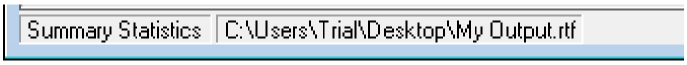
The buttons on Output Window toolbar are by and large identical to that of WordPad, with the exception of the following:

↖ **Home:** Moves the active cell to top of the file.

↘ **End:** Moves the active cell to bottom of the file.

▦ **Exit:** Returns the focus to Data Processor without closing the Output Window.

The two panels on the status bar will display the name of the last procedure executed and the output file name, if the contents have been saved to a file earlier.



Summary Statistics

C:\Users\Trial\Desktop\My Output.rtf

The first time the contents of the Output Window are to be saved, the Save As dialogue will ask for a file name. Subsequently, when the [Save] button is clicked, it will overwrite the file on disk with the current contents of the Output Window.

Unlike WordPad, however, Output Window also displays a second toolbar, the Output Medium Toolbar, which is justified to the right of the window. When one of the buttons on this toolbar is clicked, the output from the last procedure will be sent to the selected application. See 2.2.0. Output Medium Toolbar.

Default Font: On entry, 9 pt *Courier New* font is selected. Although this can be changed by the user from the Format → Font menu, UNISTAT will revert to the default font next time it is launched. To change the default font and size of Output Window permanently, enter and edit the following lines in *Documents*\Unistat65\Unistat65.ini file under the [Options] section:

```
FixedFontName=Courier New
FixedFontSize=9
```

Width of Output and Blocking: It is possible to adjust the top and left margins of output from Tools → Options dialogue's Output tab (see 2.4.1.2.3. Text Margins). More importantly, the width of output can also be set to any number from 80 to 32,000 characters. Output from UNISTAT will be re-scaled to fit the specified width. For instance, when a large correlation matrix is sent to the Output Window, the program will first work out how many columns of the matrix will fit within the specified width, and then separate the matrix into an appropriate number of blocks. Another example is character plots (i.e. character histogram, plot of residuals, fitted and actual y values, etc.) where the program determines the resolution of the graph according to the output width setting.

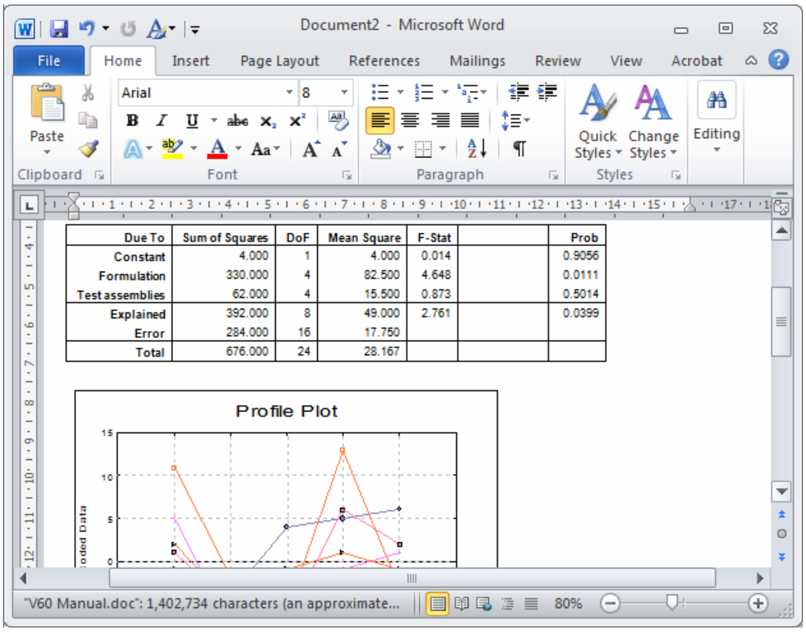
Graphics Object Size: It is possible to make the graphics objects appear smaller or bigger than the default size. The default size is 100%. To change this to a value smaller or larger than 100, enter and edit the following line in *Unistat65.ini* under the [Options] section:

```
MetafileSize=100
```

Graphics Font Size: It is possible to change the font size for all text that appear on the graph proportionately. The default size is 125%. To change this value enter and edit the following line in *Unistat65.ini* under the [Options] section:

```
FontLevelPct=125
```

2.2.2. Output to Word



Due To	Sum of Squares	DoF	Mean Square	F-Stat		Prob
Constant	4.000	1	4.000	0.014		0.9056
Formulation	330.000	4	82.500	4.648		0.0111
Test assemblies	62.000	4	15.500	0.873		0.5014
Explained	392.000	8	49.000	2.761		0.0399
Error	284.000	16	17.750			
Total	676.000	24	28.167			

"V60 Manual.doc": 1,402,734 characters (an approximate...

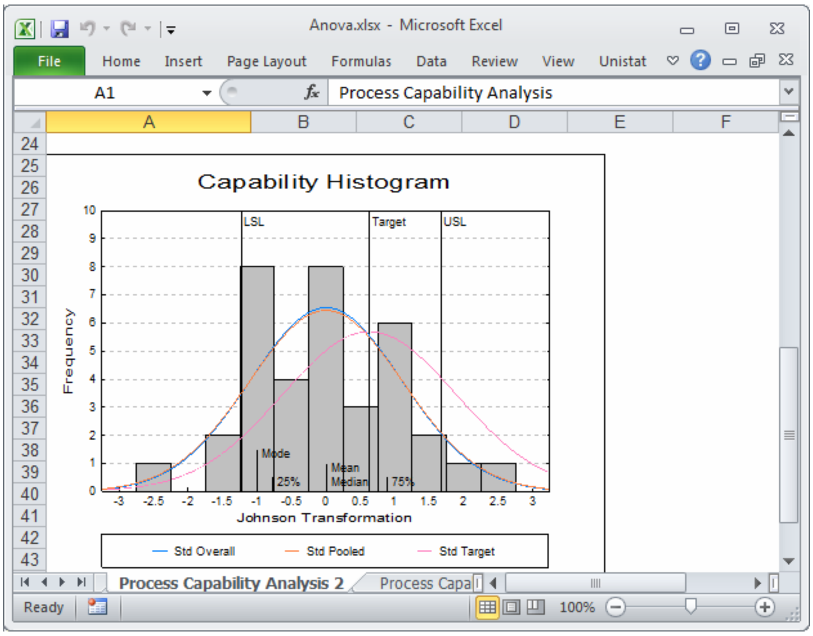
Output is sent to the current document in Word, at the current cursor position. If Word is not running already, it is launched first.

Styles: When UNISTAT output is sent to a Word document for the first time, it creates a number of Word styles for its own use. Subsequent pages of output will only refer to these styles. For more information on styles see the next section 2.2.3. Output to Excel.

Width of Output and Blocking: In Word output, large tables are parsed into blocks to facilitate easy viewing and printing. The default number of columns per block is 6, but you can change this to any number greater than 2, from Tools → Options dialogue's Output tab (see 2.4.1.2.4. Word and HTML Tables).

Graphics Object and Font Size: It is possible to make the graphics objects appear smaller or bigger than the default size. The font size can also be increased or decreased for all text on graphs proportionately. See 2.2.1. UNISTAT Output Window for details.

2.2.3. Output to Excel



Output is sent to a new worksheet in the active workbook. If Excel is not running already, it is launched first.

Styles: When UNISTAT output is sent to an Excel document for the first time, it creates some Excel styles for its own use. Subsequent pages of output will only refer to these styles and therefore will take less time to complete. Once they are automatically created by UNISTAT, you can edit these styles to your taste and all subsequent UNISTAT output will conform to your choices. The following UNISTAT styles are created:

1. **UNISTAT Main Title:** 16 pt, bold, italic, Times New Roman. In addition to the possibility of editing this style from within Excel, you can change the default title font by entering and editing the following line in *Documents*\Unistat65\Unistat65.ini file under the [Fonts] section:

   ```
   OfficeTitleFontName=Times New Roman
   ```

2. **UNISTAT Sub Title:** 12 pt, bold, italic, Times New Roman.

3. **UNISTAT Normal:** 8 pt, Arial. All text output is in this style. To change the default text font enter and edit the following line in *Unistat65.ini* under the [Fonts] section:

   ```
   OfficeBodyFontName=Arial
   ```

4. **UNISTAT Fixed:** 8 pt, Courier New. This fixed-width style is used for character plots (i.e. character histogram, plot of residuals, fitted and actual y values, etc.).

5. **UNISTAT Table Title 5:** The column and row titles of tables are defined by this style (Excel only). This style is like UNISTAT Normal, except that it has a coloured background. In addition to the possibility of editing this style from within Excel, you can also enter and edit the following line in *Unistat65.ini* under the [Options] section, to change the background colour:

   ```
   ExcelColour=19
   ```

Width of Output and Blocking: Unlike other output media, by default, Excel tables are not blocked, i.e. the entire matrix is output as one block. The main advantage of this approach is the possibility of highlighting an entire matrix and using it as input for further analysis. Under other circumstances, however, you may wish to have large matrices blocked as in other output options. To do this, enter the following line in *Unistat65.ini* under the [Options] section:

```
ExcelMaxCols=x
```

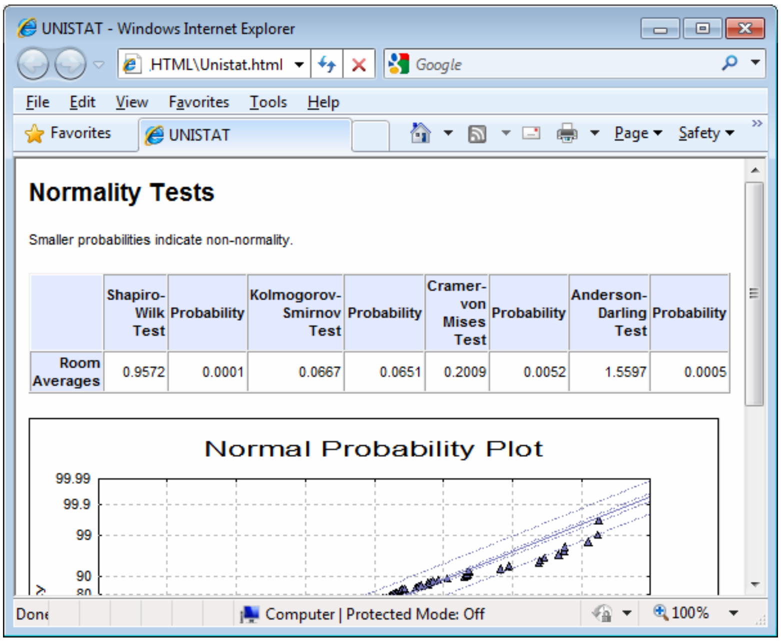
where x is the number of columns per block. This may be useful, among other things, for printing large matrices in portrait or landscape orientation, where x = 5 and x = 9 will be the appropriate values respectively.

Worksheet Titles: By default, UNISTAT will name each worksheet it creates with the title of the procedure performed, followed by a sequence number. If you do not like this approach (e.g. if these titles appear too long) you can switch titles off by entering the following line in *Unistat65.ini* under the [Options] section:

```
ExcelSheetName=0
```

Graphics Object and Font Size: It is possible to make the graphics objects appear smaller or bigger than the default size. The font size can also be increased or decreased for all text on graphs proportionately. See 2.2.1. UNISTAT Output Window for details.

2.2.4. Output to Web Browser

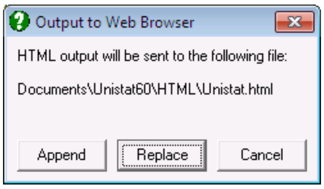

UNISTAT - Windows Internet Explorer
HTML\Unistat.html
Google
File Edit View Favorites Tools Help
Favorites
UNISTAT
Page Safety
Normal Probability Plot
99.99
99.9
99
90
80
Done
Computer | Protected Mode: Off
100%

Normality Tests

Smaller probabilities indicate non-normality.

	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	Probability	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
Room Averages	0.9572	0.0001	0.0667	0.0651	0.2009	0.0052	1.5597	0.0005

Output is sent to the default web browser in the form of HTML tables and PNG format bitmap images. If the browser is not running already, it will be launched first. When the output is sent to the web browser for the first time, it will appear in the browser automatically. Otherwise, the program will stop and ask whether you want to overwrite the existing files or append the new output to the existing page.



HTML output will be sent to the following file:

Documents\Unistat60\HTML\Unistat.html

Unlike the Output Window, Word and Excel options (which do not create a file copy of the output), this option will save all text output in the file UNISTAT.HTML in *Documents**Unistat65**HTML* folder. All graphics images are also saved in the same folder in separate PNG format image files. The .PNG files will be named sequentially as UNI00001.PNG , UNI00002.PNG , etc. until the Overwrite option is selected for new output. When the Overwrite option is selected, all .PNG files (and the UNISTAT.HTML file) are deleted and the counter is set to 1 again. Therefore, it is up to the user to maintain the .HTML and .PNG files and copy them to a different location when necessary.

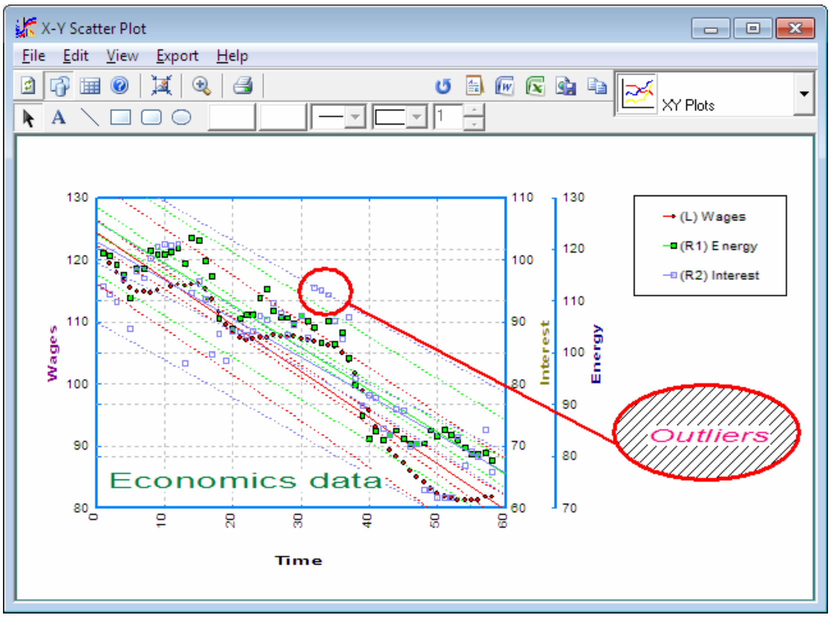
Image Size: Because PNG files contain bitmap images, they are created by means of displaying the Graphics Editor and capturing the image from screen. Therefore, it is normal to see the graphics images flashing in Graphics Editor window while HTML output is being produced. Another consequence of this process is the size of the graphics images in the HTML output being the same as the current size of the Graphics Editor window. It will be a good idea, therefore, to select the appropriate size of the Graphics Editor window before sending output to the browser.

Width of Output and Blocking: This is exactly as in Output to Word. The number of columns per block can be controlled from Tools → Options dialogue's Output tab (see 2.4.1.2.4. Word and HTML Tables).

Styles: Each HTML output page contains a CSS Cascading Style Sheet definition section, containing the following styles: main title <h3>, sub title <h4>, paragraph <p>, table heading <th> and table cell <td>. It is possible to modify the appearance by overriding the CSS classes in one place at any time after the output has been generated. Alternatively, you can change the default values of these styles by entering and editing the following lines in *Documents\Unistat65\Unistat65.ini* file under the [Options] section:

```
HTMLTitleFont=sans-serif, arial
HTMLTitleFontSize=14

HTMLSubTitleFont=sans-serif, arial
HTMLSubTitleFontSize=11

HTMLParaFont=sans-serif, arial
HTMLParaFontSize=8

HTMLHeadFont=sans-serif, arial
HTMLHeadFontSize=8
HTMLHeadFontBold=strong
HTMLHeadForeColor=black
HTMLHeadBackColor=#c0c0ff

HTMLCellFont=sans-serif, arial
HTMLCellFontSize=8
HTMLCellFontBold=weak
HTMLCellForeColor=black
HTMLCellBackColor=white
```

2.3. Graphics Editor

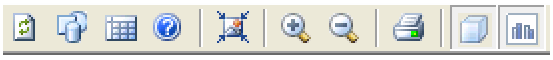


The UNISTAT Graphics Editor supports full on-screen object editing of graphs. All text, legends, and the graph itself can be drag-dropped and resized and new text, line and shape objects added. Text objects support the Rich Text format, allowing use of symbol fonts, subscripts and superscripts.

All aspects of graphs can be customised using a series of dialogues. It is possible to control the line thickness, colour and style of the picture frame, axes, grid, tick marks, line and shape objects and select the font, size and colour for any text on the graph separately. Axes are scaled automatically but options exist to override the suggested values. All new settings can be saved in a graphics template file for subsequent retrieval.

UNISTAT graphs can also be exported directly to Word or Excel, or they may be saved to the clipboard or to a file in one of bitmap (.PNG) or enhanced metafile (.EMF) formats.

2.3.1. Graphics Toolbar



The most commonly used graphics tools can be quickly accessed from the buttons on this toolbar. The Output Medium Toolbar, which is justified to the right of the same panel, was explained above.

Redraw: Draws the graph again with the current options (see 2.3.5.2. Redraw).

Drawing Toolbar: Toggles the display of the Drawing Toolbar on and off.

Back to Data: Sets the focus on data without closing the Graphics Editor (see 2.3.3.6. Exit Graph).

Help: Activates the UNISTAT help system and displays the relevant section for the current procedure.

Reset Coordinates: Restores the default coordinates of the frame, plot area, legend and all text objects (see 2.3.5.3. Reset Coordinates).

Zoom: Enlarges a part of the plot area (see 2.3.4.1. Zoom / Unzoom). This option is not available for all graph types.

Unzoom: Restores zoom. This is only visible after zoom has been used (see 2.3.4.1. Zoom / Unzoom).

Print: Prints the graph on display on the default printer without opening the Print dialogue (see 2.3.3.4. Print).

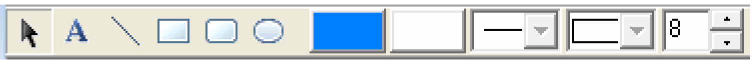
3D Effect: This is used for line, bar and pie charts to create a depth effect.

Outline: Polygons in line, bar and pie charts can be drawn with or without outline.

2.3.2. On-Screen Editing

Graphics Editor allows for full on-screen object editing of graphs. All text, legends and the plot area can be drag-dropped and resized and new text, line and shape objects added.

2.3.2.1. Drawing Toolbar



When it is shipped, the Graphics Editor displays a Drawing Toolbar. This can be hidden using the View menu or by clicking on the second icon of the Graphics Toolbar. The buttons on this toolbar are used to add new text, line, rectangle, rounded rectangle, ellipse and circle objects. Other controls are used to change aspects of objects like border colour, fill colour, border style fill style and border thickness. These buttons have the following tasks:

↖ **Select Mode:** When this button is depressed, the mouse pointer can be used to select objects either for drag-dropping or other further processing. For example you need to select an object before being able to change its border or background colour or border thickness.

A **Insert Text:** Inserts a new text object (see 2.3.2.2.4. Text Objects).

╲ **Draw Line:** Draws a line object (see 2.3.2.2.5. Line Objects).

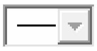
□ **Draw Rectangle:** Draws a rectangle object (see 2.3.2.2.6. Shape Objects).

▢ **Draw Rounded Rectangle:** Draws a rounded rectangle object (see 2.3.2.2.6. Shape Objects).

◯ **Draw Ellipse:** Draws an ellipse or circle object (see 2.3.2.2.6. Shape Objects).

🟦 **Border Colour:** Selects the border colour of the currently selected legend, line or shape objects.

□ **Fill Colour:** Selects the fill (inside area) colour of the currently selected legend or shape objects.

 **Border Style:** Selects the border style of the currently selected legend, line or shape objects.

Fill Style: Selects the fill (inside area) style of the currently selected legend or shape objects.

Border Thickness: This control permits the selection of the border line thickness of the currently selected legend, line or shape objects. The thickness is in printer units, which is about 1/8 of a pixel on the screen for a laser printer.

2.3.2.2. UNISTAT Graphics Objects

UNISTAT graphs consist of six types of objects; frame, plot area, legend, text, line and shape objects. The plot area, legend and some text objects are drawn by UNISTAT but additional text, line and shape objects can be added by the user. All objects, except for the plot area object, can be deleted by pressing <Delete> when the object is selected.

When an object is selected, eight small blocks will appear in the object's bounding rectangle. These blocks are called *handles*. The objects can be resized by dragging these handles (except for text objects).

2.3.2.2.1. Frame Object

By default, the frame object will be drawn on the borders of the Graphics Editor window. This object has the lowest priority (i.e. it will always lie at the background). To select it, click on an exposed region in the graphics area which does not belong to any other object. Once the frame has been selected, you may change its line or fill colour, line style, etc. The fill colour of the frame object is the background colour of the entire graph area.

2.3.2.2.2. Plot Area Object

The entire plot area, including the axis numbers, is treated as one single object. You can select this object by clicking anywhere on the plot area.

The X-Y Plots type differs from other plot types in that its plot area object contains additional vertical lines for up to four right Y-axes. These lines have only one handle allowing a right Y-axis to be moved horizontally, confined by the neighbouring axes.

Unlike other objects, which are redrawn independently, any change in the plot area object will cause the entire window to be redrawn.

2.3.2.2.3. Legend Object

All information displayed on legends is contained within a single Legend Object. This can be drag-dropped and resized like an ordinary rectangle object. However, whenever a Legend Object is resized, its contents will be rearranged so that they make best use of the new shape of the legend.

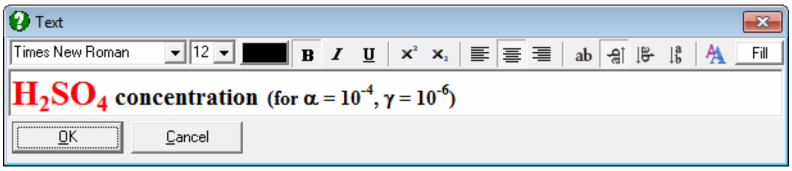
In general, the text displayed in legends can be edited in custom dialogue boxes specific to each procedure. For instance the X-Y Plots Edit → Data Series dialogue, the Plot of 2D Functions Edit → Functions dialogue and the Plot of Distribution Functions Edit → Distributions dialogue will facilitate the editing of the text for each legend item.

Other aspects of legends (like the number of items per line, legend on / off and font name, colour, size and style of the text displayed) can be controlled from a Legend dialogue by either double-clicking on the legend object or by selecting Edit → Options → Legend from the menu.

2.3.2.2.4. Text Objects

The main title, sub title and axis titles are standard text objects. Additional text objects can be inserted using the [Text] button on the Drawing Toolbar. Click this button first to insert a new text object. The mouse pointer will change into an I-beam pointer. Then click on the spot where the text is to be inserted. This will open the text editing dialogue. Enter the text and click [OK] when finished.

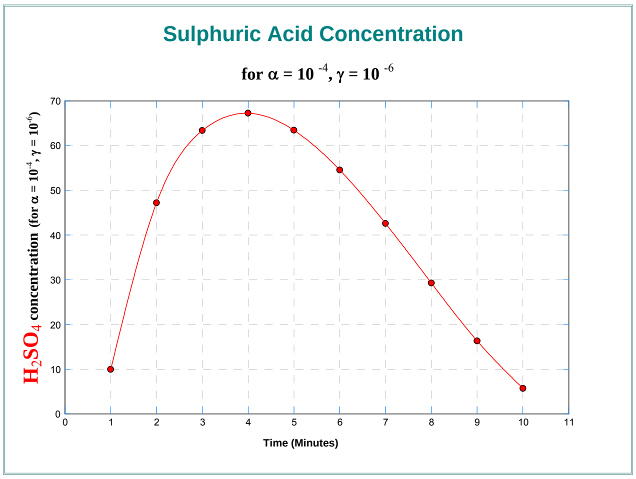
All text objects can be drag-dropped anywhere in the window and edited by double-clicking on them.



Times New Roman 12

H_2SO_4 concentration (for $\alpha = 10^{-4}$, $\gamma = 10^{-6}$)

UNISTAT text objects support Rich Text format. This means that it is possible to mix different fonts (including symbols) and subscripts and superscripts in any text object. It is also possible to copy and paste formatted text between UNISTAT text objects and Word and other Windows applications.



Buttons on the text editing dialogue's toolbar allow editing all aspects of text objects such as font name, size and colour, bold, italic, underline, subscript, superscript, justify left centre or right, horizontal, vertical up, vertical down, top-to-bottom orientations, transparent / opaque background and the background fill colour when the background is not transparent.

The main, sub and axis titles are justified with respect to the axes they represent. The default location is `Centre`. The user-inserted text objects do not have this option. All text objects can be aligned in one of 0°, 90° or 270° rotations or top-to-bottom orientation. For best results only *True Type* fonts should be used. Text rotation is only supported for *True Type* fonts.

It is possible to change the font size for all text that appear on the graph proportionately. This includes text displayed on axes, legend, as well as all text objects. The default size is 125%. To change this value enter and edit the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] section:

```
FontLevelPct=125
```

2.3.2.2.5. Line Objects

Click on the [Draw Line] button on the Drawing Toolbar to draw a line object. The mouse pointer will change from an arrow into a cross-hair. Now depress the left mouse button on the spot for one of the ends of the line and drag the pointer to the point where the line should end and then release the mouse button. The selected line objects will have two handles (small black squares) displayed on either end. These handles can be subsequently drag-and-dropped to change the position and length of the line. An unselected line can be re-selected at any time by clicking anywhere along its length.

To delete a line object simply press <Delete> when the line is selected. The colour, style and thickness of a selected line object can be determined using controls provided on the Drawing Toolbar.

2.3.2.2.6. Shape Objects

The shape objects provided are rectangle, rounded rectangle and ellipse, all of which are drawn and edited in almost exactly the same way as a line object described in the previous section. When one of these buttons on the Drawing Toolbar is clicked, the mouse pointer will change from an arrow into a cross-hair. Press the left mouse button down where the top left (or bottom right) corner of the shape is to be placed and drag the pointer to the bottom right (or top left) corner and then release the mouse button.

Although these three shapes are all different, they have the same *drag outline*, which is a rectangle. When selected, these shape objects will have eight handles (small black squares) displayed on their corners and in the middle of their edges. Dragging and dropping these handles changes the position, size and aspect ratio of objects. An unselected shape object can be re-selected at any time by clicking anywhere within its bounding rectangle. To delete a shape object simply press <Delete> whilst that object is selected.

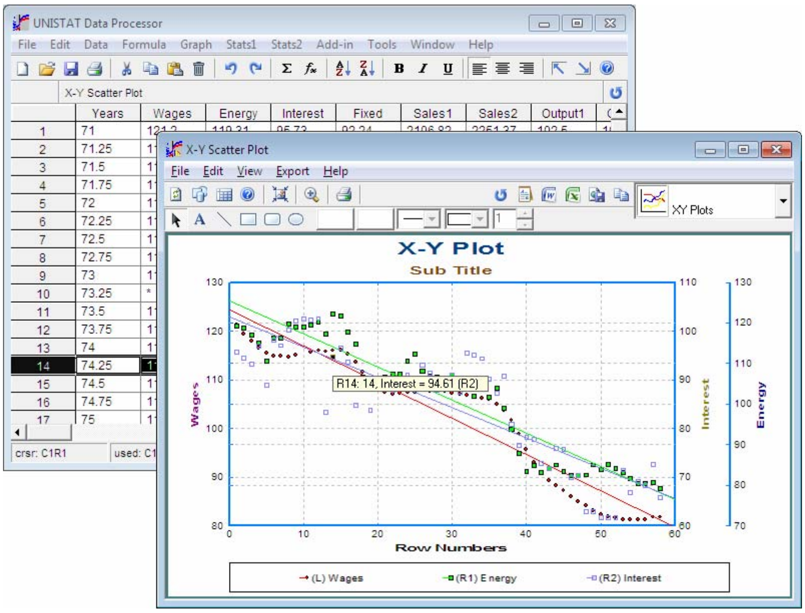
The border colour, fill colour, border style, fill style and border thickness of shape objects can be changed using the controls provided on the Drawing Toolbar.

2.3.2.3. Interactive Data Points

In the following graphics procedures, the data points displayed in Graphics Editor maintain a link with the data matrix.

Graph → 2D Plots →
X-Y Plots
Polar Plot
Graph → 3D Plots →
X-Y-Z Scatter Plot
Spin Plot
Graph → Descriptive Plots →
Normal Probability Plot

Statistics 1 → Regression Analysis →
Linear Regression
Polynomial Regression
Bioassay →
Parallel Line Method
Slope Ratio Method
Quantal Response Method



	Years	Wages	Energy	Interest	Fixed	Sales1	Sales2	Output1
1	71	121.2	110.31	95.73	92.24	2106.82	2251.37	102.5
2	71.25	1						
3	71.5	1						
4	71.75	1						
5	72	1						
6	72.25	1						
7	72.5	1						
8	72.75	1						
9	73	1						
10	73.25	*						
11	73.5	1						
12	73.75	1						
13	74	1						
14	74.25	1						
15	74.5	1						
16	74.75	1						
17	75	1						

The three regression plots where this facility is available are Plot of Actual and Fitted Values, Plot of Residuals and Normal Plot of Residuals (see 7.2.1.2. Linear Regression Output Options).

In these procedures, you can press the right mouse button on a data point and highlight it. If you do not release the mouse button for a little while, a panel displaying information about this particular point will pop up. When the mouse button is released, the highlights are switched off. This procedure is also known as *Brushing* or *Point identification*.

In Stand-Alone Mode, if the Data Processor window is exposed, you will also see that the row of the spreadsheet containing the point becomes highlighted. Conversely, it is possible to click on a row of the spreadsheet to highlight the points on the graph which belong to this row.

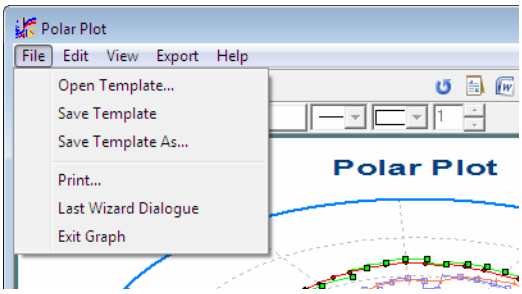
A useful feature here is the possibility of using <Delete> to exclude the highlighted case (row) from the graph. The program will first generate a Select Row column indicating which rows will be included in the graph and then the entire graph will be redrawn without this particular row. After exiting the graphics procedure, all other UNISTAT procedures will continue to exclude such points from the analysis. If you wish to enable all rows of the data matrix again, proceed as follows:

1) Stand-Alone Mode**:** Disable the Select Row column.
2) Excel Add-In Mode**:** Highlight a different block of data.

This feature is particularly useful in Regression Analysis or in X-Y Plots with a fitted regression line and confidence intervals. What is called *interactive outlier rejection* can easily be performed by omitting the cases which lie outside the desired confidence intervals.

The link between data points and the data matrix will no longer be available once the image is sent to another application (i.e. Output Window, Excel, Word, web browser).

2.3.3. File Menu



2.3.3.1. Open Template

When UNISTAT is first loaded, it will open and read the graphics template file *Graph6.usg* which holds the default settings for all graphics procedures. The information contained in this file includes all settings and text that can be edited from the Graphics Editor and its menu system.

Previously saved graphics template files can be opened by selecting File → Open Template. The default file extension for UNISTAT graphics template files is .USG. For more information on graphics template files see 2.3.3.3. Save Template As.

In order to save graphics images in a format readable by other applications, use the Export Menu.

2.3.3.2. Save Template

The program will save the current graph settings to the current graphics template file. If the file name has not been changed, then the settings will be saved to the default template file *Graph6.usg* and will be opened automatically when you next start a UNISTAT session. For more information on .USG files see the next section.

2.3.3.3. Save Template As

The graphics template files are used to store all information contained in any graph. This includes all parameters that can be edited from the pull-down menus, as well as all objects (text, line and shapes) and their edited positions. The user is responsible for maintaining the compatibility between graphics template files and their corresponding data files.

When this option is selected, a dialogue will be opened prompting for a file name. The default file extension for UNISTAT graphics template files is .USG.

If you would like UNISTAT to retrieve your own choices automatically next time you use UNISTAT, then you must save your graphics template as *Graph6.usg* file, which can be found in *Documents\Unistat65\Work* folder. Should you wish to restore the original settings, simply delete this file. The program will retrieve the original *Graph6.usg* file from the UNISTAT installation folder.

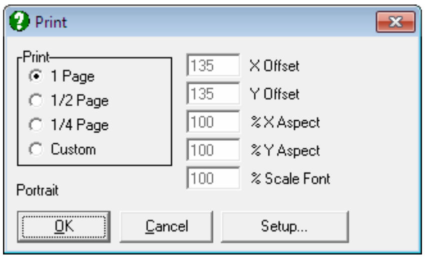
WARNING! *.USG files are not graphics files. They can only be read by UNISTAT and contain information about graphics options set within UNISTAT.*

In order to save graph images in a format readable by other applications, use one of the options under the Export Menu.

When a UNISTAT graphics template file is saved, all aspects of the graph are saved in the file. However, the data for the graph is *not* saved as part of the .USG file. Therefore, when a .USG file is saved, it is necessary to ensure that the data is also saved in an accompanying data file, preferably in .USW format. Only when the correct data file is opened can you reproduce the same graph by opening its own .USG file. Although this may seem an inconvenience which is not present in business graphics packages, this is so only because other drawing packages do not support data files containing as many as 1,000,000,000 data points.

2.3.3.4. Print

A dialogue is displayed with three pre-set size options, Full page, Half Page and Quarter Page. A fourth option Custom is provided for customised sizes. Select the desired option and click on [OK] to print the graph. The default Windows printer settings are used.



Print

Print
● 1 Page
○ 1/2 Page
○ 1/4 Page
○ Custom

Portrait

135 X Offset
135 Y Offset
100 % X Aspect
100 % Y Aspect
100 % Scale Font

The Custom option will display five text fields, two for the top-left corner, the x-aspect and y-aspect ratios and the font scale factor. These parameters can be edited to obtain the desired location and aspect ratio for the printed graph, which are in printer pixel units. They are saved as part of graphics template files (see 2.3.3.3. Save Template As).

Printer settings can be changed by clicking on the [Setup] button. This will provide access to standard Windows Printer Setup dialogues allowing the control of number of copies printed, resolution, portrait or landscape orientations, etc. If there is one, you are recommended to check the Print True Type as graphics option, since some printers cannot handle properly the rotated *True Type* fonts.

UNISTAT graphs can also be exported to other applications in one of bitmap (.PNG) or enhanced metafile (.EMF) formats (see 2.3.6. Export Menu.), which are preferable to printing a graph in a file. Alternatively use can be made of a facility provided by Windows to copy the active window to the clipboard in bitmap format. To do this, it is recommended that you first maximise UNISTAT Graphics Editor window to obtain the highest resolution. Then press <Alt> + <Print Screen> to copy the screen to clipboard. Then activate the other application and paste the image by pressing <Shift> + <Insert>.

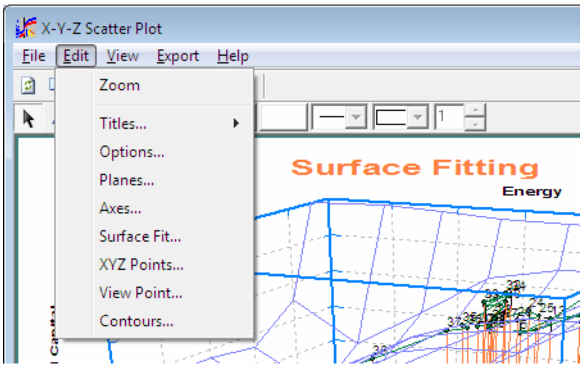
2.3.3.5. Last Procedure Dialogue

To return to the last procedure dialogue select File → Last Procedure Dialogue. Although most of the changes made in graphics settings will be retained in memory, some settings - like scaling of axes and legend text - may be lost when you go back to the previous dialogue. Therefore, if you wish to preserve *all* graphics settings, you must save a graphics template file before exiting (see 2.3.3.3. Save Template As).

2.3.3.6. Exit Graph

This option will only push the Graphics Editor to the background and is equivalent to clicking the exit button on the toolbar. If changes are made to data, or another procedure is selected from the menus, then the Graphics Editor will be closed first automatically.

2.3.4. Edit Menu



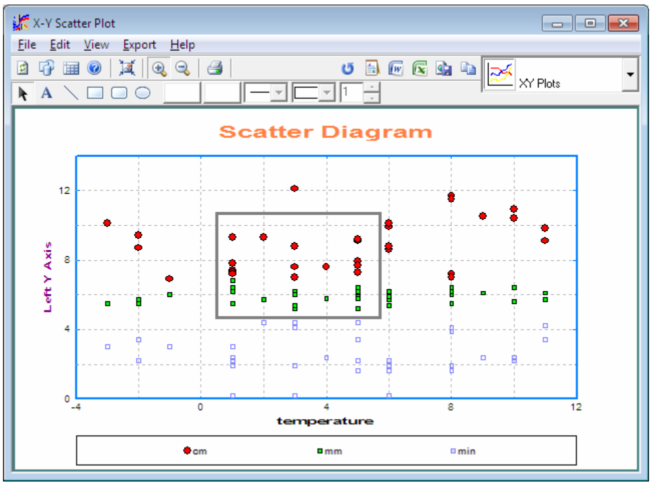
This menu provides access to dialogues used in controlling the specific features of each graphics procedure. Therefore, the menu items displayed under Edit will be different for different procedures. The two exceptions to this are the Titles and Options items, which are common to all graph types.

The following are the more commonly used Edit Menu options.

2.3.4.1. Zoom / Unzoom

Selecting this option is equivalent to clicking on the zoom button (with the picture of a magnifying glass showing the + sign). Zoom is used to display a smaller section of the graph in the plot area.

After selecting View → Zoom the mouse pointer will change into a cross-hair. Depress the left mouse button at the top left corner of the region to be zoomed (do not release the button yet), drag a box to cover the zoom region and then release the mouse button. A full size graph of the selected area will be drawn. After zoom has been used, a new button will appear to the right of the zoom button with a magnifying glass picture showing the - sign. Click on this button to restore the original graph.



The zoom facility is not available for all graphics procedures.

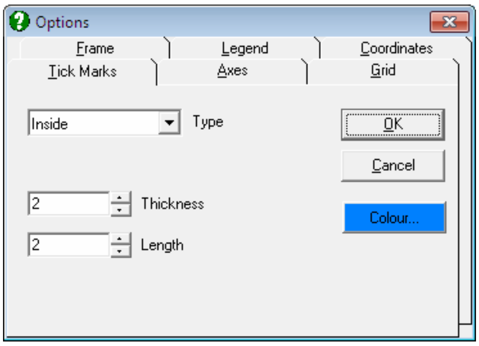
2.3.4.2. Titles

Two standard text objects, Main Title and Sub Title can be entered, edited and their properties (like font, size, orientation, etc.) can be set. If either of these two have been deleted previously, they can be reinstated by simply entering new text in the text object dialogue (see 2.3.2.2.4. Text Objects). Other standard text objects can be edited from their own dialogues, e.g. axis titles from the Edit → Axes dialogue.

2.3.4.3. Options

Common aspects of all graph types, such as style, thickness and colour of frame, axes, tick marks can be controlled using this dialogue.

2.3.4.3.1. Tick Marks



Inside Type

2 Thickness

2 Length

This dialogue is used to control the type, colour, thickness and length of the tick marks along the axes. The frequency of tick marks with or without labels can be controlled by selecting the appropriate interval values from the Edit → Axes dialogue. The controls on this dialogue have the following tasks:

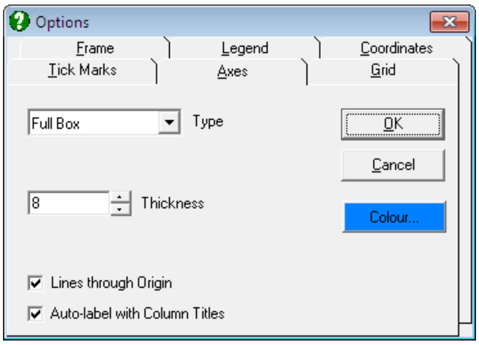
Type:

None: Tick marks are not drawn.
Outside: Tick marks are drawn outside the plot area.
Inside: Tick marks are drawn inside the plot area.
Both: Tick marks are drawn both outside and inside the plot area.

Thickness: Adjusts the thickness of the tick marks.

Length: Adjusts the length of the tick marks.

2.3.4.3.2. Axes



Full Box Type

8 Thickness

☑ Lines through Origin

☑ Auto-label with Column Titles

This tab controls the type, colour and thickness of the axis lines. To control the font and style of numbers printed alongside the tick marks and axis titles, a different dialogue, Edit → Axes is used.

Type:

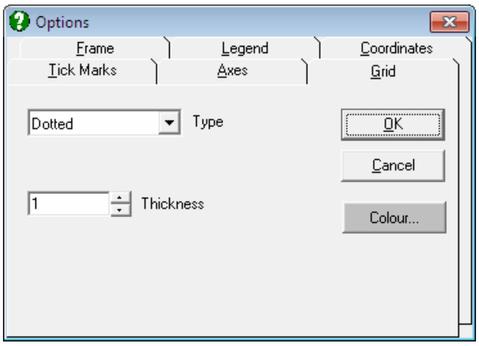
None: Axes are not drawn.
Scales only: Only the primary axes with scales are drawn.
Full box: All axes are drawn.
Open box: Axes obscuring the graph area are not drawn (for 3D graphs only).

Thickness: Adjusts the thickness of the axis lines.

Lines Through Origin: If this box is checked and if the origin is in the plot area, then an axis line will be drawn passing through the origin.

Auto-label with Column Titles: When this box is checked UNISTAT will automatically assign the column label of the variable selected for this axis as the axis title. If this box is unchecked, the axis titles entered by the user or those loaded from .USG files will remain unchanged for all subsequent graphs until manually edited or another .USG file is selected. When there is more than one variable assigned to the same axis (e.g. in X-Y Plots, Bar Chart, etc.) then the axis title will not be changed by the program even when this box is checked. A similar control is also available for the legend text (see 2.3.2.2.3. Legend Object).

2.3.4.3.3. Grid



Type: Dotted

Thickness: 1

This dialogue is used to control the type, colour and thickness of the grid lines. If any, the grid lines are drawn through the tick marks. Therefore, their frequency can be controlled by selecting the appropriate interval values from the Edit → Axes dialogue.

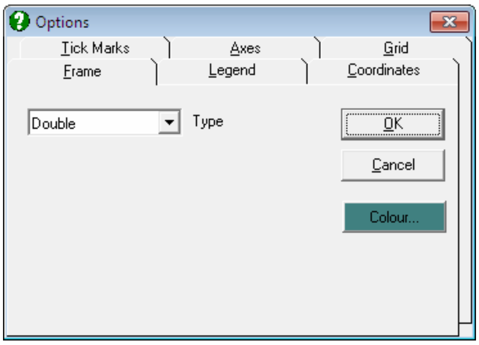
Type:

None: Grid is not drawn.
Solid: A solid line is drawn through the tick marks.
Dotted: A dotted line is drawn through the tick marks.
Intersections: A dot is plotted at the intersection of two grid lines. This will create the best results with plotters, where the pen thickness would cause over-emphasised grid lines with any other option.

Thickness: Adjusts the thickness of the grid lines.

If the Axis option above is set to None, then the grid will not be drawn either.

2.3.4.3.4. Frame



Double
Type

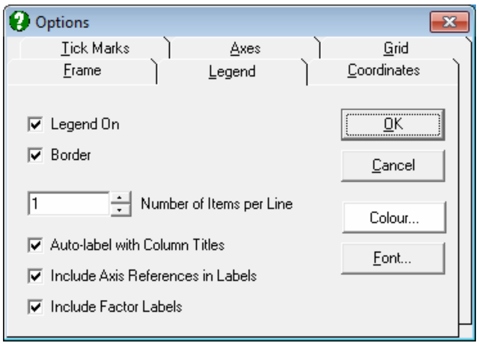
Use this tab to select the type (single or double) or colour of the frame or to redisplay the deleted frame object.

Type:

None: No frame is drawn.
Single: A single line frame is drawn.
Double: A double-line frame is drawn.

The colour, line style, thickness and other properties of the frame object can be controlled using the Drawing Toolbar.

2.3.4.3.5. Legend



☑ Legend On

☑ Border

1 Number of Items per Line

☑ Auto-label with Column Titles

☑ Include Axis References in Labels

☑ Include Factor Labels

This dialogue can also be activated by double-clicking on the Legend Object.

Colour: The fill colour of the legend background can be selected.

Font: Font, colour, style and size of all text in legends can be edited via the standard Windows font selection dialogue.

Legend On: Switches the display of legend object on and off. If the legend object has been deleted, then it can be made visible again by checking this box (see 2.3.2.2.3. Legend Object).

Border: When checked, border lines (a box) will be drawn for the legend object.

Number of Items Per Line: Although the number of items per line is automatically adjusted according to the shape of the legend object, here you can enforce your own choice.

Auto-label with Column Titles: When this is checked UNISTAT will automatically assign the Column Labels (as displayed in Data Processor) to the legend items. Otherwise the legends typed in by the user from the Edit → Data Series dialogue or those loaded from the .USG files will not be changed by the program.

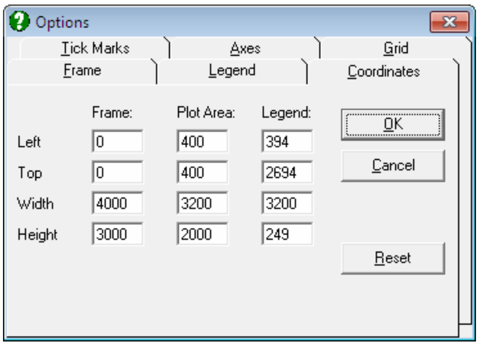
In most cases, it is desirable and also practical to represent Data Series by their column (variable) labels. UNISTAT uses the Column Labels in legend fields by default. However, sometimes the user needs to type in additional or completely new information which is not in the column label. In this case, although the graph can be plotted with the edited legend fields and this information could be saved to a .USG file, every time the user goes back to the Variable Selection Dialogue and then re-displays the graph, the edited legend fields would be replaced by Column Labels. Therefore, we recommend that this box should be unchecked if the legend fields are to contain information other than Column Labels. A similar control exists for axis titles (see 2.3.4.3.2. Axes).

Include Axis References in Labels: When a graph is drawn with more than one Y-axis, it is important to know which variable is represented on which axis. If this box and the Auto-label box above are checked simultaneously, then the program will prefix the legend texts with their axis references. For instance, if the *Wages* variable is represented on the left Y-axis and *Interest* on the third right Y-axis, then the legend will be automatically labelled with *(L) Wages* and *(R3) Interest*.

Include Factor Labels: When factor columns have been selected, you can display the factor names in the legend. See 4.1.1. X-Y Plots Categorical Plot.

The fill colour, fill style and border thickness properties of the legend object can be controlled by first selecting the legend with the mouse and then using the Drawing Toolbar.

2.3.4.3.6. Coordinates



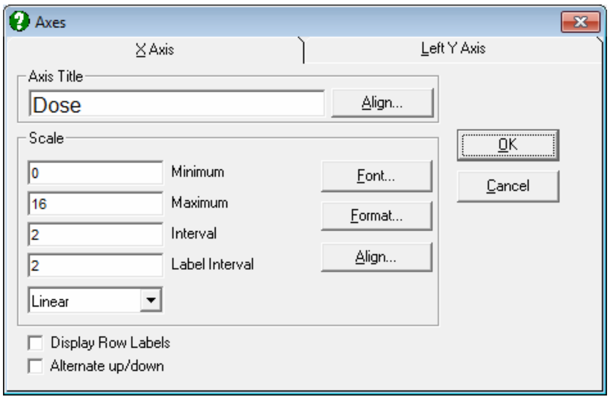
	Frame:	Plot Area:	Legend:
Left	0	400	394
Top	0	400	2694
Width	4000	3200	3200
Height	3000	2000	249

The size and location of three major graphics objects can be edited: frame object, plot area object and legend object. The logical origin is at the top left corner of the Graphics Editor and the logical size is 4000 units wide and 3000 units high.

The values displayed in this dialogue will always be the current values of these objects. To restore the original values click on the [Reset] button. These values are stored as part of the graphics template files (see 2.3.3.3. Save Template As).

The size and position of the printed image can also be customised by editing the top-left corner and the X-aspect and Y-aspect ratios given in the File → Print dialogue, without having to change any of these coordinate parameters.

2.3.4.4. Axes



Axis Title: Dose

Scale

Minimum: 0

Maximum: 16

Interval: 2

Label Interval: 2

Linear

□ Display Row Labels

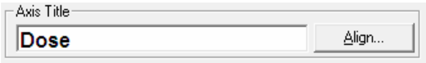
□ Alternate up/down

The minimum, maximum, interval and label interval values can be edited and font, size, colour, style and the number of digits displayed for the scale numbers can be set. For each axis, the Scale Type can be one of linear, log base 10, log base e, log based to any user-defined value (the default is 2), reciprocal, logit, probit, gompit (cloglog) or loglog independent of other axes.

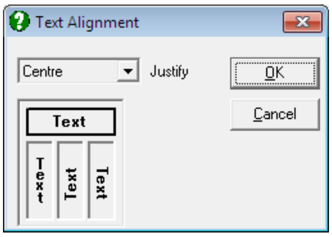
Mathematical expressions (as well as numbers) can be typed into the four text boxes for axis range and intervals. For instance, to plot a trigonometric function between *-3pi* and *pi*, *-3 * Pi()* can be entered in the Minimum field and *Pi()* in the Maximum field. Likewise, for a natural logarithmic axis, one can enter *-e()* and *e()^3*. For the syntax and rules of mathematical expressions see 3.4.0.1. Entering Formulas. Once a function is entered, the program will replace the formula with its numeric result. The formula itself will not be stored.

2.3.4.4.1. Axis Title

Editing axis title: This is a Rich Text edit box for entering and editing the selected axis title. Axis titles are standard text objects (see 2.3.2.2.4. Text Objects) and they can be copied and pasted to and from other Windows applications such as Word. They can also be edited using UNISTAT's built-in Rich Text editor. To do this you need to close the axis dialogue and double-click on the title text object. Like other text objects, axis titles can be aligned in one of 0°, 90° or 270° rotations or top-to-bottom orientation. It is possible to mix different fonts (including symbols), subscripts and superscripts.


Axis Title: Dose

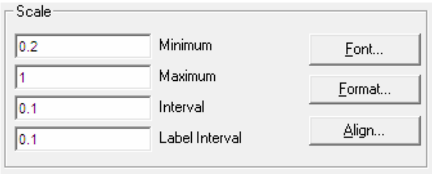
Align: A dialogue pops up allowing control of justification and orientation of the axis title.



Justify: Centre

2.3.4.4.2. Scale

The following parameters can be controlled.


Scale

Minimum: 0.2
Maximum: 1
Interval: 0.1
Label Interval: 0.1

Minimum: This is the smallest value of the variable to be displayed on the axis. By default, this value will be a round figure smaller than the actual minimum of data, in order that all points lie within the plot area. A label tick mark will not necessarily be drawn for this minimum value if it is not a round enough figure. Mathematical expressions can be entered into this field, e.g. *2*Pi(), e()^2*.

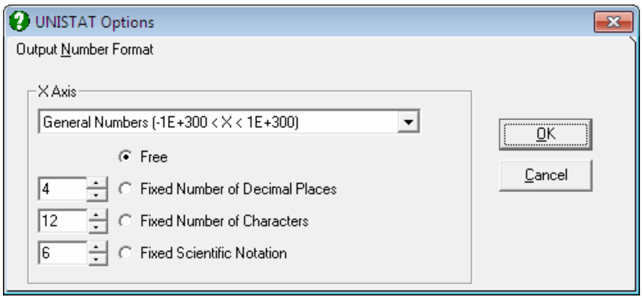
Maximum: This is the largest value of the variable to be displayed on the axis. By default, this value will be a round figure greater than the actual maximum of data, in order that all points lie within the plot area. A label tick mark will not necessarily be drawn for this maximum value, if it is not a round enough figure. Mathematical expressions can be entered into this field.

Interval: This is the interval at which minor tick marks (that is, tick marks without labels) are drawn. Mathematical expressions can be entered into this box. It is available for only linear and reciprocal scales. See Scale Type below.

Label Interval: This is the interval at which major tick marks (that is, tick marks with labels) are drawn. Major tick marks have a longer tick, a grid line and a label. Minor tick marks have a shorter tick and a grid line only. Mathematical expressions can be entered into this box. It is available for only linear and reciprocal scales. See Scale Type below.

Font: Fonts can be set for axis numbers from the standard Windows font selection dialogue.

Format: This will activate the Number Format dialogue allowing numeric formatting of scale numbers.


UNISTAT Options

Output Number Format

X Axis

General Numbers (-1E+300 < X < 1E+300)

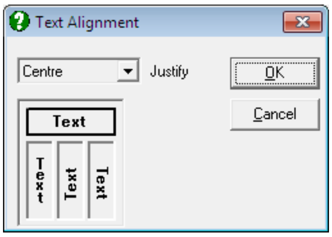
● Free

4 ○ Fixed Number of Decimal Places

12 ○ Fixed Number of Characters

6 ○ Fixed Scientific Notation

Align: This will activate the text orientation dialogue allowing you to display the numbers or labels for X-axis major ticks in one of 0°, 90° or 270° rotations or top-to-bottom orientation. This option is available for only X-axis.


Text Alignment

Justify: Centre

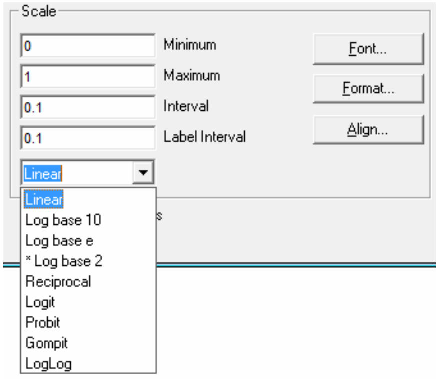
Display Row Labels: This control is available for only the X-axis. Most graph types will allow displaying alphanumeric labels for the X-axis tick marks, instead of numbers. If this box is checked, then the Row Labels will be displayed on the X-axis. In Stand-Alone Mode, these can be entered and / or edited in Data Processor. In Excel Add-In Mode, the row labels should be selected as the first column in the highlighted block of data.

The use of Row Labels in graphics requires more effort than plotting with numbers. Particularly, the X-axis interval must be chosen carefully to prevent labels overlapping.

Alternate Up/Down: This control is available for only the X-axis. It allows for printing major tick labels up and down along the X-axis, preventing overlapping.

2.3.4.4.3. Scale Type

In X-Y Plots (in fact, in almost any other 2D or 3D plot), it is possible to change the scale of an axis to one of linear, log base 10, log base e, log based to any user-defined value (the default is 2), reciprocal, logit, probit, gompit (cloglog) or loglog axes.



Scale

Minimum: 0

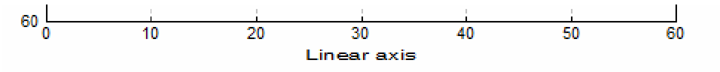
Maximum: 1

Interval: 0.1

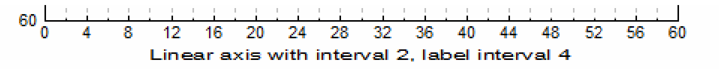
Label Interval: 0.1

Linear

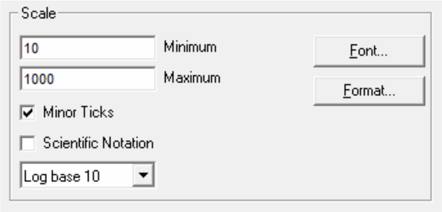
Linear: This is the default scale type. Automatic scaling is done such that a round figure which is smaller than or equal to the minimum observation and another round figure which is greater than or equal to the maximum observation in data are selected as the minimum and maximum scale numbers respectively.



Interval and Label Interval values can be edited to control the appearance of linear scale.



Log Base 10: When this option is selected, the program first scans the data for non positive values. If any negative or zero values are found, then an Illegal Interval message pops up and the selection is aborted. Otherwise, the axis is rescaled for logarithmic values (base 10) and the new (logarithmic) minimum and maximum values are displayed in the corresponding boxes. Instead of Interval and Label Interval text boxes used in linear and reciprocal scale dialogues, two check boxes Minor Ticks and Scientific Notation are featured in all logarithmic scale dialogues.


Scale

Minimum: 10

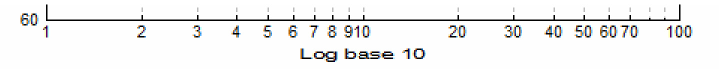
Maximum: 1000

☑ Minor Ticks

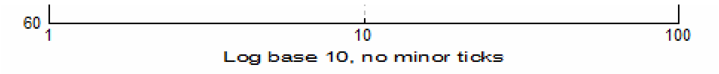
□ Scientific Notation

Log base 10

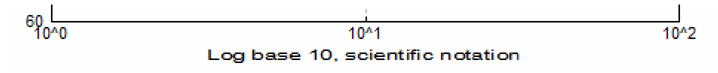
- By default, i.e. when Minor Ticks is checked and Scientific Notation is unchecked, minor tick marks are drawn between the powers of 10 and numbers are displayed in free format, as shown below. Between the powers of 10 numbers are displayed only if they do not overlap. Note that numbers 80 and 90 are not drawn below because there is not enough room for them.



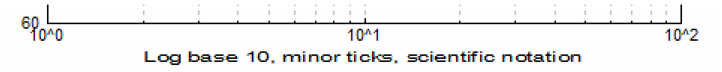
- If Minor Ticks is unchecked then minor tick marks and numbers between the powers of 10 are not displayed.



- If Scientific Notation is checked then numbers are displayed in power form.



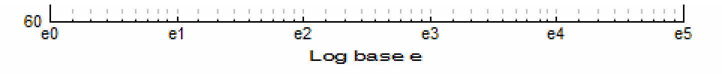
- When Minor Ticks and Scientific Notation are both checked, minor tick marks between powers of 10 are drawn but their numbers are not displayed.



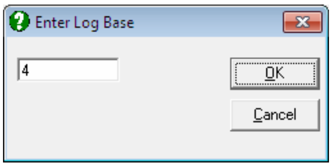
Under certain conditions, the minimum and / or maximum values may not be displayed on the scale. To force these values to appear under all circumstances, enter the following line under the [Options] section of the *Documents**Unistat65**Unistat65.ini* file:

```
ForceFirstLastTick=1
```

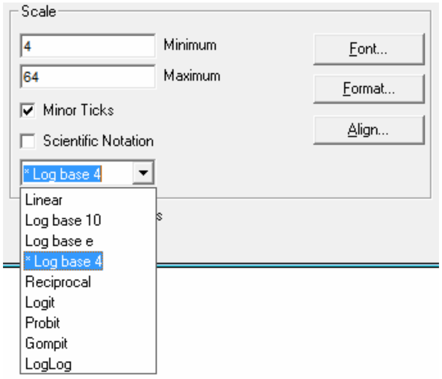
Log Base e: An axis with a natural (e based) logarithm is similar to the one with Log base 10. The only difference is that in a log base e scale numbers are displayed for only major ticks. The e constant can be entered in minimum and maximum boxes as a mathematical expression, e.g. *e()^5* for the fifth power of e.



*** Log Base 2:** Although the default base number is 2, this can be changed to any integer greater than 1. When this item is selected from the list, a dialogue pops up enabling you to enter a new base number:



The new base number will be displayed in the drop-down list, preceded by an asterisk, indicating that this is a user-defined base number.



Scale

4 Minimum

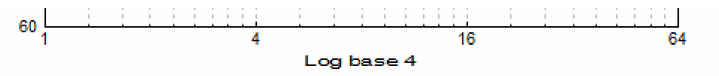
64 Maximum

☑ Minor Ticks

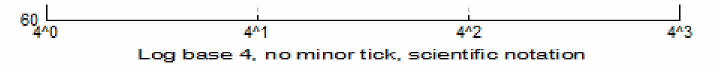
□ Scientific Notation

Log base 4

- Minor Ticks without Scientific Notation.

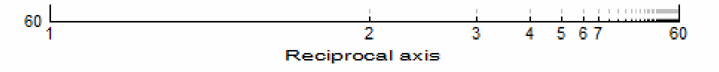


- No Minor Ticks with Scientific Notation.



Reciprocal: Variables containing strictly positive values are transformed using the reciprocal function:

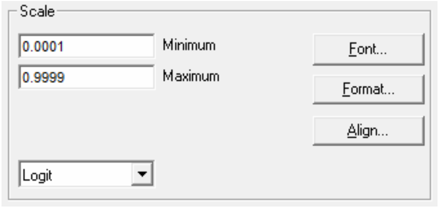
$$F(x) = 1 / x$$



As in linear axis, Interval and Label Interval text boxes can be used to control the appearance of axis scale.

Logit: The next four scaling options, logit, probit, gompit (cloglog) and loglog, can be used with any data series containing values within the range $0 \le p \le 1$. If the data contains values outside this range, an Illegal Interval message is issued and the selection is disallowed. However, as these functions are not defined for $p = 0$ and $p = 1$, the actual data range is limited to the interval $0.0001 \le p \le 0.9999$. Values outside this interval (but within the range

0 ≤ p ≤ 1) are considered missing. Logit, probit, gompit (cloglog) and loglog scale options have neither Interval, Label Interval text boxes nor Minor Ticks, Scientific Notation check boxes (with the exception of gompit (cloglog) and loglog, which have the Minor Ticks box).



Scale

0.0001 Minimum

0.9999 Maximum

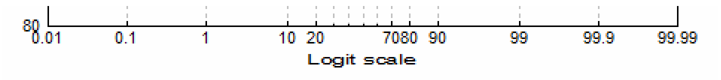
Logit

The logit function is a symmetric odds ratio for a given probability:

Logit(p) = Ln(p/(1-p))

Examples:

Logit(0.025) = -3.66
Logit(0.95) = 2.94.

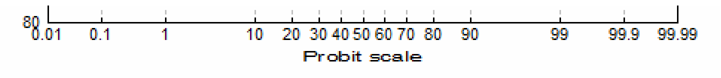


Probit: For data range see logit. Probit is the inverse standard cumulative normal distribution function for a given probability value.

Probit(p) = Φ^{-1}(p)

Examples:

Probit(0.025) = -1.96
Probit(0.95) = 1.64.



Note that a plot with a probit axis is different from a Normal Probability Plot.

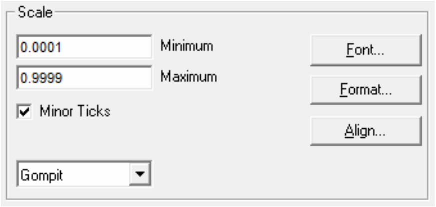
Gompit: For data range see logit. Gompit (which takes its name from the Gompertz distribution) is an extreme value function related to Weibull distribution. It is also known as the log-Weibull or the complementary log log (cloglog) function (see 7.2.5.1. Logit / Probit / Gompit Model Description). Unlike logit and probit, gompit is an asymmetric function with a long right tail.

Gompit(p) = Ln(-Ln(1-p))

Examples:

Gompit(0.1) = -2.25
Gompit(0.9) = 0.834.

An X-Y plot with a gompit Y-axis and log base 10 X-axis is known as Weibull Chart and a printed form with these axes is called Weibull paper (or Weibull probability paper).



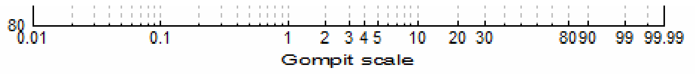
Scale

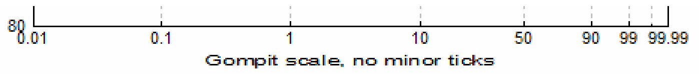
0.0001 Minimum

0.9999 Maximum

☑ Minor Ticks

Gompit



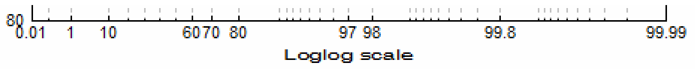


LogLog: For data range see logit. Loglog is an asymmetric function with a long left tail.

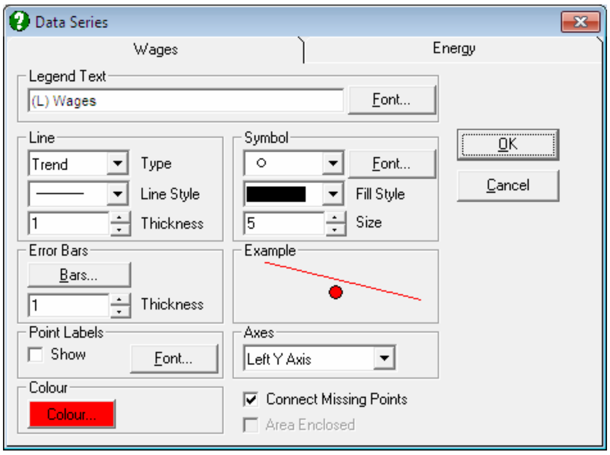
Loglog(p) = -Ln(-Ln(p))

Examples:

Loglog(0.1) = -0.834
Loglog(0.9) = 2.25.



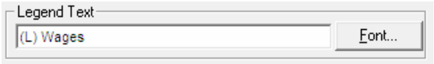
2.3.4.5. Data Series


Legend Text: (L) Wages
Line Type: Trend
Line Style: ────
Line Thickness: 1
Symbol: ○
Fill Style: ■
Size: 5
Error Bars Thickness: 1
Point Labels: ☐ Show
Axes: Left Y Axis
☑ Connect Missing Points
☐ Area Enclosed

Many types of graphs have an option under the Edit Menu for the purpose of editing the graphics properties of the data series plotted. For instance, for X-Y Plots there is an Edit → Data Series option, in Plot of 2D Functions an Edit → Functions option and in Plot of Distribution Functions an Edit → Distributions option. Selecting these options usually produce tabbed dialogues, with each tab corresponding to a different curve being plotted.

These dialogues are specific to the procedures they belong, however most of them have the following groups of controls in common.

2.3.4.5.1. Legend Text

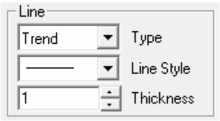

Legend Text: (L) Wages

This text box displays the text which will be displayed in the legend object for a particular curve. In most cases this will be generated by the program automatically using Column Labels (see 2.3.4.3.5. Legend).

In some procedures (like Plot of 2D Functions), however, this text box will be labelled Function and it will be used to enter the formula of the function to plot. In this case the function entered for this curve will also be displayed as legend text.

Font: This button is used to change the font, colour, style and size of the text displayed in the legend. Changing the font for one curve will change all the others, as there is only one font assigned for all curves. The legend font can also be set from the Legend dialogue under Edit → Options → Legend.

2.3.4.5.2. Line


Line

Trend Type

Line Style

1 Thickness

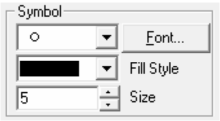
This group of controls is used to edit the properties of curves.

Type: You can select the type of curve to be drawn for each data series. The first option is None which is the default to draw scatter diagrams instead of line diagrams. Other options available are usually specific to each procedure, but will always contain a Straight line option. For X-Y Plots procedure, for instance, there are eight different line types available (see 4.1.1.1.1. Line).

Line Style: If any lines are drawn, their style can be selected via this drop-down list. Available options are solid, dash, dot, dash-dot-dot and dash dot-dot-dot.

Thickness: Thickness of the line can be assigned in printer units, which is about 1/8 of a pixel on the screen for a laser printer. The line style for a line can be only solid if it has a thickness greater than one.

2.3.4.5.3. Symbols


Symbol

○ Font...

Fill Style

5 Size

In most data plotting procedures there will be a Symbol group of controls. This provides access to a large number of symbols (up to a maximum of 260 at a time), with the ability to load any font or symbol set from file. It is possible to adjust the size of all symbols, and fill the interiors of a special set of 7 symbols with hatching patterns.

Type: By default, the drop-down list contains 10 symbols, which are generated by the program:

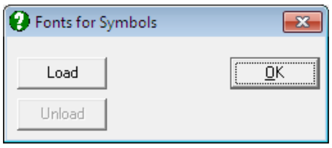
□ None
⊠ Cross
⊞ Plus
⧆ Star
⧇ Circle
⧈ Square
◇ Diamond
△ Up triangle
▽ Down triangle
▷ Right triangle
◁ Left triangle

Fill Style: Symbols 4 to 10, which define closed interiors, can be filled with solid colours or with one of the 8 hatching patterns provided by this drop-down list.

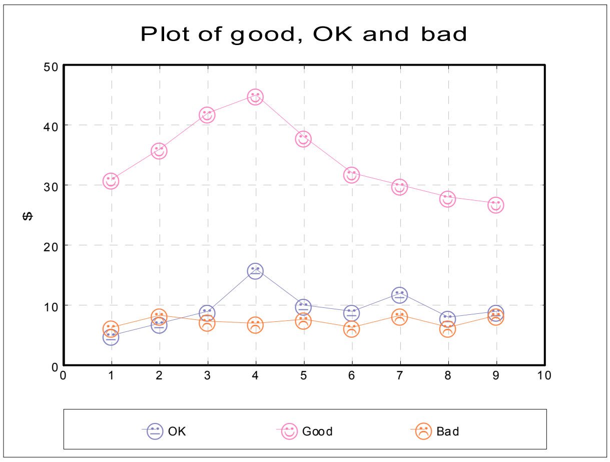
■ Solid
□ Blank
▤ Horizontal
▥ Vertical
▧ Forward diagonal
▨ Backward diagonal
▦ Cross
▩ Diagonal cross

Size: You can click on the up or down arrows to increase or decrease the line thickness of the symbols. The thickness is in printer units, which is about 1/8 of a pixel on the screen for a laser printer.

Font: Clicking on the [Font…] button within the Symbol collection of controls activates a small dialogue where the user can either load a font set into the symbols drop-down list or unload a font set that has already been loaded. In this way letters of the alphabet or other special fonts can be displayed as symbols. Font sets like *Windings* and *Symbol* are designed for this purpose.

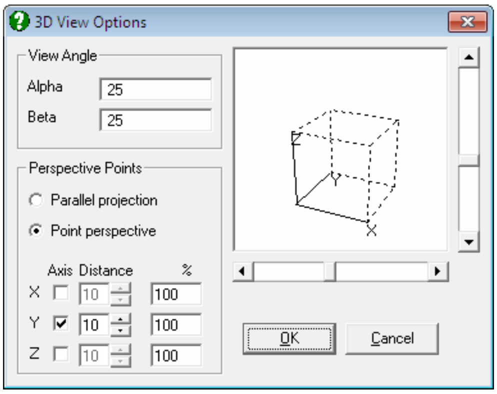


When a font set is loaded into the symbol drop-down list, the first group of symbols will always be the standard UNISTAT set as described earlier. Symbols from the font file will be appended to the end of the list.



2.3.4.6. 3D Viewpoint and Perspective

All 3D graphics procedures have an Edit → Viewpoint options dialogue. This allows a 3D graph to be rotated and viewed from different angles. It also provides the option to view the image in one of parallel, 1-, 2- or 3-point perspectives.


3D View Options

View Angle

Alpha: 25

Beta: 25

Perspective Points

○ Parallel projection

● Point perspective

		Axis Distance	%
X	□	10	100
Y	☑	10	100
Z	□	10	100

Three groups of controls are provided:

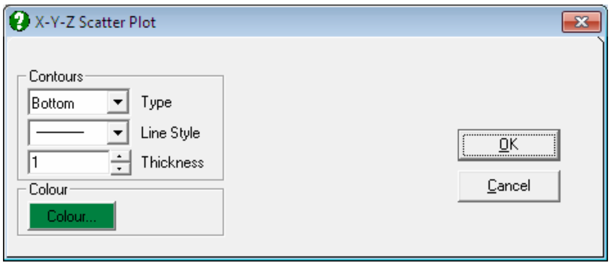
View Angle: Enter degrees of rotation for X and Y axes to select the desired viewpoint. The Z axis will always be perpendicular.

Rotating Cube: This provides a visual representation of the current selections for perspective types and angles. The cube can be rotated by clicking on the vertical and horizontal scroll bars situated next to the cube. As the cube rotates, the alpha and beta angles in the View Angle group will be updated.

Perspective Points: It is possible to select one of parallel or point perspective options. The latter permits selection of 1-, 2- or 3-point perspectives by checking the desired axes. It is also possible to change the distance from the cube to the viewpoint for each axis separately.

2.3.4.7. Contours

3D graphs with a surface display also have an Edit → Contours option. This allows the contour lines of the surface to be shown on either or both of the bottom and top planes of the 3D graph. Alternatively, a 2D contour map can be drawn with extended annotation possibilities.



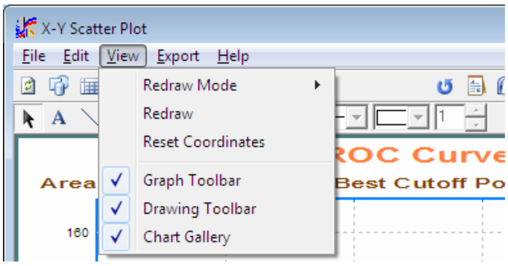
Type: Bottom
Line Style: ———
Thickness: 1

In X-Y-Z Scatter Plot procedure the Edit → Contours menu item will be disabled unless a surface is fitted on the data (see 4.2.1.5. Surface Fitting). Once a surface is fitted, selecting this menu item displays the Contours dialogue box.

The Type drop-down list provides the following choices:

None: Contours are not drawn.
Bottom: Contours are drawn on the bottom plane only (the default).
Top: Contours are drawn on the top plane only.
Both: Contours are drawn on both bottom and top planes.
2D Contours: A two dimensional projection of the contour curves is drawn with extended labelling. The frequency of contour curves can be adjusted by selecting the appropriate Z axis interval values.

2.3.5. View Menu



2.3.5.1. Redraw Mode

When a change is made to the graph settings from the menu bar and the dialogue exited by pressing <Enter/OK>, UNISTAT will automatically redraw the graph with the new settings. However, when editing complicated graphs with many data points, it may not be desirable for the program to redraw the image after every single edit operation. In order to switch off auto redrawing, and enable the manual redraw mode, select View → Redraw Mode. Two further options will appear; Auto and Manual. On entry, the Auto option will be checked, indicating that graphs are redrawn automatically. Click on Manual. Then to redraw the graph after editing a setting, either by pressing <Ctrl> + <R>, or by clicking on the first button on the toolbar which has a graph picture on a white background (see 2.3.5.2. Redraw).

To enable the program to redraw the graphs automatically again, select the Auto option from View → Redraw Mode.

For simple plots or charts the Auto option may be more convenient to use. However in the case of 3D surface plots or curves or surface fitting routines, or any other graphics procedure that requires intensive computing, switching to Manual is recommended.

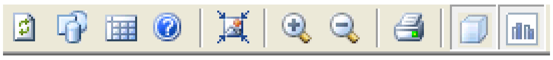
2.3.5.2. Redraw

Redraw causes the entire graph to be redrawn. This is especially useful if the Redraw Mode is set to Manual.

2.3.5.3. Reset Coordinates

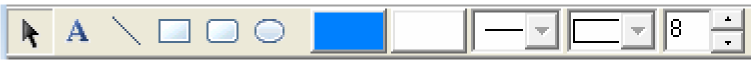
This is equivalent to selecting Edit → Options → Coordinates from the pull-down menu which will restore the default coordinates of the frame, plot area, legend and all standard text objects (main, sub and axis titles). The same effect can also be achieved by clicking on the [Reset] button on the Graphics Toolbar.

2.3.5.4. Graphics Toolbar



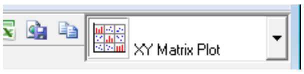
This will toggle the Graphics Toolbar on and off.

2.3.5.5. Drawing Toolbar

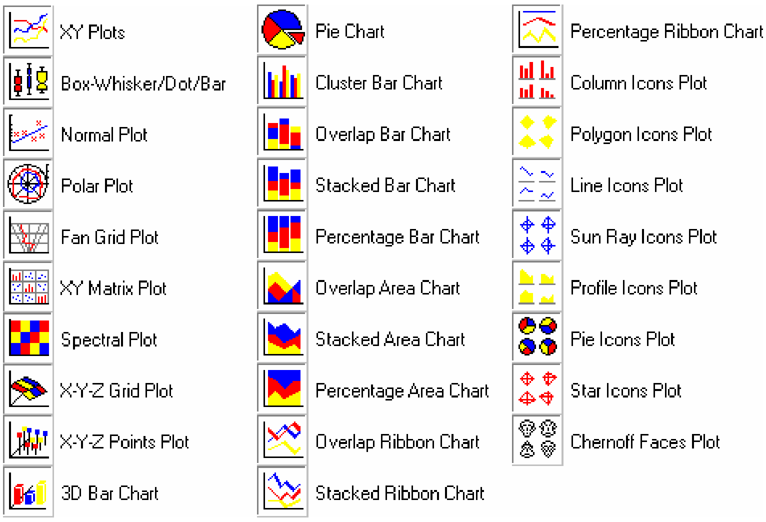


This will toggle the Drawing Toolbar on and off and is equivalent to clicking on the second button on the Graphics Toolbar.

2.3.5.6. Chart Gallery

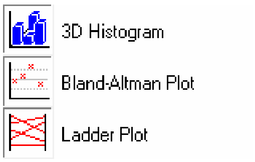


The Chart Gallery is a toolbar containing a drop-down list of graphics options. It is displayed on the top-right of the Graphics Editor window and has the following options.

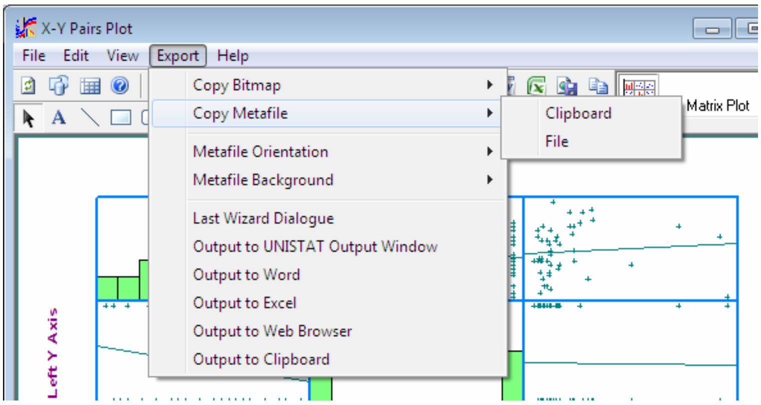


This toolbar will appear in any graphics procedure where selections are made by clicking on the [Variable] button. When a selection is made from the drop-down list, the new graph will be drawn immediately with the already selected variables, without asking for further user input.

With this version of UNISTAT a new mini Chart Gallery is introduced for the three paired data procedures (3D Histogram, Bland-Altman Plot and Ladder Plot) so that each can be visualised instantly with the same data selection.



2.3.6. Export Menu



Two copy options permit the export of images from the Graphics Editor to either the clipboard or to file in one of bitmap (.PNG) or enhanced metafile (.EMF) formats. UNISTAT graphs can also be exported directly to Word, Excel or to the default web browser.

2.3.6.1. Copy Bitmap

Graphs are copied as they are displayed on the screen. Therefore, their resolution depends on the resolution of the screen, as well as the size of the graph.

In order to obtain the best resolution from a bitmap, first maximise the Graphics Editor, switch off display of Graphics Toolbar and Drawing Toolbar and then select Export → Copy Bitmap. When the File option is selected, the program will prompt for a file name. The default file extension for bitmap files is .PNG. Ensure that files are saved with this extension. Otherwise, some drawing or graphics editing packages may not be able to recognise the file.

The image in Graphics Editor can also be copied to the clipboard in bitmap format by simply pressing <Alt> + <Print Screen>, a facility provided by the Windows environment. The only difference is that UNISTAT's Export → Copy Bitmap option will copy only the graph area (excluding the menu bar and the toolbar), whereas <Alt> + <Print Screen> copies the entire window including borders, menu bar and toolbars.

2.3.6.2. Copy Metafile

Metafiles store graphics images in vector format and thus their resolution is not limited with the screen resolution (in other words, they are device-independent). When they are replayed, metafiles redraw the image in the resolution determined by the output device.

When the Export → Copy Metafile → File option is selected, the program will prompt for a file name. The default file extension for enhanced metafiles is .EMF. Ensure that metafiles are saved with this extension. Otherwise, some drawing or graphics editing packages may not be able to recognise the metafile.

Metafiles saved from UNISTAT can be opened or pasted into Windows graphics editing packages such as MS Draw or MS Power Point. You can then edit any object on the graph, including lines, curves, text, or change positions of existing objects or add new objects like lines and text. You can also insert a metafile saved from UNISTAT into Word or Excel, and then open the metafile into the drawing utility by double-clicking on the graph.

Graphs can also be printed to file in vector graphics formats such as PCL, EPS (postscript) or PDF (Adobe Acrobat). For further information see 2.3.3.4. Print.

2.3.6.3. Metafile Orientation

As we have seen above, UNISTAT creates metafiles by selecting the Export → Copy Metafile option. The Export → Metafile Orientation option allows you to create metafiles with an original aspect ratio (scaling) of Landscape or Portrait.

Although the metafiles created by UNISTAT are fully scaleable, under certain circumstances it may be important to know whether the metafile was originally scaled in Landscape or Portrait orientation. If, for instance, the original scaling was Landscape and you wish to print the metafile in Portrait orientation, then some text (especially the left Y-axis numbers) may look displaced in the printout. If the metafile was created in portrait mode, then the correct aspect ratio of the graph would be retained and no such problem would have been encountered.

Even if the printer is configured in portrait orientation and this option is set to Portrait, graphs printed in 1/2 page will not retain their correct aspect ratio. In this case you should select the Landscape orientation using this option before printing the graph.

2.3.6.4. Metafile Background

The Export → Metafile Background option is used to control whether the metafiles saved from UNISTAT will be Transparent or Opaque when displayed over other objects. When Opaque is checked, the area covered by a metafile will be cleared first. When Transparent is checked, however, the metafile will be drawn without clearing the covered area first. In this way the user can overlay two different kinds of plot, say a histogram and a line chart.

Another point to remember is that even when the Transparent box is selected, the graphics background and the graphics area colours will not be transparent unless they are both white (see 2.3.2.2. UNISTAT Graphics Objects).

2.3.6.5. Export Options

The Export → Export Options menu duplicates the tasks of the Output Medium Toolbar buttons. For more information see 2.2. Output Medium.

When you finish editing in Graphics Editor, you can select the appropriate export option (or click on the appropriate toolbar button) to send the graph to the desired Output Medium.

Last Procedure Dialogue: The last step in the hierarchy of Procedure Dialogues will be displayed allowing you to change selections without having to run the same procedure from the menu again.

Output Window: This is the Default Output Medium for Stand-Alone Mode. Graphics images are sent to this window in the form of enhanced metafile (.EMF) objects. See 2.2.1. UNISTAT Output Window.

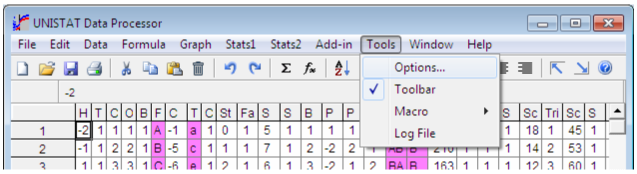
Word for Windows: Selecting this option will send the current image to Word in enhanced metafile (.EMF) format. Graphs are inserted into the current document, at the cursor location. See 2.2.2. Output to Word.

Excel for Windows: Select this option to send the current graph to Excel in enhanced metafile (.EMF) format. Graphs are inserted into the current worksheet, at the active cell. See 2.2.3. Output to Excel.

Web Browser: If output has been sent to the browser before, the program will stop and ask whether you want to overwrite the existing files or append the new output to the existing HTML page. Images are sent in PNG bitmap format. See 2.2.4. Output to Web Browser.

The Clipboard: The text will be sent to the clipboard with a fixed-width font as in the Output Window option above.

2.4. Tools

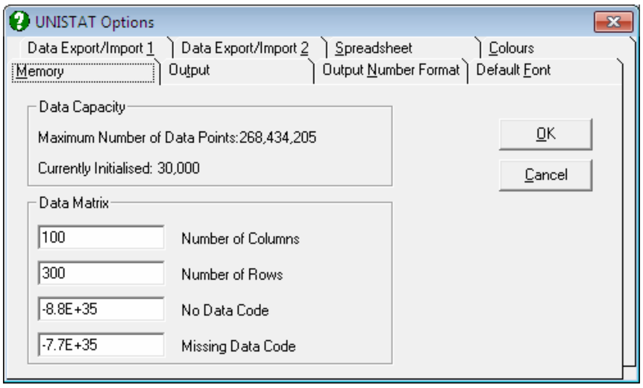


2.4.1. Options

Various global and peripheral parameters can be selected from a tabbed dialogue. In Stand-Alone Mode, this is accessible from the Data Processor menu item Tools. In Excel Add-In Mode, the [Options] button on the UNISTAT toolbar provides access to a similar dialogue, which excludes the Data Export / Import 2, Spreadsheet and Colours tabs.

If changes are made in one of the Options dialogues and the dialogue exited by clicking <Enter/OK>, all current options and parameters will be stored by the program.

2.4.1.1. Memory Management



Data Capacity

Maximum Number of Data Points:268,434,205

Currently Initialised: 30,000

Data Matrix

100 Number of Columns

300 Number of Rows

-8.8E+35 No Data Code

-7.7E+35 Missing Data Code

UNISTAT detects the free memory available in your system and displays the maximum number of data points that can be processed at any one time. The number of data points already initialised is also displayed in the group Data Capacity.

Each extra megabyte of free memory provides approximately 125,000 data points capacity (i.e. 8 bytes for each data point). There are no *ad hoc* limitations on the number of columns and rows. The dimensions of the data matrix can be set freely, provided that the number of cells in the matrix (i.e. number of columns times number of rows) does not exceed the maximum number of data points allowed.

The first two parameters in the second frame Data Matrix are used to set the dimensions of the data matrix and the latter two are for the No Data Code and Missing Data Code. Although you are allowed to change all these parameters any time during a session, the first three cannot be changed before clearing all data in the spreadsheet. If there is no data in the spreadsheet, then the new matrix will be initialised without further warning. Otherwise a prompt will ask you to confirm whether the data already in memory can be cleared first.

2.4.1.1.1. Number of Columns

This sets the maximum number of spreadsheet columns. The lower limit for this number is 4 and its upper limit depends on the total memory available on the system and the number of rows already selected. The exact number of columns that can be initialised is determined as follows:

$$4 \leq \text{MaxColNo} \leq \text{Int}(\text{MaxPoints}/(\text{MaxRowNo}+1)),$$

where *MaxPoints* is the maximum number of data points that can be processed at any one time, as reported at the top of the dialogue. Any numbers outside this range are not allowed. If this is attempted the program will display the valid range and wait for a valid entry. A higher number of columns than that allowed by the current number of rows can be set by reducing the number of rows to a sufficiently low level first, say 20, then entering the desired Number of Columns field, and going back to the Number of Rows field to see how many rows are allowed with this particular number of columns.

2.4.1.1.2. Number of Rows

This field is for setting the maximum number of rows that can be processed at a time. Its lower limit is 20. The exact number of rows that can be initialised is determined as follows:

$$20 \leq \text{MaxRowNo} \leq \text{Int}(\text{MaxPoints}/(\text{MaxColNo}-1)).$$

As in the case of setting the number of columns, a higher number of rows than that allowed by the current number of columns can be set by reducing the number of columns to a sufficiently low number first, say 4, then setting the Number of Rows field as desired, and going back to the Number of Columns field to see how many columns are allowed with this particular row number setting.

2.4.1.1.3. No Data Code

This value is used as a marker for cells that do not contain any data. When a data matrix is initialised, all its cells are assigned this value. It is highly unlikely that the need should ever arise to change the No Data Code. In the remote event of a data set containing this number, another number that will not interfere with the data can be entered. As in the case of the maximum number of data matrix columns and rows, any data in spreadsheet will be cleared when this parameter is changed.

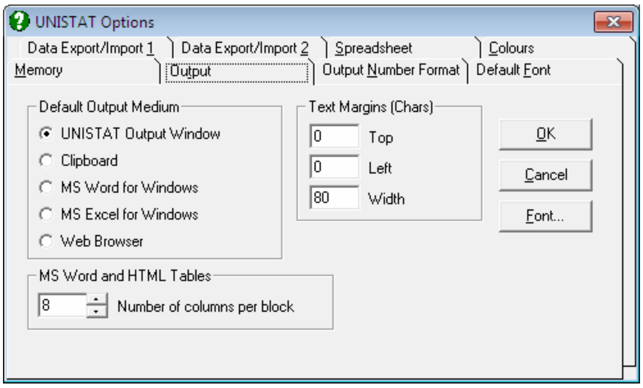
2.4.1.1.4. Missing Data Code

Any cells having this value will be considered missing. A missing data cell is different from a blank cell (though this distinction is not observed by many applications, including Excel). During normal operation of the program, this value will be invisible to the user as a number. In Stand-Alone Mode, UNISTAT's own spreadsheet represents a missing value by the asterisk (*) character. Any blank cells in a column, below which there are cells containing data, will be considered missing and filled with an asterisk automatically. In Excel Add-In Mode, you do not have to insert an asterisk into the missing data cells. UNISTAT will interpret blank cells within data columns as missing values. Ensure that such cells are truly blank and do not contain spaces or other invisible characters.

It is only when data files are exported or imported that the missing value code may be important (see 3.1.0. File Formats). For instance, in order to load a text file saved from a different application (or export data to a different application) the Missing Data Code may be changed to make it consistent with that of the external application. Also, in the remote event of data actually containing the missing value code as a data value, you will need to change it to a unique value which is not used as data.

The Missing Data Code can be changed when there is already data in the spreadsheet. In this case, the spreadsheet display will be refreshed, the cells containing the old Missing Data Code will show it as a number, and the cells containing the new Missing Data Code will show an asterisk. After the change, UNISTAT procedures will no longer recognise the cells with the previous code as missing and try to process these numbers as ordinary data. Because the default missing value is a very large negative number, in most cases the program will stop execution reporting a number overflow. To prevent this happening, change the value of the cells with the old Missing Data Code to the new one by means of the Data Processor's Data → Recode Column procedure or the **If()** function (see 3.4.2.7. Conditional Functions).

2.4.1.2. Output



Default Output Medium
● UNISTAT Output Window
○ Clipboard
○ MS Word for Windows
○ MS Excel for Windows
○ Web Browser

Text Margins (Chars)
0 Top
0 Left
80 Width

MS Word and HTML Tables
8 Number of columns per block

The Default Output Medium group on top left allows the user to set the destination for UNISTAT output. For the first two items (i.e. Output Window and the Clipboard) it is also possible to set the font and text margins using the [Font…] button and the Text Margins group of controls respectively.

The third group is used to set the number of columns per block in Word and web browser output.

2.4.1.2.1. Default Output Medium

All UNISTAT output will be sent to the selected application by default. For instance, if Word is installed and it is selected as the Default Output Medium, then all text and graphics output will be sent to Word. The output tables will be formatted in native Word table format. Similarly, when Excel is selected, all output is sent to Excel worksheets.

This dialogue is used to select the Default Output Medium, that is, the medium where output is first sent when a procedure is run. While any one of these applications is selected as the Default Output Medium, it will still be possible to send the same output to other applications in the list without having to run the procedure again. Suppose Output Window has been left as the Default Output Medium, since it is fast and it is also the accustomed type of output. In this case, after performing a procedure, the output will be sent to Output Window, but buttons on the Output Medium Toolbar will allow you to send the same output to other applications in the list. By clicking on the Word button, for instance, the same output will this time be sent to Word, in the form of a formatted Word table. In this way, Output Window - with its traditional fixed width font output - would function as the preview window, and only the final results would be sent to Word for inclusion in a report.

UNISTAT's Graphics Editor also displays a similar Output Medium Toolbar for exporting graphics output to Word, Excel or the web browser. After customising a graph in UNISTAT Graphics Editor, the image can be sent to Word in enhanced metafile format by clicking on the Word button. Clicking on the Excel button will send it to Excel.

Output Window: The statistics (text) output is created in old style *line printer* format and thus should be viewed with a fixed-width font. The margins of the output can be controlled from the Text Margins group provided in the same dialogue. Graphics images are sent as enhanced metafile objects. See 2.2.1. UNISTAT Output Window.

The Clipboard: The text will be sent to the clipboard with a fixed-width font as in the above option.

Word for Windows: This option will be available only if Word 97 or a later version of Word is installed. If Word is not running already, it will be launched first. All output will be inserted into the current cursor position in the active document. See 2.2.2. Output to Word.

Excel for Windows: Only available when Excel 97 or a later version of Excel is installed. If Excel is not running already, it will be launched first. Each output page will be placed in a new worksheet. See 2.2.3. Output to Excel.

Web Browser: This option is only available if a web browser is installed. If the browser is not running already, it will be launched first. When the output is sent to the web browser for the first time, it will appear in the browser automatically. When this option is selected again subsequently, UNISTAT will ask whether you want to overwrite the existing output or append the new output to the existing HTML file. See 2.2.4. Output to Web Browser.

2.4.1.2.2. Font

The [Font…] button is used to select the type, style and size of the Output Window font. The font selected should be non-proportional, that is, a font with fixed character widths. The best results are obtained with *Courier New* True Type font.

2.4.1.2.3. Text Margins

It is possible to adjust the top and left margins in terms of number of lines and number of characters respectively.

Top: The number of lines between output from consecutive procedures.

Left: The number of characters left blank on the left.

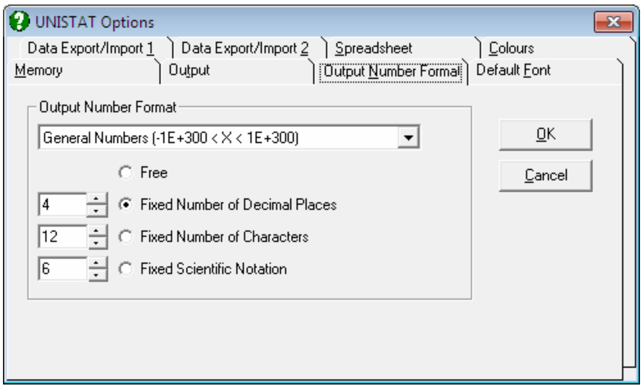
Width: This field sets the width of the printed and file output in terms of number of characters per line. This number can be minimum 80 and maximum 32,000. Output will be re-scaled and large matrix output will be blocked accordingly. The resolution of character plots (i.e. character histogram, plot of residuals, fitted and actual y values, etc.) will also be determined by this number. For further information see 2.2.1. UNISTAT Output Window.

2.4.1.2.4. Word and HTML Tables

Number of columns per block: In Word and web browser output, large tables are parsed into blocks to facilitate easy viewing and / or printing. The default number of columns per block is 6, but you can change this to any number greater than 2. For more information see 2.2.2. Output to Word and 2.2.4. Output to Web Browser.

This option does not have any effect on the width of output in UNISTAT’s Output Window or in Excel (which is not blocked by default). For more information on how to change the width of output for these media see 2.2.1. UNISTAT Output Window and 2.2.3. Output to Excel.

2.4.1.3. Number Format



Output Number Format

General Numbers (-1E+300 < X < 1E+300)

○ Free

4 ● Fixed Number of Decimal Places

12 ○ Fixed Number of Characters

6 ○ Fixed Scientific Notation

The number of digits displayed for floating point numbers in output can be controlled using this tab dialogue. You can also set the formats for the numbers displayed on the axes of graphs using a similar dialogue (see 2.3.4.4. Axes). Each axis can have its own format.

2.4.1.3.1. Options List

The numbers displayed in UNISTAT text output are classified into four groups;

General Numbers: These are the numbers that do not fall within one of the following categories and can have a range of -1E+300 to 1E+300.

Correlations and Probabilities: These are the numbers that are inherently confined to the interval of -1 to +1 and 0 to 1 respectively.

Confidence Intervals: These can be formatted separately.

ANOVA Tables: Sum of squares, mean square and F-statistic values can be formatted.

When a selection is made from this list, the options underneath will display the current settings for this selection.

2.4.1.3.2. Number Styles

A floating point number can be displayed in one of the following formats.

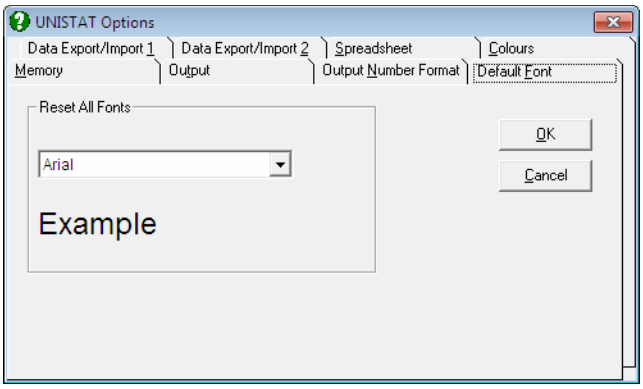
Free: Up to 15 digits will be displayed within the space allowed in the output. The position of the floating point for numbers in the same column will not necessarily be the same. Too large and too small numbers will be displayed in scientific (i.e. power) notation.

Fixed Number of Decimal Places: All numbers will be displayed with the same number of digits after the floating point. Therefore, the floating points of numbers in the same column will align properly. The numbers that do not have enough number of significant digits after the floating point will be padded with zeros.

Fixed Number of Characters: The number is displayed to fit in the selected width. The position of the floating point for numbers in the same column will not necessarily be the same. Too large and too small numbers will be displayed in scientific (i.e. power) notation.

Fixed Scientific Notation: The number is displayed in power notation (e.g. 1.00E-15). It is possible to format the mantissa to display a fixed number of digits. The number defined here is not the number of decimal places but that of the total number of characters for the mantissa, including the floating point.

2.4.1.4. Default Font

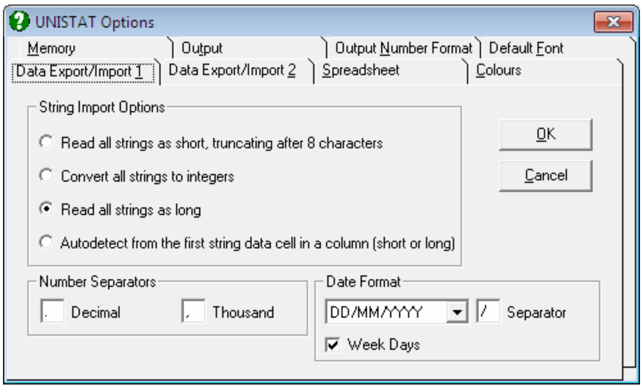


When UNISTAT is first installed, it will automatically scan the Windows system to determine which fonts are installed. It will then set a default font for all major windows of UNISTAT. This font will normally be *Arial*, except for the Output Window, which will have a fixed width font (normally *Courier New*).

However, if there are other fonts or if other applications that come with their own fonts (like Word) are installed subsequently, then some unsuitable fonts may be displayed when UNISTAT is started.

Select this option to reset all fonts in UNISTAT (i.e. Data Processor, Procedure Dialogues and all graphics fonts). The drop-down list displays available Windows fonts. Select any font from the list and click [OK]. The new font will become active immediately, and the new font settings will be stored by the program.

2.4.1.5. Data Export / Import 1



String Import Options

○ Read all strings as short, truncating after 8 characters

○ Convert all strings to integers

● Read all strings as long

○ Autodetect from the first string data cell in a column (short or long)

Number Separators

. Decimal   , Thousand

Date Format

DD/MM/YYYY   / Separator

☑ Week Days

When opening data files or pasting data from the clipboard, the user can select one of the following four options which control the type of strings:

2.4.1.5.1. String Import Options

Read All Strings as Short: When this option is selected UNISTAT will treat all string variables as short and a Long String Table will not be created (see 3.0.2.2. String Data). The maximum length of a short string data cell is eight characters.

While this option is selected, it is still possible to read the first column of the file as Row Labels, which have no limitations on the number of characters. It is up to the user to read the first column of such a file as Row Labels or as the first column of data (see 2.4.1.6.2. Labels).

Advantages of reading short strings are summarised in section 3.0.2.2.1. Short Strings. When, however, the data to be imported contains String Data that are longer than eight characters, truncation may result in an unacceptable loss of variation. For instance, a string variable which has the rows *Continent1, Continent2* will be read into UNISTAT as *Continen, Continen*, losing a vital variation in data. In such cases, the use of this option is not recommended and you will need to select one of the second or third options below.

Convert All Strings to Integers: When this option is selected, each distinct value in a string variable will be represented by an integer. The resulting column will thus be a numeric data column containing integers, but there will be no loss of variation due to truncation after the eighth character as in the above (Short Strings) option.

String values are represented by integers in the order of their appearance in the column. A Long String Table (see 3.0.2.2. String Data) is created, but no correspondence is established between the integers in a column and its strings in the table. With this option, the imported file will always contain only numeric data.

However, the user may then go to Edit → Long String Table, examine the string information, and establish the correspondence between columns containing integers and their string values by entering the function **Long()** for such columns (see 3.4.2.6.1. Data Conversion Functions).

The default dimensions for Long String Table are 200 columns and 2000 rows, meaning that the first 200 columns containing String Data, each of which containing no more than 2000 distinct values, will be translated. If these dimensions are not sufficient, select → Long String Table and increase the size of the table before opening the file.

Read All Strings as Long: This option is identical to the above option, except that the correspondence between integers and the Long String Table is established automatically, resulting in all strings in the incoming file being displayed in Data Processor.

Auto-detect: The program checks the first non-missing data in a column. If it is a string longer than eight characters, then the program will conclude that the column contains Long Strings. Once this is established, the rest of the entries in this column will be all treated as Long Strings whether they are shorter strings or they are numbers or dates.

If the first non-missing data in a column is a string and its length is less than or equal to eight characters, then the program will conclude that the column contains Short Strings. All the rest of the entries will be read as Short Strings and the ones longer than eight characters will be truncated. Truncation may result in loss of generality in a variable. In such cases, the use of this option is not recommended and the user should select one of the second or third options above.

2.4.1.5.2. Number Separators

Decimal and thousand separators can be changed here. UNISTAT will automatically detect these two parameters from the Windows environment every time it is started. Use this editing facility only when you need to override the default Windows settings. This may be necessary, for instance, when your default decimal separator is *period* and you receive a text file which was saved on a Windows system with a *comma* decimal separator.

2.4.1.5.3. Date Format

As in the Number Separators options above, UNISTAT will automatically detect these parameters from the Windows environment every time it is started. Use this facility only when you need to override the default Windows settings. This may be necessary when you need to import files saved from a Windows system with a different Date Format. The available date formats are:

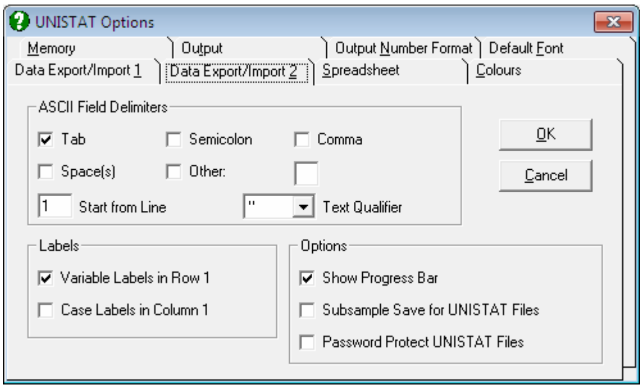
dd/mm/yyyy
mm/dd/yyyy
yyyy/mm/dd

and any character can be selected as the date separator.

When the Week Day box is checked, dates are displayed including the day of the week, such as *29/02/2000 Tue*. Otherwise, only the date is displayed.

For further information on the date and time data types see 3.0.2.3. Date Data and 3.0.2.5. Date-Time Data.

2.4.1.6. Data Export / Import 2



ASCII Field Delimiters

☑ Tab □ Semicolon □ Comma

□ Space(s) □ Other:

1 Start from Line

" Text Qualifier

Labels

☑ Variable Labels in Row 1

□ Case Labels in Column 1

Options

☑ Show Progress Bar

□ Subsample Save for UNISTAT Files

□ Password Protect UNISTAT Files

This dialogue is available in Stand-Alone Mode only (see 1.3. Modes of Running UNISTAT).

2.4.1.6.1. Text Field Delimiters

Delimiters: Tab, semicolon, comma or space(s) may be selected as the field delimiter, i.e. the character that separates different data points in the file. When more than one delimiter is selected, they will be effective simultaneously. The space character can also be selected as delimiter, but any number of contiguous spaces will be considered as one single separator. Clicking on the last check box, Other, enables the selection of any character of choice as the field delimiter. The selected field delimiter will be effective both in reading and saving text files (see 3.1.0.4. Delimited Text Files and 3.1.0.5. Free Format Text Files) and pasting the contents of the clipboard into Data Processor (see 3.2.5. Paste).

Start from Line: Some text files may contain some header information at the top of the file. By entering the line number to start from, you can read such files without having to edit them first.

Text Qualifier: Any characters enclosed between two characters shown in this control will be read as text (rather than numbers). There may not be a text qualifier, or it can be a single or double quote.

2.4.1.6.2. Labels

These check boxes are used by all data import and export procedures to control whether the Column Labels or Row Labels are included as part of files saved or loaded.

UNISTAT internal (.USW) files will always save and load all labels.

WARNING! *When importing files, you should ensure that current settings of label options are consistent with the actual file. Otherwise the file may not be read correctly.*

Column Labels in Row 1: When this box is checked, Column Labels will be saved as part of exported files. When loading or merging files, the first row of the imported file will be assumed to contain Column Labels.

Row Labels in Column 1: When this box is checked, Row Labels will be saved as part of exported files. When loading or merging files, the first column of the imported file will be assumed to contain Row Labels.

2.4.1.6.3. Options

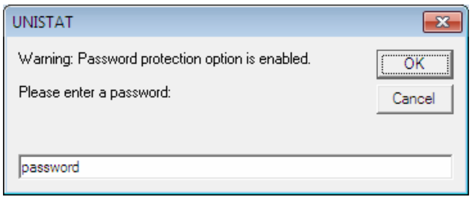
Show Progress Bar: When opening or saving files, a progress bar on the Status Panel shows the percentage of the task performed. Uncheck this box in order not to display the progress bar.

Subsample Save for UNISTAT Files: UNISTAT internal files (.USW) will normally save all data contained in the Data Processor. However, this option will allow you to save only those rows of the data matrix which conform to a given logical condition. When this box is checked, you need to move the active cell to a factor column (i.e. a categorical variable) first. When the save option is selected the program will prompt you to enter a level (i.e. a value) of this factor column. Then only those rows of the whole data matrix containing this particular value in the factor column will be saved to the file. There is also an option to save all rows by entering an asterisk (*) in the input field, which is equivalent to leaving this subsample selection box unchecked.

There are several different ways of selecting subsamples from a data set. For instance, you can use the Data Processor's Data → Select Row facility to mark rows to be included in subsequent analyses. You can use the **If()** (see 3.4.2.7. Conditional Functions) function to use a logical condition to generate a Select Row column. The Data → Recode Column facility can also be used.

It also possible to run UNISTAT procedures on subsamples of data, without having to edit the actual data matrix (see 2.1.2. Categorical Data Analysis).

Password Protect UNISTAT Files: When this option is checked and an attempt is made to save UNISTAT internal (.USW files), a dialogue will pop up asking for a password. The password protected .USW files can only be retrieved by entering the correct password.

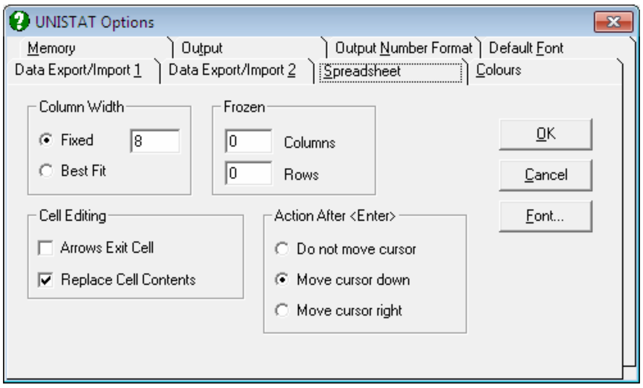


Warning: Password protection option is enabled.

Please enter a password:

password

2.4.1.7. Spreadsheet



Column Width

● Fixed 8

○ Best Fit

Frozen

0 Columns

0 Rows

Cell Editing

□ Arrows Exit Cell

☑ Replace Cell Contents

Action After <Enter>

○ Do not move cursor

● Move cursor down

○ Move cursor right

This dialogue is available in Stand-Alone Mode only (see 1.3. Modes of Running UNISTAT).

Various aspects of the spreadsheet, font type, style, size, column widths, frozen (non scrolling) columns and / or rows, and Cell Editing options can be selected from this dialogue.

2.4.1.7.1. Font

Click on the [Font…] button to activate the standard Windows font dialogue. Any fonts in any size can be selected and font attributes like bold, italic, underline, strikethrough can be set. All changes made will be saved as part of .USW files. The following rules will apply:

1) If no block of cells has been highlighted then the font of the current column will be set.
2) If a block of cells has been highlighted, then the font of the columns in the block will be set.
3) Any font selections will be valid for the printer if the font is *True Type* or it has a printer equivalent.

2.4.1.7.2. Column Width

A fixed column width (which can be determined by the user) can be set to the selected range of columns. Alternatively, the Best fit option will work out the optimum width for the selected columns automatically. You can also change the widths of highlighted columns by mouse, from the borders between Column Labels. All changes made in column widths will be saved as part of .USW files. The following rules will apply:

1) If no block of cells has been highlighted then the width of the current column will be set.
2) If a block of cells has been highlighted, then the widths of the columns in the block will be set.
3) Widths of non adjacent columns (multiple selections) cannot be set in one go.

2.4.1.7.3. Frozen

This option causes a fixed number of columns or rows to be displayed permanently on the screen. The so called *frozen* or *nonscroll* columns or rows will always be the first n columns or rows. These parameters will be saved as part of .USW files.

2.4.1.7.4. Cell Editing

The following options are available:

Arrows Exit Cell: When this box is checked, pressing an arrow key will terminate editing, and the active cell will move in the direction of arrow. Otherwise, arrow keys will move the blinking cursor along the edited text.

Replace Cell Contents: When this box is checked, any new text typed into a cell will replace the contents of the cell. Otherwise, the new text will be appended to the old text.

2.4.1.7.5. Action After <Enter>

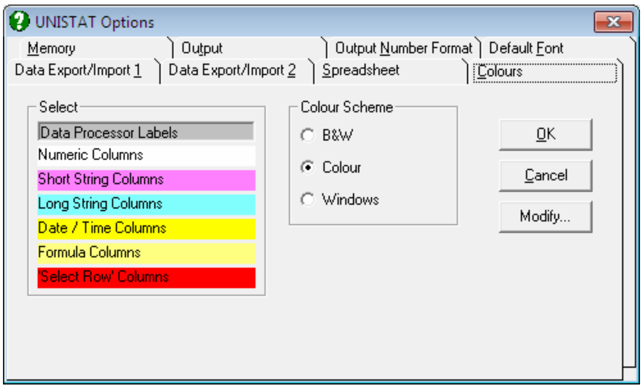
This allows selecting the action to be taken after inputting data into a cell. The following options are available:

Do Not Move Cursor: The current cell remains to be the active cell.

Move Cursor Down: The next cell in the same column becomes the active cell.

Move Cursor Right: The next cell in the same row becomes the active cell.

2.4.1.8. Colours



Colour Scheme

○ B&W
● Colour
○ Windows

This dialogue is available in Stand-Alone Mode only (see 1.3. Modes of Running UNISTAT).

It is possible to select background colours for the most commonly used windows and switch colours off entirely.

2.4.1.8.1. Select Colours

Click on one of the items in the Select list to make it active, then click on the [Modify…] button. A standard Windows colour selection dialogue will be opened containing a palette of colours and tools for composing your own colours.

When all the colours are selected exit the dialogue pressing <Enter/OK>. New choices will become effective immediately. If the dialogue is exited by pressing <Esc/Cancel> the old colour will not be changed.

2.4.1.8.2. Colour Scheme

Black and White: When this option is selected, all colours will be displayed as black or white or as shades of grey.

Colour: This option will make the user-selected colour options active.

Windows: This option will detect the Windows environment's colour settings and make them active in UNISTAT.

2.4.2. Macros

This feature is also available in Excel Add-In Mode (see 1.3. Modes of Running UNISTAT).

As a spreadsheet-based, menu-driven statistical package, UNISTAT does not feature an explicit programming language. However, it does have a powerful macro record/playback facility that enables the user to automate repetitive tasks.

The macro recorder does not save keystrokes or mouse movements. Instead, it records all settings in Procedure Dialogues, namely, variable selections, intermediate data input (text and numbers), output options and command button actions. In graphics procedures, all editable properties of graphs that can be saved in a graphics template file are saved as part of macro files (see 2.3.3.3. Save Template As). Therefore, while saving macro files, you should not worry about having made too many changes or trials, as only the final configuration of the settings are saved in the macro file.

Macro files are in binary format and therefore they are not suitable for editing by the user.

When a macro contains graphical procedures, the Graphics Editor window will not be displayed during playback. To switch on the display of Graphics Editor window, enter the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

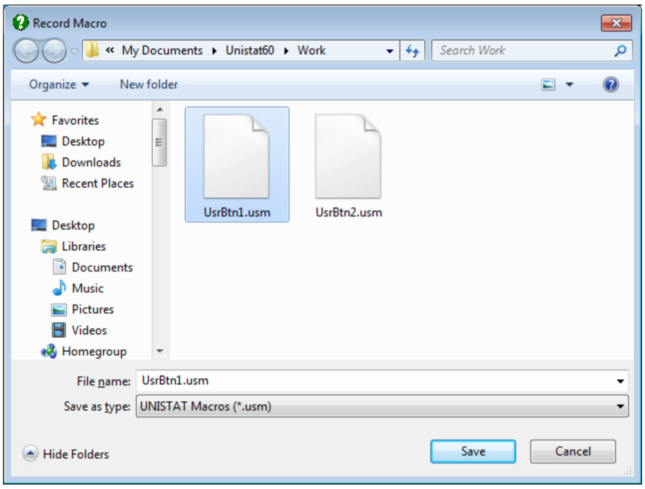
```
MacroDisplayGraph=1
```

In both cases, the image will be sent to the Output Medium as an enhanced metafile object. When the Output Medium is the web browser, the Graphics Editor will always be displayed, irrespective of the value of this parameter.

Normally, UNISTAT graphics axes will be scaled automatically according to the minimum and maximum values in the data series plotted. Under certain circumstances, however, it may be desirable to keep the axis scaling unchanged, as it was during the recording stage. To do this, enter the following line in *Unistat65.ini* file under the [Options] group:

```
MacroScaleFixed=1
```

2.4.2.1. Recording Macros

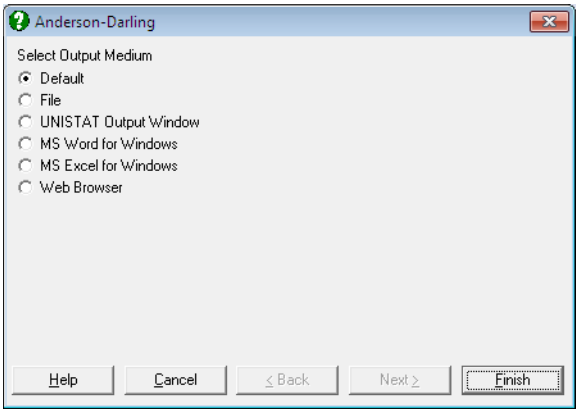


File name: UsrBtn1.usm

Save as type: UNISTAT Macros (*.usm)

To record a macro file select Tools → Macro → Record Macro. The standard file dialogue will open and prompt for a file name (the default extension is .USM). After clicking [OK] proceed as usual with performing a procedure (graphics or statistics). A single macro file may contain an unlimited number of procedures. It will continue recording until the Tools → Macro → Stop Recording Macro is selected.

2.4.2.2. Running Macros



Select Output Medium

● Default
○ File
○ UNISTAT Output Window
○ MS Word for Windows
○ MS Excel for Windows
○ Web Browser

A previously saved macro file can be opened from Tools → Macro → Run Macro. The standard file dialogue will prompt for the selection of the macro file. Once the macro file has been selected, a dialogue will ask you to select the Output Medium to use. The options are as follows:

Default: All output from the macro will be sent to the Default Output Medium

File: The standard file dialogue will prompt you to choose a file name. All text output from the macro will be appended to this file. Each graph generated by the macro will be saved as a metafile in the same folder as this text file. These metafiles will have the names UNISTAT1.EMF, UNISTAT2.EMF, ..., UNISTA10.EMF, ... depending on which files are already in the folder. So, if UNISTAT1.EMF is already exists in the folder, the first metafile saved from the current macro will be named UNISTAT2.EMF. No files are deleted or overwritten in this way.

If you wish to have a particular graphics output saved with same name each time the same macro is run, then it is necessary to ensure that all metafiles in the folder are deleted before running the macro.

Output Window: All text and graphics output will be sent to UNISTAT's Output Window. The text output will be in old style *line printer* format.

Word for Windows: This option only appears if Word is installed. Both text and graphics output are sent to Word. See 2.2.2. Output to Word.

Excel for Windows: This option only appears if Excel is installed. Both text and graphical output are sent to Excel. See 2.2.3. Output to Excel.

Web Browser: This option only appears if a browser is installed. Both text and graphics output are sent to the default browser. See 2.2.4. Output to Web Browser.

It is important to ensure that the data present in the spreadsheet is compatible with the macro to be run.

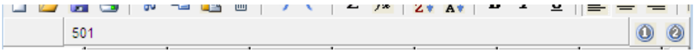
2.4.2.3. Combining Macros

This option is used for combining two or more macro files into a single macro file. This is often desirable because it is not so easy to record long macro files.

A file open dialogue pops up with a default file extension .USM. It is important to select files in the correct order because this will be the order in which they will succeed each other in the combined macro file. The selected files may appear in the File Name box in reverse order, i.e. the first file clicked becomes the last and the last one clicked becomes the first file in the list.

The output file name is always *CombinedMacro.usm* and it will be saved to the same folder where the input macro files are. Therefore, you will need to rename this file as soon as it is created, in order to prevent it from being overwritten next time you combine macros.

2.4.2.4. Macro Shortcut Buttons



This feature is only available in Excel Add-In Mode (see 1.3. Modes of Running UNISTAT).

One or two shortcut buttons can be assigned to frequently used macros on the second Data Processor toolbar, next to the Output Medium buttons. To do this, a macro should be recorded first.

1) Select Tools → Macro → Record Macro from Data Processor.
2) Assign the following path and file name:
   Documents\Unistat65\UsrBtn1.usm
   The location and name of this file should not be changed.
3) Run the procedures you want to include in the macro.
4) When finished, select Tools → Macro → Stop Recording Macro.
5) A button with an icon (1) should appear on the second toolbar.
6) When you move the mouse pointer over this button the tool tip will show UserButton1. You can change this text and display any information you want by entering the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

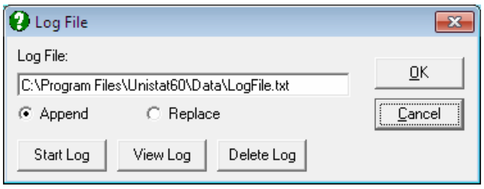
   ```
   UsrBtn1=My Tooltip
   ```

   The text after the equal sign will be displayed as tool tip for User Button 1.

You may create a second button by replacing 1 above with 2.

Before clicking on macro shortcut buttons ensure that the data present in the spreadsheet is compatible with the macro to be run.

2.4.3. Log File



C:\Program Files\Unistat60\Data\LogFile.txt

● Append

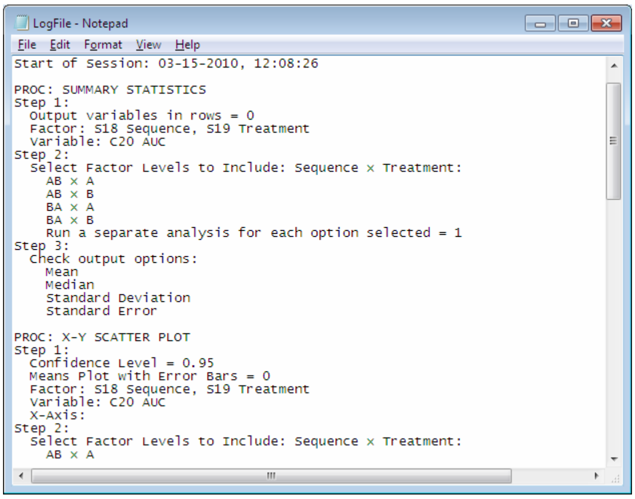
This option facilitates keeping a record of all selections made from UNISTAT's Graph, Statistics 1 and Statistics 2 menus.

This feature is also available in Excel Add-In Mode (see 1.3. Modes of Running UNISTAT).

To start keeping a log select Tools → Log File. The above dialogue pops up. The default saving option is Append, which means that the logs will be written to the same file and nothing in the Log File will be deleted. If the Replace option is selected, the old Log File will be deleted when UNISTAT is launched again.

The default Log File path and name can be changed by entering and editing the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

```
LogFileName=..\Documents\Unistat65\LogFile.txt
```



```
Start of Session: 03-15-2010, 12:08:26

PROC: SUMMARY STATISTICS
Step 1:
  Output variables in rows = 0
  Factor: S18 Sequence, S19 Treatment
  Variable: C20 AUC
Step 2:
  Select Factor Levels to Include: Sequence × Treatment:
    AB × A
    AB × B
    BA × A
    BA × B
    Run a separate analysis for each option selected = 1
Step 3:
  Check output options:
    Mean
    Median
    Standard Deviation
    Standard Error

PROC: X-Y SCATTER PLOT
Step 1:
  Confidence Level = 0.95
  Means Plot with Error Bars = 0
  Factor: S18 Sequence, S19 Treatment
  Variable: C20 AUC
  X-Axis:
Step 2:
  Select Factor Levels to Include: Sequence × Treatment:
    AB × A
```

It is easy to forget that one has entered the above lines in *Unistat65.ini*. When the Append option is on, the Log File will keep on growing. It is the user's responsibility to maintain this file.

The spreadsheet operations are not logged.

2.4.4. Example Files

The data sets used in examples can be accessed from Tools → Example Files (which are located in *Program Files**Unistat Ltd**Unistat 6.5**Examples* folder). The data sets are presented in two different formats: Excel (.XLS) and UNISTAT internal format (.USW).

For further information see section 1.2.2. Reproducing the Examples.

2.4.5. Help

Help: The help system is programmed as a single PDF file, which is identical to the printed manual. Context sensitive help can be invoked in any procedure, either by clicking on <Help> or pressing the <f1> key. This will open the help file and display the top of the page containing the title of the currently selected procedure. You can search for keywords and click on the hyperlinks to jump between topics.

Using UNISTAT help: This will activate the help system, displaying section 1.2. Using UNISTAT User's Guide.

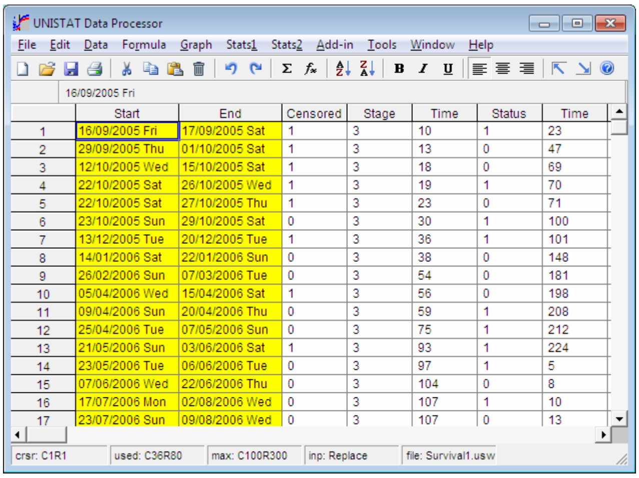
About UNISTAT: This displays information about your particular installation of UNISTAT Statistical Package.

UNISTAT Statistical Package

Chapter 3
Data Processor

3.0. Overview

The Data Processor is a column-based alphanumeric spreadsheet, which is used for entering and handling data.


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help
16/09/2005 Fri

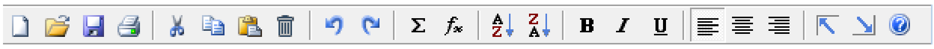
	Start	End	Censored	Stage	Time	Status	Time
1	16/09/2005 Fri	17/09/2005 Sat	1	3	10	1	23
2	29/09/2005 Thu	01/10/2005 Sat	1	3	13	0	47
3	12/10/2005 Wed	15/10/2005 Sat	1	3	18	0	69
4	22/10/2005 Sat	26/10/2005 Wed	1	3	19	1	70
5	22/10/2005 Sat	27/10/2005 Thu	1	3	23	0	71
6	23/10/2005 Sun	29/10/2005 Sat	0	3	30	1	100
7	13/12/2005 Tue	20/12/2005 Tue	1	3	36	1	101
8	14/01/2006 Sat	22/01/2006 Sun	0	3	38	0	148
9	26/02/2006 Sun	07/03/2006 Tue	0	3	54	0	181
10	05/04/2006 Wed	15/04/2006 Sat	1	3	56	0	198
11	09/04/2006 Sun	20/04/2006 Thu	0	3	59	1	208
12	25/04/2006 Tue	07/05/2006 Sun	0	3	75	1	212
13	21/05/2006 Sun	03/06/2006 Sat	1	3	93	1	224
14	23/05/2006 Tue	06/06/2006 Tue	0	3	97	1	5
15	07/06/2006 Wed	22/06/2006 Thu	0	3	104	0	8
16	17/07/2006 Mon	02/08/2006 Wed	0	3	107	1	10
17	23/07/2006 Sun	09/08/2006 Wed	0	3	107	0	13

crsr: C1R1 | used: C36R80 | max: C100R300 | inp: Replace | file: Survival1.usw

3.0.1. Screen Layout

The top line of the screen displays the pull-down menu bar. The next line is the Data Processor toolbar which contains a number of buttons. The Input Panel is situated under the toolbar.

3.0.1.1. Toolbar



The toolbar provides direct access to commonly used Data Processor functions. You can toggle the display of toolbar on and off from Tools → Toolbar.

New: Clears the data in spreadsheet. If data has been changed you will be prompted whether you wish to save the existing data first.

Open: Activates the file open dialogue (see 3.1.2. Open).

Save: Updates the file on disk with the version in spreadsheet. No dialogues will pop up. If the data has been entered from the keyboard or pasted, and therefore there does not exist a file name, a file name will be asked first.

Print: Activates the Print dialogue.

Cut: Copies the highlighted range to the clipboard and then erases it. If there is data in lower rows then the erased range is filled with missing values.

Copy: Copies the highlighted range to the clipboard.

Paste: Copies the contents of the clipboard to the spreadsheet. The top left corner of the incoming block is placed at the active cell (see 3.2.5. Paste).

Clear: The highlighted block is cleared. If there is data in lower rows then the erased range is filled with missing values.

Undo: Restores the state of the spreadsheet before the last edit.

Repeat: Repeats the last editing action.

Σ **Range Statistics:** Data Processor's Range Statistics procedure is activated.

Formula Editor: Activates the Formula Editor dialogue.

Sort Ascending: Sorts the current column (i.e. where the active cell is) in ascending order.

Sort Descending: Sorts the current column (i.e. where the active cell is) in descending order.

B *I* U **Column Font Style Group: Bold, Italic, Underline:** The text in selected columns will be displayed bold, italic or underlined, or in any combination of the three styles. These selections are saved as part of .USW files.

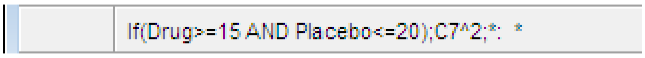
Column Alignment Group: Align Left, Centre, Align Right: These three buttons will left-justify, centre or right-justify the text in highlighted range of columns. The default format is left-justify. This information is saved as part of .USW files.

Home: Simulates pressing <Ctrl> + <Home>. The active cell will be located at (1, 1).

End: Simulates pressing <Ctrl> + <End>. The active cell will be located at the last column and the last row containing data.

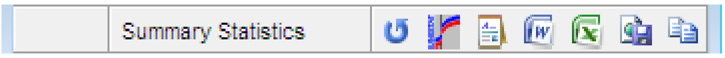
Help: Activates the UNISTAT help system.

3.0.1.2. Input Panel



If(Drug>=15 AND Placebo<=20);C7^2;*: *

The line below the toolbar is used to display the contents of the active cell, and to input various parameters or formulas. When the spreadsheet is in ready mode, contents of the active cell are displayed on the left of the panel. This can be a number, string expression, date, time, a missing value marker (*) or a formula, depending on the type of the column (see 3.0.2. Data Types).

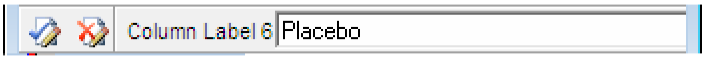


Summary Statistics

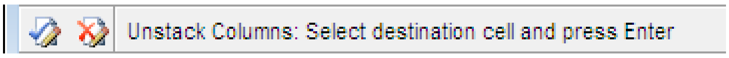
After a procedure is performed, some other buttons are added to the Input Panel. The button with a curly arrow on the left is used to activate the Last Procedure Dialogue for the current procedure (the name of which is displayed on the left). The next six icons are used to send the output from this procedure to another Output Medium (see 2.2.0. Output Medium Toolbar). The available options are:

1) Data Processor
2) Output Window
3) Word (if installed)
4) Excel (if installed)
5) Default web browser (if installed)
6) Windows clipboard

If they have been created by the user before, Macro Shortcut Buttons (displaying numbers 1 and 2) will be placed at to the right of the Windows clipboard button.



Column Label 6 Placebo



Unstack Columns: Select destination cell and press Enter

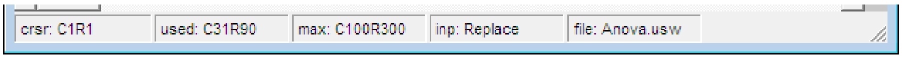
The Input Panel is also used for various prompts and inputs. When the user is prompted for input, for instance, the panel will look slightly different. The two buttons that appear on the left have the following tasks.

Check: OK Simulates pressing <Enter>.

Cross: Cancel. Simulates pressing <Esc>.

3.0.1.3. Status Panel

The bottom line is called the Status Panel and displays information about the Data Processor parameters:



crsr: C1R1 | used: C31R90 | max: C100R300 | inp: Replace | file: Anova.usw

Crsr: The two numbers displayed are the coordinates of the active cell in the data matrix; first the column number and then the row number.

Used: This gives the largest column number containing data or formula, followed by the largest row number in use.

Max: Current dimensions of the data matrix, i.e. the largest number of columns and rows that can be used. These dimensions can be reset from Tools → Options → Memory Management.

Inp: This refers to the current input mode. It can be either Replace or Append which is set from Tools → Options → Spreadsheet.

File: This shows the name of the data file that is currently in Data Processor.

3.0.2. Data Types

It is possible to enter numeric, string, date or time data and missing values. Numeric data can be integers or double precision floating point numbers, including exponential numbers. String Data can be any string expression with practically no restrictions on the number of characters in each cell. Dates can be entered in any international format and the program automatically works out the day of the week.

Columns containing numeric, string and date / time data (as well as function columns) can be displayed in different colours (see 2.4.1.8. Colours).

To input data from keyboard into a cell, position the active cell on the cell and type in the data. The keystrokes are typed into the cell directly.

To enter a column or row label, double-click on the label. A text editor will be placed on the Input Panel. Enter or edit the label, and then click <Enter/OK>.

Data already entered into a cell can be edited if the Cell Edit mode is set to Append (see 2.4.1.7.4. Cell Editing). In this way, numeric, string, date or time data can be edited without having to retype the whole expression.

Click anywhere outside the cell or press <Enter/OK> and the newly typed data will become a part of the spreadsheet. The cell may not display the data in the form in which it was typed-in. For instance, when you enter a date, the program will instantly interpret this and print the day of the week alongside it (see 2.4.1.5.3. Date Format). When you enter positive numbers, they are displayed with a leading space.

If <Escape/Cancel> is pressed input (or editing) will be terminated and the previous content of the cell will be restored.

3.0.2.1. Numeric Data

It is possible to enter numbers in any format, including decimal, integer and exponential formats. However, expressions including mathematical operations cannot be entered. For this purpose use the Formula → Calculate procedure. If the number entered is too small or too large, then it will be displayed in exponential format.

The number typed in may be a double precision number with up to fifteen digits. Although any floating point numbers or integers between -1E+300 and 1E+300 can be entered, the user should be careful to process data within -1E+30 and 1E+30 range in order to prevent some unexpected number overflow problems.

In some cases (e.g. exponential numbers) fewer digits may be displayed in a cell than are internally stored, owing to the format of the display.

3.0.2.2. String Data

In order to input string (or character) data, bring the active cell to the desired position and type the expression. If the first letter typed is not a number (a floating point, or a plus or a minus sign), then the program will assume that the cell will contain String Data. If this cell is the first one typed into a blank column, it will define the column's type as a string column. If the string typed contains less than eight characters (including spaces) then the column will be defined as a short string column (see 3.0.2.2.1. Short Strings). If it has more than eight characters then the column will be defined as a long string column (see 3.0.2.2.2. Long Strings). If the type of column was already defined as one of number, date or time, then its type definition will be changed to string, and all existing cells in the column will be converted into their equivalent string representations. These conversions do not always produce meaningful results. (see 3.4.2.6.1. Data Conversion Functions).

Because Data Processor is a column based spreadsheet, once String Data is entered in a cell, the entire column will be assumed to contain String Data. A column can contain only one of numeric, short string, long string, date or time data, but not a mixture.

To tell Data processor that you wish to enter a formula rather than string data, press <=> first when Data Processor is in *Ready* mode. This is equivalent to selecting Formula → Quick Formula from the menu bar.

It is also possible to define a column as a String Data column by entering a special spreadsheet function. To do this move the active cell to the desired column and select Formula → Quick Formula (or simply press <=>), as if entering a formula (see 3.4. Formula). Then, to define the column as a short string column type either **String** or **Character** and press <Enter/OK> (see 3.4.2.6.1. Data Conversion Functions). It is sufficient to enter the first four characters **Stri** or **Char**. To define the column as a long string column enter the function **Long(n)**, where n is an integer linking the current spreadsheet column with the nth column in the Long String Table. A string column can be redefined as a numeric column by entering either **Number** or **Data**. Again, the first four characters will be sufficient.

The String Data is case-sensitive; i.e., lower case letters are distinguished from upper case letters. In Variable Selection Dialogues, string variables are distinguished from numeric variables in that the letter *C* in their column references is replaced by the character *S* for short strings and *L* for long strings. The colour of String Data columns can be changed from Tools → Options → Colours.

3.0.2.2.1. Short Strings

If no cells in a String Data column contain more than eight characters, then this data type should be preferred to long strings. Use of a short string variable (instead of a long string variable) has the following advantages:

1) Faster execution speed
2) Smaller data file size
3) Reduced memory usage
4) No need to keep track of correspondence with a string table.

Once a column is defined as a short string column, subsequent entries into this column will be truncated to eight characters, if they are longer. In case you have to convert a short string column into a long one, you can use the Data Processor's **Long** function (see 3.4.2.6.1. Data Conversion Functions).This will create a Long String Table entry for the existing strings and convert the type of the column to *long*. Subsequent long string entries can be made manually.

3.0.2.2.2. Long Strings

This data type is inherently different from other data types. While the information for all other data types (including short strings) are held in 8-byte data cells, the long strings are stored in a separate Long String Table. The long data string cells only contain integers referring to entries in this table. The program constantly maintains and updates this correspondence and displays the long strings, instead of the underlying integers, in all parts of the user interface, as well as in output. Therefore, under normal circumstances, you do not need to worry about all this, and may consider long strings as a data type just like others.

On the other hand, the advanced user is provided with means of maintaining absolute control over long strings. The Edit → Long String Table dialogue allows you to edit, enter and rearrange long strings manually.

The main advantage of this approach is to prevent unnecessary loss of memory and storage space by storing all occurrences of repeating sequences of String Data. In an overwhelming majority of cases in statistical analysis, the use of String Data is confined to categorical (factor) variables, where each distinct String Data point (level) is likely to be repeated many times in the same column.

You can convert a long string column into a short one using the Data Processor's **Short** function (see 3.4.2.6.1. Data Conversion Functions). This will truncate all entries longer than 8 characters, and you may thus loose vital variability in data. The Long String Table entry for the existing strings will not be deleted.

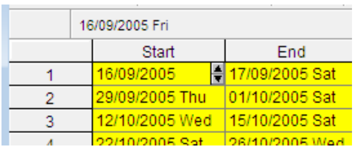
3.0.2.3. Date Data

Dates can be entered in any international format. When a date is entered, UNISTAT automatically works out the day of the week and, optionally, displays it alongside the date (see 2.4.1.5.3. Date Format). It is possible to subtract two columns containing dates to obtain the number of days between the date pairs.

In order to enter a date into the Data Processor the user needs to know the Windows international date format setting and the date separator character. This information should have been automatically configured during the setup of your Windows system. If you are not sure what these settings are, you can have a look at the Tools → Options → Data Export / Import 2 dialogue's Date Format frame. The following are some typical date formats:

UK:	day/month/year
Germany:	day.month.year
US:	month/day/year

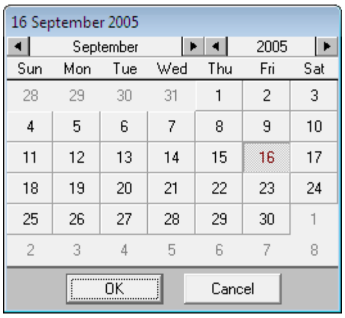
When a string of characters typed into a cell contains the date separator character twice, then the cell will be considered as a date cell. Because the Data Processor is a column based spreadsheet, once Date Data is entered in a cell, then the entire column will be assumed to contain Date Data.



16/09/2005 Fri

	Start	End
1	16/09/2005	17/09/2005 Sat
2	29/09/2005 Thu	01/10/2005 Sat
3	12/10/2005 Wed	15/10/2005 Sat
4	22/10/2005 Sat	26/10/2005 Wed

Double-clicking on a date cell will put the cell in edit mode. Once in edit mode, dates can be incremented or decremented by two scroll buttons. If the small triangle pointer is under the month then the month will increment / decrement when the scroll buttons are clicked. Double-clicking on a date cell in edit mode will invoke a calendar with the current date highlighted. Any date can be selected by clicking on the calendar. Clicking <OK>, control will return to edit mode with the highlighted date entered into the cell.


16 September 2005

Sun	Mon	Tue	Wed	Thu	Fri	Sat
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

In Variable Selection Dialogue, date variables are distinguished from numeric variables in that the letter *C* in their column reference is replaced by *D*.

It is possible to fill in a column with dates automatically using the DAYS function (see 3.4.2.6.2. Date and Time Functions), which has options for the number of days to increment, 5-day working week, etc. The colour of the Date Data columns can be changed from Tools → Options → Colours.

3.0.2.4. Time Data

Time Data is entered using the colon character (:) as the separator. If only one colon is used then it is assumed to separate hours and minutes. If two are used, then they are assumed to separate hours, minutes and seconds. If three colon characters are used, then they are assumed to separate days, hours, minutes and seconds. For instance:

1:15	01:15:00
1:15:17	01:15:17
1:1:15:17	1:01:15:17

In Variable Selection Dialogues, time variables are distinguished from numeric variables in that the letter *C* in their column reference is replaced by *T*.

A column can be filled in with Time Data automatically using one of **Secs()**, **Mins()** or **Hour()** functions (see 3.4.2.6.2. Date and Time Functions). The colour of the Time Data columns can be changed from Tools → Options → Colours.

3.0.2.5. Date-Time Data

In addition to Date Data and Time Data, UNISTAT also supports the Date-Time Data type:

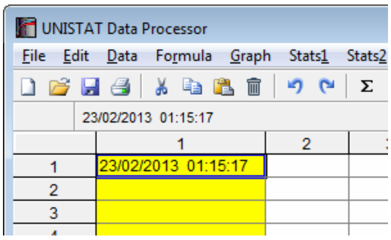
23/02/2013 01:15:17
23.02.2013 19:53:35
02/23/2013 23:19:03

The date and time parts are as described above and they are separated by a space. UNISTAT does not provide a separate date-time data type command, but automatically recognises such data and stores it as Date Data.

Internally, all date and time data are stored as floating point numbers. The integer part represents the day and the decimal part the time. For instance, if you enter the following data into Data Processor:

23/02/2013 01:15:17

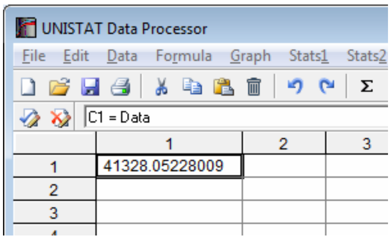
it will recognise this column as a **Date** column and by default, mark it with a yellow background.



23/02/2013 01:15:17

	1	2
1	23/02/2013 01:15:17	
2		
3		

If you press the equal key and enter **Data**, the format of the column will change to numeric and the floating point equivalent of this particular date will be displayed.



C1 = Data

	1	2	3
1	41328.05228009		
2			
3			

To display the same column as a date column, enter:

= Date

To strip the time component from a Date-Time Data column, display the numeric values and enter:

= Int(c1)

and re-define the column as Date. Similarly, you can enter the following function to strip the date component, and re-define the column as Time:

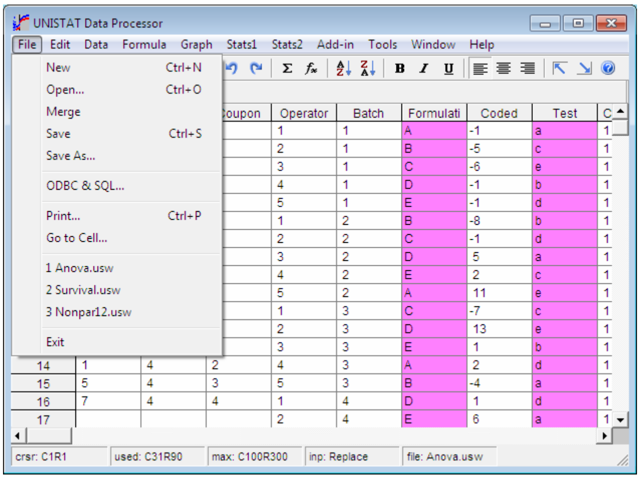
= c1-Int(c1)

3.0.2.6. Missing Data

Any blank cells in a column, below which there are cells containing data, will be considered missing. For more information see 2.4.1.1.4. Missing Data Code.

3.1. File Menu

The first option on the menu bar File provides access to file and printing operations.



			oupon	Operator	Batch	Formulati	Coded	Test	C
				1	1	A	-1	a	1
				2	1	B	-5	c	1
				3	1	C	-6	e	1
				4	1	D	-1	b	1
				5	1	E	-1	d	1
				1	2	B	-8	b	1
				2	2	C	-1	d	1
				3	2	D	5	a	1
				4	2	E	2	c	1
				5	2	A	11	e	1
				1	3	C	-7	c	1
				2	3	D	13	e	1
				3	3	E	1	b	1
14	1	4	2	4	3	A	2	d	1
15	5	4	3	5	3	B	-4	a	1
16	7	4	4	1	4	D	1	d	1
17				2	4	E	6	a	1

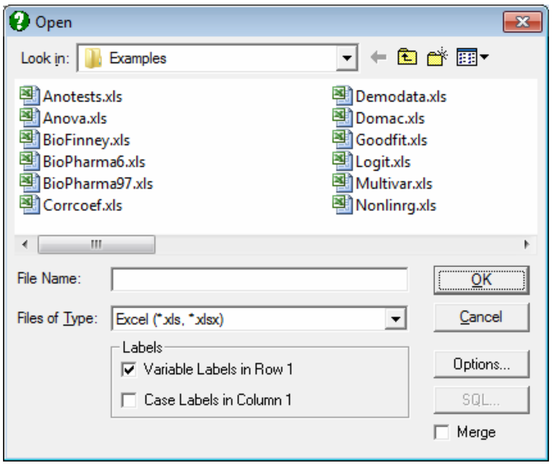
3.1.0. File Formats

Whenever one of Open, Merge or Save As is selected from File, a standard Windows file dialogue will pop up to open or save files in popular PC file formats.

While exporting or importing data, it is possible to include or exclude Column Labels and Row Labels as part of the file. The following two check boxes provided in Open and Save As dialogues, as well as in the Tools → Options → Data Export / Import 2 dialogue (see 2.4.1.6.2. Labels), should be set correctly:

Column labels in row 1
Row labels in column 1

When importing a file, it is essential to know whether the data in row 1 and column 1 of such files contain Column Labels and Row Labels respectively. This information should be consistent with the current settings of the above check boxes. Otherwise the result may be a Device or File Error message and the data may not be read in correctly.



Look in: Examples

Anotests.xls
Anova.xls
BioFinney.xls
BioPharma6.xls
BioPharma97.xls
Corrcoef.xls
Demodata.xls
Domac.xls
Goodfit.xls
Logit.xls
Multivar.xls
Nonlinrg.xls

File Name:

Files of Type: Excel (*.xls, *.xlsx)

Labels
☑ Variable Labels in Row 1
☐ Case Labels in Column 1

☐ Merge

WARNING! *When importing files, you should ensure that current settings of label options are consistent with the actual file. Otherwise the file may not be read correctly (see* 2.4.1.6.2. Labels*).*

UNISTAT can recognise the missing values in files imported from other scientific applications (see 2.4.1.1.4. Missing Data Code).

When importing files, UNISTAT will distinguish between numbers and alphanumeric information. A data point in file can be one of the following:

1) Number (in the range of –E300 to +E300)
2) Column label (unlimited number of characters)
3) Row label (unlimited number of characters)
4) String variable (unlimited number of characters)
5) Date variable (format as in Windows international settings)
6) Time variable (format as in Windows international settings)
7) Missing Data Code (as in Tools → Options → Memory Management)
8) No Data Code (as in Tools → Options → Memory Management)

Being a column-based spreadsheet, UNISTAT's Data Processor expects a variable to contain one of numeric, string, date or time formats, but not a mixture of them. Different data types cannot be mixed within a single column.

If the missing and no data codes in the incoming file are not consistent with the corresponding values in UNISTAT, you can change UNISTAT's codes in Tools → Options → Memory Management dialogue first (see 2.4.1.1.4. Missing Data Code and 2.4.1.1.3. No Data Code).

Parentheses (and) within labels will be filtered out by the program during text file loading or merging. Also, the leading and trailing spaces are trimmed from labels and string variables.

3.1.0.1. UNISTAT Internal Files

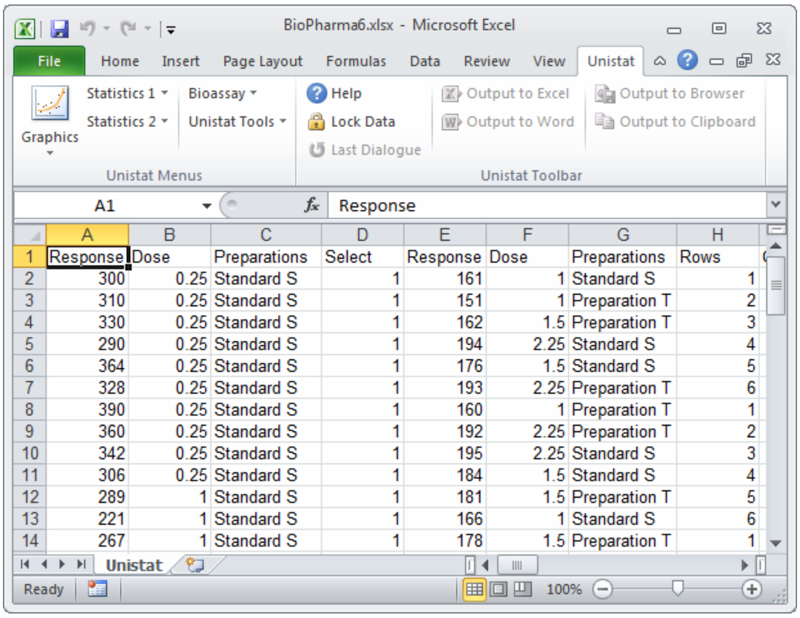
UNISTAT Internal (.USW) files store almost all information the Data Processor contains. This includes all numeric, short and long String Data, missing observations, Column Labels and Row Labels, formulas, formatting information such as column widths, fonts, style, colour, size and justification. UNISTAT internal files are in binary form and thus their contents are not visible in a text editor. However, compared with any other file format supported here (including Excel files), UNISTAT internal files occupy much less disk space and they can be saved, loaded and read more quickly. The user is strongly recommended to use the internal UNISTAT (.USW) file format for routine data storing / retrieving operations.

When loading or merging a UNISTAT internal file, the program will first compare the number of columns and rows in the file with the dimensions of the data matrix. If the dimensions of the file are too large for the present matrix, the program will display the minimum number of columns and rows required for this data file. In this case, select Tools → Options → Memory Management, set the spreadsheet dimensions to more than that required by the data file and try loading it again.

When saving a UNISTAT internal file it is possible to select a subsample of rows to be saved in file rather than saving all rows in the Data Processor. A check box is provided in Tools → Options → Data Export / Import 1 → Options dialogue to enable this option.

3.1.0.2. Excel Files

UNISTAT can read Excel files, retrieving all numeric, string, date and time data, cell formats, and values of formula cells, but not the formulas themselves. Files can be saved in Excel 97 – 2003 (.XLS) and 2007-2010 (.XLSX, .XLSM, .XLSB) formats.



	A	B	C	D	E	F	G	H
1	Response	Dose	Preparations	Select	Response	Dose	Preparations	Rows
2	300	0.25	Standard S	1	161	1	Standard S	1
3	310	0.25	Standard S	1	151	1	Preparation T	2
4	330	0.25	Standard S	1	162	1.5	Preparation T	3
5	290	0.25	Standard S	1	194	2.25	Standard S	4
6	364	0.25	Standard S	1	176	1.5	Standard S	5
7	328	0.25	Standard S	1	193	2.25	Preparation T	6
8	390	0.25	Standard S	1	160	1	Preparation T	1
9	360	0.25	Standard S	1	192	2.25	Preparation T	2
10	342	0.25	Standard S	1	195	2.25	Standard S	3
11	306	0.25	Standard S	1	184	1.5	Standard S	4
12	289	1	Standard S	1	181	1.5	Preparation T	5
13	221	1	Standard S	1	166	1	Standard S	6
14	267	1	Standard S	1	178	1.5	Preparation T	1

As workbook files may contain more than one worksheet, when loading such a file UNISTAT will list the worksheets in the file and allow loading one worksheet at a time. If the workbook contains multiple worksheets, and you wish to load more than one, then you may select the File → Merge option to place the Excel worksheets into different parts of the Data Processor.

WARNING! *When importing files, you should ensure that current settings of label options are consistent with the actual file. Otherwise the file may not be read correctly (see* 2.4.1.6.2. Labels*).*

String and numeric data should not be mixed in columns, with the exception of Column Labels in row 1. Please read the beginning of this section File Formats and sections 2.4.1.5. Data Export / Import 1 and 2.4.1.6. Data Export / Import 2 for the rules regarding reading and writing Excel Files.

If the file to be imported contains string variables, they will be read according to the settings in Tools → Options → Data Export / Import 1 → String Import Options.

Blank cells within data and cells containing an asterisk (*) character will be read as missing data. If the incoming file has a numeric Missing Data Code, you can change UNISTAT's internal missing data code to recognise the missing values in the file (see 2.4.1.1.4. Missing Data Code).

3.1.0.3. Lotus 1-2-3 Files

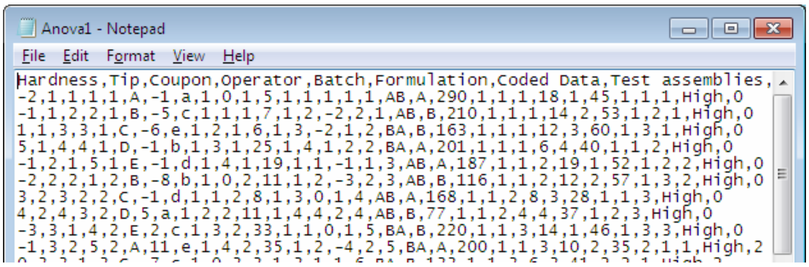
UNISTAT can write the entire contents of the Data Processor into a .WK1 file, including Column Labels and Row Labels, numeric and String Data and cell formats, but excluding formulas. It can also load or merge .WK1, .WKS and .WK3 (Release 3) files, retrieving all numeric and String Data, cell formats, and values of formula cells, but will not load formulas themselves.

As Release 3 files may contain more than one worksheet, when loading such a file UNISTAT will prompt for the number of the worksheet. Enter the number of worksheet you wish to load. If the Release 3 file contains only one worksheet, then its number will be 1. If the file contains multiple worksheets and you wish to load more than one, you are recommended to use the File → Merge option to place worksheets into different parts of the Data Processor.

Other aspects of reading and writing Lotus files are the same as in Excel Files.

3.1.0.4. Delimited Text Files

Delimited text files (also known as CSV files) are commonly used for exporting and importing data to and from database and spreadsheet applications. In English language Windows systems data values are separated by a comma (the field delimiter) and String Data are enclosed within double quotation marks if they contain a comma. In non-English Windows systems the semi-colon character ';' may be used as a field delimiter.



```
Hardness,Tip,Coupon,Operator,Batch,Formulation,Coded Data,Test assemblies,
-2,1,1,1,1,A,-1,a,1,0,1,5,1,1,1,1,1,AB,A,290,1,1,1,18,1,45,1,1,1,High,0
-1,1,2,2,1,B,-5,c,1,1,1,7,1,2,-2,2,1,AB,B,210,1,1,1,14,2,53,1,2,1,High,0
1,1,3,3,1,C,-6,e,1,2,1,6,1,3,-2,1,2,BA,B,163,1,1,1,12,3,60,1,3,1,High,0
5,1,4,4,1,D,-1,b,1,3,1,25,1,4,1,2,2,BA,A,201,1,1,1,6,4,40,1,1,2,High,0
-1,2,1,5,1,E,-1,d,1,4,1,19,1,1,-1,1,3,AB,A,187,1,1,2,19,1,52,1,2,2,High,0
-2,2,2,1,2,B,-8,b,1,0,2,11,1,2,-3,2,3,AB,B,116,1,1,2,12,2,57,1,3,2,High,0
3,2,3,2,2,C,-1,d,1,1,2,8,1,3,0,1,4,AB,A,168,1,1,2,8,3,28,1,1,3,High,0
4,2,4,3,2,D,5,a,1,2,2,11,1,4,4,2,4,AB,B,77,1,1,2,4,4,37,1,2,3,High,0
-3,3,1,4,2,E,2,c,1,3,2,33,1,1,0,1,5,BA,B,220,1,1,3,14,1,46,1,3,3,High,0
-1,3,2,5,2,A,11,e,1,4,2,35,1,2,-4,2,5,BA,A,200,1,1,3,10,2,35,2,1,1,High,2
```

UNISTAT employs a simple rule in deciding which character is to be used as a delimiter in CSV files. If the decimal numbers use the dot character as the floating point, then comma is used as the CSV field delimiter. Otherwise, the semicolon character is used. You can override this automatic choice by entering the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

```
CSVdelimiter=x
```

where x can be any character to be used as the field delimiter.

If the file to be imported contains string variables, they will be read according to the settings in Tools → Options → Data Export / Import 1 → String Import Options.

Another characteristic of the CSV files is that one row in the file is read as one row of the spreadsheet. In other words, the carriage return is used as a row delimiter. UNISTAT supports other text file formats where this limitation is not present (see 3.1.0.5.1. Free Format Text Open).

3.1.0.5. Free Format Text Files

As users of scientific applications know well, the delimited text format is not sufficiently powerful to read and write all kinds of text files.

In UNISTAT, we classify free format text files further according to the sequence of data written into them. Here we make a distinction between files written *by row* (i.e. the first rows of all columns first and then the second rows of all columns etc.) and *by column* (i.e. all rows in the first column first and then all rows in the second column, etc.). These two ways of sequencing the data are also referred to as *by observation* and *by variable*. The *by row* format can sometimes be referred to as the *matrix* format.

The user can select tab, comma, semicolon or any other character as a field delimiter from Tools → Options → Data Export / Import 2 → Text Field Delimiters dialogue. However, as distinct from the CSV format, the end of line character (carriage return) does not mean the end of a row of data. This will be explained in more detail in the following sections.

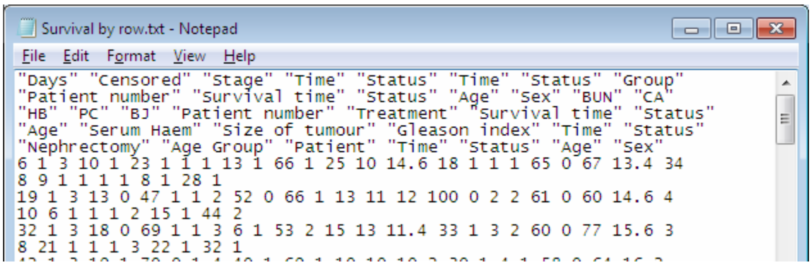
UNISTAT can load and save Column Labels and Row Labels as part of text files, both in *by row* and *by column* formats. Whether Column Labels or Row Labels will be read or saved as part of text files is controlled from Tools → Options → Data Export / Import 2 → Labels.

UNISTAT assumes that a free format file will have a .TXT extension. However, this can be changed to any other extension during loading or saving (see 3.1.2. Open).

If the file to be imported contains string variables, they will be read according to the settings in Tools → Options → Data Export / Import 1 → String Import Options.

3.1.0.5.1. Free Format Text Open

By Row: If the file is being loaded by row, the number of columns must be specified by the user so that the program will know when to finish reading the first row and continue with the second. This is a superior method than used by the CSV format, which can only read one line of the file as one row of the data matrix. When the rows of the incoming file are sufficiently long, they may take several lines each in the text file. In this case, users of packages which support the CSV format only are faced with the extremely unpleasant task of moving rows to their correct positions in the spreadsheet.



```
"Days" "Censored" "Stage" "Time" "Status" "Time" "Status" "Group"
"Patient number" "Survival time" "Status" "Age" "Sex" "BUN" "CA"
"HB" "PC" "BJ" "Patient number" "Treatment" "Survival time" "Status"
"Age" "Serum Haem" "Size of tumour" "Gleason index" "Time" "Status"
"Nephrectomy" "Age Group" "Patient" "Time" "Status" "Age" "Sex"
6 1 3 10 1 23 1 1 1 13 1 66 1 25 10 14.6 18 1 1 1 65 0 67 13.4 34
8 9 1 1 1 1 8 1 28 1
19 1 3 13 0 47 1 1 2 52 0 66 1 13 11 12 100 0 2 2 61 0 60 14.6 4
10 6 1 1 1 2 15 1 44 2
32 1 3 18 0 69 1 1 3 6 1 53 2 15 13 11.4 33 1 3 2 60 0 77 15.6 3
8 21 1 1 1 3 22 1 32 1
```

If the Column Labels in Row 1 box is checked in Tools → Options → Data Export / Import 2 → Labels dialogue, the program will read the first n entries (i.e. the number of columns as specified by the user) as Column Labels. This is irrespective of whether these entries are enclosed within text qualifiers or not.

If the Row labels in column 1 box is checked (see 2.4.1.6.2. Labels), the program will read the next entry as the first row label. Again, this is irrespective of whether this entry is enclosed within text qualifiers or not.

WARNING! *When importing files, you should ensure that current settings of label options are consistent with the actual file. Otherwise the file may not be read correctly (see 2.4.1.6.2. Labels).*

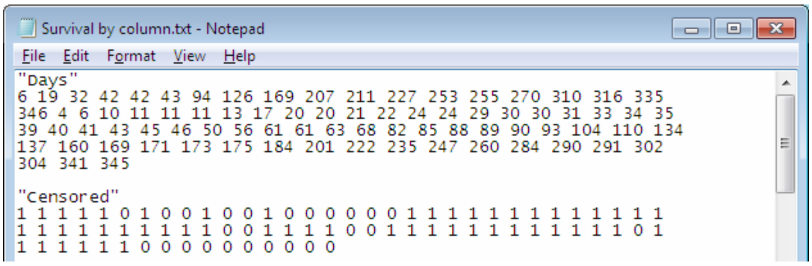
The program then goes on reading the next n entries as the first row of data. UNISTAT will work out the type of data in row 1 (or row 2, if the first row contains Column Labels) and it will assume that the rest of the rows will conform to this. If a column starts with missing values, then UNISTAT will determine the type of such columns from their first non-missing value.

IMPORTANT! *When importing files, UNISTAT determines the type of column (i.e. numeric, string, date or time) from the first row of data in the file.*

Special attention must be paid to text files which are to be read by row, and where the field delimiter is Space(s). While UNISTAT can recognise a blank cell as a missing cell with any other field delimiter, it is impossible to distinguish a blank cell when the field delimiter is Space(s). Also, for the same reason, in such files all columns must have an equal number of observations. If the file contains columns with unequal lengths, then the data will not be read into the correct cells after the shortest column is fully read into the data matrix. In order to prevent this happening, shorter columns should be padded either with the No Data Code or the Missing Data Code.

By Column: When the *by column* option is selected, data is read into a column cell by cell. Variables are assumed to be separated by a pair of double quotes. If the Column Labels in Row 1 box is checked in Tools → Options → Data Export / Import 2 → Labels dialogue, the program will read the String Data enclosed within the specified text qualifiers as Column Labels.

It is not possible for the Data Processor to read *by column* text files containing string variables, that is, String Data apart from Column Labels and Row Labels.



```
"Days"
6 19 32 42 42 43 94 126 169 207 211 227 253 255 270 310 316 335
346 4 6 10 11 11 11 13 17 20 20 21 22 24 24 29 30 30 31 33 34 35
39 40 41 43 45 46 50 56 61 61 63 68 82 85 88 89 90 93 104 110 134
137 160 169 171 173 175 184 201 222 235 247 260 284 290 291 302
304 341 345

"Censored"
1 1 1 1 1 0 1 0 0 1 0 0 1 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1
1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
```

If the Row labels in column 1 box is checked (see 2.4.1.6.2. Labels), the program will read all alphanumeric data at the beginning of the file as Row Labels first, as if the Row Labels were the column zero of the data matrix. When a numeric input is found then the program will put the last label read into the first column and start reading numbers. When a double quote is encountered the program will assume that the current column is read completely and put the contents of the string into the label of the next column. The subsequent numeric information is placed into the new column until another double quote is found.

3.1.0.5.2. Free Format Text Save As

The maximum width of lines in a text file saved by UNISTAT is determined by the Width parameter defined in Tools → Options → Output → Text Margins. That is, if the output width is set to 80, then the width of the file will also be 80 characters. The maximum allowed width is 32,000 characters per line.

In order to obtain the desired Number Format in the text file, Data → Format Columns option can be used before selecting File → Save.

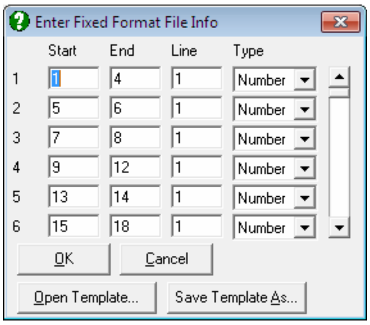
In numeric data columns, cells containing missing or *no data* will be represented by their numeric values, i.e. the codes as displayed in Tools → Options → Memory Management dialogue. For string, date or time data columns, the missing and *no data* cells will be saved as blank, i.e. “”.

By Row: If the data is to be saved *by row*, and column lengths are not equal, all shorter columns will be added a sufficient number of No Data Codes to complete the matrix elements.

By Column: A file saved in column format can be read directly back into the Data Processor, as long as it does not contain string variables.

If the Column Labels in Row 1 box is checked in Tools → Options → Data Export / Import 2 → Labels dialogue, then all Row Labels will be written before any numeric data. Each column will start with its column label followed by all data in that column. In case the length of a column is less than the maximum then *no data* markers will not be written to the end of the column.

3.1.0.6. Fixed Format Text Files



	Start	End	Line	Type
1	1	4	1	Number
2	5	6	1	Number
3	7	8	1	Number
4	9	12	1	Number
5	13	14	1	Number
6	15	18	1	Number

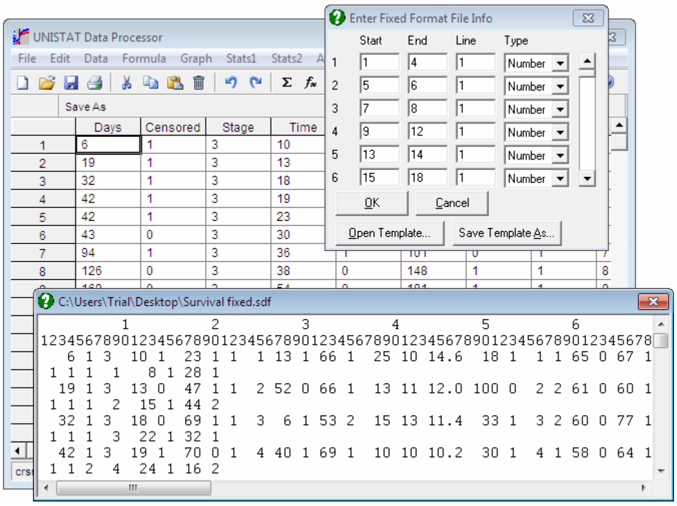
Fixed format text files are similar to free format text files saved *by row*, in that they follow the same sequence, i.e., the first rows of all columns, then the second rows of all columns, etc. However, they are different in that data points need not necessarily be separated by special characters (delimiters). Each column (variable) is allocated a fixed number of characters and thus the characters which belong to a certain column can always be distinguished according to their position in the file. Fixed format text files are also known as non-delimited text files and they can be more compact than free format text files.

UNISTAT assumes that a fixed format file will have a .SDF extension. However, this can be changed to any other extension during loading or saving.

Each line in a fixed format text file is called a record. If there are too many variables to fit in a single line then data belonging to the same row can be continued in the subsequent lines. For example, a file which contains 10 rows of data each of which is in three records will have 30 lines.

3.1.0.6.1. Fixed Format Text Open

After selecting this format, a window for entering and / or editing the column positions will be opened on the upper half of the screen. The view file window on the lower part will remain open. At any stage, you can click on the View File window to see the contents of the file.



	Days	Censored	Stage	Time
1	6	1	3	10
2	19	1	3	13
3	32	1	3	18
4	42	1	3	19
5	42	1	3	23
6	43	0	3	30
7	94	1	3	36
8	126	0	3	38

Enter Fixed Format File Info

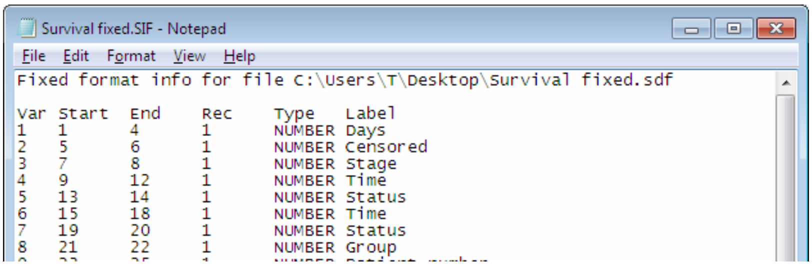
	Start	End	Line	Type
1	1	4	1	Number
2	5	6	1	Number
3	7	8	1	Number
4	9	12	1	Number
5	13	14	1	Number
6	15	18	1	Number

C:\Users\Trial\Desktop\Survival fixed.sdf

```
         1         2         3         4         5         6
123456789012345678901234567890123456789012345678901234567890123456789012345678
   6 1 3  10 1  23 1 1  1 13 1 66 1  25 10 14.6  18 1  1 1 65 0 67 1
 1 1 1  1   8 1 28 1
  19 1 3  13 0  47 1 1  2 52 0 66 1  13 11 12.0 100 0  2 2 61 0 60 1
 1 1 1  2  15 1 44 2
  32 1 3  18 0  69 1 1  3  6 1 53 2  15 13 11.4  33 1  3 2 60 0 77 1
 1 1 1  3  22 1 32 1
  42 1 3  19 1  70 0 1  4 40 1 69 1  10 10 10.2  30 1  4 1 58 0 64 1
 1 1 2  4  24 1 16 2
```

The first column in the upper window indicates the column number of the variable. The subsequent three columns are numeric input fields for the start and end column positions of the variable and its record number respectively. The last field indicates the type of the column and it can be one of Label, Number, String, Date or Time.

When loading a fixed format file which was previously saved by UNISTAT, the program will first look for a fixed format template file, that is, a file containing the information needed to load a fixed format file. Whenever a fixed format file is saved from UNISTAT, a template file is created containing column positions, types and labels of all columns in the main file. Fixed format template files have the same name, but a .SIF extension. If UNISTAT can find an accompanying .SIF file during loading an .SDF file, it will get the necessary information from this file and fill in the input fields of the upper window automatically. The default values may be either accepted or edited. If there is a corresponding template file with a different name or extension, then you can browse and open any template files by clicking on [Open Template].



```
Fixed format info for file C:\Users\T\Desktop\Survival fixed.sdf

Var Start  End    Rec    Type   Label
1   1      4      1      NUMBER Days
2   5      6      1      NUMBER Censored
3   7      8      1      NUMBER Stage
4   9      12     1      NUMBER Time
5   13     14     1      NUMBER Status
6   15     18     1      NUMBER Time
7   19     20     1      NUMBER Status
8   21     22     1      NUMBER Group
9   23     25     1      NUMBER Patient number
```

If a fixed format text file is loaded which was saved from a different application, then there will not be an accompanying template (.SIF) file and, on entry, the fields of the upper window will be empty. You can fill in all the required fields, if necessary switching between the lower and upper windows. Unlike saving a fixed format file, however, the last column in the upper window Type is vitally important here, since it will instruct the program to read data columns as Row Labels, string, numeric, date or time data.

Once the field specifications are entered, you can save this information in a fixed format template (.SIF) file clicking on the [Save Template As] button. Saving a fixed format file will also create an accompanying template file automatically.

It is always possible to create a .SIF file for any fixed format files prior to entering UNISTAT, using a text editor. The best way to do this is to edit an already existing UNISTAT fixed format template (.SIF) file, as the column and row positions of all information in .SIF files are absolutely crucial. If you prefer to fill in the upper window fields interactively, it will be a good idea to save this information in a template file. Alternatively, you can also save the same file immediately after reading it (better with a different name), in order to create a template (.SIF) file for it automatically.

3.1.0.6.2. Fixed Format Text Save As

Saving a fixed format text file is similar to loading one. Again, a window for entering and / or editing the column positions will be opened on the upper half of the screen, and a window will be opened on the lower half displaying contents of the file to be saved. At any stage, you can click on the view file window and view its contents.

The maximum width of lines in a text file saved by UNISTAT is determined by the Width parameter defined in Tools → Options → Output → Text Margins. That is, if the output width is set to 80, then the width of the file will also be 80 characters. The maximum allowed width is 32,000 characters per line.

UNISTAT will first format the data to be saved in fixed format, i.e., either the whole matrix or the highlighted block of cells. It will determine the minimum width of each column in which all data could be presented without a loss of significant digits, and open two windows, one for giving you the opportunity to change column positions interactively and the other to preview the appearance of the file to be saved.

Row Labels will be saved as a part of the file, only if the Column Labels in Row 1 box is checked in Tools → Options → Data Export / Import 2 → Labels dialogue. Column Labels are not saved as part of .SDF files. They will, however, be saved in the accompanying template (.SIF) file (see 3.1.0.6.1. Fixed Format Text Open).

The first column in the upper window indicates the column number of the variable. The next three columns are numeric input fields for the start and end column positions and the record number of the variable respectively. The last column is a drop-down list and indicates the type of the column. It can take one of the following five values: Label, Number, String, Date or Time. To save the file as formatted by the Data Processor just press <Enter/OK>.

The numeric fields in the upper window can be edited to achieve a fixed file format of choice. The field Type, however, is there only for information, and changing it will not make any impact on the format of the file. This information is predetermined by the nature of the data in the Data Processor.

When a fixed format text file is saved, the Data Processor will save the file with a .SDF extension, if one was not supplied. However, the Data Processor will also create a second (template) file with the same name but with a .SIF extension, which contains information on the position, the record number and the label of each variable; in effect, a record of the upper window. The fixed format template (.SIF) files are not only useful for reference, to keep track of which variable is where in the file, but they are also used by the Data Processor to load back a fixed format text file without any need to type the column and record information once again.

3.1.0.7. Data Interchange Format (.DIF) and Symbolic Link Format (.SLK) Sylk Files

These two file types are similar in that they store information in text format (text files) and store individual data points in separate lines. These are usually long, narrow files which can be, in theory, edited by the user. .DIF and .SLK files will save or load almost all information in the Data Processor, including all Column Labels and Row Labels, numeric and String Data and column formats, but excluding formulas.

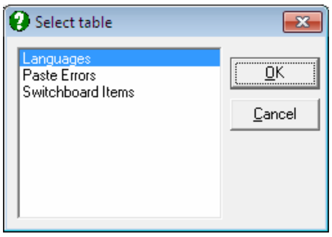
WARNING! *When importing files, you should ensure that current settings of label options are consistent with the actual file. Otherwise the file may not be read correctly (see* 2.4.1.6.2. Labels*).*

If the file to be imported contains string variables, they will be read according to the settings in Tools → Options → Data Export / Import 1 → String Import Options.

3.1.0.8. Database Files

UNISTAT's File → Open and File → Save As dialogues support the Access 97, 2000-2003 (.MDB), 2007-2010 (.ACCDB) and dBase (.DBF) database formats directly. It is possible to open, merge and save files directly in these database formats.

It is also possible to open any database files which are not directly supported by UNISTAT, but for which ODBC drivers are installed in the Windows system. The File → ODBC and SQL option will list all available ODBC drivers in your system.



Languages
Paste Errors
Switchboard Items

If the database contains more than one table, a dialogue will pop up asking for the table to be imported. The list of tables normally excludes the hidden ones. If you wish to have all tables displayed, enter and edit the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

```
ODBCAllTables=1
```

By default, UNISTAT will attempt to read the entire table opened from the File → Open or File → Merge dialogues, i.e. it will try to read all fields and all cases in the table. If the database is too large, then UNISTAT will read only those fields and cases which can fit into the current dimensions of the Data Processor (see 2.4.1.1. Memory Management). Should it be necessary to extract a subset of fields or cases from a large table then it is possible to click on the [SQL] button (see 3.1.6. ODBC and SQL) to execute SQL statements.

Older dBase files support a limited number of fields (columns) and they do not recognise exponential numbers. When saving data into dBase files, all exponential numbers will be truncated. Missing and *no data* cells are saved as -999.

When loading a database table, the field names will only be imported if the Row labels in column 1 box is checked in Tools → Options → Data Export / Import 2.

If the file to be imported contains string variables, they will be read according to the settings in Tools → Options → Data Export / Import 1 → String Import Options.

When creating a database file, by default the field width for Row Labels and long string variables is set to 20. To increase this limit enter and edit the following line in *Unistat65.ini* file under the [Options] group:

```
DBLabelWidth=20
```

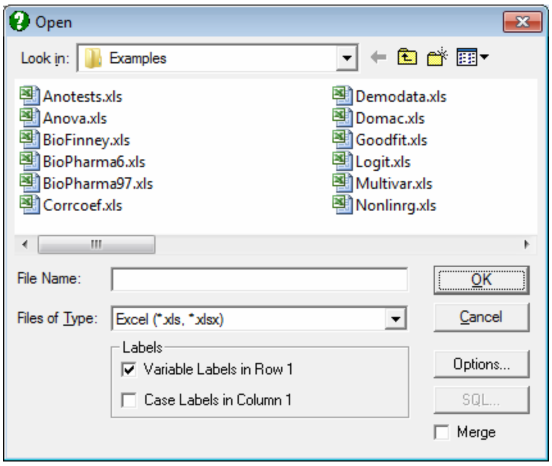
3.1.1. New

In order to clear all data in the Data Processor, all data as well as Column Labels and Row Labels, formulas, etc. select this option.

If the data present in Data Processor was not saved, the program will offer the option to save it first. If no changes were made in data since it was last saved, it will be cleared without a warning.

3.1.2. Open

UNISTAT's file open dialogue is similar to the standard Windows Open dialogue. The Files of Type drop-down list at the bottom left is used to list the supported file formats with their standard extensions.



Look in: Examples

Files of Type: Excel (*.xls, *.xlsx)

☑ Variable Labels in Row 1

□ Case Labels in Column 1

□ Merge

1) UNISTAT (.USW)
2) Excel
3) Lotus
4) Text Delimited
5) Text Free
6) Text Fixed
7) Sylk
8) DIF
9) Access
10) dBase
11) FoxPro

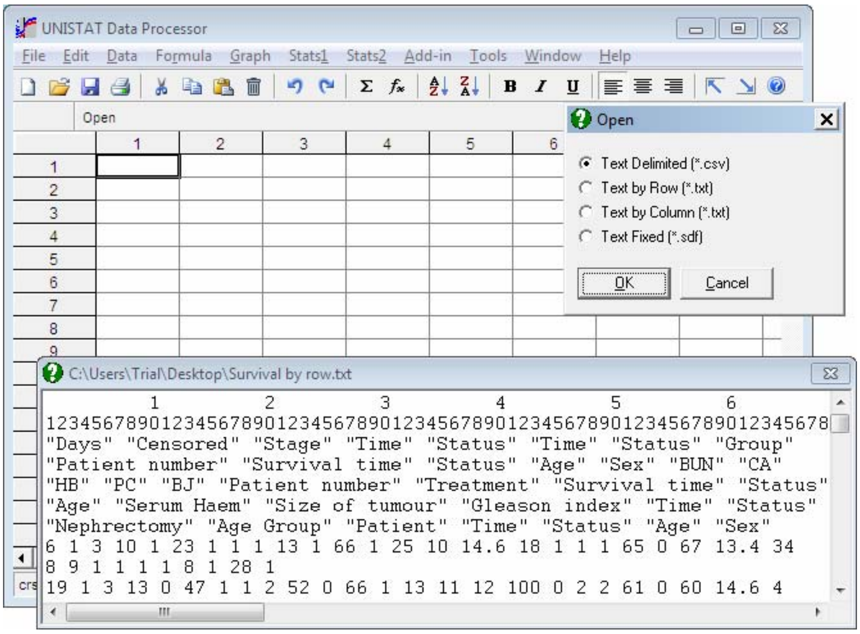
12) Paradox
13) HTML

When this dialogue is opened, it will display UNISTAT internal files by default. By entering and editing the following line in *Unistat65.ini* file under the [Options] group, you can change the default file format displayed:

```
DefaultFileFormat=1
```

where the file format numbers are as in the above list.

When a file is opened, UNISTAT first checks the first few bytes of the file to determine its format. Therefore, if you are not sure about the format of a data file, you can select the All Files (*.*) option on top of the list and in most cases UNISTAT will be able to read the file without any problems. If UNISTAT cannot determine the format of the file from its few bytes, or the file is in text format, then the contents of the file is displayed in a window for inspection. Another window pops up simultaneously allowing selection of one of the four text file formats, i.e. delimited, by row, by column or fixed (see 3.1.0. File Formats).


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help
Open
1 2 3 4 5 6
1 2 3 4 5 6 7 8 9
crs

Open

● Text Delimited (*.csv)
○ Text by Row (*.txt)
○ Text by Column (*.txt)
○ Text Fixed (*.sdf)

OK Cancel

C:\Users\Trial\Desktop\Survival by row.txt

```
              1         2         3         4         5         6
12345678901234567890123456789012345678901234567890123456789012345678
"Days" "Censored" "Stage" "Time" "Status" "Time" "Status" "Group"
"Patient number" "Survival time" "Status" "Age" "Sex" "BUN" "CA"
"HB" "PC" "BJ" "Patient number" "Treatment" "Survival time" "Status"
"Age" "Serum Haem" "Size of tumour" "Gleason index" "Time" "Status"
"Nephrectomy" "Age Group" "Patient" "Time" "Status" "Age" "Sex"
6 1 3 10 1 23 1 1 1 13 1 66 1 25 10 14.6 18 1 1 1 65 0 67 13.4 34
8 9 1 1 1 1 8 1 28 1
19 1 3 13 0 47 1 1 2 52 0 66 1 13 11 12 100 0 2 2 61 0 60 14.6 4
```

If there is already data in the spreadsheet (i.e. if the used display on the Status Panel is not 0, 0) then a prompt will ask whether to clear the existing data first.

WARNING! *Opening a file will clear all data present in the spreadsheet.*

In order to keep the existing data and merge the incoming file, check the [Merge] button under the command buttons (see below). Press <Enter/OK> to proceed with loading. <Escape/Cancel> will abort the Open procedure.

WARNING! *When importing files, you should ensure that current settings of label options are consistent with the actual file. Otherwise the file may not be read correctly (see* 2.4.1.6.2. Labels*).*

Options: This command button will invoke the Data Export / Import 2 dialogue which is also available under the Tools → Options menu.

Merge: Check this box to read the incoming file without clearing existing data first. This is equivalent to selecting File → Merge from the menu.

3.1.3. Merge

A file on disk can be merged into the Data Processor without clearing the data that already exist in the worksheet. The position of data to be merged from the file is controlled by means of the position of the active cell. The top left hand corner of the data matrix in file will be placed at the active cell. Therefore, the user needs to make the desired cell active before selecting the merge option.

If there are empty cells above any column read from the file then these will be treated as missing and filled with missing data markers. In case of a clash of existing data and the new data from the file, the existing data will be overwritten.

When a file is merged, all its formulas will be converted to **Data** (see 3.4.2.6.1. Data Conversion Functions).

3.1.4. Save

This will save the existing file in Data Processor without asking for a file format or a new file name. The file will be saved in the last format selected. If the data were entered from keyboard, rather than file, i.e. when there is no file name assigned yet, then the Save As dialogue will be activated. For details of saving different format files see 3.1.0. File Formats.

3.1.5. Save As

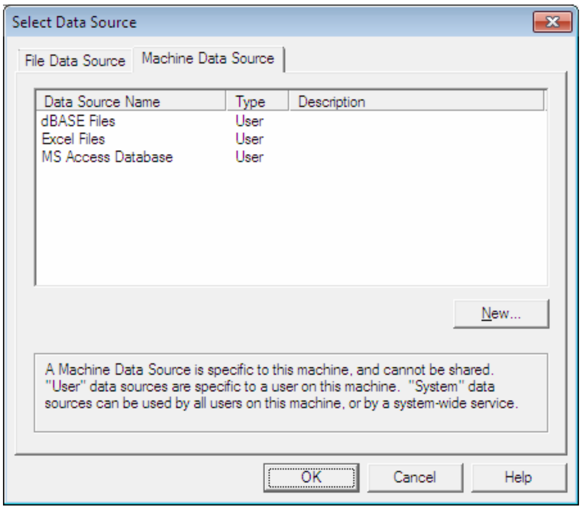
A dialogue for saving files will be opened. Either select an already existing file from the list or enter a new one. The program will check whether there already exists a file with the same name. If there is one, then you will be asked to choose between going ahead with saving and replacing the contents of the old file, or cancelling the procedure.

When the Save As option is selected, the user must select one of the formats available in the Files of Type drop-down list. The default is UNISTAT internal (.USW) format. For details of saving different format files see 3.1.0. File Formats.

3.1.6. ODBC and SQL

It is possible to open any database files which are not directly supported by UNISTAT, but for which ODBC drivers are installed in the Windows system.

Like the database export / import options available in File → Open and File → Merge dialogues (see 3.1.0.8. Database Files), an SQL dialogue will be available to extract a subset of fields or cases from a large database.



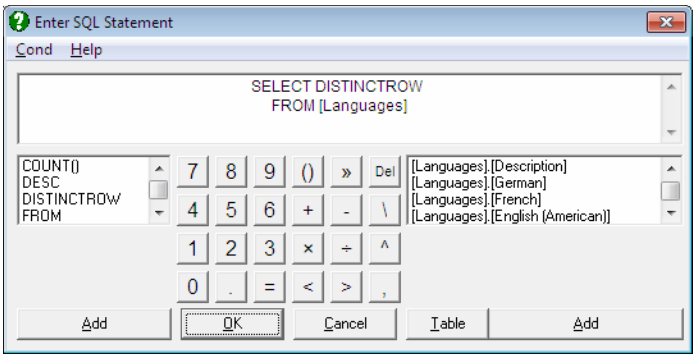
Data Source Name	Type	Description
dBASE Files	User	
Excel Files	User	
MS Access Database	User	

A Machine Data Source is specific to this machine, and cannot be shared. "User" data sources are specific to a user on this machine. "System" data sources can be used by all users on this machine, or by a system-wide service.

When the File → ODBC and SQL option is first selected, a dialogue will display a list of all registered ODBC drivers in your system. Upon selecting a particular entry, the program may respond in one of the following two ways:

1) If the ODBC driver selected has a binding with a particular database file, the database will be opened immediately and UNISTAT's SQL dialogue will pop up.

2) If the ODBC driver does not have a binding with any files, then the standard file open dialogue will pop up prompting you to select a database file. If the selected file has the correct format, then it will be opened and UNISTAT's SQL dialogue will pop up.



```
SELECT DISTINCTROW
FROM [Languages]
```

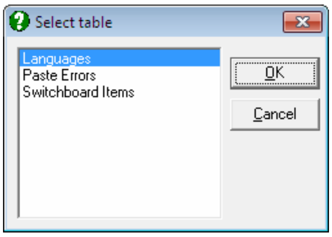
The SQL dialogue allows construction of SQL statements to select a subsample of fields and cases from a database. The dialogue consists of the following components:

SQL Input Box: The user can type any SQL statements directly into this box. If the statement has already been constructed elsewhere, it is also possible to copy it to the clipboard and then paste into this box. You should have a prior knowledge about the field names of the database to type in an SQL statement directly. If this is not the case, use the [Table] button to display the field names. See List of Field Names below.

SQL Commands List: This list on the left is provided to facilitate entering SQL commands into the text box without having to re-type them. Items selected will be pasted into the input box at the cursor position. Although only a few commonly used commands are listed, all commands supported by the particular driver can be typed in by hand.

Operation Buttons: These buttons are used to paste numbers and arithmetic operators into the input box at the current cursor position.

List of Field Names: This list will be empty on entry. Clicking the [Table] button, a dialogue will pop-up displaying the tables available in the database. Note that this may not work for all types of external databases.

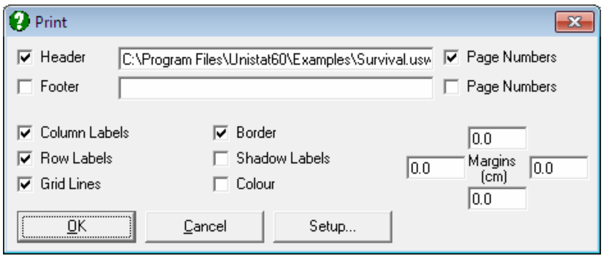


Select a table and click [OK]. The field names of the selected table will appear in the list. They will be enclosed within square brackets by the program, in order to prevent *bad* names causing problems. A highlighted list item can be pasted into the input box at the current cursor location by clicking the [Add] button. It is possible to open more than one table simultaneously.

Please note, once again, that it is possible to either type the SQL statement in the input box directly, or compose it by double-clicking on the desired list items. If the statement entered is valid (or compatible with the particular database you have opened) the outcome will be imported into Data Processor.

3.1.7. Print

To print the entire data matrix either switch off any highlighted block of cells, or highlight the whole spreadsheet by clicking on the top left corner. In order to print a block of cells, highlight the block first. Then select File → Print. The print dialogue will be displayed. The following options are available:



☑ Header C:\Program Files\Unistat60\Examples\Survival.usw ☑ Page Numbers
☐ Footer ☐ Page Numbers
☑ Column Labels ☑ Border
☑ Row Labels ☐ Shadow Labels
☑ Grid Lines ☐ Colour
Margins (cm) 0.0 0.0 0.0 0.0

Setup Button: This provides access to the Windows Printer Setup dialogue.

Header: A line of text can be typed in. This will be displayed on top of every page. By default, UNISTAT suggests the current file name, including its path.

Footer: A line of text can be typed in. This will be displayed on the bottom margin of every page.

Page Numbers: If the box which is to the right of the header text field is checked, page numbers will be printed at the top right corner of the page. Likewise, if the box which is to the right of the footer text field is checked, page numbers will be printed at the bottom right corner of the page.

Column Labels: Check this box *off* not to print Column Labels.

Row Labels: Check this box *off* not to print Row Labels.

Grid Lines: Check this box *off* not to print grid lines.

Border: Check this box *off* not to print borderlines.

Shadow Labels: Check this box *on* to print Column Labels and Row Labels on a tinted background.

Colour: Check this box *on* to print data in colour on a colour printer.

Margins: Print margins can be set via the group of four text fields on bottom right of the Print dialogue. The values displayed are in terms of the default Windows units (i.e. cm, inches, etc.). The default margins (0, 0, 0, 0) will give a print area as defined in Windows Printer Setup dialogue. If you wish to print on a smaller area, edit these fields as necessary.

If the width of the specified data range is greater than the width of the page as specified in Windows Printer Setup dialogue, then the program will divide the range into a number of blocks so that each one fits in one page. Similarly, output will be separated into an appropriate number of rows so that there is sufficient margin left at the top and the bottom of the page.

All information displayed on the File → Print dialogue can be stored by the program so that changes made will become default automatically in subsequent sessions. The program will ask whether you wish to save the changes if any have been made. If changes are not saved, they will be effective during the current UNISTAT session only.

3.1.8. Goto

By means of this the active cell can be moved to a distant position faster than by using the arrow keys or the mouse. After selecting File → Goto type in the column number and press <Enter/OK>, and then type in the row number and press <Enter/OK>. The active cell will then be located at the specified position. If the cell is not within the bounds of the present display screen, the display will be repositioned with the required cell being placed as near as possible to the centre.

3.1.9. Auto Load

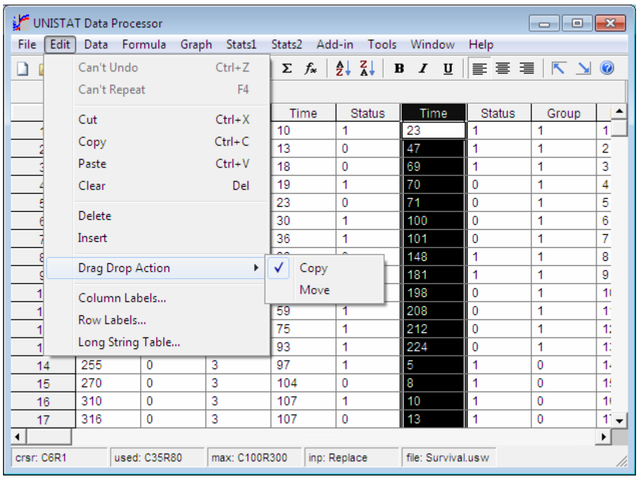
On entry, there will be no items displayed between Goto and Exit in the File menu. As you open and save data files, up to nine most recently used file names will be displayed between Goto and Exit, which are assigned short keys from 1 to 9. More recently used files will be placed higher in the list. It is possible to load a file from the list by simply pressing its short key, or by clicking on it, regardless of its format or location.

The file names in the list are stored in a text file FILELOG.TXT, which resides in the user's private working folder. To reset the contents of Auto Load list, simply delete this file.

3.1.10. Exit

This will terminate the current UNISTAT session and close the program. If you have entered new data or edited the contents of a spreadsheet, the program will first ask whether you wish to save your data.

3.2. Edit Menu



Most of the Edit menu options are also accessible from a pop up menu that can be activated by clicking the right mouse button.

3.2.1. Undo

A one level undo facility is available for all Data Processor operations that modify data. This option will restore the state of the spreadsheet before the last edit. The undo buffer is cleared when a graphics or statistics procedure is executed.

3.2.2. Redo

The last data editing operation is repeated. The redo buffer is cleared when you exit Data Processor to execute a procedure.

3.2.3. Cut

The highlighted range is copied to the Windows clipboard and the source range erased.

When the length of a column is greater than the end row number of the range, the erased portion of these columns will be filled with missing data.

Data copied to the clipboard is tab delimited; i.e. columns are separated by tab characters (ASCII 9) and rows are separated by carriage return characters (ASCII 13 + 10).

3.2.4. Copy

This is like cut, in that the highlighted range is copied to the clipboard, but the source range is not erased.

Data copied to the clipboard is tab delimited; i.e. columns are separated by tab characters (ASCII 9) and rows are separated by carriage return characters (ASCII 13 + 10).

3.2.5. Paste

The rules for pasting the contents of the clipboard are similar to those for opening CSV files (see 3.1.0.4. Delimited Text Files). The field and text delimiters selected in the Tools → Options → Data Export / Import 2 dialogue (see 2.4.1.6.1. Text Field Delimiters) will be in effect and they can be changed to accommodate different types of text in the clipboard. The default field delimiter is the tab character.

To obtain the best results observe the following rules:

1) If the data in the clipboard contains Column Labels then highlight the columns where this data is to be pasted. Otherwise Column Labels will be put into the first row and the data will be corrupted.

   Likewise, if you wish to place the first column of the clipboard data into the Row Labels, highlight a few rows first.

   If you wish to place the first row and the first column of the clipboard data into Column Labels and Row Labels simultaneously, highlight the entire spreadsheet by clicking on the top-left corner (cell 0, 0) first.

2) When using space as a column separator, either alone or alongside other delimiters, take care that Column Labels or String Data do not contain any spaces. For instance, if you attempt to copy a table into the clipboard with Row Labels like *Fixed Capital*, then *Fixed* and *Capital* will be put into separate cells unless they are enclosed within text qualifiers. This may cause a misalignment of columns of the table.

3) Do not cut or copy horizontal lines, etc. of a table in order to paste it into UNISTAT's spreadsheet subsequently. These may cause confusion on the part of the program whether a column contains string or numeric data.

4) Although it is always possible to cut or copy text from UNISTAT's own Output Window and paste it into the spreadsheet for further analysis, remember that these numbers will probably have been formatted for the output. We therefore strongly recommend that output tables are saved to spreadsheet by clicking on the UNISTAT button provided on the Data Processor's Output Medium Toolbar. In this case the floating point numbers will be saved with the full 15 digits of precision. Also, in this way large tables will be saved in a single block - which may have been divided into several blocks in the Output Window in order to fit into the specified output width (see 2.4.1.2.3. Text Margins).

3.2.6. Clear

Highlighted range will be erased and the data will be lost. When the length of a column is greater than the end row number of the range, the erased parts of these columns will be filled with missing data.

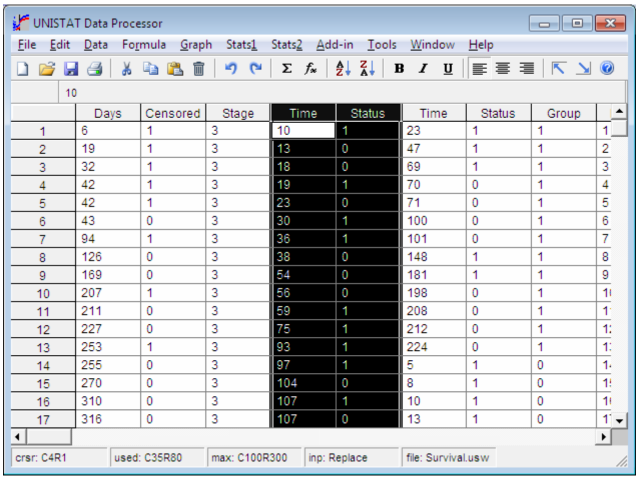
3.2.7. Delete

A block of columns or rows can be deleted. The following actions are possible, depending on what type of highlighting you have had prior to selecting this option:

No block is highlighted: In this case, when Edit → Delete is selected, a daughter menu will provide two further options: Columns and Rows. If you select Columns, then the current column will be deleted without further notice. Likewise, the current row can be deleted by selecting Rows.

A block of cells is highlighted: As above, it is possible to select one of the Columns or Rows options to delete the range of columns or rows which the highlighted area covers.

A block of columns is highlighted: In this case the program understands that you wish to delete columns and accordingly it will not request further information Columns or Rows. The column range will be deleted and all columns to the right of the range will be moved to the start of the range. The highlight will not be switched off.



10

	Days	Censored	Stage	Time	Status	Time	Status	Group	
1	6	1	3	10	1	23	1	1	1
2	19	1	3	13	0	47	1	1	2
3	32	1	3	18	0	69	1	1	3
4	42	1	3	19	1	70	0	1	4
5	42	1	3	23	0	71	0	1	5
6	43	0	3	30	1	100	0	1	6
7	94	1	3	36	1	101	0	1	7
8	126	0	3	38	0	148	1	1	8
9	169	0	3	54	0	181	1	1	9
10	207	1	3	56	0	198	0	1	1
11	211	0	3	59	1	208	0	1	1
12	227	0	3	75	1	212	0	1	1
13	253	1	3	93	1	224	0	1	1
14	255	0	3	97	1	5	1	0	1
15	270	0	3	104	0	8	1	0	1
16	310	0	3	107	1	10	1	0	1
17	316	0	3	107	0	13	1	0	1

crsr: C4R1 | used: C35R80 | max: C100R300 | inp: Replace | file: Survival.usw

As in the case of inserting a range of columns, all formulas will be updated so that they will refer to the correct columns in their new positions. If, however, some of the columns which are to be deleted are referred to in functions, then their column numbers will be replaced by an exclamation mark (!). If such a formula is recomputed by the Compute Matrix procedure, or it is used without editing first, it will generate a Syntax Error message.

A block of rows is highlighted: In this case the program understands that it must delete rows and accordingly it will not request further information, Columns or Rows. The highlighted range of rows will be deleted and all rows below (including Row Labels) will be shifted up. All column lengths and thus the maximum number of rows used will be updated. Formulas will not be lost.

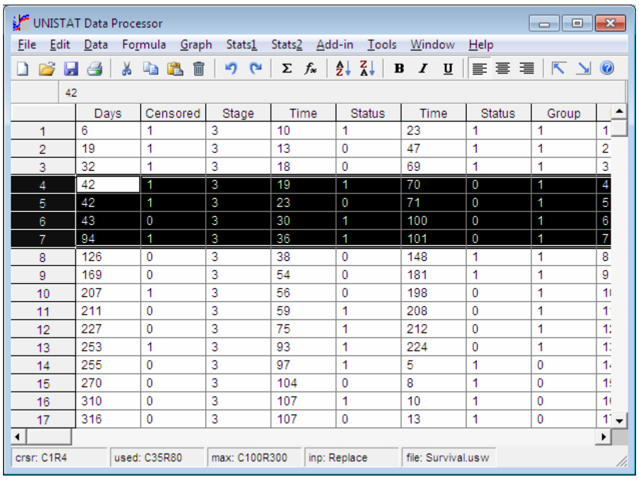
3.2.8. Insert

A block of columns or rows can be inserted. The following actions are possible, depending on what type of highlighting you have had prior to selecting this option:

No block is highlighted: When Edit → Insert is selected, a daughter menu will provide two further options; Columns or Rows. If the Columns option is selected, then a blank column will be inserted at the active cell location without further notice. Likewise, if Rows is selected, then a blank row will be inserted at the active cell location

A block of cells is highlighted: As above, select one of the Columns or Rows options to insert as many columns or rows as covered by the highlighted area.

A block of columns is highlighted: In this case the program understands that it must insert columns and accordingly it will not request further information, Columns or Rows. The column range will be inserted and all columns to the right of the start of the highlighted range (including Column Labels) will be shifted right by the number of columns inserted. The highlight will not be switched off.


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help
42

	Days	Censored	Stage	Time	Status	Time	Status	Group	
1	6	1	3	10	1	23	1	1	1
2	19	1	3	13	0	47	1	1	2
3	32	1	3	18	0	69	1	1	3
4	42	1	3	19	1	70	0	1	4
5	42	1	3	23	0	71	0	1	5
6	43	0	3	30	1	100	0	1	6
7	94	1	3	36	1	101	0	1	7
8	126	0	3	38	0	148	1	1	8
9	169	0	3	54	0	181	1	1	9
10	207	1	3	56	0	198	0	1	1
11	211	0	3	59	1	208	0	1	1
12	227	0	3	75	1	212	0	1	1
13	253	1	3	93	1	224	0	1	1
14	255	0	3	97	1	5	1	0	1
15	270	0	3	104	0	8	1	0	1
16	310	0	3	107	1	10	1	0	1
17	316	0	3	107	0	13	1	0	1

crsr: C1R4 | used: C35R80 | max: C100R300 | inp: Replace | file: Survival.usw

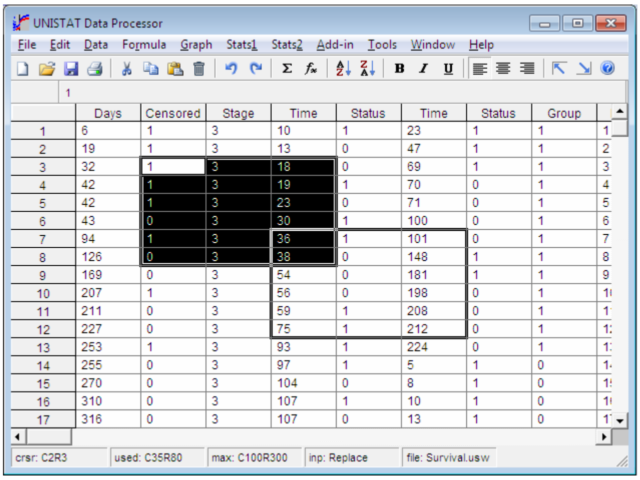
As in the case of deleting a range of columns, all formulas will be updated so that they will refer to the correct columns in their new positions. If there are insufficient empty columns in the Data Processor, then a prompt will be issued and the insert procedure will be aborted.

A block of rows is highlighted: In this case the program understands that it must insert rows and accordingly it will not request further information, Columns or Rows. The highlighted range of rows will be inserted and all rows below (including Row Labels) will be shifted down. All column lengths and thus the maximum number of rows used will be updated. Formulas will not be lost.

If the Data Processor does not have a sufficient number of empty rows, then a message will pop up and the insert procedure will be aborted.

3.2.9. Drag-Drop Action

When a range is highlighted (which can be a block of cells, a range of columns or a range of rows) and the mouse pointer (which has a plus shape) is placed on the boundary of the highlighted range, you will see that it changes into an arrow shaped pointer. At this point, you may press the left mouse button and drag the pointer. The outline of the block will move with the mouse pointer. Drag the block until it reaches its destination and then release the mouse button. The source will be copied to the new position.



	Days	Censored	Stage	Time	Status	Time	Status	Group	
1	6	1	3	10	1	23	1	1	1
2	19	1	3	13	0	47	1	1	2
3	32	1	3	18	0	69	1	1	3
4	42	1	3	19	1	70	0	1	4
5	42	1	3	23	0	71	0	1	5
6	43	0	3	30	1	100	0	1	6
7	94	1	3	36	1	101	0	1	7
8	126	0	3	38	0	148	1	1	8
9	169	0	3	54	0	181	1	1	9
10	207	1	3	56	0	198	0	1	1
11	211	0	3	59	1	208	0	1	1
12	227	0	3	75	1	212	0	1	1
13	253	1	3	93	1	224	0	1	1
14	255	0	3	97	1	5	1	0	1
15	270	0	3	104	0	8	1	0	1
16	310	0	3	107	1	10	1	0	1
17	316	0	3	107	0	13	1	0	1

crsr: C2R3 | used: C35R80 | max: C100R300 | inp: Replace | file: Survival.usw

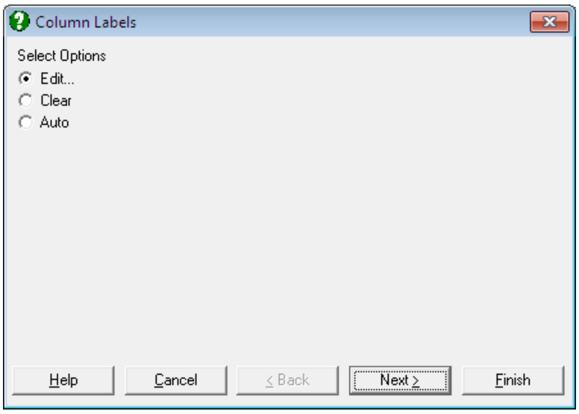
The Edit → Drag-Drop Action option may take two values: Copy or Move. The first option Copy is the default and will work as described above. If the Move option is selected, then copying will take place as described above. However, additionally, the source range will be erased.

3.2.10. Column Labels

The Column Labels can be entered, edited and cleared. It is also possible to enter a specified range of Column Labels automatically, by incrementing a given base number by integer values. You can also load Column Labels from a text file (see 3.1.0.5. Free Format Text Files and 3.1.0.6. Fixed Format Text Files).

To enter or edit a single column or row label at a time, double-click on the label. A text editor will be placed on the Input Panel. Enter or edit the label, and then click <Enter/OK>. To enter several Column Labels manually, select this item from the menu.

Like other Data Processor parameters, Column Labels are also saved as part of UNISTAT internal format files, and optionally as part of text, Excel, Lotus, dBase, DIF and Sylk files (see 3.1.0. File Formats).



Select Options

● Edit...

○ Clear

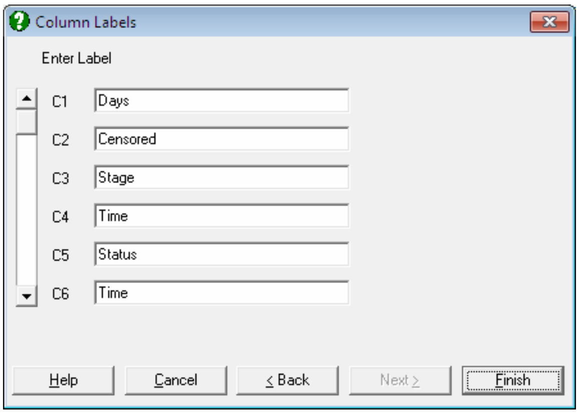
○ Auto

Edit: A single label can be edited or entered simply by double-clicking on a column label, without having to go through the Edit menu.

When the Edit → Column Labels → Edit option is selected from the menu, a window with six text fields and a vertical scroll bar will be displayed. Any string of characters, excluding parentheses (and), can be typed in.

WARNING! *Pure numbers or string expressions starting with numbers cannot be entered as column labels.*

If a number is entered, the program will automatically prefix it with the character *A* so that the label is not confused with a number in formulas. There are no limitations on the number of characters in labels (except for system limitations).



Enter Label

C1	Days
C2	Censored
C3	Stage
C4	Time
C5	Status
C6	Time

The display of Column Labels may be scrolled by cursor pad keys or clicking the mouse on the scroll bar. When a significant amount of editing has been done, do not forget to save the data in spreadsheet to a file.

Clear: All Column Labels will be erased. First a warning will be issued.

Auto: A range of Column Labels can be assigned automatically. Place the active cell at any row of the starting column first, then select Edit → Column Labels and select Auto from the list. An input field will be placed on the Input Panel. Type the base string, let us say *Var*, followed by a hash sign #, and followed by the number of columns to be labelled, say *5*. The Input Panel would look like this:

ENTER COLUMN LABELS: Var#5

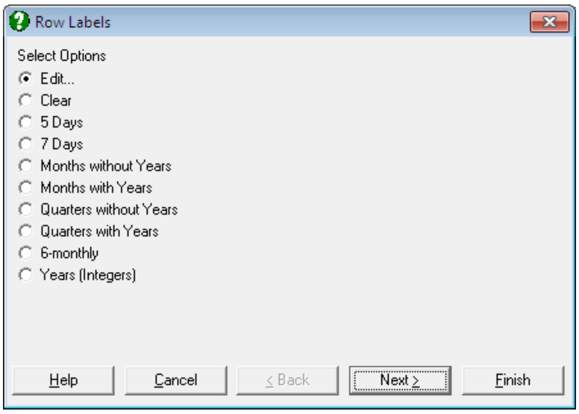
Pressing <Enter/OK>, the current column and the following four columns will be automatically assigned labels *Var1, Var2, ..., Var5*. Remember that pure numbers cannot be assigned as Column Labels.

3.2.11. Row Labels

It is possible to enter or edit individual Row Labels or to assign one of daily, monthly, quarterly, six monthly or annual annotations automatically to all rows, up to the maximum row in use. It is also possible to load Row Labels from a text file (see 3.1.0.5. Free Format Text Files). There are no limitations on the number of characters in labels (except for system limitations).

To enter or edit a column or row label individually, double-click on the label. A text editor will be placed on the Input Panel. Enter or edit the label, and then click <Enter/OK>. To enter a number of Row Labels manually, select Edit → Row Labels from the menu.

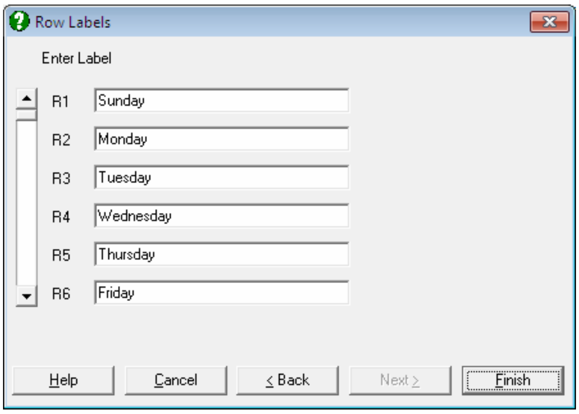
Like other Data Processor parameters Row Labels are also saved as part of UNISTAT internal format files, and optionally, of text, Excel, Lotus, dBase, DIF and Sylk files.



Select Options
● Edit...
○ Clear
○ 5 Days
○ 7 Days
○ Months without Years
○ Months with Years
○ Quarters without Years
○ Quarters with Years
○ 6-monthly
○ Years (Integers)

A dialogue will provide access to the following options:

Edit: Individual labels may be edited or entered simply by double-clicking on a label, without having to go through the Edit menu.



Enter Label
R1 Sunday
R2 Monday
R3 Tuesday
R4 Wednesday
R5 Thursday
R6 Friday

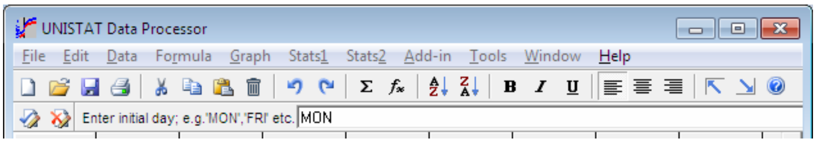
When the Edit → Row Labels → Edit option is selected from the menu, a window with six text fields and a vertical scroll bar will be displayed. Unlike

Column Labels, pure numbers or string expressions starting with numbers can be entered as Row Labels.

It is possible to scroll the display of Column Labels by arrow keys or clicking the mouse on the scroll bar. When a significant amount of editing has been done, do not forget to save the data in spreadsheet to a file.

Clear: This option will erase all Row Labels. A warning will be issued first.

Working Days (5 Days): An input field will be placed on the Input Panel. Enter the first three letters of the working day (e.g. MON, WED) for the first label (i.e. for row 1). The program will then work out and fill all Row Labels up to the maximum row number in use with consecutive working days.



Enter initial day; e.g.'MON','FRI' etc. MON

Days of Week (7 Days): Enter the first three letters of the day (e.g. SUN, SAT) for the first label. The program will then work out and fill all Row Labels up to the maximum row number in use with consecutive days of the week.

Months without Years: Enter the first three letters of the month (e.g. JAN, AUG) for the first label. The program will then work out and fill all Row Labels up to the maximum row number in use with consecutive months.

Months with Years: Type in the year (e.g. 1989, 1817) followed by a semicolon and then either the first three letters of the month (e.g. JUN, DEC etc.), or the number of the month (1-12). The date entered is for the first row. The rest of the labels up to the maximum used will be assigned by the program automatically.

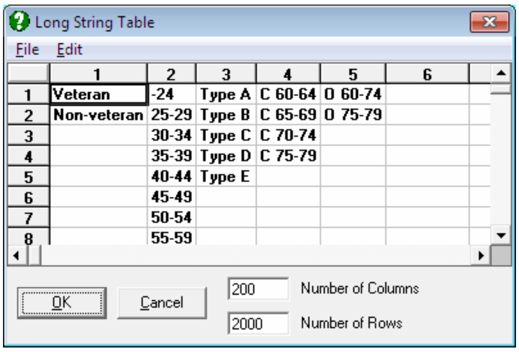
Quarters without Years: Enter the first three letters of the quarter (WIN, SPR, SUM, AUT) for the first label. The program will then work out and fill all Row Labels up to the maximum row number in use with consecutive quarters.

Quarters with Years: Type in the year (e.g. 2010, 1953), followed by a semicolon, and then either the first three letters of the quarter (WIN, SPR, SUM, AUT) or the number of the quarter (1 to 4) for the first row of the display.

Six-Monthly: Type in the year, followed by a semicolon, and then the number of the half-year (1 or 2) for the first row. There are no string alternatives for the numbers 1 and 2.

Annual (or Integers): Type in the year for the first row. This option may also be used to assign row numbers as Row Labels, by entering 1 for the first row. This may be useful in labelling points in 2D Plots and 3D Plots procedures.

3.2.12. Long String Table



	1	2	3	4	5	6
1	Veteran	-24	Type A	C 60-64	O 60-74	
2	Non-veteran	25-29	Type B	C 65-69	O 75-79	
3		30-34	Type C	C 70-74		
4		35-39	Type D	C 75-79		
5		40-44	Type E			
6		45-49				
7		50-54				
8		55-59				

Number of Columns: 200
Number of Rows: 2000

The Edit → Long String Table dialogue consists of a grid containing the String Data and two text boxes displaying the dimensions of the table. If no files containing long strings have been opened before, or no long strings have been typed into the Data Processor manually, the string table will be blank on entry. At this stage, you can type any strings into the table. Row numbers of the table correspond to the integers by which these strings are represented in Data Processor. Suppose, for instance, we type the values *25 C*, *30 C*, *Amb RH*, *05 - 30 C* and *05 - 30 C Ambient* into the first four rows of the second column of the string table respectively. We then click [OK] and accept the changes and go back to the Data Processor. Also suppose that we have a column in our data sheet that contains a succession of integers from 1 to 4, say, 1, 1, 3, 2, 4, 3, 2, 3, 3, 4, 3, 2.

But how are we going to make UNISTAT display the strings entered into the table instead of these integers? In other words how shall we establish a correspondence between the columns of the data sheet and the Long String Table? This task is performed by entering a special data conversion function **Long(n)** (see 3.4.2.6.1. Data Conversion Functions). All we need to do is to click somewhere on the column containing the integers, press <=> (the equal sign) and then enter **Long(2)**. The data sheet will be redrawn immediately and the long strings will be displayed instead of integers. Cells containing *1* will now display the string *25 C*, cells containing *2* the string *30 C, Amb RH*, etc. The correspondence established in this way can be cancelled by entering the special data conversion function **Data**, in which case the column will display the integers again.

When you wish to change all strings represented by an integer, you need to activate Edit → Long String Table and make the change in the string table. If you edit a long string cell in Data Processor, this will be accepted as a new manual String Data entry and it will be added to the corresponding column in the string table at the next free row.

You can also convert a long string column to a short string column using the **Short** function. However, you must not forget that this function will truncate the string values longer than eight characters.

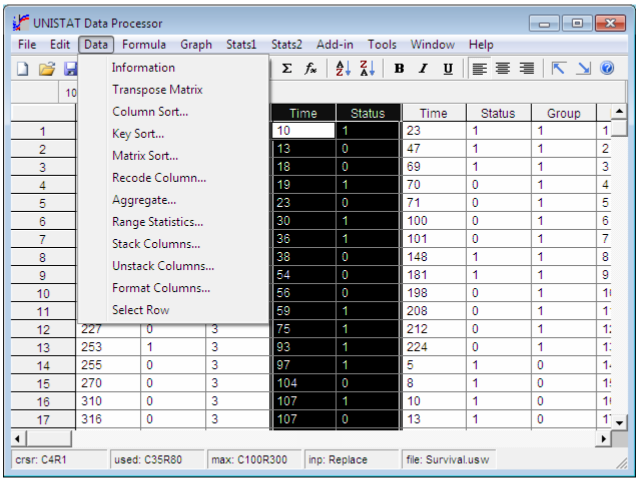
Changing the type of - or even deleting - a long String Data column will not change its corresponding Long String Table entries. This is because there may be other long string columns in Data Processor that also reference the same Long String Table entries.

The default dimensions of the Long String Table are 200 columns by 2000 rows, but this can be changed if desired. Any changes made to these numbers will be valid during a UNISTAT session. Next time UNISTAT is launched, the default values will be restored. If you want to change the default values permanently, enter and edit the following lines in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

```
MaxStringVars=200

MaxFactorLevel=2000
```

The Number of Rows parameter also determines the maximum number of levels (i.e. distinct values) in factor (i.e. categorical) variables. If you attempt to run, for instance, an ANOVA model with a factor variable that has more than 2000 levels, the program will stop and ask you to increase this parameter.

3.3. Data Menu



				Time	Status	Time	Status	Group	
1				10	1	23	1	1	1
2				13	0	47	1	1	2
3				18	0	69	1	1	3
4				19	1	70	0	1	4
5				23	0	71	0	1	5
6				30	1	100	0	1	6
7				36	1	101	0	1	7
8				38	0	148	1	1	8
9				54	0	181	1	1	9
10				56	0	198	0	1	1
11				59	1	208	0	1	1
12	227	0	3	75	1	212	0	1	1
13	253	1	3	93	1	224	0	1	1
14	255	0	3	97	1	5	1	0	1
15	270	0	3	104	0	8	1	0	1
16	310	0	3	107	1	10	1	0	1
17	316	0	3	107	0	13	1	0	1

Various utilities are provided for transforming and formatting data. It is also possible to sort, key sort and recode data matrix columns or perform operations like Range Statistics and Aggregate.

Some of the options will not be active (i.e. they will be greyed out) unless a range of columns has been highlighted first.

3.3.1. Information

This procedure displays a basic statistics table for all data columns in Data Processor. The statistics displayed are the current values of Scalar Functions for each column. These are **Len()** (total number of observations including missing observations), **Mis()** (number of missing observations), **Avg()** (arithmetic mean), **Std()** (sample standard deviation), **Var()** (sample variance), **Sum()**, **Ssq()** (sum of squares), **Min()** (minimum observation), **Max()** (maximum observation) and the formulas, if any. In case a column contains missing values, the values of **Avg()**, **Var()** and **Std()** are adjusted for the number of missing observations.

If a Select Row variable is in effect, the number of rows omitted will also be reported and all statistics will be adjusted for the valid number of cases.

Example

Information

file: Anova.usw

	LEN	MIS	AVG	STD	VAR
C1 Hardness	16	0	1.2500	2.9326	8.6000
C2 Tip	16	0	2.5000	1.1547	1.3333
C3 Coupon	16	0	2.5000	1.1547	1.3333
C4 Operator	25	0	3.0000	1.4434	2.0833
C5 Batch	25	0	3.0000	1.4434	2.0833
C6 Formulation	25	0	String	String	String
C7 Coded Data	25	0	0.4000	5.3072	28.1667
C8 Test assemblies	25	0	String	String	String

	SUM	SSQ	MIN	MAX
C1 Hardness	20.0000	154.0000	-3.0000	7.0000
C2 Tip	40.0000	120.0000	1.0000	4.0000
C3 Coupon	40.0000	120.0000	1.0000	4.0000
C4 Operator	75.0000	275.0000	1.0000	5.0000
C5 Batch	75.0000	275.0000	1.0000	5.0000
C6 Formulation	String	String	A	E
C7 Coded Data	10.0000	680.0000	-8.0000	13.0000
C8 Test assemblies	String	String	a	e

Information on string, date and time variables will also be displayed under this option, though many of the statistics displayed are irrelevant. The displayed statistics are **Len()**, **Mis()**, **Min()**, **Max()** and the remaining entries are indicated as one of **String**, **Date** or **Time**.

3.3.2. Transpose Matrix

The whole data matrix - including Column Labels and Row Labels - can be transposed. The procedure will be executed immediately without further warning.

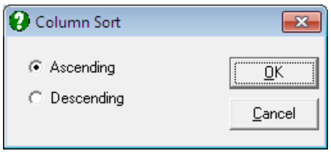
This procedure is particularly useful for performing operations with rows, instead of columns. By transposing the matrix first, computing any functions using columns as arguments, and then transposing the matrix once again, it is possible to compute functions using rows as arguments.

Remember that in order for this procedure to work there must be at least as many columns available in Data Processor as the number of rows used and vice versa. If not, save your data first, change dimensions of the data matrix from Tools →

Options → Memory Management, load the file again and perform the transpose operation.

3.3.3. Column Sort

First position the active cell at any row of the column to be sorted. Then select Data → Column Sort. A dialogue will appear, giving the choice of Ascending or Descending sorts.



String variables can be selected and will be sorted according to their alphabetical ordering. Date and time variables are sorted according to their chronological order.

If the column contains missing values, they will be treated just like any other number. That is, if the missing value code is a very large negative number as suggested in Tools → Options → Memory Management, in ascending sort all missing values will move to the top of the column and in descending sort they will move to the bottom of the column.

WARNING! *The* Column Sort *procedure sorts a column in its own place. If you need a copy of the unsorted form of the column, then you must save or copy it beforehand.*

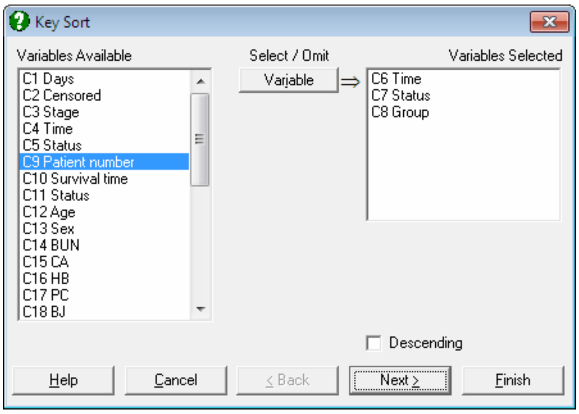
An alternative way of sorting a column is using the function Sort(*Cn*) (see 3.4.2.5. Statistical Functions).

3.3.4. Key Sort

The standard variable selection dialogue will allow you to select the key columns by clicking on [Variable]. When the selection process is complete click on [Finish] to perform the sort.

WARNING! *The* Key Sort *procedure sorts the selected columns in their own place. If you need a copy of the unsorted form of the columns, then you must save or copy them beforehand.*

String variables can be selected and will be sorted according to their alphabetical ordering. Date and time variables are sorted according to their chronological order.



Key Sort

Variables Available

C1 Days
C2 Censored
C3 Stage
C4 Time
C5 Status
C9 Patient number
C10 Survival time
C11 Status
C12 Age
C13 Sex
C14 BUN
C15 CA
C16 HB
C17 PC
C18 BJ

Variables Selected

C6 Time
C7 Status
C8 Group

□ Descending

The Key Sort will sort the first column in the right list first, keeping the relative row positions of all other marked columns. If the first column contains groups of equal values then Key Sort will proceed to the second column selected and sort recursively only those groups of numbers which correspond to the equal value groups of the first column. This process is repeated until all choice columns are sorted.

In most cases, all columns involved in Key Sort will be of the same length. In the unlikely event of the marked columns having unequal lengths, Key Sort will consider empty cells of all columns below the maximum column length as missing and proceed with sorting. As in the case of single Column Sort, missing values are treated like ordinary numbers.

3.3.5. Matrix Sort

This procedure will sort the rows of the data matrix according to a single key column, keeping the rows themselves intact.

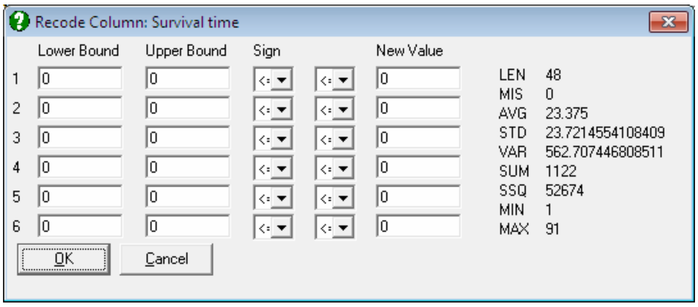
Place the active cell on the key column first and then select Data → Matrix Sort. A dialogue box, similar to the one in Column Sort procedure, will offer a choice of Ascending or Descending sorts. All rows of the data matrix will be sorted according to this column.

WARNING! *The* Matrix Sort *procedure sorts the selected columns in their own place. If you need a copy of the unsorted form of the columns, then you must save or copy them beforehand.*

3.3.6. Recode Column

This procedure is used to assign new values to the specified value ranges of a range of columns. First highlight the range of columns to recode. Non contiguous blocks of columns can be selected by pressing [Ctrl] first. After selecting Data → Recode Column a dialogue will be displayed.

All types of data (numeric, string, date and time) can be recoded. The program will select editing fields according to the type of the variable selected.



Recode Column: Survival time

	Lower Bound	Upper Bound	Sign		New Value
1	0	0	<=	<=	0
2	0	0	<=	<=	0
3	0	0	<=	<=	0
4	0	0	<=	<=	0
5	0	0	<=	<=	0
6	0	0	<=	<=	0

LEN	48
MIS	0
AVG	23.375
STD	23.7214554108409
VAR	562.707446808511
SUM	1122
SSQ	52674
MIN	1
MAX	91

Each row of the dialogue allows for entering information about a particular interval. The first field is for the lower bound and the second is for the upper bound of the interval. The next two controls are drop-down lists for the signs of the lower and upper bounds (lb and ub). The default values of these signs are both ≤. The possible combinations are:

1) lb ≤ x < ub,
2) lb ≤ x ≤ ub,
3) lb < x ≤ ub and
4) lb < x < ub.

The last text field is for the new value of the interval. This can be an asterisk (*) for missing values.

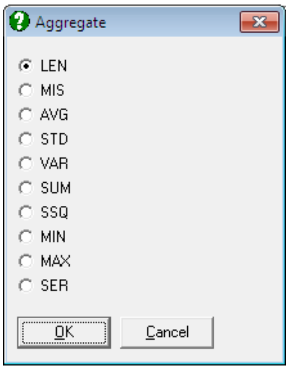
Any row that displays a valid interval (i.e. lb ≤ ub) will be recoded. Therefore, any row left with lb = 0 and ub = 0 values (or for a string variable lb and ub left empty) will remain inactive. Information about the first column of the range to recode is displayed on the right of the screen to help with the process of entering interval limits.

WARNING! *The Recode procedure changes the values of a column in their own place. If you need a copy of the original form of the column, then you must save or copy it beforehand.*

Up to six intervals can be specified at a time. Note, however, that since the intervals not recoded will retain their original values, the Data → Recode Column procedure can be run more than once on the same range of columns.

3.3.7. Aggregate

Data sets with a large number of rows can be compressed for further processing. A row-wise compression is made according to the levels of a factor column. One of the following ten Scalar Functions can be used for end values: **Len()**, **Mis()**, **Avg()**, **Var()**, **Std()**, **Ser()**, **Sum()**, **Ssq()**, **Min()** and **Max()**.



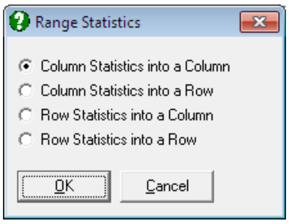
The way this procedure works is similar to that of the Matrix Sort procedure. First, move the active cell to a factor column. A factor column can be a character or numeric data column and should have a limited number of distinct values. Then select Data → Aggregate and select the function to be used in compression. First a Matrix Sort is performed. Then rows of the matrix corresponding to each distinct value of the factor column are collapsed into a single row, using the scalar function selected. For instance, if **Avg()** is selected then the arithmetic mean of the original rows corresponding to the same factor value will constitute a single row of the new matrix.

Columns containing non numeric data cannot be aggregated by taking their average, sum, etc. They can be aggregated only when one of the following functions is selected: **Len()**, **Mis()**, **Min()** and **Max()**. The first two functions will convert the column into a numeric column. If any other function is used the resultant column will be filled with missing values.

3.3.8. Range Statistics

It is possible to highlight a block of cells, generate row-wise or column-wise summary statistics and place the results in a column or row of the spreadsheet. The following options are offered:

1) Column statistics into a column
2) Column statistics into a row
3) Row statistics into a column
4) Row statistics into a row



Range Statistics

● Column Statistics into a Column
○ Column Statistics into a Row
○ Row Statistics into a Column
○ Row Statistics into a Row

The selected statistic can be one of the ten Scalar Functions, i.e., **Len()**, **Mis()**, **Avg()**, **Var()**, **Std()**, **Ser()**, **Sum()**, **Ssq()**, **Min()** and **Max()**.

First, highlight a block of cells and select Data → Range Statistics, then select one of the four options mentioned above. Then move the active cell to the start of the destination range and press <Enter/OK>. A dialogue will appear allowing you to choose from these ten statistics.

The procedure will not terminate after entering a single statistic. Assuming you will want to get more than one statistic on the same block of cells, it will work in a circular fashion. Immediately after selecting a statistic from the dialogue, you can move the active cell to a different position and press <Enter/OK> again. The Range Statistics dialogue will appear again, allowing you to select a different statistic. To break out of the loop, press <Escape/Cancel>.

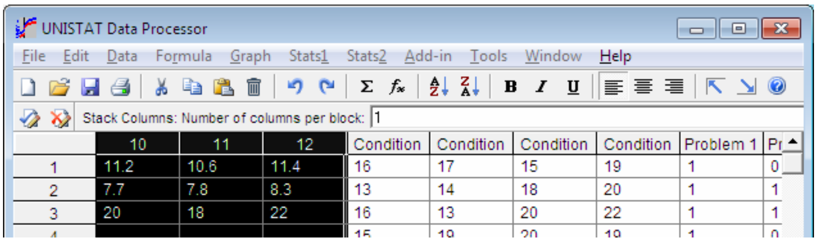
3.3.9. Stack Columns

Data → Stack Columns and the accompanying Data → Unstack Columns are highly specialised but powerful procedures. The Stack Columns procedure is used to stack a number of blocks of data columns. An extra (factor) column will be created which keeps track of the blocks stacked together.

The Stack Columns procedure is particularly useful if there is data on different subjects and / or treatments and it is decided to perform Analysis of Variance. Because the Analysis of Variance procedure requires a factor column to indicate the levels of treatment for a dependent variable, the user is often faced with the cumbersome task of editing data files into this format. Stack Columns will automatically arrange data suitable for various procedures that require factor columns, such as Analysis of Variance, Break-Down, Table of Means or some nonparametric tests.

First highlight the range of columns to be stacked, then select Data → Stack Columns.

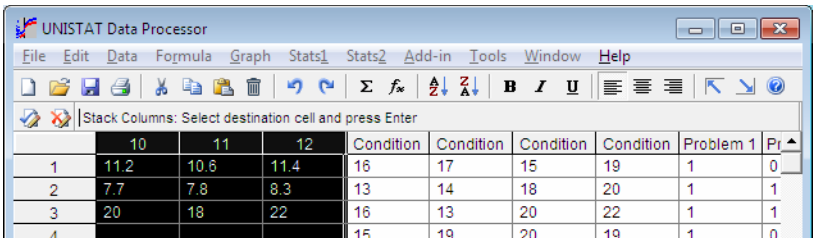
WARNING! *The Stack procedure cannot be used unless a column or a range of columns has been highlighted first.*


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help

Stack Columns: Number of columns per block: 1

	10	11	12	Condition	Condition	Condition	Condition	Problem 1	Pr
1	11.2	10.6	11.4	16	17	15	19	1	0
2	7.7	7.8	8.3	13	14	18	20	1	1
3	20	18	22	16	13	20	22	1	1
4				15	19	20	19	1	0

The program will ask for the number of columns per block. Since blocks of columns (i.e. not only single columns) can be stacked together, you must specify the number of columns per block. If only single columns are to be stacked, then accept the program's choice 1 by pressing <Enter/OK>. Otherwise enter your own choice. The total number of highlighted columns must be divisible by the number of columns per block. If this condition is not met, then the program will repeat the number of columns input until a valid number is entered or <Escape/Cancel> is pressed.


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help

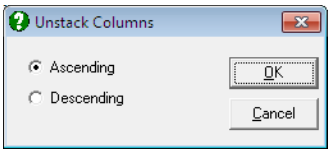
Stack Columns: Select destination cell and press Enter

	10	11	12	Condition	Condition	Condition	Condition	Problem 1	Pr
1	11.2	10.6	11.4	16	17	15	19	1	0
2	7.7	7.8	8.3	13	14	18	20	1	1
3	20	18	22	16	13	20	22	1	1
4				15	19	20	19	1	0

Then, move the active cell to the desired location and press <Enter/OK>. An extra column containing factor levels (i.e. the number of source blocks) will be added to the front of the stacked block of columns. The length of the stacked column block will be equal to the sum total of lengths of all blocks. With large data sets, this may exceed the number of rows available in the Data Processor. If such is the case, save your data first, increase the number of rows from Tools → Options → Memory Management, load back the data and then perform the Stack Columns procedure.

3.3.10. Unstack Columns

This is the reverse of Stack Columns procedure. A block of columns will be divided into separate blocks according to the groups of similar values in the first column of the source block. Therefore, the first column of the block is expected to be a factor column with a limited number of levels (i.e. distinct values). Factor levels can be any floating point numbers, string, date or time data and they need not be sorted. The Unstack Columns procedure will first sort the block according to the (first) factor column and then create a new block for each distinct value of the factor column. Missing values are treated like ordinary numbers and a separate block is created for them.

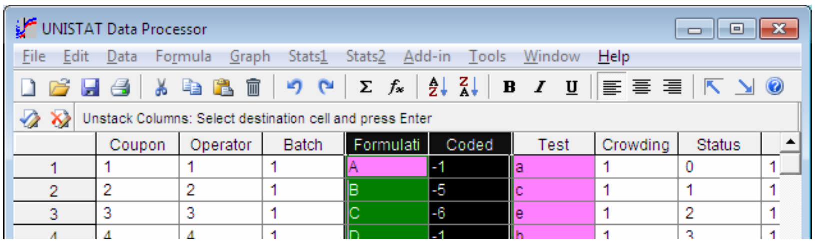


First highlight the range of columns to be unstacked, then select Data → Unstack Columns.

WARNING! *The Unstack Columns procedure cannot be used unless a range of columns, the first of which is a factor column, has been highlighted first.*

A dialogue will offer a choice of Ascending or Descending sorts. If the lengths of columns in the highlighted range are not equal, shorter columns will be padded up with missing values.

WARNING! *The Unstack Columns procedure will sort the source columns in their own places. If you need a copy of the original form of the columns, then you must save or copy them beforehand.*


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help

Unstack Columns: Select destination cell and press Enter

	Coupon	Operator	Batch	Formulati	Coded	Test	Crowding	Status	
1	1	1	1	A	-1	a	1	0	1
2	2	2	1	B	-5	c	1	1	1
3	3	3	1	C	-6	e	1	2	1
4	4	4	1	D	-1	h	1	3	1

The number of columns generated is directly proportional to the number of levels of the factor column and this can be substantial even with moderately large data sets. If this number exceeds the number of columns available in the Data Processor, then save your data first, increase the number of columns from Tools → Options → Memory Management, load back the data and then perform the Unstack Columns procedure.

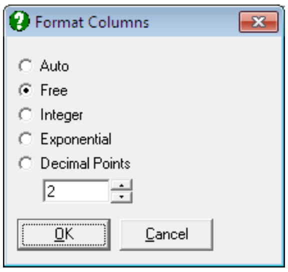
3.3.11. Format Columns

It is possible to display all numbers in a range of columns in a specified format (e.g. decimal, exponential, integer) such that all decimal points (if there are any) are displayed at a fixed position.

All format types will affect only the display of numbers but not their significant digits in memory. To change the actual numbers use the **Round()** function (see 3.4.2.5. Statistical Functions).

WARNING! *The format procedure cannot be used unless a column or range of columns has been highlighted first.*

Highlight the range of columns to be formatted first and then select Data → Format Columns. A dialogue will appear providing access to the following format options.



○ Auto
● Free
○ Integer
○ Exponential
○ Decimal Points
2

Auto: This will display as many significant digits as possible, at the same time maintaining a fixed position for the decimal point. If this cannot be done, i.e. if the column contains too small and too large numbers which are not exponential, then all numbers will be displayed in free format (see the next paragraph). If there is at least one exponential number in the column then the format will be exponential.

Free: This will in effect undo all other format types and revert to the unformatted display of numbers. This means that no zeros after the last significant digit are displayed, only too small or too large numbers are displayed in exponential format and thus decimal points are not necessarily lined up at a fixed position.

When fresh data is entered in a column which was previously formatted, then the initial format will be retained. In this case you can reformat the column selecting one of Auto or Free options.

Integer: No floating points are displayed. Decimal numbers are rounded to the nearest integer.

Exponential: One digit will be displayed before the floating point, and two digits after it. This is followed by *E*, the sign of the power term (- or +) and then the power term (maximum three digits).

Decimal Points: You can choose the number of digits to be displayed after the floating point. However, the maximum number of digits to be displayed depends on the largest positive or the smallest negative number in the column. If the choice is unacceptable, then a message will be issued.

3.3.12. Select Row

This menu item is used to mark a column containing categorical data as a row-wise selection criterion for the subsample of rows (cases) to include in all subsequent analyses. This functionality is also known as *select-if*. You can enter a logical condition (an **If()** function) to fill a new column with *True* and *False* values. In all subsequent graphical or statistical procedures, only those rows that contain a *True* value will be included in the analysis. UNISTAT considers the value 0 as *False* and any other value as *True*.

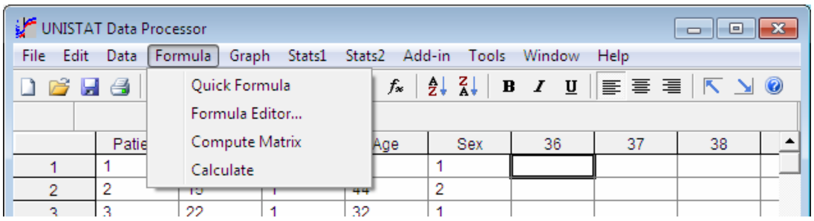
Once a column is selected using this option, it will remain active until another column is selected by Select Row or it is deleted. To deselect a Select Row column without deleting it, place the active cell on this column and select Data → Select Row once again. Alternatively, you can also use the Data Processor’s **Select** function to select and deselect a Select Row column (see 3.4.2.6.3. UNISTAT

Functions). At any one time, there can only be one Select Row column in the Data Processor. The colour of the Select Row column can be changed from Tools → Options → Colours.

A Select Row column may be generated by the program automatically when outliers are deleted from a regression graph (see 2.3.2.3. Interactive Data Points), or a rectangular block of cells is highlighted in Data Processor before selecting a procedure.

In Excel Add-In Mode, when a Select Row column is created automatically by deleting outliers from a regression graph interactively, the only way to switch it off is to highlight a different block of data in Excel.

3.4. Formula



Columns may be defined as functions of other columns. A function returns in the n^th^ row of the current column the function evaluation of the n^th^ rows of columns in its argument.

In order to enter a simple formula simply press <=> or select Formula → Quick Formula. For more complex formulas select Formula → Formula Editor, which will provide full information on the options available.

3.4.0. Overview

3.4.0.1. Entering Formulas

To enter a formula, position the active cell at any row of the destination column (empty or used) and select Formula → Quick Formula. If the formula is to be defined, say, for column 4, then a text box will be placed on the Input Panel already containing *C4* = . Type in the right hand side of the formula followed by <Enter/OK>. The formula will be computed immediately and results placed in column 4. Any combination of upper and lower case letters can be used.

In the above example, although the text box appears with *C4* = , the current column placed on the left hand side of the equation, you can edit this to send the results to any other column you choose. For instance, if you enter *C10* = *C1* * *C2* then the results will be sent to column 10 regardless of the initial active cell position. Examples given in the following sections will omit the left hand side of formulas, which means that the results are put into the current column.

There are two ways of referring to a column in functions. If column n has the label *Label*, then it may be referred to as either *Cn* or *Label*. For example, to define every row of column 4 as the square of every corresponding row of column 2, the formula can be written either as:

C2^2

or as:

Label^2

3.4.0.2. Computing a Range of Formulas

When you need to compute the same formula over a range of columns, first enter the formula as normal, and then append the number of times it should be repeated preceded by a hash sign #. The column numbers in the formula will be incremented by one each time the formula is computed for a new column. For instance, if you enter the following when the active cell is on column 4:

C2#3

then column 2 will be copied to column 4, column 3 to column 5, and column 4 to column 6. In this example, the final contents of columns 2, 4 and 6 will be the same, as the column numbers in the function will be incremented irrespective of this being the end result of a previous function evaluation or not. Incrementing will work regardless of the original position of the active cell, such that it is possible to redefine columns in their own places. Suppose there is data from *C1* to *C6* in the Data Processor and you wish to take their logarithms. Suppose also that there is no need to retain the original numbers. Then all you have to do is to place the active cell on any row of column 1, select Formula → Quick Formula and enter the following:

Log(*C1***)**#6

You may not always want all column numbers in a function incremented. For instance, to multiply one column with all other columns in a range you must be able to keep this column fixed while the others are incremented. This is done by prefixing such columns by a *$* sign:

C4+$*C2**__Log(__*C1*__)__#6

In this example columns *C4* and *C1* will be incremented but *C2* will remain fixed.

Any type of formulas which will be discussed in the following sections can be computed over a range in this way.

3.4.0.3. Using Column Labels in Formulas

UNISTAT imposes no restrictions on the use of Column Labels in formulas, but you must realise that you are not absolutely free to use any column label in a formula. Suppose, for example, you want to compute the following:

C1/**Sqr(***C1*+3**)**

If the label of *C1* is *S* and you want to use this label rather than the column number in the formula, then it should look like this:

S/**Sqr(***S*+3**)**

In this case the program will consider the letter *S* in **Sqr()** as another occurrence of the column label *S* and then report a Syntax error. You must ensure that no Column Labels used in formulas are a subset of the function strings used by the Data Processor. The function compilation is case sensitive; that is, lower and upper case letters are distinguished by the program. If, for instance, in the above example the column label were *s* (i.e. lower case) and the formula were written in the same way, i.e.:

s/**Sqr(***s*+3**)**

there would be no confusion. Therefore, if you avoid entering both labels and function strings in the same case (i.e. both lower or both upper case) the problem will never arise. Of course another alternative is to avoid using Column Labels shorter than four characters. Then there will be no scope for confusion.

3.4.0.4. Formula Syntax

If a formula is entered improperly, then the program will display the prompt Syntax error and then reactivate the formula input field to enable you to edit the input string immediately. It is possible to leave the input field by pressing <Escape/Cancel>. Note, however, that the existence of wrong formulas in the data matrix will generate further error messages when the Compute Matrix procedure is used subsequently. Thus it is recommended that the wrong formulas are either corrected or converted to **Data** (see 3.4.2.6.1. Data Conversion Functions) or deleted.

All spaces can be omitted from a formula string. This is in fact what the program does before processing a formula input. Formulas are calculated immediately after they are entered. If a change is made in a data column which is used as one of the arguments of a formula, the formula values will not be updated automatically. All formulas in the data matrix can be updated by selecting Formula → Compute Matrix.

Scalar Functions: If a column is defined as a function of scalars only, then the program will fill all cells with the same scalar value of the function up to the maximum row used (see 3.4.2.4. Scalar Functions).

Missing Values: If there is missing data in at least one of the argument columns then the corresponding row of the function value will be missing.

Unequal Column Lengths: If a combination of columns with unequal lengths is used in a formula, the rows of the resultant column above the shortest length will be treated as missing.

3.4.0.5. Converting Formula Columns into Data

When a column is defined as a formula, then data entry, editing, deleting, etc. are not allowed in this column. To do this, you can redefine the column as data by setting its formula to **Data** (see 3.4.2.6.1. Data Conversion Functions). Suppose, for instance, column 2 is defined as natural logarithm of column 1:

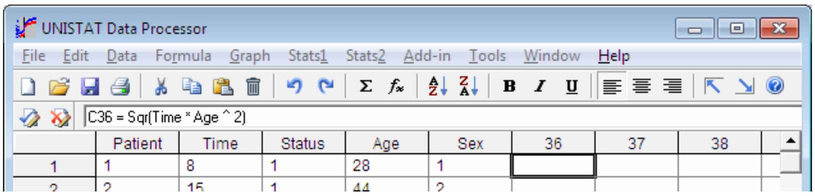
C2 = **Lne(***C1***)**

To convert *C2* into a numeric data column you can press the equal key <=> again while the active cell is somewhere on *C2* and enter one of the following two functions:

Data
Number

This will delete the formula in column 2 and its entries can be overwritten, deleted or new data points appended.

3.4.1. Quick Formula


UNISTAT Data Processor
File Edit Data Formula Graph Stats1 Stats2 Add-in Tools Window Help

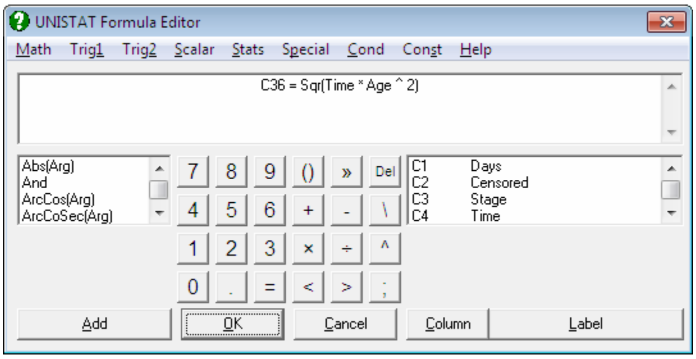
C36 = Sqr(Time * Age ^ 2)

	Patient	Time	Status	Age	Sex	36	37	38
1	1	8	1	28	1			
2	2	15	1	44	2			

If you are sure about the syntax of the formula you wish to enter, then select this option. Otherwise, use the Formula → Formula Editor option which is described below.

The short key for Quick Formula is the *equal* key <=>.

3.4.2. Formula Editor



C36 = Sqr(Time * Age ^ 2)

Formulas can be typed into Input Panel or the Formula Editor can be used to build a formula. In Data Processor, you can access this dialogue from Formula → Formula Editor. In other parts of the program, where you are asked to supply a formula (like Plot of Functions or Nonlinear Regression) double-clicking on a formula text box will activate the Formula Editor. The options available on this dialogue will be different in different parts of the program.

3.4.2.0. Overview

You can build a formula with the Formula Editor without using the keyboard. All numbers, operators, formulas and variable names are available on buttons, in lists or from the pull down menus. If the formula is too long it will be split over a number of lines which are displayed at the top of the Formula Editor.

1) The menu bar provides access to all available functions which are grouped into relevant categories. When you make a selection from a menu list as usual, the text for the selected item will be pasted into the formula text field.

2) A list situated on the left of the window also provides access to all functions which are sorted in alphabetical order. When you highlight an item and then click the [Add] button under the list, the text for the selected item will be pasted into the formula at the cursor position.

3) Another list situated on the right of the window displays all spreadsheet columns containing data (i.e. the variables). First highlight the desired column in the list. In order to paste the column reference by number (i.e. *C1, C5,* etc.) click on the [Column] button. If the column has a label, you may click on the *Label* button to paste the label of the column (e.g. *Time*, *Fixed Capital*) into the formula.

3.4.2.1. Mathematical Operators and Functions

Any expressions (including columns and arithmetic expressions) can be supplied as arguments of functions. String variables cannot be arguments of such functions.

Operators: The following mathematical operators can be used in formulas:

+	add
-	subtract
*	multiply
/	floating point divide
^	power
\	integer divide; divides two numbers and returns only the integer part
Mod	divides two numbers and returns only the remainder

Exp(): Returns the exponential of its argument. This is e raised to the power of *arg*, which is the inverse of log to base e. If *arg* takes on a value which is outside the range -700 to 700, then an Argument Error message is issued and the procedure is terminated.

Syntax:
Exp(*arg***)**

Examples:
Exp(*C1**2***Sqr(***C2***))**
Exp(*Label1*^2/**Lne(***Label2***))**

Log(): Returns the log to base 10 of its argument. If *arg* takes on a non positive value, then an Argument Error message is issued and the procedure is terminated.

Syntax:
Log(*arg***)**

Examples:
Log(*C1**2* **Sqr(***C2***))**
Log(*Label1*^2/10^*Label2***)**

Lne(): Returns the log to base e of its argument. The log to base *N* of x is given by **Lne(***x***)/Lne(***N***)**. If *arg* takes on a non positive value, then an Argument Error message is issued and the procedure is terminated.

Syntax:
Lne(*arg***)**

Examples:
Lne(Exp(*C1***)**/2**)**
Lne(*Label1*^2**)**

Sqr(): Returns the square root of its argument. If *arg* takes on a negative value, then an Argument Error message is issued and the procedure is terminated.

Syntax:
Sqr(*arg***)**

Examples:
Sqr(*C1*C2***)**
Sqr(*Label1*^3**)**

Fct(): Returns the factorial of its argument when the argument is a positive integer m, where 0 ≤ m ≤ 169. If *arg* is a positive decimal number, then the gamma function of *arg* +1 is returned. The gamma function of x+1 is equal to the factorial of x when x is a positive integer.

Syntax:
Fct(*arg***)**

Examples:
Fct(Rno)
Fct(25**)**/(**Fct(**12**)*****Fct(**13**)**)
Fct(3.5**)**

The last example will return the scalar 11.631728.

Int(): Returns the integer part of its argument, rounding it down.

Syntax:
Int(*arg***)**

Examples:
Int(Sqr(*C1*/2**))**
Int(*Label1*^1.234**)**

Abs(): Returns the magnitude of its expression. For positive values this is the value itself and for negative values -1 times the value.

Syntax:
Abs(*arg***)**

Examples:
Abs(*C1*/*C2***)**
Abs(Int(*Label1***))**

Sgn(): Returns +1 if *arg* is positive, 0 if *arg* is zero and -1 if *arg* is negative.

Syntax:
Sgn(*arg***)**

Examples:
Sgn(*C1***C2***)**
Sgn(*Label1***)**

3.4.2.2. Trigonometric Functions 1

Angular arguments of all trigonometric functions must be expressed in radians.

1 radian = 180/Pi() degrees.

Sin(): Returns the sine of its argument.

Cos(): Returns the cosine of its argument.

Tan(): Returns the tangent of its argument.

Sec(): Returns the secant of its argument. The secant of *arg* is equal to the inverse cosine of *arg*:
Sec(*arg***)** = 1/**Cos(***arg***)**.

CoSec(): Returns the cosecant of its argument. The cosecant of *arg* is equal to the inverse sine of *arg*:
CoSec(*arg***)** = 1/**Sin(***arg***)**.

CoTan(): Returns the cotangent of its argument. The cotangent of *arg* is equal to the inverse tangent of *arg*:
CoTan(*arg***)** = 1/**Tan(***arg***)**.

ArcSin(): Returns the inverse sine of its argument.
-1 ≤ *arg* ≤ 1.

ArcCos(): Returns the inverse cosine of its argument.
-1 ≤ *arg* ≤ 1.

ArcTan(): Returns the inverse tangent of *arg*.
-1E300 ≤ *arg* ≤ 1E300.

ArcSec(): Returns the inverse secant of *arg*.
-1E300 ≤ *arg* ≤ 1E300.

ArcCoSec(): Returns the inverse cosecant of *arg*.
-1E300 ≤ *arg* ≤ 1E300.

ArcCoTan(): Returns the inverse cotangent of *arg*.
-1E300 ≤ *arg* ≤ 1E300.

3.4.2.3. Trigonometric Functions 2

These are the hyperbolic trigonometric functions.

Angular arguments of all trigonometric functions must be expressed in radians.

1 radian = 180/Pi() degrees.

HSin(): Returns the hyperbolic sine of its argument.

HCos(): Returns the hyperbolic cosine of its argument.

HTan(): Returns the hyperbolic tangent of its argument.

HSec(): Returns the hyperbolic secant of its argument. The secant of *arg* is equal to the inverse hyperbolic cosine of *arg*:
HSec(*arg***)** = 1/ **HCos(***arg***)**.

HCoSec(): Returns the hyperbolic cosecant of its argument. The cosecant of *arg* is equal to the inverse hyperbolic sine of *arg*:
HCoSec(*arg***)** = 1/**HSin(***arg***)**.

HCoTan(): Returns the hyperbolic cotangent its argument. The cotangent of *arg* is equal the inverse hyperbolic tangent of *arg*:
HCoTan(*arg***)** = 1/**HTan(***arg***)**.

HArcSin(): Returns the inverse hyperbolic sine of its argument.

HArcCos(): Returns the inverse hyperbolic cosine of its argument.

HArcTan(): Returns the inverse hyperbolic tangent of its argument.

HArcSec(): Returns the inverse hyperbolic secant of its argument.

HArcCoSec(): Returns the inverse hyperbolic cosecant of its argument.

HArcCoTan(): Returns the inverse hyperbolic cotangent of its argument.

3.4.2.4. Scalar Functions

These functions return scalar values about their argument columns. If only Scalar Functions are used in the definition of a column then the column will be filled with the same scalar value up to the maximum row used. The argument of a scalar function can only be a single column, i.e. either a column number *Cn* or a column label.

If the type of the argument column is one of **String**, **Date** or **Time**, then only the **Len()**, **Mis()**, **Min()**, **Max()** functions will return nonmissing values.

Len(): Number of observations in column n (including missing observations).

Syntax:
Len(*Cn***)**
Len(*Label***)**

Mis(): The number of missing observations in column n.

Syntax:
Mis(*Cn***)**
Mis(*Label***)**

Avg(): The mean of column n (adjusted for missing values).

Syntax:
Avg(*Cn***)**
Avg(*Label***)**

Std(): The sample standard deviation of column n (adjusted for missing values) with **Len(***Cn***)**- **Mis(***Cn***)**-1 degrees of freedom.

Syntax:
Std(*Cn***)**
Std(*Label***)**

Ser(): The standard error, i.e. standard deviation divided by the square root of the valid number of cases (adjusted for missing values) with **Len(***Cn***)**- **Mis(***Cn***)**-1 degrees of freedom.

Syntax:
Ser(*Cn***)**
Ser(*Label***)**

Var(): The sample variance of column n (adjusted for missing values) with **Len(***Cn***)**- **Mis(***Cn***)**-1 degrees of freedom.

Syntax:
Var(*Cn***)**
Var(*Label***)**

Sum(): The sum of observations in column n.

Syntax:
Sum(*Cn***)**
Sum(*Label***)**

Ssq(): The sum of squares of observations in column n.

Syntax:
Ssq(*Cn***)**
Ssq(*Label***)**

Min(): The smallest observation in column n.

Syntax:
Min(*Cn***)**
Min(*Label***)**

Max(): The largest observation in column n.

Syntax:
Max(*Cn***)**
Max(*Label***)**

3.4.2.5. Statistical Functions

These are some of the most commonly used Statistical Functions. They must be used as individual functions and should not form a part of a larger formula. Also, their argument can only be a single column (i.e. either a column number *Cn* or a column label).

These functions are only available in the Data Processor.

Sort(): Sorts a column *Cn* and places the results in the current column. If *D* is appended to the end (following a semicolon) then the sort will be in descending order.

Syntax:
Sort(*Cn*)[;D]

Examples:
Sort(*C1*)
Sort(*C1*);D
Sort(*Label*)
Sort(*Label*);D

Missing data are treated as ordinary numbers, having the value displayed in Tools → Options → Memory Management. This function is in effect identical to Data → Column Sort option. Its main advantage is, however, that it keeps the source column unchanged when the source and destination columns are different.

Rank(): Computes the ascending ranks of a column of numbers or strings and places the results in the current column. The source column and the destination column cannot be the same. If the character *D* is appended to the end (following a semicolon) then the descending ranks are computed. Equal observations are assigned their average rank.

Syntax:
Rank(*Cn*)[;D]

Examples:
Rank(*C1*)
Rank(*C1*);D
Rank(*Label*)
Rank(*Label*);D

Cumul(): Computes the cumulative values of a column. If *R* is appended to the end (following a semicolon) then the relative cumulative values are calculated; i.e., values will be divided by the column sum.

Syntax:
Cumul(*Cn*)[;R]

Examples:
Cumul(*C1*)
Cumul(*C1*);R
Cumul(*Label*)
Cumul(*Label*);R

Mova(): Computes the m order moving averages of column *Cn* and places the results in the current column.

Syntax:
Mova(*Cn*;m)

Examples:
Mova(*C3*;5)
Mova(*Label*;8)

Round(): Rounds off the values of column *Cn* to m decimal places and places the results in the current column. This is different from the Data → Format Columns option in that it actually changes the values stored in memory. Format affects only the display of numbers.

Syntax:
Round(*Cn*;m)

Examples:
Round(*C3*;2)
Round(*Label*;0)

Stan(): Standardises the values of column *Cn* using its mean and sample standard deviation and places the results in the current column. The resultant column will have a zero mean and unit variance.

Syntax:
Stan(*Cn*)

Example:
Stan(*C3*)
Stan(*Label*)

Level(): This function is used to generate a balanced factor column with integer entries. The argument n is a positive integer and the column is filled with recurring sequences of integers from 1 to n, up to the maximum column length, using the formula ((n - 1) Mod J + 1). If the optional argument *B* is used, the column is filled a sequence of n-integer blocks using the formula (**Int(**(n - 1) / J**)** + 1).

Syntax:
Level(n**)[;B]**

Examples:
Level(3**)**: Fills the column with 1 2 3 1 2 3 1 2 3 1 2 3 …
Level(2**);B**: Fills the column with 1 1 2 2 3 3 4 4 …

Dummy(): This function is used to create n new columns (dummy variables) for a factor column with n levels (its argument), each of which corresponding to a level. A case in a dummy column will have the value of 1 if the factor contains the corresponding level in the same row, and 0 otherwise.

The **Dummy()** function also accepts two options, *F* and *L*, which will omit the first or last level respectively and generate n - 1 new columns. This may be desirable to remove the multicollinearity caused by inclusion of all levels, since dummy variables created in this way will always add up to the unit vector.

Dummy variables can also be created for interaction terms of up to three factors. Options *F* and *L* are also applicable to interaction terms.

The first of the dummy columns generated will be put in the current column (i.e. where the active cell is), which should not be one of the argument columns. Ensure that there are enough empty columns in the Data Processor, as the number of columns generated may be very large, especially when dummies are created for interactions.

Syntax:
Dummy(*Cm***[*****Cn***[*****Co***]])[;F][;L]**

Examples:
Dummy(*C3***)**
Dummy(*Region***);L**
Dummy(*Region*Type*Country***);F**

Freq(): Creates two new columns of length n for a (factor) column, which contains n distinct values (levels) and places the results in the current and the following column. The first new column contains the levels of the factor column and the second the number of times each level occurs.

Syntax:
Freq(*Cn***)**

Example:
Freq(*C3*)
Freq(*Label*)

MdRk(): The exact median ranks are generated from the binomial function:

$$\sum_{k=i}^{N} \binom{N}{k} R_i^k \left(1 - R_i\right)^{N-k} = 0.5$$

where N, the population size, should be supplied by the user and Ri is the median rank of the ith row. For large values of N not all exact median ranks can be computed and will be reported as missing values.

Syntax:
MdRk(*N*)

Example:
MdRk(*85*)

3.4.2.6. Special Functions

3.4.2.6.1. Data Conversion Functions

In order to understand the way these functions work, it is essential to remember that cells in Data Processor contain eight bytes of information. UNISTAT can interpret these eight bytes as a double precision floating number, eight characters of String Data (see 3.0.2.2.1. Short Strings), dates with days of the week and time including hours, minutes and seconds. The long String Data type (see 3.0.2.2.2. Long Strings) is a fundamentally different one requiring the long strings to be stored in a separate table (see 3.0.2.2. String Data and Long String Table).

It should be noted here that although these functions are extremely useful in some advanced data manipulations, their use is not needed under normal operating conditions.

Data or Number: When a column is defined as a formula column, its cells cannot be edited. Use this function to convert a formula column into a data column. Values in the cells will not change.

Data and **Number** can be used as alternatives but they cannot be used as part of a larger function. The first four characters are sufficient.

Syntax:
Data
Numb

Examples:
Data
Number

This function can also be used to convert a short string column into numbers. In this case, the 8-byte character data in each cell will be converted to its 8-byte floating point equivalent. Therefore, the outcome may not have any resemblance to original strings. To restore the short strings use the **String** function.

When used on a long string column, this function will convert Long Strings into their underlying integer values. The strings in Long String Table will not be deleted. To restore the long String Data use the **Long()** function.

String or Character: Converts numeric data columns into short strings. Strings longer than 8 characters are truncated. **String** or **Character** can be used as alternatives but they cannot be used as part of a larger function. The first four characters are sufficient. Also see 3.0.2.2. String Data.

Syntax:
Stri
Char

Examples:
String
Character

This function only converts 8-byte numbers to their 8-byte string equivalents. Therefore, the outcome may not have any resemblance to the original numbers. This function is only useful in restoring a short string column which has been converted into numbers using the **Data** function. You can also convert a long string column into short strings using the **Short** function, and a short string column into Long Strings using the **Long** function.

Short: Converts a numeric or long string column into a short string column. It cannot be used as part of a larger function. and the first four characters are sufficient.

Syntax:
Short

Examples:
Short

Although the Long String Table entries of the original long string column are not deleted, its underlying integers are lost.

Long: Converts a numeric or short string column into a long string column, creating a Long String Table column in the process. It cannot be used as part of a larger function.

Syntax:
Long

Examples:
Long

The short string columns converted into long using this function can be restored using the **Short** function.

Long(n): Converts numeric data columns containing integers into long string columns, establishing a correspondence between the current column and the column of the Long String Table referenced in its argument. The Long String Table is assumed to be populated earlier. This function cannot be used as part of a larger function. Also see 3.0.2.2. String Data.

Syntax:
Long(n**)**

Examples:
Long(3**)**

This function will not convert a short string column into a long string one. To do this use the **Long** function.

Date: Converts function or numeric data columns into Date Data. This function cannot be used as part of a larger formula.

Syntax:
Date

Example:
Date

Time: Converts function or numeric data columns into Time Data. This function cannot be used as part of a larger formula.

Syntax:
Time

Example:
Time

3.4.2.6.2. Date and Time Functions

Days(): This function is used to fill a column with dates. The user is expected to supply an initial date. The function will then automatically increment rows at one day intervals. Optionally, a second argument can be supplied separated by a semicolon. When *N* is a positive integer days will be incremented by *N* days at a time. *N* may also take negative values, -1, -2, -3, to create the following effects:

-1: Working days, Monday to Friday.
-2: 6-day week, Monday to Saturday.
-3: Weekend days, Saturday and Sunday.

Syntax:
Days(*Date***)**
Days(*Date*;*N***)**

Examples:
Days(12/12/1990**)**
Days(5/11/1984;7**)**
Days(7/4/82;-1**)**

Hour(): This function is used to fill a column with Time Data. The user is expected to supply an initial time. The function will then automatically increment rows at one hour intervals. Optionally, a second argument can be supplied separated by a semicolon. When *N* is a positive integer, time will be incremented by *N* hours at a time.

The **Hour()** function returns the data type *Time* and it cannot be used as part of a larger formula. If you want a step length which is a fraction of an hour, then use one of **Mins()** or **Secs()** functions.

Syntax:
Hour(*Time***)**
Hour(*Time*;*N***)**

Examples:
Hour(0:0**)**
Hour(10:00:00;2**)**

Mins(): This function is used to fill an entire column with Time Data. The user is expected to supply an initial time. The function will then automatically increment rows at one minute intervals. Optionally, a second argument can be supplied separated by a semicolon. When *N* is a positive integer, time will be incremented by *N* minutes at a time.

The **Mins()** function returns data type *Time* and it cannot be used as part of a larger formula. If you want a step length which is a fraction of a minute, then use the **Secs()** function.

Syntax:
Mins(*Time***)**
Mins(*Time*;*N***)**

Examples:
Mins(0:0**)**
Mins(0:1:0;3**)**
Mins(1:00:00;10**)**

Secs(): This function is used to fill an entire column with Time Data. The user is expected to supply an initial time. The function will then automatically increment rows at one second intervals. Optionally, a second argument can be supplied separated by a semicolon. When *N* is a positive integer, time will be incremented by *N* seconds at a time.

The **Secs()** function returns data type *Time* and it cannot be used as part of a larger formula.

Syntax:
Secs(*Time***)**
Secs(*Time;N***)**

Examples:
Secs(0:0**)**
Secs(0:1;30**)**
Secs(1:10:10;100**)**

3.4.2.6.3. UNISTAT Functions

Select: Selects and deselects a Select Row column. The first four characters are sufficient. This is equivalent to selecting Data → Select Row from the Data Processor menu (see 3.3.12. Select Row).

Reg: Computes the fitted values for an estimated regression equation.

Syntax:
Reg

Example:
Reg

This is a powerful prediction and interpolation tool and is used in conjunction with the following procedures:

Data and Function plots	X-Y Curve fitting options	Polynomial Geometric Exponential
	X-Y-Z Surface fitting options	Plane Polynomial surface
Regression and ANOVA:	Linear regression	
	Polynomial regression	

This function can only be used immediately after fitting a curve or running a regression in one of the procedures listed above. If you return to the Data Processor after running a regression, then the program will hold the regression variable list and the regression coefficients in its memory. Placing the active cell on an empty column and entering the function **Reg**, the fitted values can be recomputed.

The **Reg** function allows performance of prediction and *what if* scenarios by adding new observations to the regression variables or changing their values respectively. The **Reg** function will not work properly if the positions of the regression variables in the data matrix are changed or other procedures are executed after running a regression.

Geometric and exponential fits or any other curve fitting options with one or more logarithmic axes will generate coefficients for the transformed variables. In such cases it is left to the user to transform the fitted results back to the original coordinates by using the necessary Data Processor functions (like **Exp()**). Also remember that coefficients for the fitted equation will be saved in the file POLYCOEF.TXT (see 4.1.1.2. Curve Fitting and 4.2.1.5. Surface Fitting).

Rnd(): Generates Random Numbers between 0 and 1, with seed m.

Syntax:
Rnd(m**)**

Example:
Int(100***Rnd(**-1**))**

You can also generate Random Numbers conforming to a number of Distribution Functions in Descriptive Statistics module.

Rno: Returns the row number of the cell.

Syntax:
Rno

Example:
1899+ **Rno**

Row(): Returns the value of the mth previous row of the current column.

Syntax:
Row(-m**)**

This function allows the user to perform a variety of row operations. It is possible to mix more than one lag within one formula, including **Row(**0**)** which returns the value of the current row. Positive integers cannot be used. The user must provide as many initial values as the highest number of lags. The function will start computing with the row number corresponding to the highest lag value.

Examples:
Row(-1**)**

If the first cell of the current column contains 1, then this will return the row numbers just like the function **Rno**.

Row(-3**)*****Sqr(Row(**-1**)**/2**)**+ **Row(**0**)**

In this case the first three rows of the current column must contain numbers.

WARNING! *The* **Row()** *function must be used with care. Operations containing exponentiation or multiplication may quickly result in a number overflow.*

3.4.2.7. Conditional Functions

Complex Conditional Functions can be computed by means of the **If()** function.

Syntax:
If(*arg1***)**; *arg2*; *arg3*

where:
arg1 is a logical condition,
arg2 is returned when arg1 is true, and
arg3 is returned when arg1 is false.

Any combination of the following logical operators can be used to construct the logical condition *arg1*:

And: Bitwise **And** operator. e.g. 7 **And** 14 = 6, True **And** False = False.

Or: Bitwise **Or** operator. e.g. 7 **Or** 14 = 15, True **Or** False = True.

Not(*arg***):** Returns the bitwise **Not()** of *arg*. e.g. **Not(**1**)** = -2, **Not(**True**)** = False. Use of parentheses in **Not()** is compulsory (without a space between **Not** and the left parenthesis).

arg2 and *arg3* can be mathematical expressions containing functions, scalars or missing values.

The **If()** function and use of logical operators are only available in the Data Processor.

Examples:
If(*C1*>*C2***)**;*C1*;*C2*
If(*C1*+*C2*>100**)**;*C1*^2-2;*C1*^3+3
If(*C3*>0 **Or Not(***C2*<5**))**;*;**Rno**
If(*Label*>1 **And** *Label*<2**)**;*Label*;*

Columns containing String Data can also be used as arguments of an **If()** function. When a particular value is referred to, it should be enclosed within single quotes:
If(*C3*>'*England*' **Or** *C2*='*Scotland*'**)**;*;*C2*
If(*Label*>1 **And** *Label*<2**)**; '*Yes*';'*No*'

String Data are compared according to their alphabetical ordering.

3.4.2.8. Constants

These functions return the commonly used constants in formulas.

Missing Value: This returns UNISTAT's internal missing value code as displayed in Data Processor's Tools → Options dialogue (see 2.4.1.1. Memory Management).

Pi(): Returns the value 3.14159265358979.

e(): Returns the value 2.71828182845905.

3.4.3. Compute Matrix

This option is used for updating formulas after changes have been made in their argument data columns. If changes have been made in columns containing data and if these are used as arguments of a formula (directly or indirectly), then values of the function can be updated by selecting Formula → Compute Matrix.

3.4.4. Calculate

Any calculations involving scalars can be performed. It is also possible to include the Data Processor's Scalar Functions **Len()**, **Mis()**, **Avg()**, **Var()**, **Std()**, **Ser()**, **Sum()**, **Ssq()**, **Min()** and **Max()** in expressions.

After writing the expression press <Enter/OK> once to display the result. Pressing it once again will enter this value in the active cell. Pressing any other key will clear the Input Panel.

UNISTAT Statistical Package

Chapter 4
Data and Function Plots

4.0. Overview

The graphics procedures accessible from the Graph menu are explained in this chapter. For graphics tools common to all graphics procedures see 2.3. Graphics Editor.

The data plotting options will first ask for the variables to be plotted, by means of the usual Variable Selection Dialogue. After this, an automatically scaled and annotated graph is displayed. Choices made by the program include scaling of axes (see Scale Type) and estimating initial values for the minimum, maximum and minor and major interval values for each axis, and, where appropriate, filling in the legend fields with the label of the selected variables. In most cases the program's choices will generate a satisfactory graph. However, the graphs can be fully customised by the user.

Some inputs are transportable between different graph procedures. For instance, font selections, the main title, sub title, axis titles and Row Labels are not deleted when a particular graphics procedure is exited. This feature is particularly useful when the same data set is to be plotted using different graph types.

UNISTAT's Graphics Editor supports full on-screen object editing of graphs. All text labels, legends, and the graph itself can be drag-dropped and resized and new text and shape objects added (see 2.3.2. On-Screen Editing).

A second toolbar contains controls for adding new text and line, rectangle, rounded rectangle and ellipse (circle) objects. There are also controls for changing all aspects of shape objects. These include border colour, fill colour, border style, fill style and border thickness.

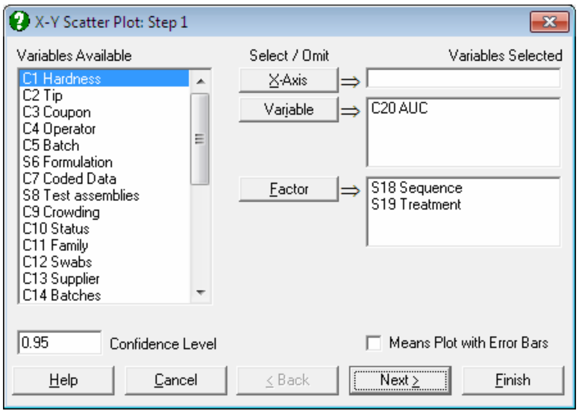
Another useful feature of the Graphics Editor is the Chart Gallery, which provides access to 29 graph types using the same data set. The selected graph types from the gallery will be drawn immediately with the already selected variables, without going through the variable selection process again.

Any information entered or edited by the user can be saved in graphics template files, which can store all graphics objects (text and shapes) and their positions. The graphics template files can be opened subsequently to apply a particular style to different data sets. Graphs can also be exported to other applications in either bitmap or enhanced metafile formats, either via the Windows clipboard or by saving to a file first.

4.1. 2D Plots

4.1.1. X-Y Plots

The Variable Selection Dialogue for this procedure allows plotting an unlimited number of data series. Each data series can be a column of the data matrix, or alternatively, a subsample of a data column defined by one or more factor columns. At least one data column must be selected by clicking on [Variable].



X-Axis ⇒

Variable ⇒ C20 AUC

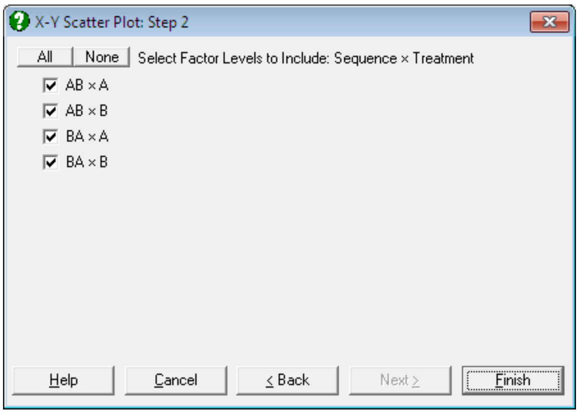
Factor ⇒ S18 Sequence
S19 Treatment

0.95 Confidence Level

□ Means Plot with Error Bars

It is optional to select an X-axis column by clicking on [X axis]. If an X-axis column is not selected then the program will plot the Y-axis variables against the index (i.e. the row numbers). Each axis can have the Scale Type Log base 10, Log base e, log based to any user-defined value, reciprocal, logit, probit, gompit (cloglog) or loglog.

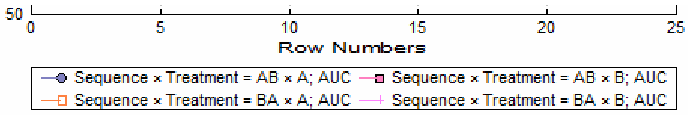
Categorical Plot: Selecting a [Factor] column is optional. In case one is selected, this will define the subgroups of the data column and each subgroup will be plotted as a separate series. If more than one [Factor] is selected, then combinations of factor levels will define the subgroups. An unlimited number of data and factor variables can be selected simultaneously. For more information on these data types see 5.0.1. Multisample Data Types. When at least one factor column is selected, a further dialogue pops up displaying a check list of all combinations of levels. You can then select the ones to be included in the plot.



Select Factor Levels to Include: Sequence × Treatment

- ☑ AB × A
- ☑ AB × B
- ☑ BA × A
- ☑ BA × B

Factor labels are now included in legends and *Point identification* (see 2.3.2.3. Interactive Data Points).



If these labels take too much space, you can switch them off by selecting Edit → Options → Legend from the menu or by double-clicking on the Legend Object..

Means Plot with Error Bars: If this box on the Variable Selection Dialogue is checked, each point on the plot will represent the mean of a data series rather than an individual data point. For a detailed description of this option see 4.1.1.3. Means Plot. Note that Polar Plot, Bar Chart, Area Chart and Ribbon Chart procedures also support this feature.

Interactive Data Points: When the graph is still in Graphics Editor (i.e. before it is sent to an output medium such as Excel or Word), the data points plotted by this procedure are linked with the data in data matrix (see 2.3.2.3. Interactive Data Points). Move the mouse pointer over a data point and press down the right mouse button. A tooltip-like information panel will be displayed about that point until you release the right mouse button. If you are running UNISTAT in Stand-Alone Mode, the row of Data Processor containing this point will also be highlighted. If the delete key is pressed while highlighting a point, this point will be excluded from the plot and the graph will be redrawn. In Stand-Alone Mode, it is also possible to select a row of the spreadsheet to highlight the points on the graph which belong to this row.

Missing Values: Any x-y pairs with at least one missing value are treated as missing. If symbols are drawn without lines then a missing point will simply not exist on the graph. If however, the Line field is set to one of Line or Curve values, then there are two options provided for two points which have missing data between them: (i) leave a gap between them or (ii) connect them.

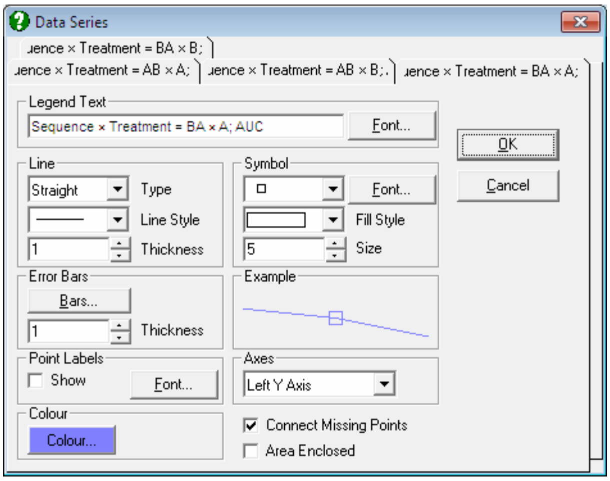
In the first case the lines will stop before a contiguous group of missing values and start again with the first non missing observation. This gives a much better understanding of missing data in a series compared with connecting the two points just before and after a block of missing values.

Unequal Column Lengths: Columns with different lengths can be selected for both X and Y axes. Any pair with at least one *no data* value will be considered as missing.

Date / Time X-Axis: When a date or time variable is selected for the X-Axis, the data points will be separated according to the time difference between them, taking care of leap years, if any. This is also called a *true time axis*. For further information see 3.0.2.5. Date-Time Data.

The number of options available under the Edit Menu depends on the number of columns selected for the graph. The Curve Fitting option will be available when only one Y-axis variable is selected.

4.1.1.1. Data Series



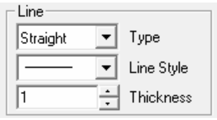
This dialogue provides control over all aspects of individual data series. To display settings for a data series click on the tab index for it. All controls are updated to display the settings for the selected series. The Example box will display the effect of current selections. It will be updated instantly for any changes in controls.

Although an unlimited number of data series can be plotted, properties of only the first nine can be individually controlled using this dialogue. The rest of the series will repeat the properties of the first nine in a circular fashion.

The following aspects of any data series can be controlled independently.

4.1.1.1.1. Line

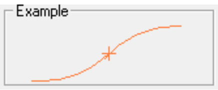
Points can be connected by lines or curves, or a trend line can be fitted on each data series separately. These options are independent of the curve fitting options described below (see 4.1.1.2. Curve Fitting). It is possible, for instance, to draw a trend line, using this option and to fit a polynomial on the same data series using the Edit → Curve Fitting facility.



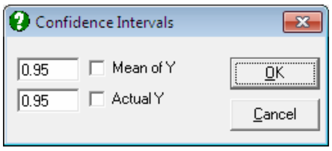
None: No lines or curves are drawn.

Line: Two consecutive x-y points (belonging to the same series) are connected by a straight line. The style of the line (e.g. dashed line, dotted line) can be selected from the Style list in the same frame. Lines will not be drawn for any x-y pairs with at least one missing value.

Curve: Cubic spline interpolation coefficients are computed for each data series. A curve passing through all points is drawn. This option will only work when the X-axis values are in strictly increasing order.



Trend and Confidence Intervals: A line of best fit (linear least squares) is drawn for the selected data series. When this option is selected from the Line list, a further dialogue will pop up allowing you to draw confidence interval curves for the mean of Y and / or actual Y values at any confidence level.


Confidence Intervals

0.95 □ Mean of Y

0.95 □ Actual Y

Either or both confidence intervals or none can be drawn by checking the control boxes on the left as desired. The text fields on the left can be edited for any confidence level between 0 and 1. The two confidence intervals are computed as follows:

1) Confidence interval for the mean of Y:

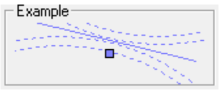
$$(C_0 + C_1 X_0) \pm t_{\alpha/2} \sigma_2 \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i{}^2}}$$

2) Confidence interval for actual values of Y:

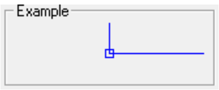
$$(C_0 + C_1 X_0) \pm t_{\alpha/2}\sigma_2 \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}}$$

where X_0 is any given value of X, the first term in brackets is the fitted Y value, the next term is the critical t-value for an α / 2 level of significance with n - 2 degrees of freedom. The next term is the estimate for the standard error of the disturbance term and x_i is the difference between X_i and the mean of X_i.

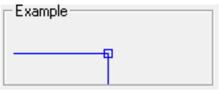
Coefficients for the fitted line, standard errors or the R-squared values are not displayed. If you want to display these parameters on the graph, use the polynomial fit option with a degree one (see 4.1.1.2. Curve Fitting). You can also run the Linear Regression procedure for full output.



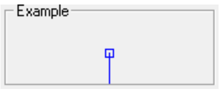
Step Right: A horizontal line is drawn from the current point to the X coordinate of the next point. Then a vertical line is drawn connecting to the next point.



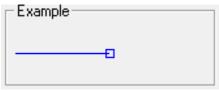
Step Down: A vertical line is drawn from the current point to the Y coordinate of the next point. Then a horizontal line is drawn connecting to the next point.



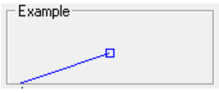
X-connect: A vertical line is drawn from the current point to the X-axis.



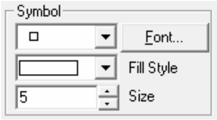
Y-connect: A horizontal line is drawn from the current point to the Y-axis.



O-connect (Vector lines): A line is drawn from the current point to the origin.

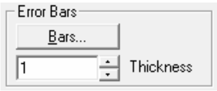


4.1.1.1.2. Symbols



Hundreds of different types of Symbols can be selected for the X-Y points.

4.1.1.1.3. Error Bars

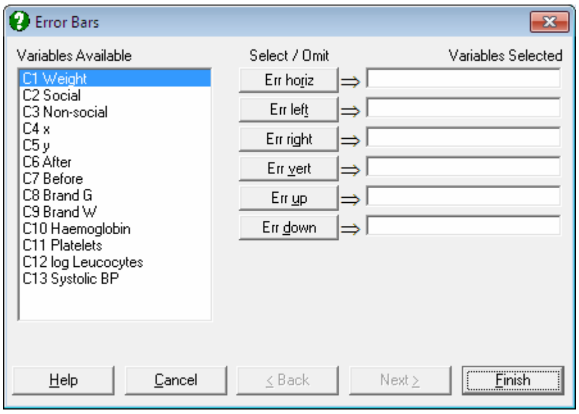


In most cases, a means plot can be generated automatically by checking the Means Plot with Error Bars box on the Variable Selection Dialogue of X-Y Plots (see 4.1.1.3. Means Plot). However, the error bars feature provided in the Data Series dialogue is much more powerful, allowing for horizontal and asymmetric error bars.

The [Bars…] button in the Error Bars group provides access to a Variable Selection Dialogue, where it is possible to select the following:

- Horizontal Error Bars
    - Symmetric
    - Left
    - Right

- Vertical Error Bars
  - Symmetric
  - Up
  - Down



Any columns in the data matrix can be selected for the values of error bars. Horizontal error bars can be symmetric, in which case only one column is selected by clicking on [Err horiz], or they can be asymmetric in which case the column containing left-pointing bars is selected by clicking on [Err left] and the column containing right-pointing bars is selected by clicking on [Err right]. Simultaneously, and independent of the horizontal error bars, vertical error bars can also be symmetric, in which case only one column is selected by clicking on [Err vert], or they can be asymmetric in which case the column containing up-pointing bars is selected by clicking on [Err up] and the column containing down-pointing bars is selected by clicking on [Err down]. Error bars in any direction can be selected independently for all data series.

This method of displaying error bars assumes that you already have data column(s) in the spreadsheet (e.g. standard errors, standard deviations) to be displayed as error bars, prior to selecting the graphics procedure from the menu. If you are running UNISTAT in Stand-Alone Mode, you can easily generate columns containing means and standard errors for a range of data columns using the Range Statistics procedure in Data Processor. In Excel Add-In Mode, you can use one of Summary Statistics or Sample Statistics procedures with Output variables in rows option to create data for means and standard errors.

When calculating the new minimum and maximum axis values the program will take error bars into consideration. Therefore, after selecting error bars from this dialogue, it is normal to be warned by the program that the axes will be rescaled.

The size of the tip of the error bars can be controlled by entering and editing the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

```
ErrorBarSize=9
```

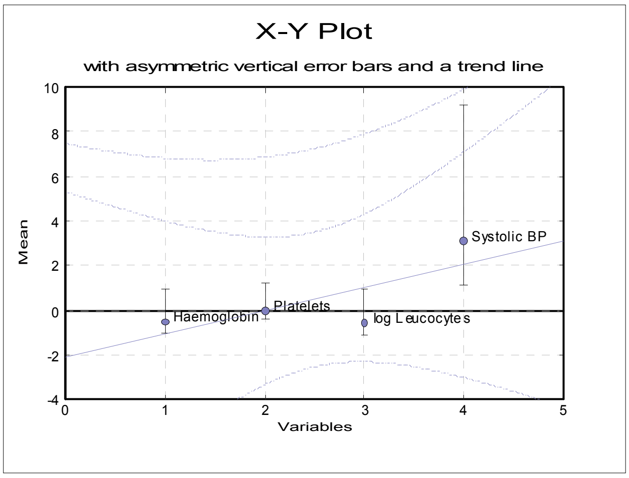
Example

Open PARTEST and select Statistics 1 → Descriptive Statistics → Summary Statistics. Select *Haemoglobin*, *Platelets*, *log Leucocytes*, and *Systolic BP* (*C10* to *C13*) as [Variable]s, check Output variables in rows and from the Output Options Dialogue select only the Mean, Standard Deviation and Standard Error options and click [Finish].

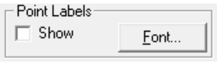
Summary Statistics

	Mean	Standard Deviation	Standard Error
Haemoglobin	-0.5300	1.4629	0.4626
Platelets	-0.0300	1.2193	0.3856
log Leucocytes	-0.5900	1.5524	0.4909
Systolic BP	3.1000	6.1545	1.9462

In Excel Add-In Mode, select the output matrix as data (including its row and column labels) and select Graph → 2D Plots → X-Y Plots. From the Output Options Dialogue select Mean as [Variable], leave the Standard Means Plot with Error Bars box unchecked and click [Finish]. When the graph is displayed, select Edit → Data Series select Line Type as Trend, check Mean of Y and Actual Y boxes and click [OK]. Also, check the Point Labels Show box. Then, click on [Bars…], select *Standard Deviation* (*C2*) as [Err up] and *Standard Error* (*C3*) as [Err down] and click [Finish].



4.1.1.1.4. Point Labels



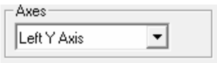
This option is useful for tracing the locations of individual x-y points. When the Show box is checked, Row Labels will be drawn next to the x-y points. If there are no Row Labels then the row numbers will be printed.

In Stand-Alone Mode, Row Labels may be entered and edited using the Data Processor's Edit → Row Labels facility. In Excel Add-In Mode, Row Labels are assumed to be in the first column of the highlighted block and the first column should be selected as *row labels* rather than data (see 1.3.2. Excel Add-In Mode).

By default, the colour of point labels is selected from the Font dialogue and the colour selected applies to all point labels. If you wish to display point labels in the same colour as the series line and symbols, enter the following line in *Documents**Unistat65**Unistat65.ini* file under the [Options] group:

```
PointLabelSeriesColour=1
```

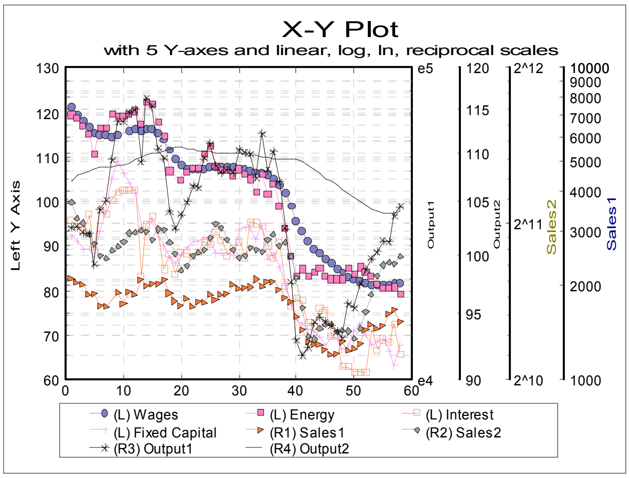
4.1.1.1.5. Right Y-Axes



Axes

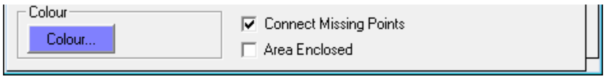
Left Y Axis

Each Y-axis variable can be displayed independently on the left Y-axis or on one of four right Y-axes, by means of the Axes drop-down list.



The program will scale each axis separately, for the variables selected for that particular axis. The legend for each Y-axis variable will also contain either an *L* or *R1*, *R2*, …, indicating to which axis this variable belongs.

4.1.1.1.6. Area Enclosed



Colour

Colour...

☑ Connect Missing Points

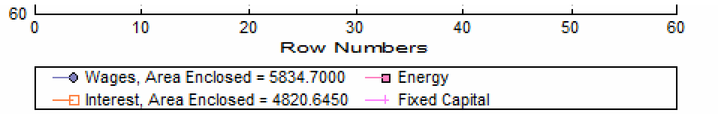
□ Area Enclosed

The area enclosed between a data series and the Y = 0 line (not the minimum of Y-axis) can be computed for each variable separately. The area enclosed is displayed in the legend when all of the following conditions are met:

- X-Axis variable is strictly increasing,
- Y-Axis variable is nonnegative,
- line type is Straight for the variable,

- no Curve Fit option is selected.

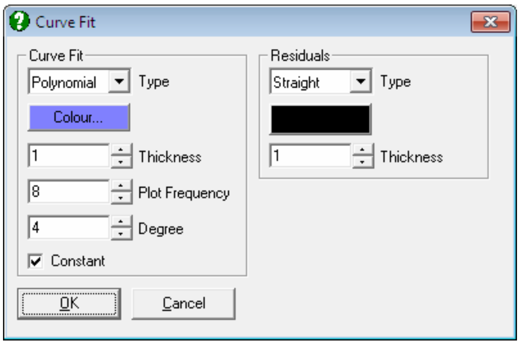
The area enclosed under the curve is displayed in the legend for each variable.



If the Connect missing points box is not checked, only the area under the lines drawn between data points is computed.

4.1.1.2. Curve Fitting

Five different types of curves can be fitted on a bivariate plot, that is, when only one Y-axis variable is selected. This option will not be available if more than one Y-axis variable has been selected.



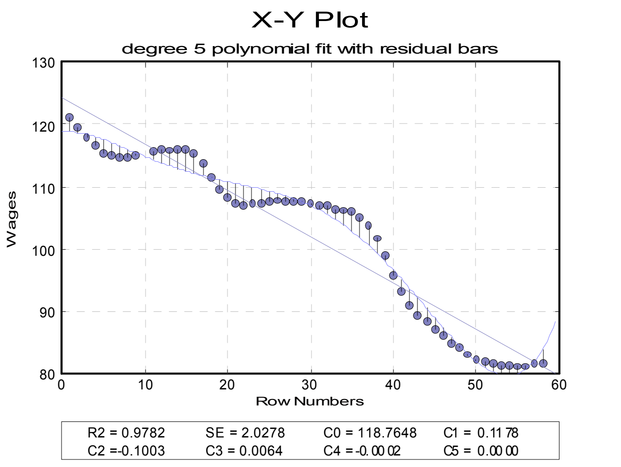
Curve Fit

Curve Fit
Type: Polynomial
Colour...
Thickness: 1
Plot Frequency: 8
Degree: 4
☑ Constant

Residuals
Type: Straight
Thickness: 1

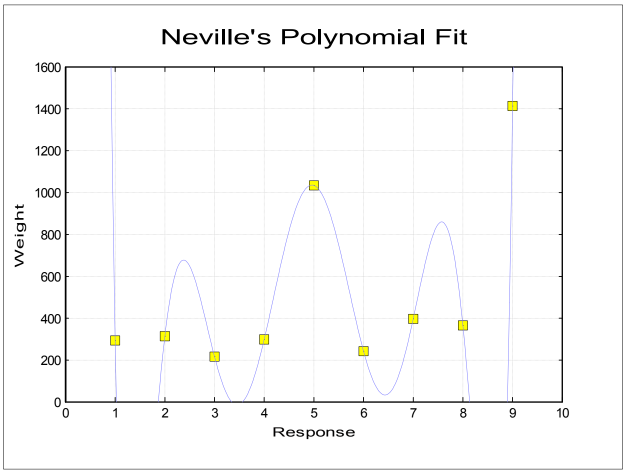
OK Cancel

A new feature with this version of UNISTAT is the facility to display residual bars on the series. When the Type is Straight, vertical lines connecting each data point to the fitted curve will be drawn. It is possible to control the colour and thickness of residual bars and whether they are to be displayed or not.

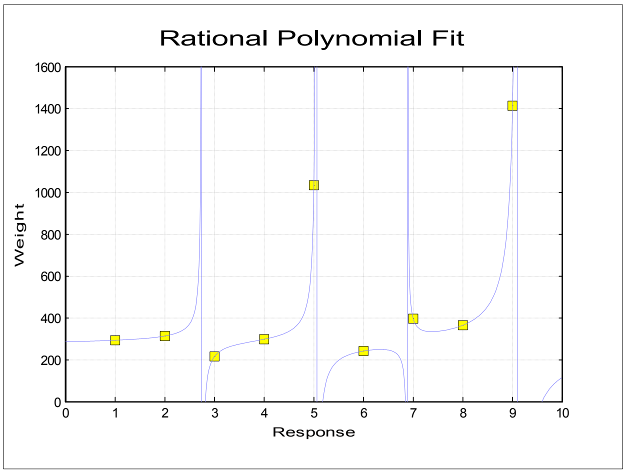


In addition to the R-squared and standard error values, coefficients of the fitted equation are displayed in the legend for polynomial, geometric and exponential fits. In Stand-Alone Mode, it is also possible to run interpolations on the fitted curve, without having to retype these coefficients, using the Data Processor's **Reg** function (see 3.4.2.6.3. UNISTAT Functions). The same coefficients will also be saved automatically in the file POLYCOEF.TXT, in the order of constant term (if any), *X^1, X^2, ..., X^r* for a degree r polynomial.

Neville: For a variable containing *r* observations, coefficients of a degree *r* - 1 polynomial passing through all points is calculated. X-axis values must be strictly increasing. Typically, this is a polynomial wildly oscillating at extreme x values. The amount of computing time will increase quickly with increasing *r*.

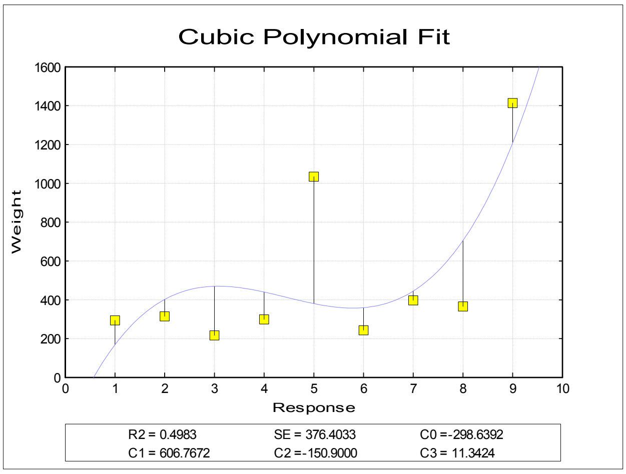


Rational: Rational functions are quotients of polynomials. Like the Neville's polynomial, this procedure will also draw a curve passing through all points, but it will probably have many points of discontinuity. These are the points where the denominator of the rational function approaches zero. Like Neville's polynomial, it is not practical to fit rational functions on large series due to intensive computing requirements. X-axis values must be in strictly increasing order.



Polynomial: When this option is selected, two new controls Degree and Const will be placed in the Curve Fit dialogue. In this way, you are provided with the possibility of fitting polynomials of any degree and with or without a constant term. A line of best fit (i.e. the plot of bivariate regression) is equivalent to fitting a first degree polynomial.

A dedicated Polynomial Regression algorithm is used to estimate the coefficients of the least squares fit. The estimated coefficients, R-squared and standard error of regression are displayed. If values of the column selected for X-axis are too large and a high degree polynomial fit is attempted, then a number overflow may occur. Although this error will be trapped by the program in most cases, it cannot be guaranteed that all overflow errors can be trapped. Some errors may result in a crash causing loss of data.



WARNING! *You must ensure that the combination of X and Y-axis values and the degree of the polynomial fitted are low enough not to cause a number overflow.*

Even if a crash does not occur, the precision of fits will be poorer with large X-axis values. All this can be easily overcome by scaling down the values of the column to be chosen as X-axis before fitting a polynomial. For instance, if the X-axis consists of years 1950 to 1999 then much better results can be obtained by changing these values to 50 to 99.

The estimated coefficients are saved in memory so that they can be shared by the Plot and Roots of Polynomials procedure. Therefore, once a polynomial is fitted on data it is possible to plot the estimated polynomial in any interval and also to determine its roots by choosing the Plot and Roots of Polynomials procedure.

In Stand-Alone Mode, it is also possible to run interpolations on the fitted curve, without having to retype these coefficients. This is done using the Data Processor's **Reg** function (see 3.4.2.6.3. UNISTAT Functions). The same coefficients will also be saved automatically in the file POLYCOEF.TXT, in the order of constant term (if any), *X^1, X^2, ..., X^r.*

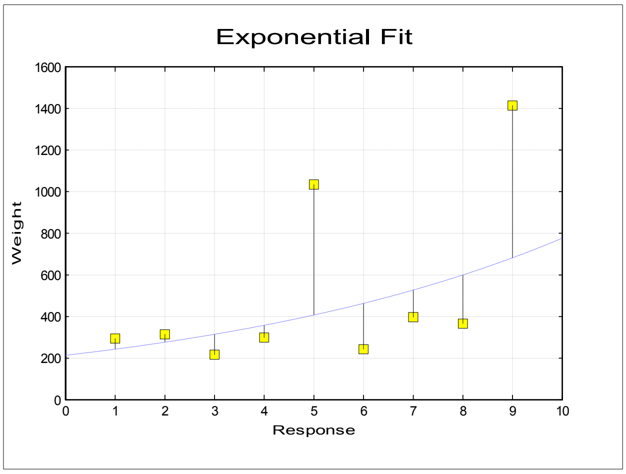
Exponential: The following least squares model (exponential regression) is fitted on data:

$$y = c_0 \, \mathrm{Exp}(c_1 x_1) \ldots \mathrm{Exp}(c_n x_n)$$

The equation is first linearised as:

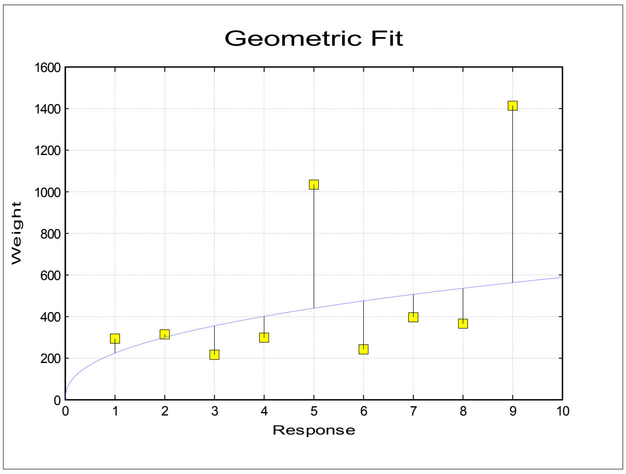
$$\mathrm{Ln}(y) = \mathrm{Ln}(c_0) + c_1 x_1 + \ldots + c_n x_n$$

The constant term can be omitted. If the y variable contains non positive values, then the program reports data as unsuitable for exponential fit.



Geometric: The following least squares model (geometric regression) is fitted on data:

$$y = c_0 \, x_1^{c_1} \ldots x_n^{c_n}$$



The equation is first linearised as:

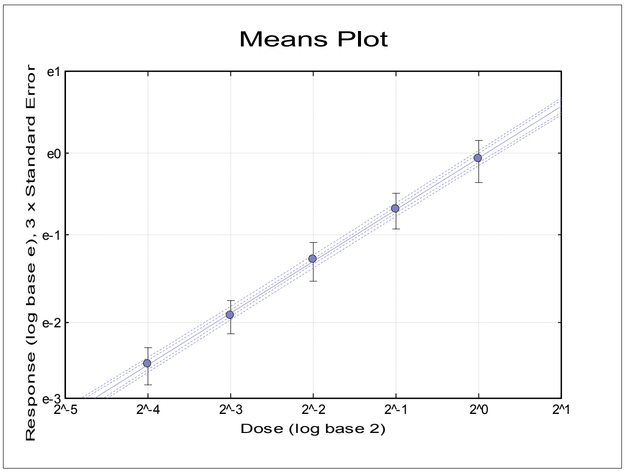
$$\mathrm{Ln}(y) = \mathrm{Ln}(c_0) + c_1\mathrm{Ln}(x_1) + \ldots + c_n\mathrm{Ln}(x_n)$$

The constant term can be omitted. If one of the X or Y variables contains non positive values, then the program reports data as unsuitable for geometric fit.

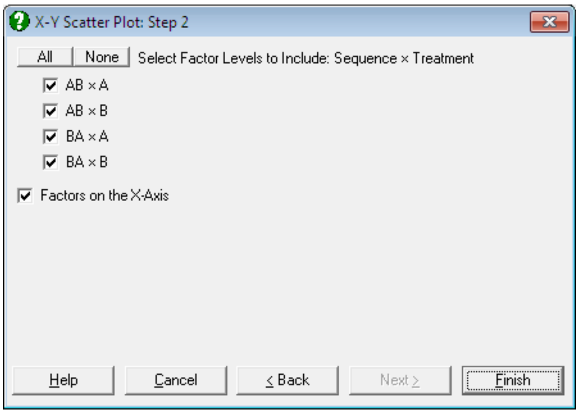
4.1.1.3. Means Plot

The Variable Selection Dialogue of X-Y Plots, Polar Plot, Bar Chart, Area Chart and Ribbon Chart procedures support a Means Plot with Error Bars check box. When it is checked, each point on the plot will represent the mean of a data series rather than an individual data point. By default, the program also plots the standard error of mean for each point in the form of a symmetric vertical error bar. It is possible to switch off the display of error bars or select other measures of dispersion.

It is also possible to select a continuous variable for the X-Axis, where one or more Y-Axis variables have multiple values corresponding to the same X-Axis variable value. A typical case is the dose-response plot where there are several response variable values for each dose level. When one more factor variables are also selected, the X-Axis selection will be ignored.



If one or more factor columns are selected, then the means of subgroups defined by combinations of factor levels are plotted. In this case a further dialogue pops up, displaying a check list of all combinations of levels. This dialogue also contains a check box Factors on the X-Axis, which is used to determine whether the variables or the factors will be represented on the X-Axis.



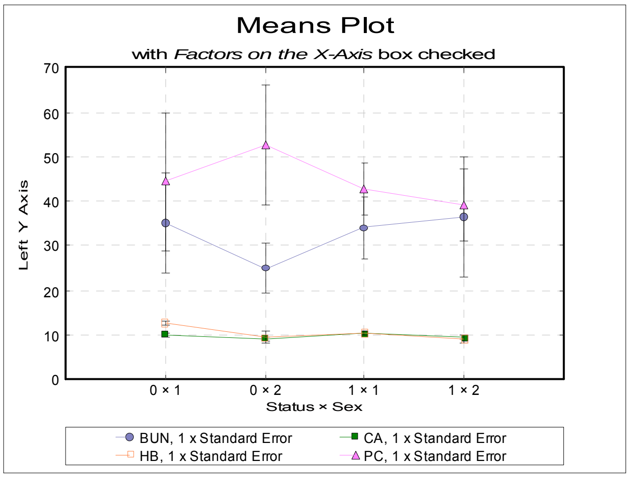
Select Factor Levels to Include: Sequence × Treatment

☑ AB × A
☑ AB × B
☑ BA × A
☑ BA × B

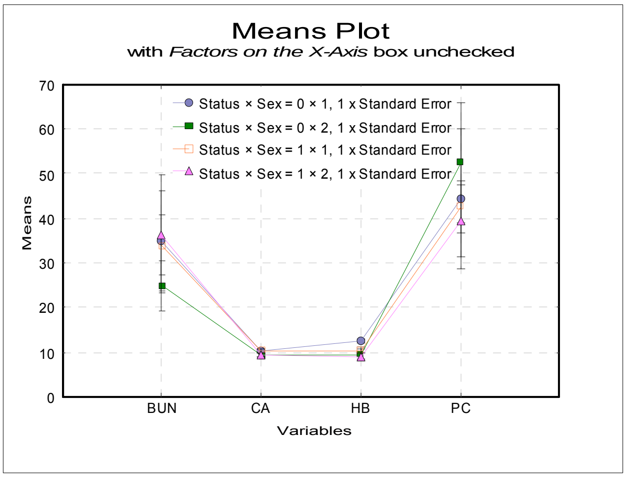
☑ Factors on the X-Axis

When only one factor is selected, Means Plot looks similar to Profile Plot of GLM procedure (see 7.3.2.3. GLM Output Options).

The following example illustrates a Means Plot with four variables, two factors and the Factors on the X-Axis box checked. Here, the combination of factor levels are represented on the X-Axis and variables in different lines.

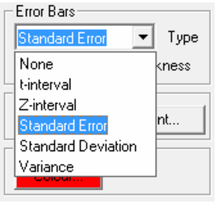


This example is for the same set of variables with the Factors on the X-Axis box unchecked. Here variables are represented on the X-Axis and factor levels in separate lines.



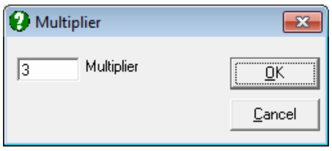
The Error Bars control on the Edit → Data Series dialogue for Means Plot allows selecting one of the following dispersion measures.

- None
- t-interval
- Z-interval
- Standard Error
- Standard Deviation
- Variance



Selecting a new error bar type from the list will enforce a re-scaling of the relevant axis. The confidence level for t- and Z- intervals can be set from the Variable Selection Dialogue.

When one of Standard Error or Standard Deviation options is selected, a dialogue pops up asking for a multiplier.



Multiplier: 3

Error bars for standard error will then be calculated as:

$$\text{Lower limit} = \text{Mean} - \text{k x SE}$$
$$\text{Upper limit} = \text{Mean} + \text{k x SE}$$

and for standard deviation:

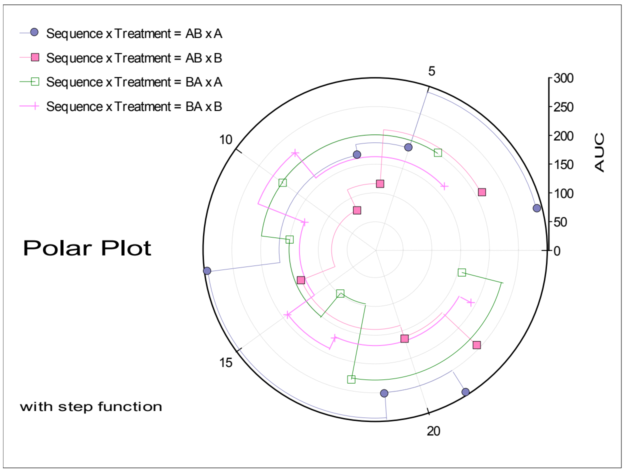
$$\text{Lower limit} = \text{Mean} - \text{k x s}$$
$$\text{Upper limit} = \text{Mean} + \text{k x s}$$

where k is the multiplier defined by the user.

4.1.2. Polar Plot

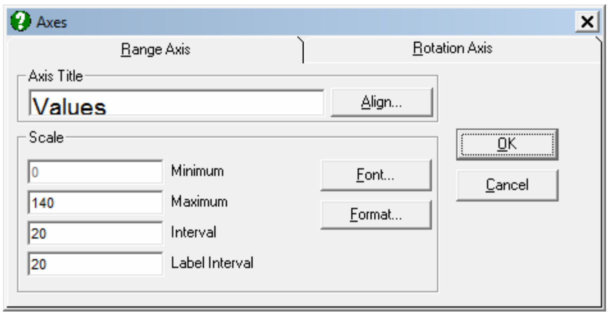
The Variable Selection Dialogue for this procedure is similar to that of X-Y Plots. It is possible to plot an unlimited number of data series. Each data series can be a column of the data matrix, or alternatively, a subsample of a data column defined by one or more factor columns. The Means Plot with Error Bars check box allows plotting means of series with their error bars (see 4.1.1.3. Means Plot).

In both cases it is optional to select a data column for the Rotation axis (which is equivalent to X-axis in X-Y Plots) by clicking on [X axis]. If an X-axis variable is selected then the default is a Polar Plot in degrees. If an X-axis variable is not selected then the default is a star diagram. In this way, the Polar Plot procedure can draw true polar plots or star diagrams. The number of units per revolution is controlled using the rotation scale option. The maximum value is the value per revolution. Hence, to set up a Polar Plot in degrees, the maximum value should be 360. To set up a Polar Plot in radians the maximum value should be 2 x Pi().



An unlimited number of data series can be plotted, but properties of only the first nine can be controlled from the Edit → Data Series dialogue. The rest of the series will repeat the properties of the first nine in a circular fashion.

The check box Draw Circular controls whether the Polar Plot should always be drawn as a circle. Otherwise the plot will be drawn as an oval filling up the available space. The Connect Last Point check box for each line controls whether a line should be drawn from the last point in the series to the first point in the series. This allows star diagrams to be drawn. A line will only be drawn if a line style is selected for the particular data series.



Axis Title: Values

Scale

0 Minimum

140 Maximum

20 Interval

20 Label Interval

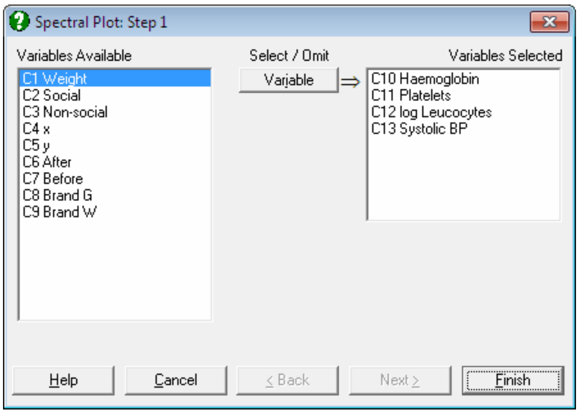
Interactive Data Points: When the graph is still in Graphics Editor (i.e. before it is sent to Excel or Word), the data points plotted by this procedure are linked with the data in data matrix (see 2.3.2.3. Interactive Data Points). Move the mouse pointer over a data point and press down the right mouse button. A tooltip-like information panel will be displayed about that point until you release the right mouse button.

Missing Values: Any pairs with at least one missing value are treated as missing. If symbols are drawn without lines then a missing point will simply not exist. If however, the Line field is set to one of Line or Curve values, then there are two options provided for two points which have missing data between them: (i) leave a gap between them or (ii) connect them.

In the first case the lines will stop before a contiguous group of missing values and start again with the first non missing observation. This gives a much better understanding of missing data in a column than that obtained by connecting the two points just before and after a block of missing values.

Unequal Column Lengths: Columns with different lengths can be selected for both X and Y axes. Any pair with at least one *no data* value will be considered as missing.

4.1.3. Spectral Diagram



Spectral Plot: Step 1

Variables Available

C1 Weight
C2 Social
C3 Non-social
C4 x
C5 y
C6 After
C7 Before
C8 Brand G
C9 Brand W

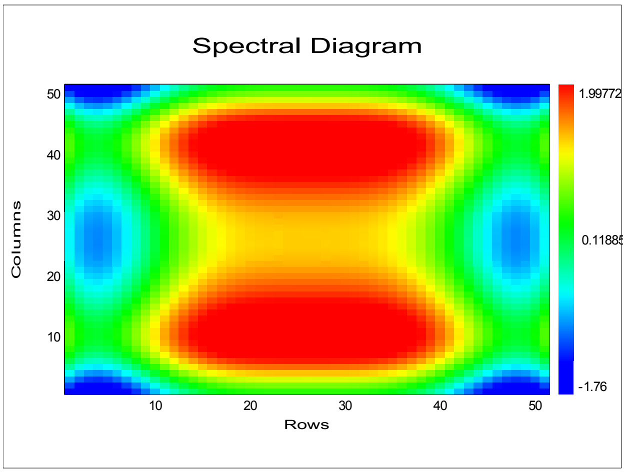
Select / Omit

Variable

Variables Selected

C10 Haemoglobin
C11 Platelets
C12 log Leucocytes
C13 Systolic BP

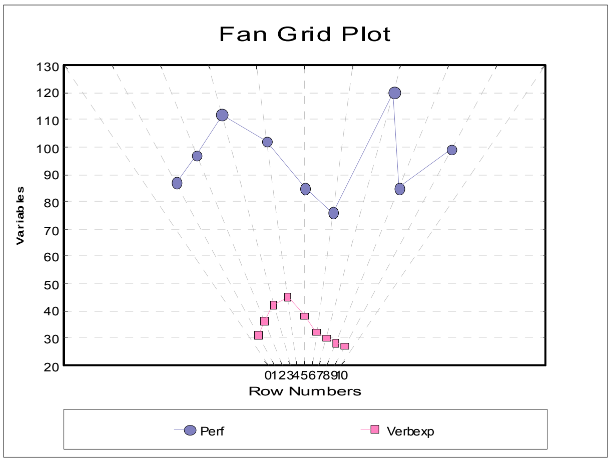
Any number of columns can be selected by clicking [Variable]. The columns may contain missing values and the column lengths may be unequal. The program will assume that the observations at the end of the shorter columns are missing.



A grid is drawn for each cell. The colour of the grid depends on the relative value of the cell. The cells with the lowest values will be coloured blue and the cells with the highest values will be coloured red. A scale of the colours used is shown on the right of the plot.

4.1.4. Fan Grid Plot

This procedure is similar to X-Y Plots except that the vertical grid lines converge at X = *centre*, Y = 0.

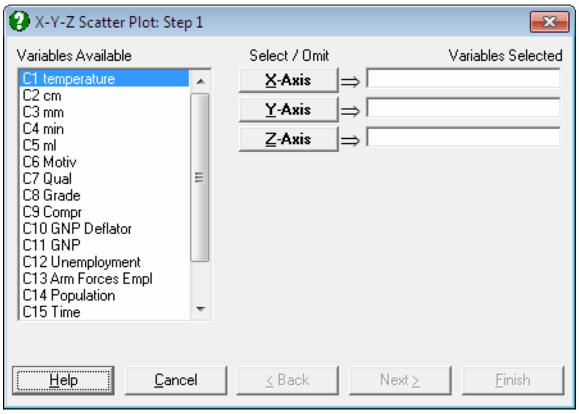


It is possible to control the axes as in X-Y Plots, but error bars, nonlinear axis scaling and right Y-axes are not available in Fan Grid plots.

4.2. 3D Plots

X-Y-Z scatter and grid plots can be drawn. It is also possible to spin a 3-D scatter plot around three axes.

4.2.1. X-Y-Z Scatter Plot

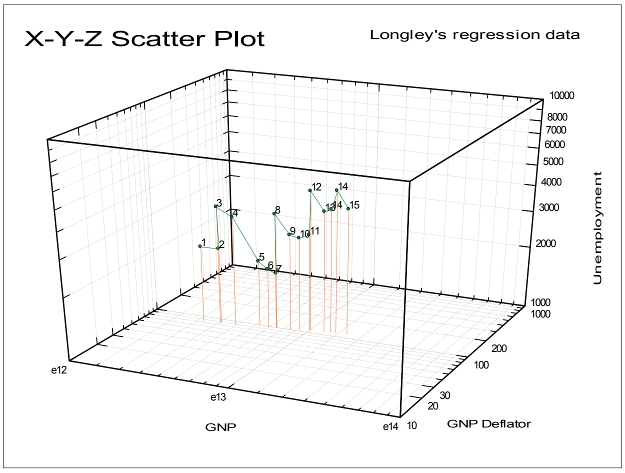


Three columns are selected by clicking on [X axis], [Y axis] and [Z axis]. Each axis can have the Scale Type Log base 10, Log base e, log based to any user-defined value, reciprocal, logit, probit, gompit (cloglog) or loglog. It is also possible to draw bivariate projection plots on X-Y, X-Z and Y-Z planes, error bars in any direction, fit linear regression planes and polynomial surfaces. Contours can be drawn for fitted surfaces.

Interactive Data Points: When the graph is still in Graphics Editor (i.e. before it is sent to Excel or Word), the data points plotted by this procedure are linked with the data in data matrix (see 2.3.2.3. Interactive Data Points). Move the mouse pointer over a data point and press down the right mouse button. A tooltip-like information panel will be displayed about that point until you release the right mouse button. If you are running UNISTAT in Stand-Alone Mode, the row of Data Processor containing this point will also be highlighted. If the delete key is pressed while highlighting a point, this point will be excluded from the plot and the graph will be redrawn. In Stand-Alone Mode, it is also possible to select a row of the spreadsheet to highlight the points on the graph which belong to this row.

Missing Values: Any point with at least one missing value is treated as missing.

Unequal Column Lengths: Columns with different lengths can be selected for X, Y and Z axes. Any triplet with at least one *no data* value will be considered as missing.



The Edit options specific to this procedure are as follows:

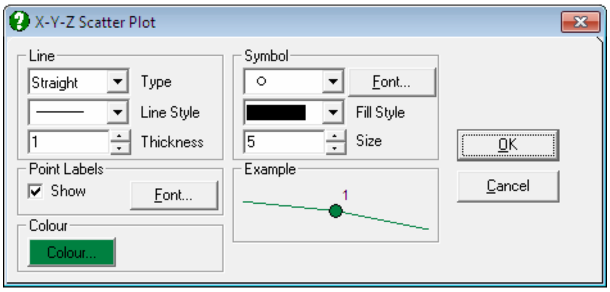
4.2.1.1. Viewpoint

As in all 3D Plots, you can alter the viewpoint and perspective options for the unit cube (see 2.3.4.6. 3D Viewpoint and Perspective).

4.2.1.2. Contours

When a surface is fitted on the data, contour curves can also be drawn (see 2.3.4.7. Contours).

4.2.1.3. X-Y-Z Points



X-Y-Z Scatter Plot

Line

Straight Type

Line Style

1 Thickness

Symbol

○ Font...

Fill Style

5 Size

Point Labels

☑ Show Font...

Colour

Colour...

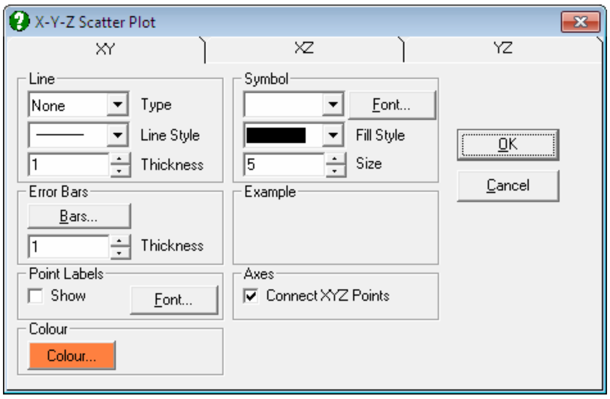
Example

Line: If the Line Type option is set to Line, then the consecutive x-y-z points will be connected with straight lines. There is another Line field in the Planes dialogue, which is used for drawing bivariate projection plots on reference planes (see 4.2.1.4. Planes).

Symbol: Hundreds of different types of symbols can be selected for the x-y-z points (see 2.3.4.5.3. Symbols). Again, this is not to be confused with the Symbol field for the Planes dialogue.

Point Labels: Point Labels will be drawn alongside the x-y-z points. As in X-Y Plots, the text for Row Labels are used. If no Row Labels have been entered then the row numbers will be displayed.

4.2.1.4. Planes



X-Y-Z Scatter Plot

XY XZ YZ

Line

None Type

Line Style

1 Thickness

Symbol

Font...

Fill Style

5 Size

Error Bars

Bars...

1 Thickness

Example

Point Labels

□ Show Font...

Axes

☑ Connect XYZ Points

Colour

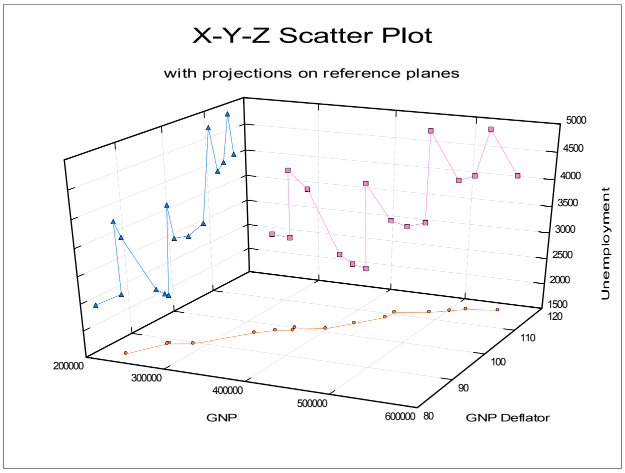
Colour...

This dialogue controls how UNISTAT will draw bivariate plots on X-Y, X-Z and Y-Z planes, connect x-y-z points to reference Planes, and draw symmetric and asymmetric error bars on x-y-z points in any one of six directions.

Click on the desired tab to control the parameters for each plane.

Line: This frame is similar to the one in X-Y Plots (see 4.1.1.1.1. Line), except that the Curve and Frame options are not available for Line Type.

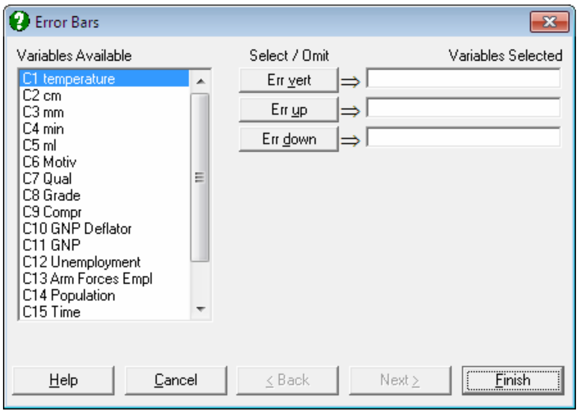
Symbol: Hundreds of different types of Symbols can be selected for the x-y points.



Point Labels: Point Labels will be drawn alongside the x-y points for the selected plane. As in X-Y Plots, the text for Row Labels are used. If no Row Labels have been entered then the row numbers will be displayed.

Connect X-Y-Z Points: If this box is checked, then each x-y-z point will be connected to the currently selected plane with a perpendicular line. On entry, this box will be checked for the X-Y plane.

Error Bars: This will control error bars in the direction perpendicular to the selected plane. For instance, when the Symmetric Bars option is selected for the X-Y Plane, then the up and down error bars will be drawn in the positive and negative Z directions.



One or two data columns containing a dispersion measure (e.g. standard error, standard deviation) can be selected from the Variable Selection Dialogue. Error bars can be symmetric, in which case only one column is selected by clicking on [Err vert], or they can be asymmetric in which case the column containing up-pointing bars is selected by clicking on [Err up] and the one containing down-pointing bars is selected by clicking on [Err down]. Although the *up* and *down* options make sense for the Z direction, they must be interpreted as *left* and *right* or *in* and *out* for the X and Y directions.

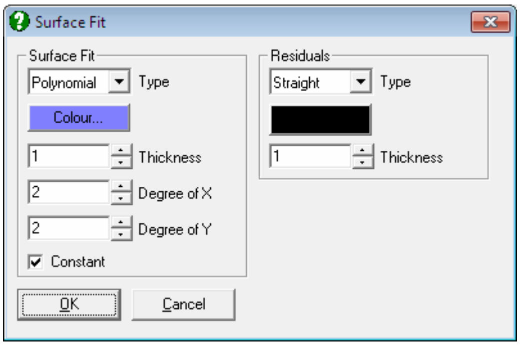
In Stand-Alone Mode, columns containing means and standard errors for a range of data columns can be generated using the Range Statistics procedure in Data Processor. In Excel Add-In Mode, you can use one of Summary Statistics or Sample Statistics procedures with Output variables in rows option to create data for standard errors or standard deviations.

When calculating the minimum and maximum axis values the program will take error bars into consideration.

4.2.1.5. Surface Fitting

The Edit → Surface Fitting option provides access to three surface fitting options: Linear Regression Plane, Polynomial Surface and Weighted Averages Surface. In the first two cases you will have the option of fitting with or without a constant term.

Residual bars, i.e. the lines connecting each data point to the fitted surface can also be drawn. It is possible to control the colour and thickness of residual bars and whether they are to be displayed or not.



Surface Fit

Polynomial Type

Colour...

1 Thickness

2 Degree of X

2 Degree of Y

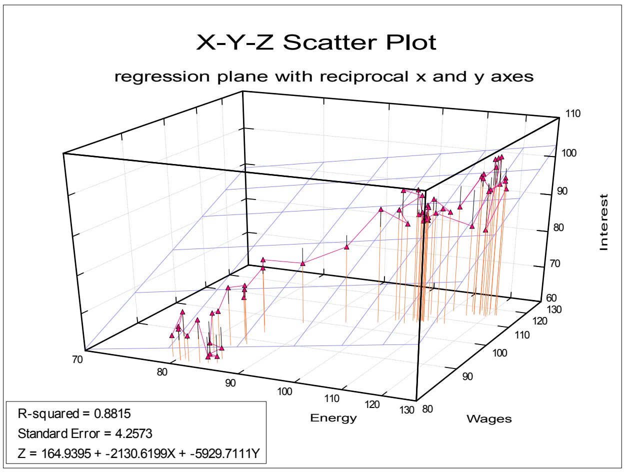
☑ Constant

Residuals

Straight Type

1 Thickness

Coefficients of the fitted plane, as well as its R-squared and standard error values will be displayed in the legend object. However, since there may be up to 15 coefficients for the polynomial surface fit, they will not be displayed in the legend. In Stand-Alone Mode, it is also possible to run interpolations on the fitted curve, without having to retype these coefficients, using the Data Processor's **Reg** function (see 3.4.2.6.3. UNISTAT Functions). The same coefficients will also be saved automatically in the file POLYCOEF.TXT, in the order of constant term (if any), *X^1, X^2, ...,Y^1, Y^2, ..., Y^r* for a degree *r* polynomial.

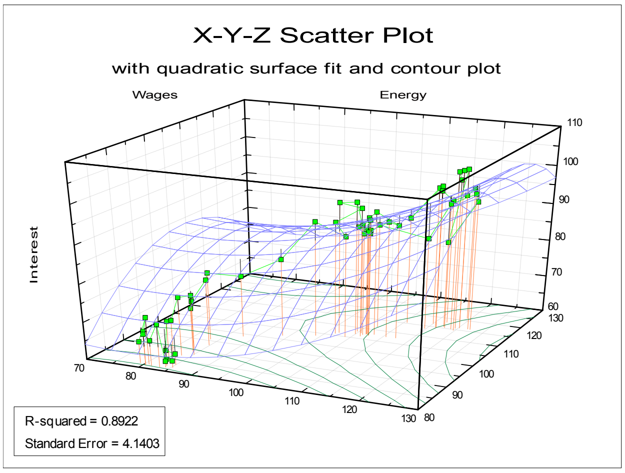


R-squared = 0.8815
Standard Error = 4.2573
Z = 164.9395 + -2130.6199X + -5929.7111Y

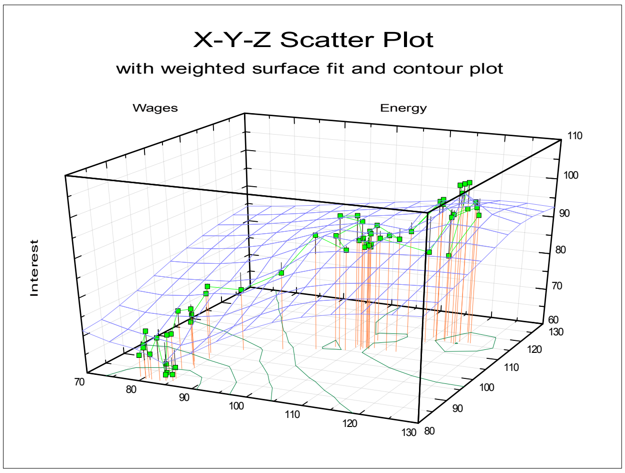
Linear Regression Plane: Output includes R-squared, standard error of regression and the equation of the plane fitted. When a log option is selected for an axis, the fitted values calculated using the Data Processor's **Reg** function (see 3.4.2.6.3. UNISTAT Functions) must be transformed back to the original coordinates.

Polynomial Surface: The degree of X and Y variables can be determined independently and the constant term included or omitted. You must ensure that values of X and Y variables are not too large to cause a number overflow, particularly with higher degree polynomials.

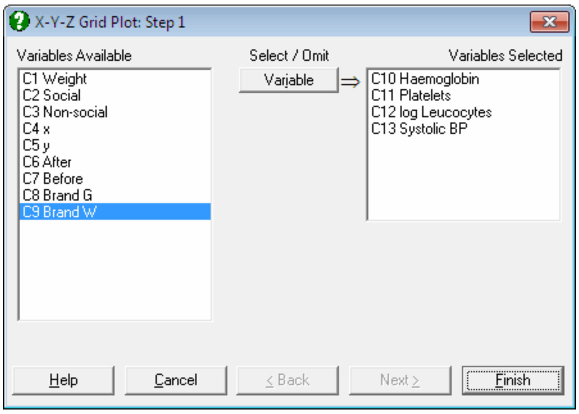
Logarithmic and other nonlinear scaling options on X, Y and Z axes are available for polynomial surface fitting, but the contour curves will not be drawn. Coefficients of the fitted equation will be saved to the file POLYCOEF.TXT. It is also possible to run interpolations using the Data Processor's **Reg** function (see 3.4.2.6.3. UNISTAT Functions).



Weighted Average: For each vertex of the mesh, the Z value of each point is weighted by the inverse square of its distance from the vertex, to form an average value for the fitted surface at this vertex.


X-Y-Z Scatter Plot
with weighted surface fit and contour plot
Wages
Energy
Interest
110
100
90
80
70
60
70
80
90
100
110
120
130
80
90
100
110
120
130

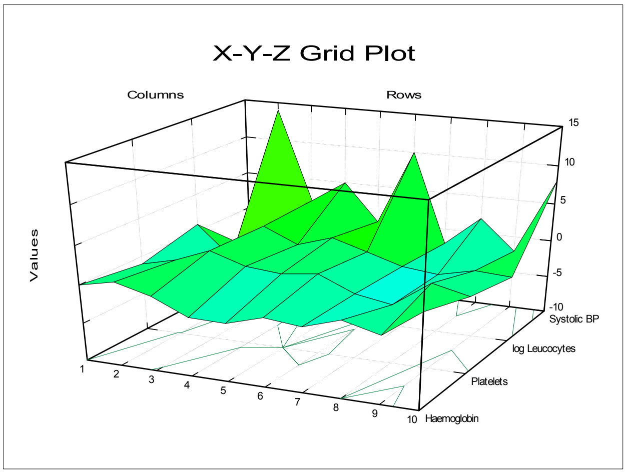
4.2.2. X-Y-Z Grid Plot



Variables Available: C1 Weight, C2 Social, C3 Non-social, C4 x, C5 y, C6 After, C7 Before, C8 Brand G, C9 Brand W

Variables Selected: C10 Haemoglobin, C11 Platelets, C12 log Leucocytes, C13 Systolic BP

An unlimited number of data columns may be selected by clicking on [Variable]. Surface, bar and point graphs can be drawn. Data must be readily produced in the form of a regular grid. This procedure assumes that the elements of the grid are the Z axis values obtained from a function F(x,y) by means of evaluating it at regular X and Y intervals. The grid is assumed to be placed in the data matrix such that columns correspond to values of the Y variable and rows correspond to values of the X variable. For instance, the cell (5,2) of the data block must contain the value F(x(5),y(2)), i.e. value of the bivariate function evaluated at the fifth step of the X variable and the second step of the Y variable.

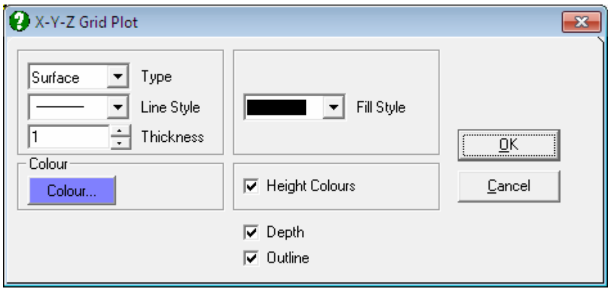


Missing Values: Any polygons with at least one missing data vertex value will be considered as missing.

Unequal Column Lengths: Columns with different lengths can be selected. Any polygons with at least one *no data* vertex value will be considered as missing.

This procedure is intended for three-dimensional plotting of data points. If you want to plot an explicit function of two variables F(x,y) this can be done in the Plot of 3D Functions procedure more easily, without having to generate a grid first.

Z axis values are automatically scaled. For X and Y axes only the minimum and maximum value fields are provided. The values displayed in these fields are the minimum and maximum X and Y grid numbers rather than minimum and maximum X and Y values. They are intended for labelling purposes only and will not affect the appearance of graphs.



X-Y-Z Grid Plot

Type: Surface

Line Style: ———

Thickness: 1

Fill Style: ■

☑ Height Colours

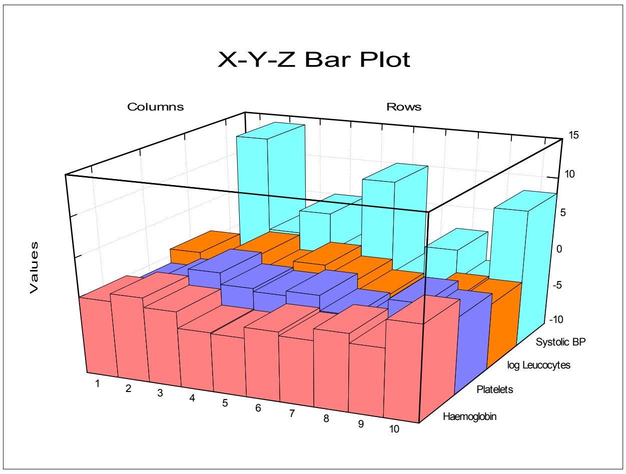
☑ Depth

☑ Outline

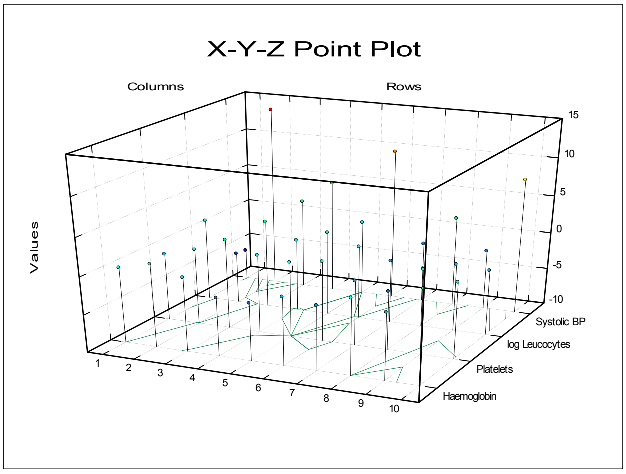
The Edit → Plot Type dialogue provides access to various aspects of the plot. The Type drop-down list is used to set the following three plot types:

Surface Plot: A polygon is drawn for every four neighbouring grid elements. Plotting starts from the far end of the cube so that the most recently drawn surfaces are nearest to the view point. Polygons with one or more missing values will not be drawn. It must be remembered that n grid columns will produce n - 1 polygon rows.

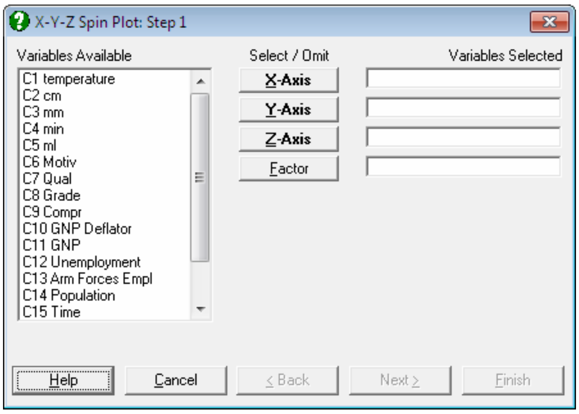
Bar Plot: A three-dimensional bar is drawn for each grid element. The missing elements are not drawn.



Point Plot: A circle symbol is drawn for each grid element and it is connected with the base plane by a vertical line. It is advisable to plot large grids in point form first in order to get some idea of the nature of the data. When the base plane happens to lie above some grid elements, these elements will be connected to the plane by vertical lines pointing upward.



4.2.3. Spin Plot



X-Y-Z Spin Plot: Step 1

Variables Available

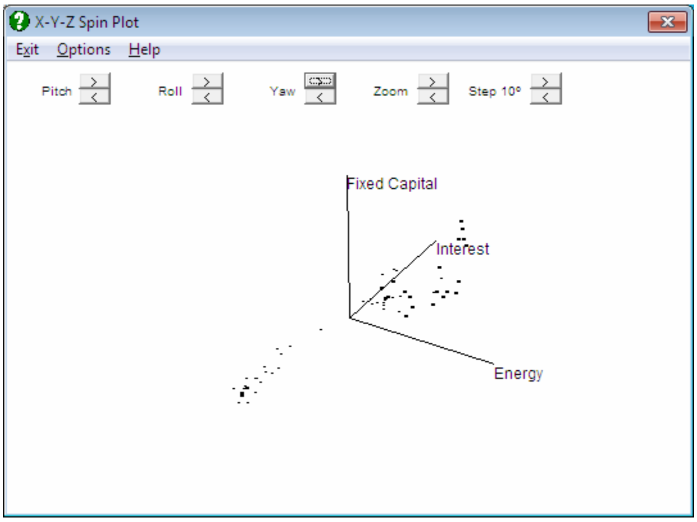
C1 temperature
C2 cm
C3 mm
C4 min
C5 ml
C6 Motiv
C7 Qual
C8 Grade
C9 Compr
C10 GNP Deflator
C11 GNP
C12 Unemployment
C13 Arm Forces Empl
C14 Population
C15 Time

This plot provides an animated 3D visualisation of data. Three variables must be selected (one for each axis) by clicking on [X axis], [Y axis] and [Z axis] from the Variable Selection Dialogue. The program then draws a 3D scatter diagram. Optionally, a factor (categorical) variable can be selected by clicking on [Factor], to plot different subgroups in different colours.

Interactive Data Points: When the graph is still in Graphics Editor (i.e. before it is sent to Excel or Word), the data points plotted by this procedure are linked with the data in data matrix (see 2.3.2.3. Interactive Data Points). Move the mouse pointer over a data point and press down the right mouse button. A tooltip-like information panel will be displayed about that point until you release the right mouse button. If you are running UNISTAT in Stand-Alone Mode, the row of Data Processor containing this point will also be highlighted. If the delete key is pressed while highlighting a point, this point will be excluded from the plot and the graph will be redrawn. In Stand-Alone Mode, it is also possible to select a row of the spreadsheet to highlight the points on the graph which belong to this row.

Missing Values: Any x-y-z triplets with at least one missing value are treated as missing.

Unequal Column Lengths: Columns with different lengths can be selected for X, Y and Z axes. Any point with at least one *no data* value will be considered as missing.



On entry, the view point will be in positive X direction, and all axes will be labelled with their Column Labels (if any). The spin buttons provided at the top of graph can be used to rotate the scatter diagram in six directions. The points nearer to the view point will be larger.

The buttons available on the graph have the following tasks:

Pitch: This rotates the points along the imaginary horizontal axis which lies across the screen.

Roll: This rotates the points along the imaginary axis which passes through the viewpoint.

Yaw: This rotates the points along the imaginary vertical axis which lies on the screen.

Zoom: This moves the view point closer to or further away from the origin.

Step 10°: By default, the rotation is in steps of 10 degrees, which can be increased or decreased by clicking on these buttons.

Also, the menu bar will provide access to the following further options:

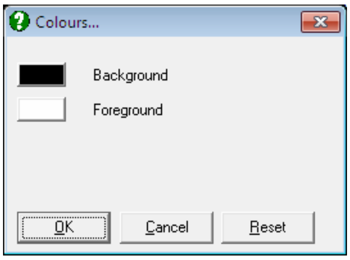
Axes: Uncheck this to remove axes from the plot.

Centre: Uncheck this to move the origin of the graph to (0, 0, 0) in data units. When checked, the origin will be at the centre of all points.

Shake: This shakes the plot slightly from one side to another. After the shake has finished the plot will be in the same position as it began.

Cycle: When this option is selected nothing will happen immediately. However, when one of Pitch, Roll or Yaw buttons is clicked once subsequently, the plot will be rotated a full 360 degree cycle.

Colours: Background and foreground colours can be selected independently.

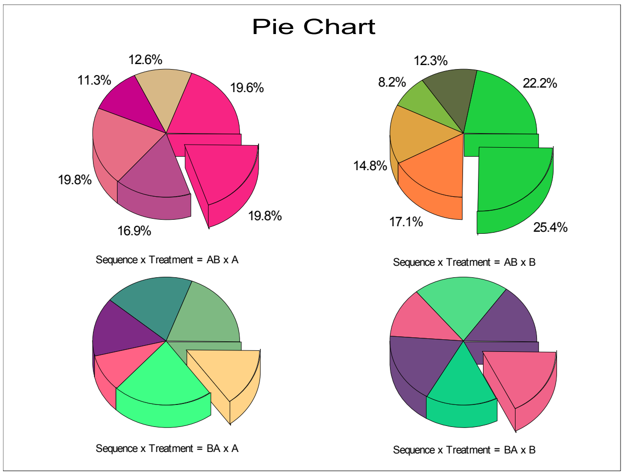


4.3. Charts

Pie, bar, area and 3D bar charts can be drawn. Each procedure provides a series of controls allowing for many different variants of these charts to be plotted.

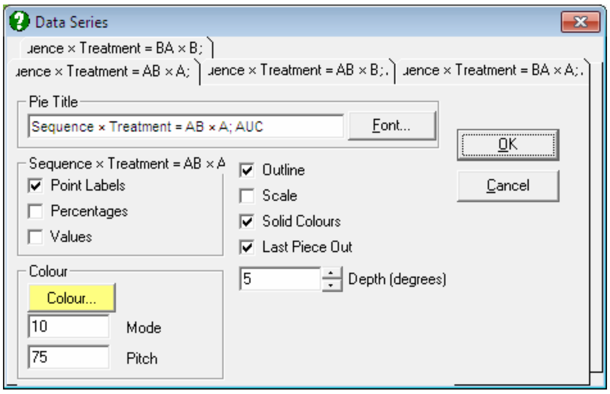
4.3.1. Pie Chart

The Variable Selection Dialogue for this procedure is similar to that of X-Y Plots. Each data series can be a column of the data matrix, or alternatively, a subsample of a data column defined by one or more factor columns. Any number of columns can be selected but only the first six series are plotted. If you need to plot more than six pies, you can use the Pie Icon Plot procedure (see 8.8.3. Icon Plots).



There are no restrictions on the size of the data columns. However, Pie Charts for columns with more than 50 rows look cluttered. In particular, the pie labels must be chosen with care to prevent overlapping. Missing values are simply skipped. Non positive values are considered as missing.

To edit various chart properties select Edit → Data Series. The following controls will apply to all pies.


Data Series
ɹence × Treatment = BA × B;
ɹence × Treatment = AB × A;
ɹence × Treatment = AB × B;.
ɹence × Treatment = BA × A;.
Font...
OK
Cancel
Colour...

Pie Title: Sequence × Treatment = AB × A; AUC

Sequence × Treatment = AB × A

☑ Point Labels

☐ Percentages

☐ Values

☑ Outline

☐ Scale

☑ Solid Colours

☑ Last Piece Out

5 Depth (degrees)

Colour

10 Mode

75 Pitch

Outline: When this box is checked, the outline of a filled region will be emphasised by a black line.

Scale: When two or more pies are drawn, it is possible to scale pie sizes according to the sum of their values.

Solid Colours: When this box is checked the pie segments will be filled with solid colours. Uncheck this box to fill pies with alternating patterns.

Last Piece Out: The last slice will be exploded.

Depth (degrees): This control is used to set the degree of inclination of 3D pies. When the value is zero pies will be drawn in 2D.

The following controls are available for each pie separately. Select the pie to be edited by clicking on its tab.

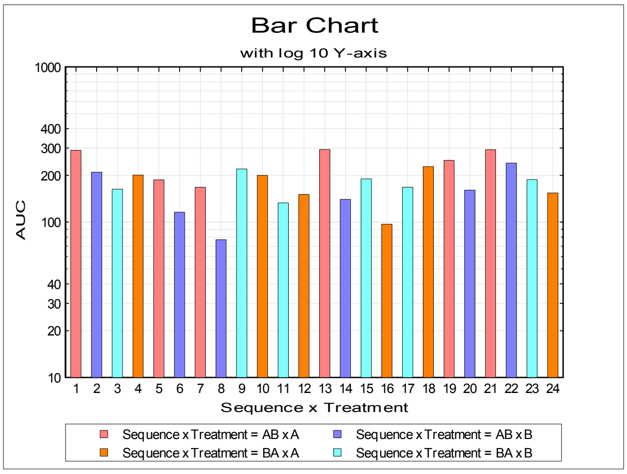
Point Labels: When this box is checked Point Labels will be drawn for pie segments instead of their value. The text for labels are entered into Row Labels. If no Row Labels have been entered then the row numbers will be printed.

Percentages: When this box is checked the percentages will be printed underneath the segment labels.

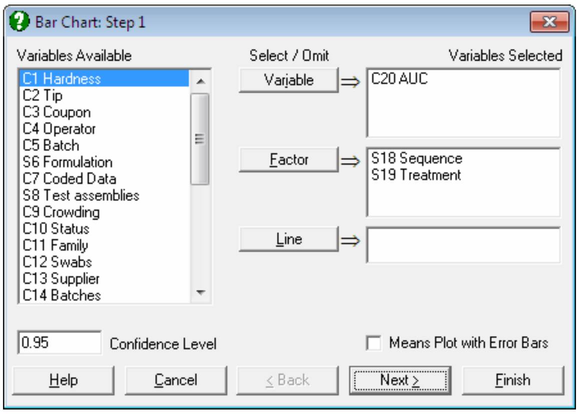
Values: When this box is checked the values of pie segments will be printed. Point Labels and Values cannot be displayed simultaneously.

Colour: The [Colour] button is used to select a base colour for pie segments. The Mode and Pitch values can be edited to obtain a wide variety of colour spectra.

4.3.2. Bar Chart



Bar Chart Variable Selection Dialogue is similar to that of X-Y Plots procedure. Each data series can be a column of the data matrix, or alternatively, a subsample of a data column defined by one or more factor columns. Although an unlimited number of data series can be plotted, properties of only the first nine can be individually controlled from the Edit → Data Series dialogue. The rest of the series will repeat the properties of the first nine in a circular fashion.



Variables Available: C1 Hardness, C2 Tip, C3 Coupon, C4 Operator, C5 Batch, S6 Formulation, C7 Coded Data, S8 Test assemblies, C9 Crowding, C10 Status, C11 Family, C12 Swabs, C13 Supplier, C14 Batches

Variable ⇒ C20 AUC

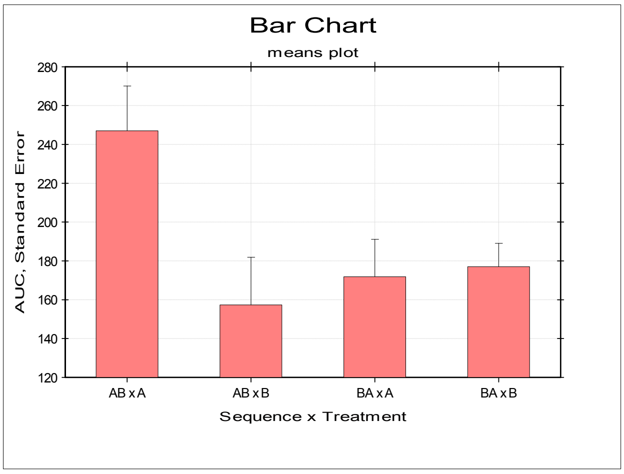
Factor ⇒ S18 Sequence, S19 Treatment

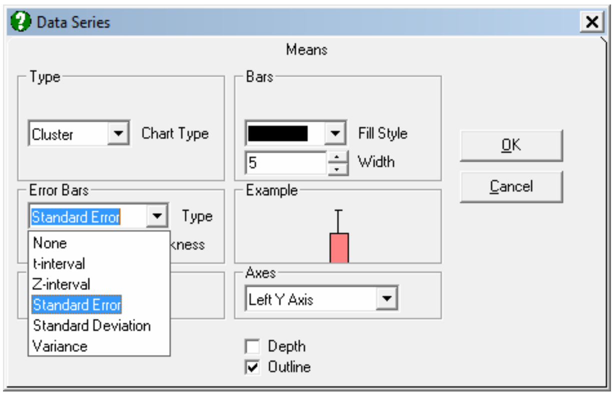
Line ⇒

0.95 Confidence Level

□ Means Plot with Error Bars

Means Plot with Error Bars: If this box on the Variable Selection Dialogue is checked, each bar on the plot will represent the mean of a data series rather than an individual data point. For a detailed description of this option see 4.1.1.3. Means Plot. X-Y Plots, Polar Plot, Area Chart and Ribbon Chart procedures also support the same capability.

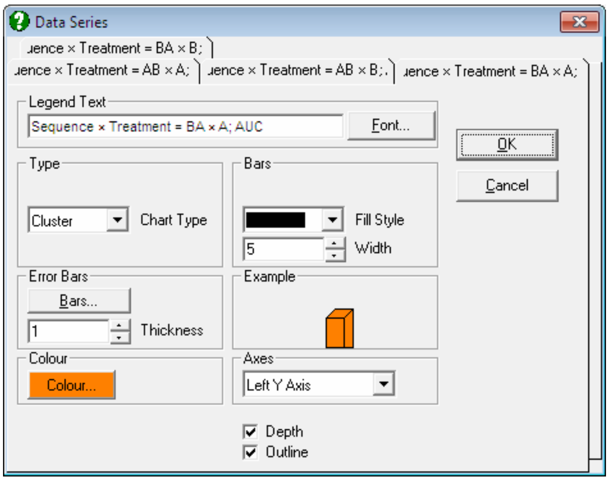




Cluster, overlap, stacked and percentage bar charts can be drawn with or without depth, with symmetric / asymmetric error bars and with two independent Y-axes.

The bars can be left-justified or centred in the plot area by entering and editing the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

```
BarsCentre=1
```



Data Series

ıence × Treatment = BA × B;

ıence × Treatment = AB × A; | ıence × Treatment = AB × B;. | ıence × Treatment = BA × A;

Legend Text: Sequence × Treatment = BA × A; AUC

Type: Cluster — Chart Type

Bars: Fill Style; Width: 5

Error Bars: Thickness: 1

Example

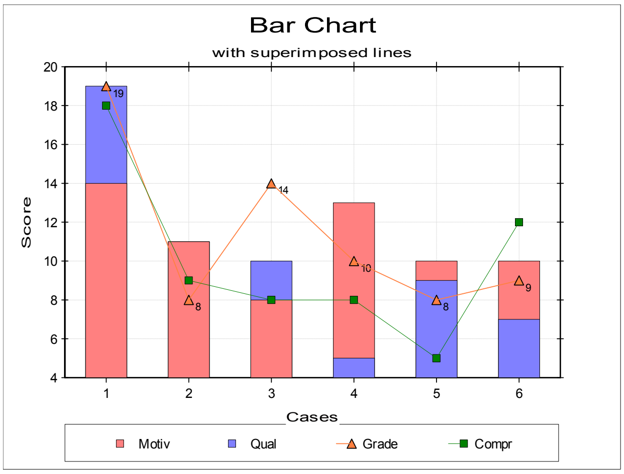
Axes: Left Y Axis

☑ Depth

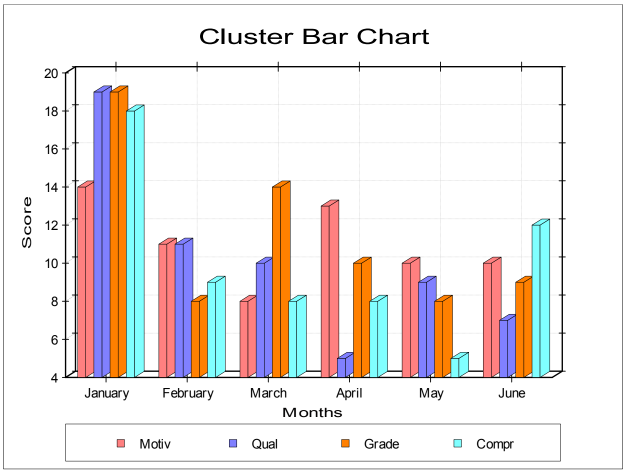
☑ Outline

One difference from X-Y Plots is that there are no interval inputs for the X-axis here since bar charts are always drawn against the index (row numbers). A second difference is that bar charts can only have one right Y-axis. Other specific features are as follows:

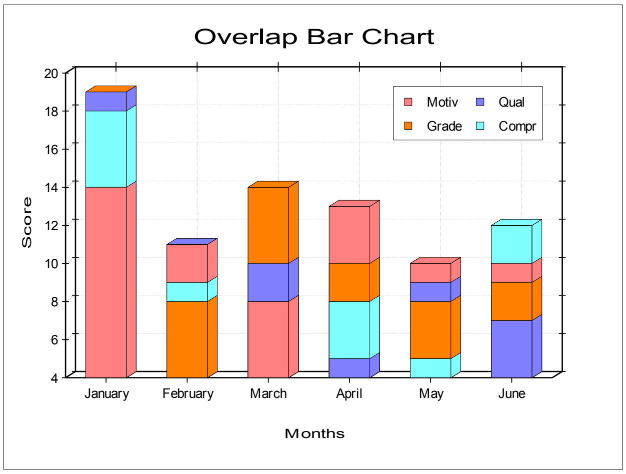
Lines: Up to three columns can be selected from the Variables Available list to draw as lines alongside the bars by clicking on [Line]. When at least one column is selected as line then an Edit → Lines option will be available. All line controls provided for X-Y Plots procedure (e.g. Line options None, Straight, Curve, Trend, Symbol, independent Right Y-axis) will be available.



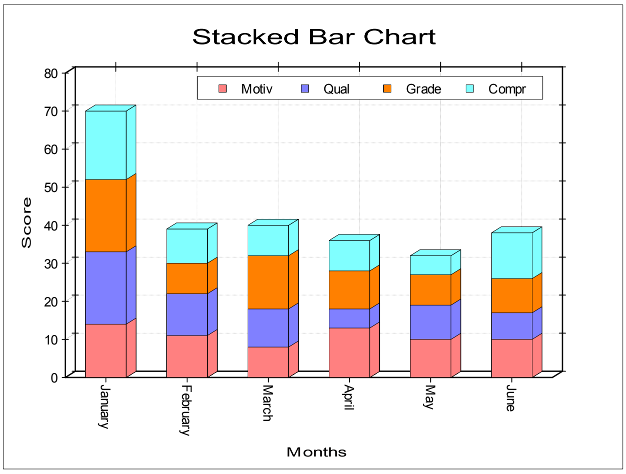
Cluster: Bars having the same row number (but belonging to different series) are drawn next to each other. One such group is called a cluster. There is a small gap between clusters. If there are negative numbers in data then the bars will be drawn upside-down. The missing value positions will be left blank.



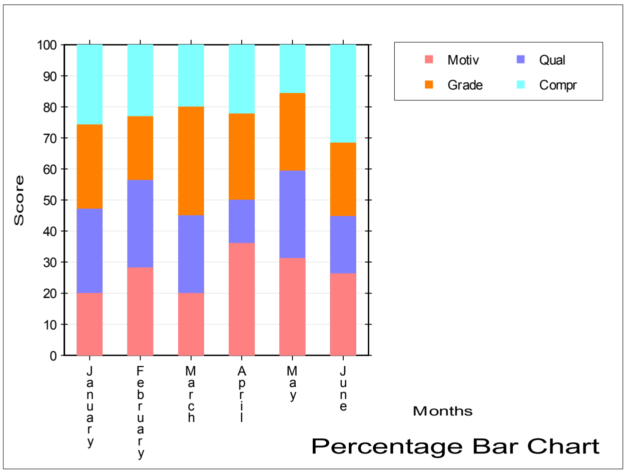
Overlap: Observations having the same row number are sorted and drawn in the same location in descending order. In this way, all observations will have visible parts on the same bar (distinguished either by different colours or patterns). When the column sizes are relatively large the overlap bar charts give better results than the cluster bar charts.



Stacked: This is similar to the overlap bar chart but bars belonging to the same row of the different series will be stacked. Y-axis (or axes) will be rescaled with a maximum value large enough to display all stacked bars.



Percentage: Individual observations are scaled such that the sum of values for a row is 100. Bars reflect the relative magnitudes of individual values within the same row.



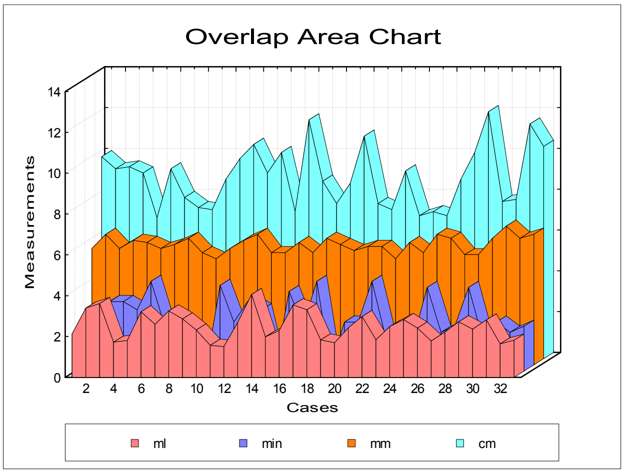
4.3.3. Area Chart

This procedure is similar to Bar Chart procedure, in that it has exactly the same Variable Selection Dialogue and allows drawing charts with two independent Y-axes and with up to three additional columns as lines. The latter have the usual options of connecting with straight lines, curves (cubic spline) or trends. Error bars can also be drawn on any data point. The Means Plot with Error Bars check box allows plotting means of series with their error bars (see 4.1.1.3. Means Plot).

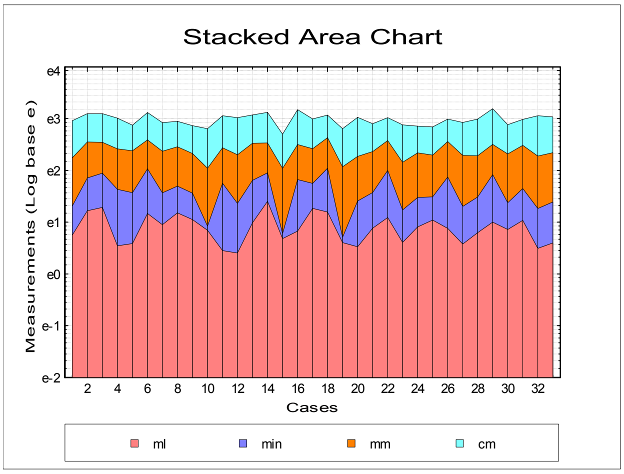
There are two differences from bar charts. The first is that here the Cluster option is not present, since it does not have any meaning in the context of Area Chart.

Overlap: Choose the sequence of the columns to be plotted that will maximise the visibility of all the data series by selecting them from the Variable Selection Dialogue in the desired order. Further control over the appearance of the chart can be achieved by means of allocating columns between Left and Right Y-axes.

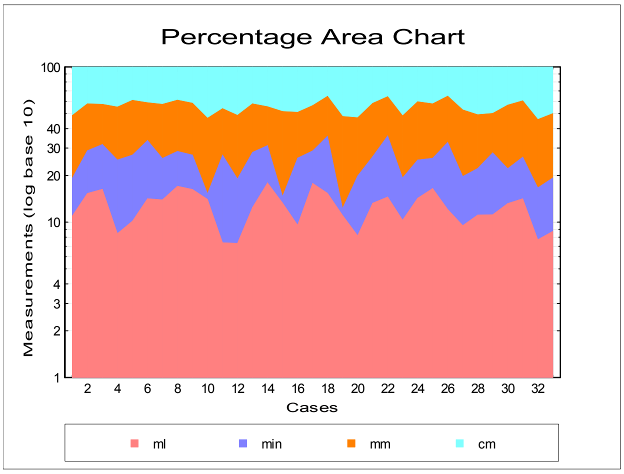
Overlap Area Chart with and without depth will look radically different. When the depth option is on, the overall depth of the chart will depend on the number of columns.



Stacked: Data series will be drawn additively with a new Y-axis scaled to display the maximum row sum.

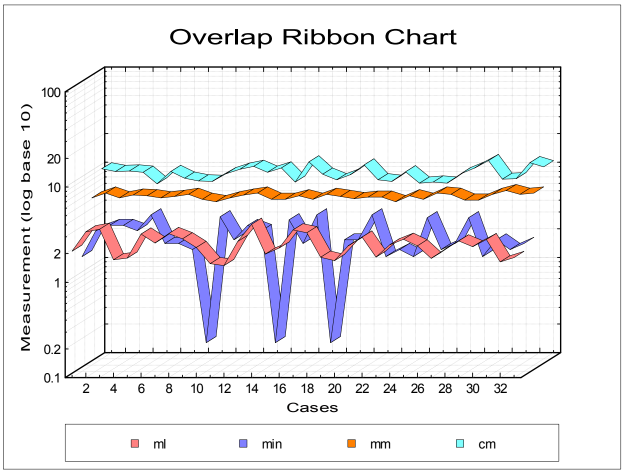


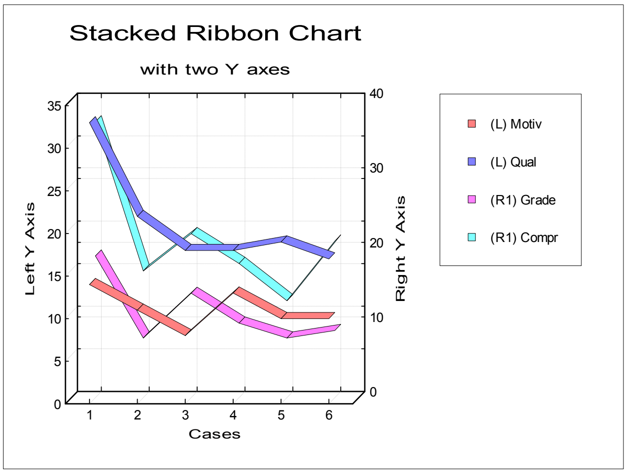
Percentage: Individual observations are scaled such that the sum of values for a row is 100.



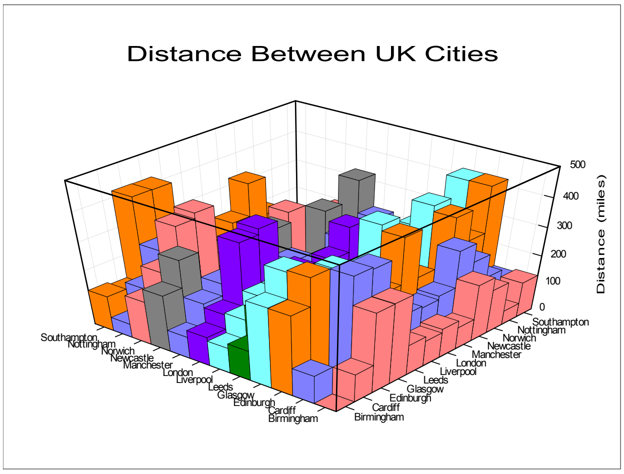
4.3.4. Ribbon Chart

All controls for Ribbon Chart are identical to that for Area Chart. The only difference between the Area Chart and Ribbon Chart is that for the former the area enclosed between the data values and the X-axis will be coloured (or shaded) whereas the Ribbon Chart leaves the area enclosed blank.

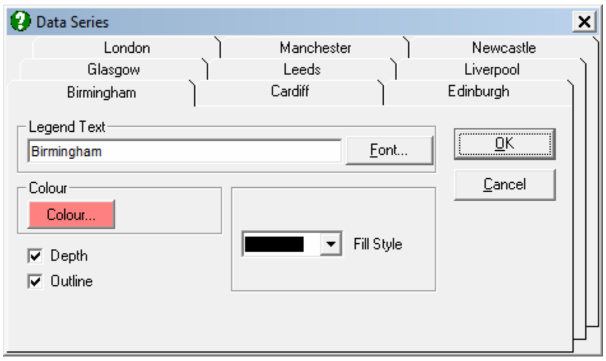




4.3.5. 3D Bar Chart



The Variable Selection Dialogue for this procedure is similar to that for X-Y Plots. Each data series can be a column of the data matrix, or alternatively, a subsample of a data column defined by one or more factor columns. Although an unlimited number of data series can be plotted, properties of only the first nine can be individually controlled from the Edit → Data Series dialogue. The rest of the series will repeat the properties of the first nine in a circular fashion.


Data Series

London | Manchester | Newcastle
Glasgow | Leeds | Liverpool
Birmingham | Cardiff | Edinburgh

Legend Text: Birmingham

Colour: Colour...

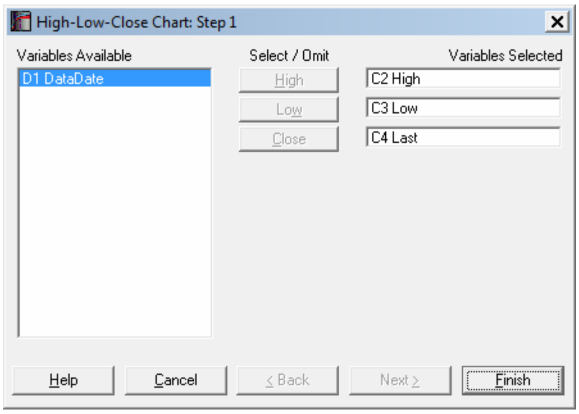
☑ Depth
☑ Outline

Fill Style

The output of this procedure is similar to that of X-Y-Z plot with the Type field set to Bars in Edit → Plot Type dialogue. The main difference is that here more extensive labelling facilities are provided. When the Depth field is unchecked vertical planes will be drawn instead of bars.

4.3.6. High-Low-Close Chart

This chart is used (usually) to plot prices of shares on the stock market. A (compulsory) column is selected as the high value of the share by clicking on [High]. Another (compulsory) column is selected as the low value of the share by clicking on [Low]. A final (optional) column may be selected as the closing value of the share by clicking on [Close].

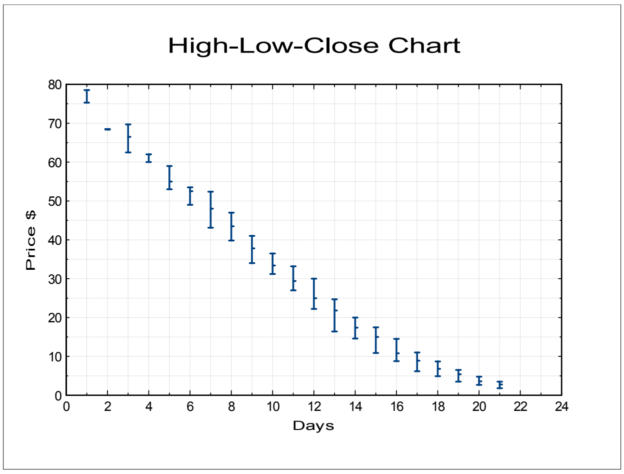


High-Low-Close Chart: Step 1

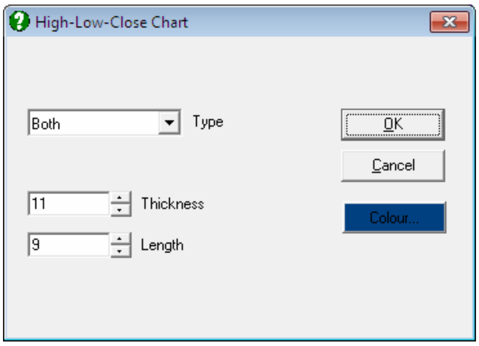
Variables Available: D1 DataDate

Variables Selected: C2 High, C3 Low, C4 Last

The selected variables may contain missing values but they should all have the same length. Each row is expected to have values such that *Low* ≤ *Close* ≤ *High*. If this is not the case, a warning message is issued and then the chart is displayed.



A bar is drawn for each row in the data. The top and bottom of the bar show the high and low values. The horizontal bar in the middle denotes the closing value. The appearance of the bar can be controlled using the Edit → Bars dialogue. The following controls are provided:



Type: It is possible to draw ticks for high and low values only, for close values only, for both (the default), or none.

Colour: You can select the colour of bars.

Thickness: Line thickness of the bars can be selected in printer units, which is about 1/8 of a pixel on the screen for a laser printer.

Length: The length of optional ticks can be controlled.

4.4. Plot of Functions

These are the graphics procedures which do not require data.

The first two procedures, i.e. plot of functions with one and two variables, share the same routine to evaluate the user-defined functions. This is a stripped-down version of the function evaluation algorithm used in the Data Processor and supports the following operations and functions:

^, **/**, *****, **-**, **+**, **Mod**, **Exp()**, **Log()**, **Lne()**, **Sqr()**, **Fct()**, **Int()**, **Abs()**, **Sgn()**, **Pi()**, **e()** and all trigonometric functions.

Function Syntax: Functions of one variable are expressed in terms of X and functions of two variables in terms of X and Y. Functions must conform to the computational syntax rules. For instance:

3*X + 5, Sin(X*Pi()), X ^ 3 + 2

are valid, but

3X + 5, SinX, X ^ A + 2

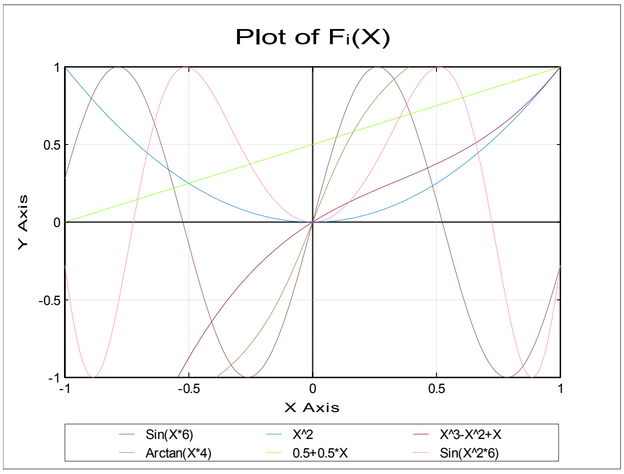
are invalid.

The use of spaces within functions is allowed, but they are not necessary. Before proceeding with the plot, the program will evaluate the supplied function once to check for its syntax. If a syntax error is found an error message will be issued and the offending part of the formula indicated.

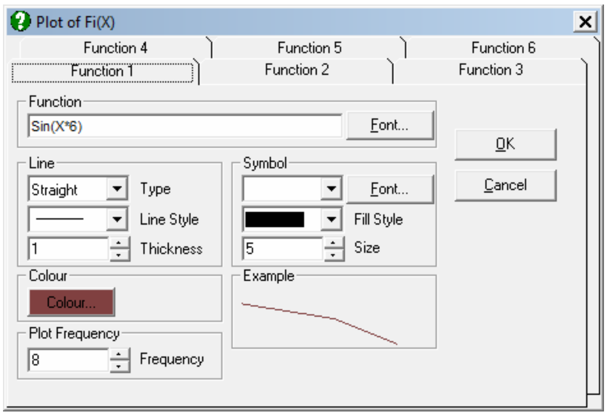
Argument Errors: Any X or Y intervals can be entered for functions which are not defined at certain intervals (like **Sqr()**, **Lne()** or **Log()**). The program will simply not plot the function at an illegal interval. However, it is important to ensure that the functions entered do not give rise to number overflow within the specified interval. The most likely cases are functions like **Tan()**, **Lne()**, **Log()**, or whenever an X or Y variable is in the denominator. Even though in most cases the program will draw a graph (particularly when the Plot Frequency is low) a crash resulting in loss of data cannot be ruled out.

WARNING! *Save any useful data before plotting a function.*

4.4.1. Plot of 2D Functions



Up to six functions can be plotted simultaneously. Using the Edit → Functions dialogue, you can either type the formula directly into the Function box or double-click on it to open the Formula Editor. Functions can be plotted in any interval. However, neither X nor Y-axis values will be automatically scaled. The values suggested by the program on entry have nothing to do with the functions supplied by the user. The X and Y interval values may be edited to regulate the frequency of ticks and numbers displayed on X and Y axes.



Function: Sin(X*6)

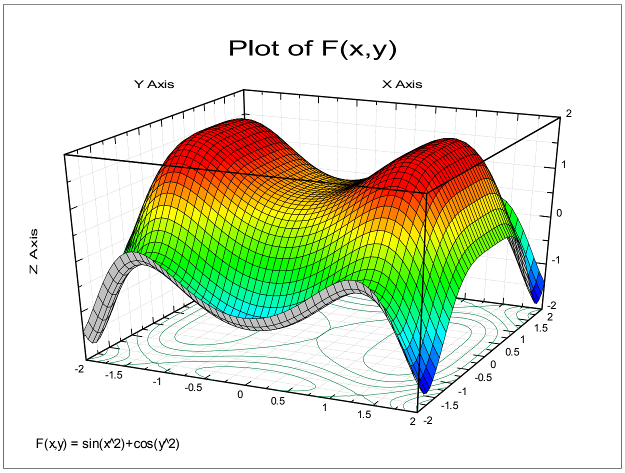
Line
Type: Straight
Line Style: ———
Thickness: 1

Symbol
Fill Style: ■
Size: 5

Plot Frequency
Frequency: 8

Plot Frequency values range from 1 to 10. The measure of frequency is in terms of the number of pixels at which the function will be evaluated. For instance, if the frequency is 10, then functions are evaluated at every 10th pixel. If the field displays 1 then functions will be evaluated at each and every pixel. Therefore, it is advisable to plot functions first at the lowest resolution (i.e. at a Plot Frequency of 10) in order to get some idea of their behaviour in the specified interval. Although a Plot Frequency of 1 will provide the highest possible resolution, the plotting speed may be low when complex functions are evaluated.

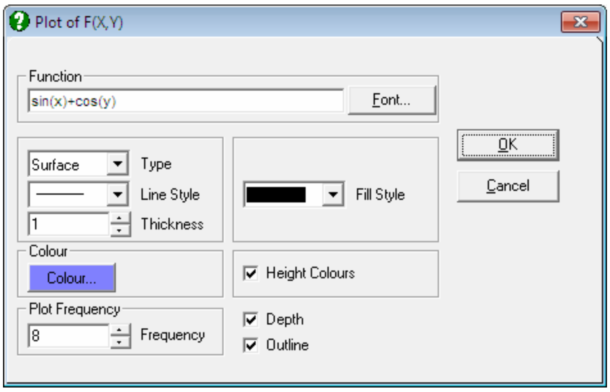
4.4.2. Plot of 3D Functions



This procedure is similar to X-Y-Z Grid Plot except that instead of reading the grid values from the data matrix, it will produce its own grid from a function and intervals supplied by the user.

All controls except F(X,Y) and Plot Frequency will work in exactly the same way as they do in 3D Plots procedure (see 4.2.2. X-Y-Z Grid Plot). The F(X,Y) field is used for entering a formula in terms of the variables X and Y. No other variable names can be used. Either type the formula directly into the text box or double-click on it to open the Formula Editor dialogue. For the syntax of functions see the beginning of section 4.4. Plot of Functions.

The maximum grid size is 51 × 51. This allows for plotting 50 × 50 surface plots and 51 × 51 bar or point plots.



Plot of F(X,Y)

Function: sin(x)+cos(y)

Type: Surface

Line Style: ———

Thickness: 1

Fill Style: ■

☑ Height Colours

Plot Frequency: 8 Frequency

☑ Depth

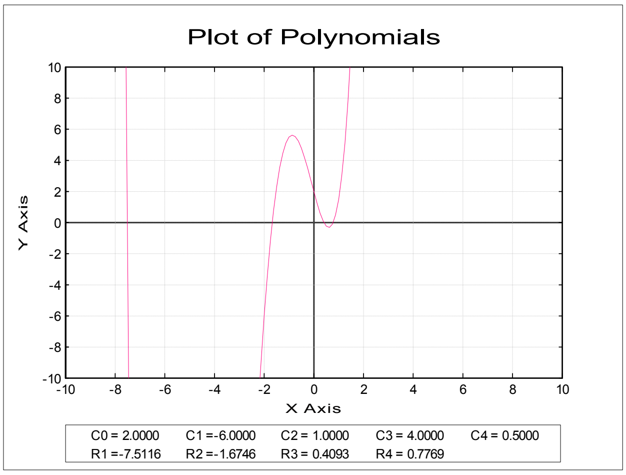
☑ Outline

The Plot Frequency control will regulate the resolution of plots. The highest resolution is obtained when this field is set to 1. In this case the function will be evaluated 50 times for each axis giving a total of 2,500 function evaluations. The lowest resolution is 10, which means that the maximum grid size 50 will be divided by 10, giving 5 polygons (or 5 bars or 5 points) for each axis.

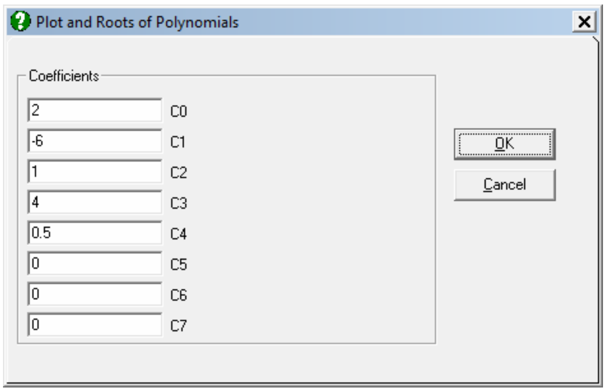
Since computing the grid elements from the supplied function may be a time consuming process, it will not be repeated whenever a graph property is changed. For instance the grid will not be recomputed if changes are made in the main title, sub title, X, Y and Z axis labels, minimum, maximum and interval values of the Z axis, type, rotation, grid and colours fields. If changes are made in one of minimum, maximum X or Y-axis values, the definition of function or the Plot Frequency, then the whole grid will be recomputed.

4.4.3. Plot and Roots of Polynomials

The Edit → Coefficients dialogue will allow for entering coefficients for up to degree 7 polynomials. There is no need to enter the degree of the polynomial separately as the program will determine this from the field with the highest number which has a non zero value.



If a polynomial has been fitted in procedure X-Y Plots (see 4.1.1.2. Curve Fitting) just before selecting this procedure, then the estimated coefficients for this polynomial will be placed in the coefficient fields. You can then plot the same polynomial in different intervals and examine its roots.



Coefficients

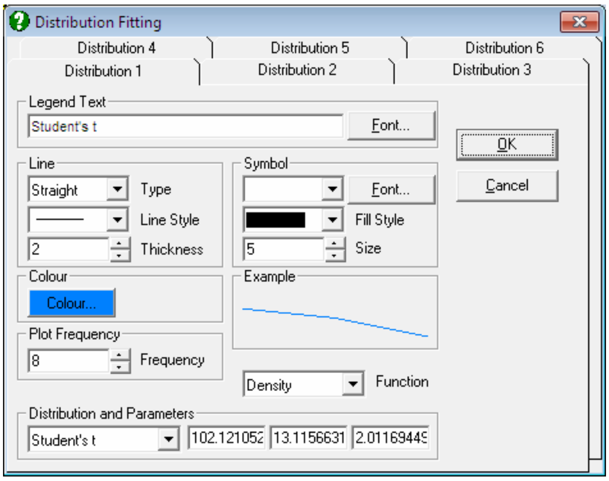
2	C0
-6	C1
1	C2
4	C3
0.5	C4
0	C5
0	C6
0	C7

As the program plots a polynomial within the given interval, it will also search for a change of sign in the value of the polynomial. If a change occurs then a half interval search algorithm is called to determine the value of the root. Search is stopped and the root is reported when the accuracy is better than 1E-8.

4.4.4. Plot of Distribution Functions

Like the other three Plot of Functions procedures, this procedure does not require selection of a data column. It is possible to display up to six Distribution Functions simultaneously. Density, cumulative density, survivor function, log-survivor function and hazard function curves can be drawn for 19 continuous and discrete distribution functions.

The Edit → Distributions dialogue is similar to that for the Histogram procedure (see 5.3.3.3. Fitting Distribution Functions). Here, an additional drop-down list, Function, allows you to plot one of density, cumulative, survival, log survival and hazard functions.



Legend Text: Student's t

Line: Straight Type; Line Style; 2 Thickness

Symbol: Fill Style; 5 Size

Colour

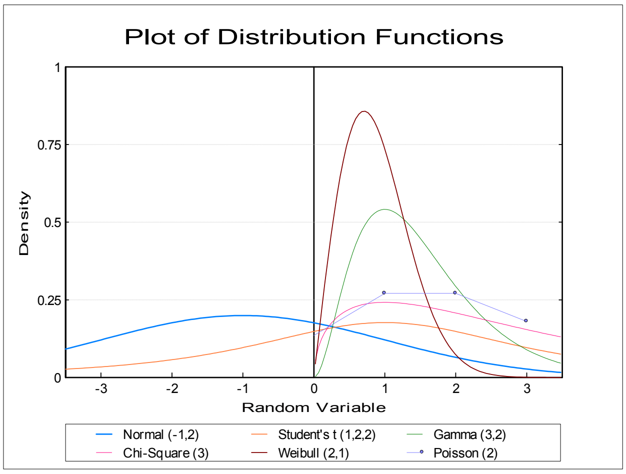
Plot Frequency: 8 Frequency

Density Function

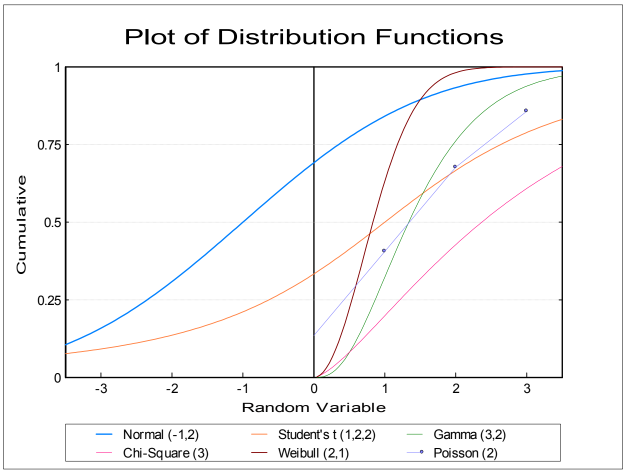
Distribution and Parameters: Student's t | 102.121052 | 13.1156631 | 2.01169449

The Distribution Functions are selected in the same way as in the Histogram procedure. The only difference is that the values displayed in the distribution parameter fields are the ones either last entered by the user or the ones estimated by the program for a data column, that is, if either Expected Frequencies or Histogram procedures have been used previously. Therefore, if this procedure is entered before the latter two, all parameter fields will be -99. In this case, in order to plot a distribution function, you must enter the distribution parameters manually (see Appendix).

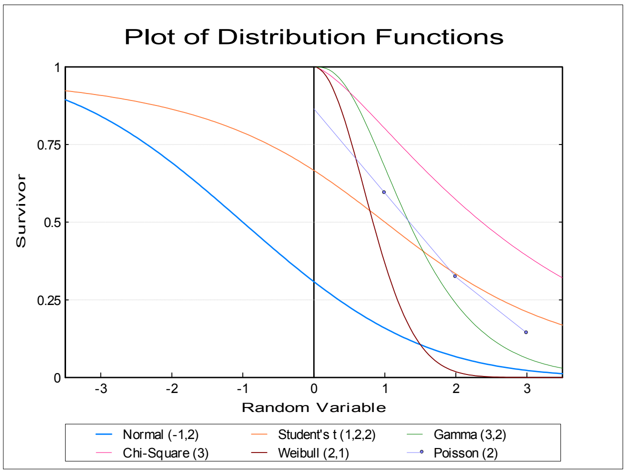
Density Function: The density functions of the selected distributions (see Appendix) will be plotted with the given parameters.



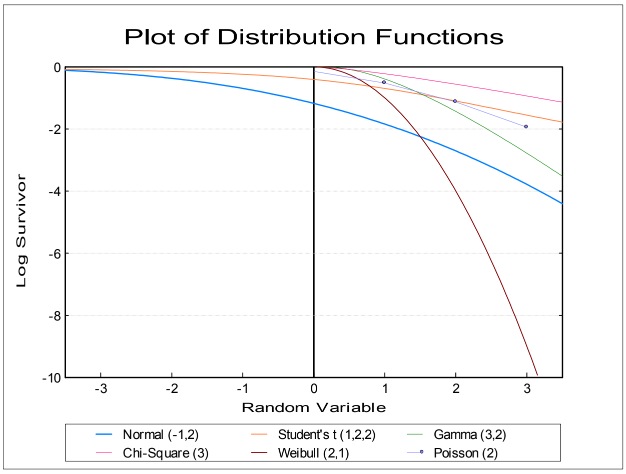
Cumulative Density Function: Area enclosed under the density function from negative infinity to the x values are plotted.



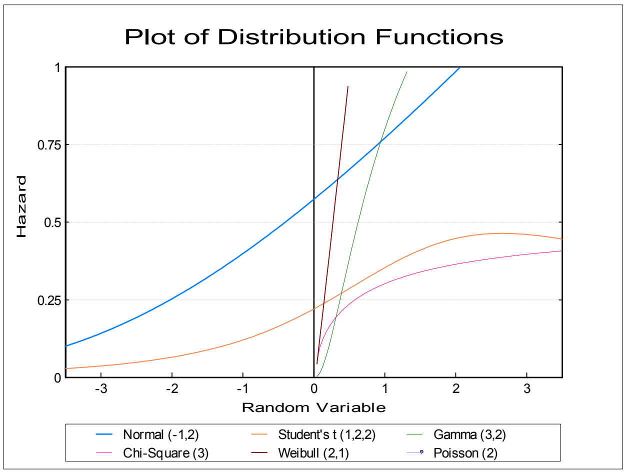
Survivor Function: The complement of the cumulative density function (i.e. 1 - cumulative df) is plotted.



Log Survivor Function: The logarithm (to base e) of survivor functions are drawn.



Hazard Function: Density functions divided by their survivor functions are drawn (available only for continuous distributions).



UNISTAT Statistical Package

Chapter 5
Descriptive Statistics and Distributions

5.0. Overview

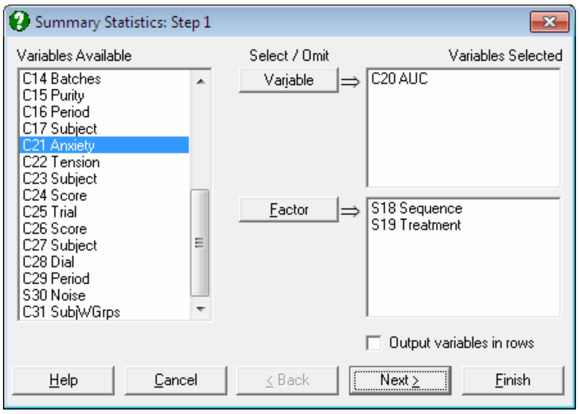
Procedures described in this chapter work with one sample, but some of them accept more than one sample at a time and report the results on the same table.

5.0.1. Multisample Data Types

The procedures that can work with multiple one-sample data are as follows:

Statistics 1 → Descriptive Statistics →
Summary Statistics
Confidence Intervals
Quantiles (Percentiles)
Statistics 1 → Descriptive Plots →
Box-Whisker, Dot and Bar Plots
Normal Probability Plot
Histogram

With these procedures, it is possible to analyse samples in separate columns or subsamples of data defined by one or more factor columns. An unlimited number of variables and factors can be selected. Selection of a factor variable is optional. Factors can be numeric or String Data columns, but should contain a limited number of distinct values.



Summary Statistics: Step 1

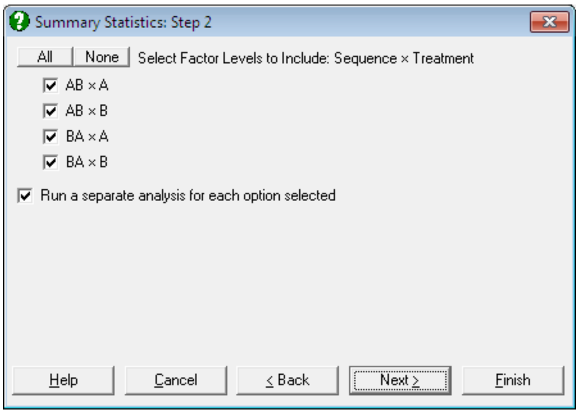
Variables Available
C14 Batches
C15 Purity
C16 Period
C17 Subject
C21 Anxiety
C22 Tension
C23 Subject
C24 Score
C25 Trial
C26 Score
C27 Subject
C28 Dial
C29 Period
S30 Noise
C31 SubjWGrps

Variables Selected
C20 AUC

S18 Sequence
S19 Treatment

□ Output variables in rows

If at least one factor column is selected, then a further dialogue will pop up, displaying a check list of all combinations of levels in selected factors. There will also be a check box Run a separate analysis for each option selected, which is used to determine whether the variables or the factors will have the priority in the output. Note that Confidence Intervals and Box-Whisker, Dot and Bar Plots procedures do not have this check box, since the ordering of variables and factors is not significant for these procedures.


Summary Statistics: Step 2

Select Factor Levels to Include: Sequence × Treatment

☑ AB × A
☑ AB × B
☑ BA × A
☑ BA × B

☑ Run a separate analysis for each option selected

a) **Samples are in separate columns:** If no factors and at least one data column is selected, then the program will treat each column as one sample. Samples are not required to have equal lengths.

b) **Samples are in separate columns and subsets are defined by one or more categorical variables:** If at least one factor and one variable are selected from the Variable Selection Dialogue, and the option Run a separate analysis for each option selected is checked on the second dialogue, then the program will perform the procedure on all variables, for each level (or combinations of levels) of the factor column(s). The selected columns should all have the same length.

For instance, if two data variables and one factor containing three levels are selected, three check boxes will be displayed representing each level. The test will be performed three times for the two variables (the outer loop), each time for only those rows containing the selected level of the factor column (the inner loop).

c) **Samples are stacked in columns and categorical variables define the samples:** If at least one factor and one variable are selected, and the option Run a separate analysis for each option selected is unchecked, then the program will perform the procedure on all selected levels (or combinations of levels) of the factor columns, for each data column separately. The selected columns should have the same length.

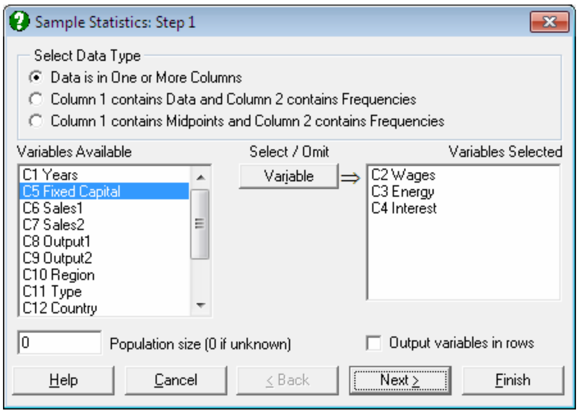
For instance, if two variables and one factor containing three levels is selected and all are checked, then the test will be performed once for each variable (the outer loop), for all of factor levels (the inner loop).

For an example demonstrating these data options see 5.1.1. Summary Statistics.

5.0.2. One-Sample Data Types

Three types of data can be analysed:

1) Any number of columns containing raw data can be selected.
2) A pair of columns, one containing data and the other frequency counts (frequency data).
3) A pair of columns, one containing class mid-points and the other frequencies (grouped data).



Sample Statistics: Step 1

Select Data Type

● Data is in One or More Columns
○ Column 1 contains Data and Column 2 contains Frequencies
○ Column 1 contains Midpoints and Column 2 contains Frequencies

Variables Available: C1 Years, C5 Fixed Capital, C6 Sales1, C7 Sales2, C8 Output1, C9 Output2, C10 Region, C11 Type, C12 Country

Variables Selected: C2 Wages, C3 Energy, C4 Interest

Population size (0 if unknown): 0

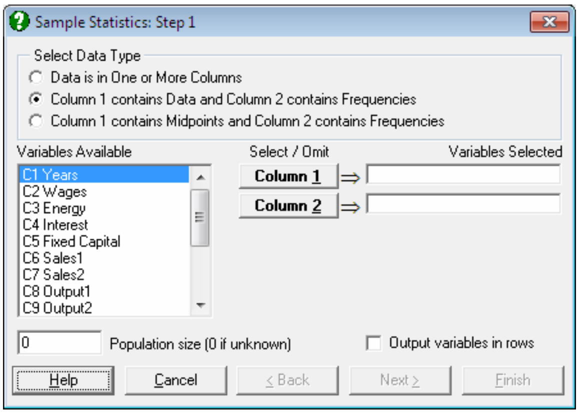
□ Output variables in rows

The following procedures have a Variable Selection Dialogue with three data types:

Statistics 1 → Descriptive Statistics →
Sample Statistics
Frequency Distributions
Statistics 1 → Distribution Functions →
Expected Frequencies

Ungrouped Raw Data: Any number of columns containing raw data can be selected by clicking on [Variable]. In this case, the program will perform the procedure on each selected column separately. The column lengths need not be equal. The graphical procedures with this type of data (see 5.3.3. Histogram) will plot only one column at a time.

Data with Frequency Counts: This option assumes that the first column contains data values and the second column contains frequencies corresponding to the values in the first column. First click on this data type option and then select two columns clicking on [Column 1] for the column of values and [Column 2] for the column of frequencies. The program keeps track of the value-frequency correspondence in all relevant procedures.



Sample Statistics: Step 1

Select Data Type

○ Data is in One or More Columns
● Column 1 contains Data and Column 2 contains Frequencies
○ Column 1 contains Midpoints and Column 2 contains Frequencies

Variables Available: C1 Years, C2 Wages, C3 Energy, C4 Interest, C5 Fixed Capital, C6 Sales1, C7 Sales2, C8 Output1, C9 Output2

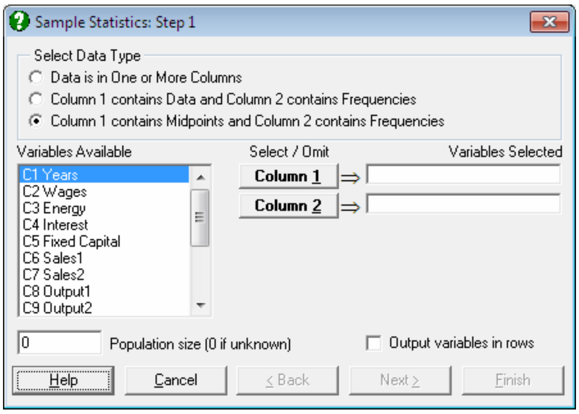
Select / Omit: Column 1 ⇒, Column 2 ⇒

Variables Selected:

0 Population size (0 if unknown)

□ Output variables in rows

Grouped Data: It is assumed that the first column contains class midpoint values and the second column contains frequencies corresponding to these midpoints. As in the case of data with frequency counts, two columns must be selected by clicking on [Column 1] and [Column 2].



Sample Statistics: Step 1

Select Data Type

○ Data is in One or More Columns

○ Column 1 contains Data and Column 2 contains Frequencies

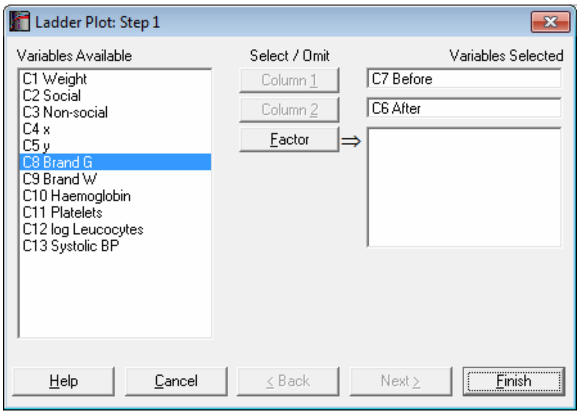
● Column 1 contains Midpoints and Column 2 contains Frequencies

0 Population size (0 if unknown)

□ Output variables in rows

5.0.3. Paired Data Types

A pair of data columns must be selected as [Column 1] and [Column 2]. Optionally, one or more factors can be selected, in which case the program will display a list of all possible combinations of factor levels. The unchecked levels will be excluded from the analysis.



Ladder Plot: Step 1

Column 1: C7 Before

Column 2: C6 After

Graph → Descriptive Plots →
5.3.4. 3D Histogram
5.3.5. Bland-Altman Plot
5.3.6. Ladder Plot

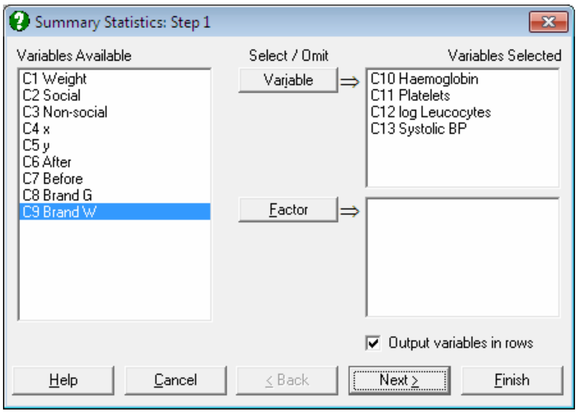
5.1. Descriptive Statistics

Summary Statistics, Confidence Intervals and Quantiles (Percentiles) are multisample procedures that work with continuous and categorical data. The Quantiles (Percentiles) procedure enables you to calculate any quantile values (or percentiles) with their confidence intervals for up to nine quantiles at a time. Six alternative quantile methods and four confidence interval methods are provided.

Sample Statistics procedure provides a wide range of statistics for frequency and grouped data, as well as continuous data. The Frequency Distributions procedure works on the same types of data and allows for selecting lower and upper bounds and class intervals. Its output can be displayed in the form of a Table or a Character Histogram. Four types of character plots, namely, histograms, stem and leaf plots, sequence and scatter diagrams, can be drawn.

5.1.1. Summary Statistics

This procedure is also known as the *Code Book* procedure. Multisample data can be entered either in the form of multiple columns or data columns classified by factor columns (see 5.0.1. Multisample Data Types). If at least one factor column is selected, then a further dialogue will pop up asking for the combination of factor levels to be included.

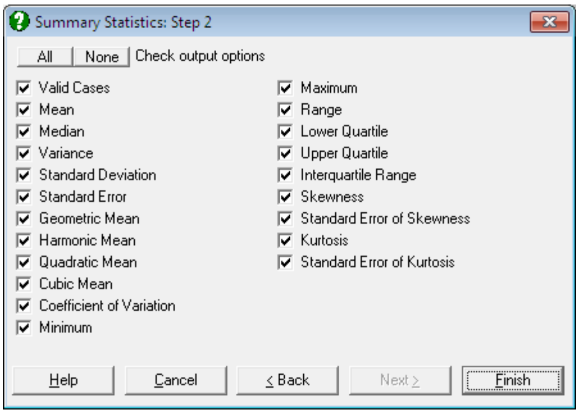

Summary Statistics: Step 1

Variables Available: C1 Weight, C2 Social, C3 Non-social, C4 x, C5 y, C6 After, C7 Before, C8 Brand G, C9 Brand W

Variables Selected: C10 Haemoglobin, C11 Platelets, C12 log Leucocytes, C13 Systolic BP

☑ Output variables in rows

The Output variables in rows check box allows you to transpose the output matrix. This will be useful when you wish to use the output from this procedure (such as means and standard errors) for further analysis in other procedures.



☑ Valid Cases
☑ Mean
☑ Median
☑ Variance
☑ Standard Deviation
☑ Standard Error
☑ Geometric Mean
☑ Harmonic Mean
☑ Quadratic Mean
☑ Cubic Mean
☑ Coefficient of Variation
☑ Minimum
☑ Maximum
☑ Range
☑ Lower Quartile
☑ Upper Quartile
☑ Interquartile Range
☑ Skewness
☑ Standard Error of Skewness
☑ Kurtosis
☑ Standard Error of Kurtosis

Most of the statistics reported here are similar to those in the Sample Statistics procedure. The current procedure should be preferred if the data is broken down by one or more factor columns, whereas Sample Statistics should be preferred if the data has weights, frequency counts or if it is grouped.

An Output Options Dialogue allows for including some or all available statistics.

Valid Cases: This is the number of cases in the sample, excluding any missing values.

Mean:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Median: The method used in computing the median is indicated in the header. This can be one of the six methods described in section 5.1.3.1. Quantile Methods.

Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Standard Deviation:

$$s = \sqrt{s^2}$$

Standard Error:

$$SE = \frac{s}{\sqrt{n}}$$

Geometric Mean:

$$G = \sqrt[n]{\prod_{i=1}^{n} X_i} \quad Xi > 0, i = 1, \ldots, n$$

Harmonic Mean:

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}} \quad Xi > 0, i = 1, \ldots, n$$

The following relationship should hold if $X_i > 0$, $i = 1, \ldots, n$:

$$H \le G \le \overline{X}$$

Quadratic Mean (Root mean square):

$$M_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

Cubic Mean:

$$M_3 = \sqrt[3]{\frac{1}{n}\sum_{i=1}^{n} X_i^3}$$

Coefficient of Variation:

$$CoefVar = \frac{s}{\overline{X}}$$

Minimum: Smallest observed value in data.

Maximum: Largest observed value in data.

Range: Difference between maximum and minimum values.

Lower Quartile: As in median above, but for the 25% quantile.

Upper Quartile: As in median above, but for the 75% quantile.

Interquartile Range: Difference between upper and lower quartiles.

Skewness: This should not be confused with the moment coefficient of skewness which is covered under section 5.1.4. Sample Statistics.

$$\text{Skewness} = \frac{nm_3}{(n-1)(n-2)s^2}$$

where m_j is the j^th^ moment about the mean (see Bliss 1967, p. 144):

$$m_j = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^j$$

Standard Error of Skewness:

$$\text{SE Skewness} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

Kurtosis: This should not be confused with the moment coefficient of kurtosis which is covered under section 5.1.4. Sample Statistics.

$$\text{Kurtosis} = \frac{n(n+1)m_4 - 3m_2^2(n-1)}{(n-1)(n-2)(n-3)s^2}$$

Standard Error of Kurtosis:

$$\text{SE Kurtosis} = \sqrt{\frac{4(n^2-1)\text{SE}^2_{\text{Skewness}}}{(n-3)(n+5)}}$$

Example 1: Variables in columns of output table

Open PARTEST and select Statistics 1 → Descriptive Statistics → Summary Statistics. Select *Haemoglobin*, *Platelets*, *log Leucocytes*, and *Systolic BP* (*C10* to *C13*) as [Variable]s, leave the Output variables in rows box unchecked and click [Next]. From the Output Options Dialogue select all output options and click [Finish].

Summary Statistics

Quantile Method: Simple Average

	Haemoglobin	Platelets	log Leucocytes	Systolic BP
Valid Cases	10.0000	10.0000	10.0000	10.0000
Mean	-0.5300	-0.0300	-0.5900	3.1000
Median	-0.6000	0.1000	-0.6500	2.0000
Variance	2.1401	1.4868	2.4099	37.8778
Standard Deviation	1.4629	1.2193	1.5524	6.1545
Standard Error	0.4626	0.3856	0.4909	1.9462
Geometric Mean	*	*	*	*
Harmonic Mean	*	*	*	*
Quadratic Mean	1.4856	1.1572	1.5865	6.6106
Cubic Mean	-1.2371	-0.8205	-1.6133	7.4680
Coefficient of Variation	-2.7602	-40.6445	-2.6312	1.9853
Minimum	-2.4000	-2.2000	-3.2000	-6.0000
Maximum	2.3000	1.9000	1.7000	14.0000
Range	4.7000	4.1000	4.9000	20.0000
Lower Quartile	-1.5000	-1.0000	-1.6000	-2.0000
Upper Quartile	0.0000	0.6000	0.9000	8.0000
Interquartile Range	1.5000	1.6000	2.5000	10.0000
Skewness	0.5846	-0.3308	-0.0573	0.4151
Standard Error of Skewness	0.6870	0.6870	0.6870	0.6870
Kurtosis	0.0911	-0.0998	-0.7446	-0.4940
Standard Error of Kurtosis	1.3342	1.3342	1.3342	1.3342

Example 2: Variables in rows of output table

Open PARTEST and select Statistics 1 → Descriptive Statistics → Summary Statistics. Select *Haemoglobin*, *Platelets*, *log Leucocytes*, and *Systolic BP* (*C10* to *C13*) as [Variable]s, this time check the Output variables in rows box and click [Next]. From the Output Options Dialogue select the following options only and click [Finish].

Summary Statistics

	Valid Cases	Mean	Standard Deviation	Standard Error
Haemoglobin	10	-0.5300	1.4629	0.4626
Platelets	10	-0.0300	1.2193	0.3856
log Leucocytes	10	-0.5900	1.5524	0.4909
Systolic BP	10	3.1000	6.1545	1.9462

Example 3: Variables in columns of output table, one table for each factor level

Open ANOVA and select Statistics 1 → Descriptive Statistics → Summary Statistics. Select *AUC* (*C20*) as [Variable] and *Treatment* (*S19*) as [Factor], uncheck the Output variables in rows box and click [Next]. On the next dialogue, leave the Run a separate analysis for each option selected box checked and from the Output Options Dialogue select the following options only and click [Finish].

Summary Statistics

Subsample selected by: Treatment = A

	AUC
Valid Cases	12.0000
Mean	209.4167
Geometric Mean	199.8379
Harmonic Mean	189.4269
Quadratic Mean	218.0193
Cubic Mean	225.6315

Subsample selected by: Treatment = B

	AUC
Valid Cases	12.0000
Mean	167.1667
Geometric Mean	160.4173
Harmonic Mean	152.6247
Quadratic Mean	172.9480
Cubic Mean	177.9293

Example 4: Factor levels in columns of output table, one table for each variable

Continuing from Example 3 above, click on the [Last Procedure Dialogue] button and go back to the Variable Selection Dialogue. Select *AUC* and *Subject* (C17 - *C20*) as [Variable]s and *Sequence* and *Treatment* and (S18 - *S19*) as [Factor]s, on the next dialogue uncheck the Run a separate analysis for each option selected box and click [Finish].

Summary Statistics

Data variable: AUC
Subsample selected by: Sequence × Treatment

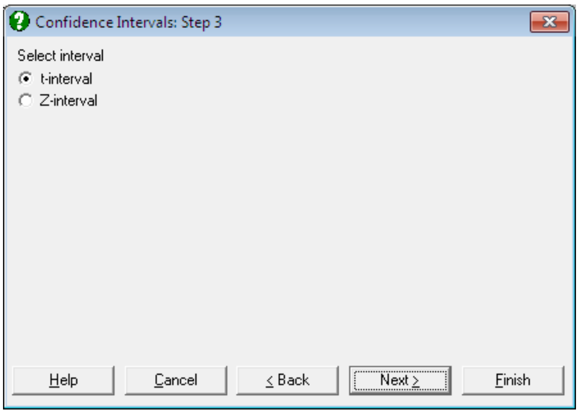
	AB × A	**AB × B**	**BA × A**	**BA × B**
Valid Cases	6.0000	6.0000	6.0000	6.0000
Mean	247.0000	157.3333	171.8333	177.0000
Geometric Mean	241.0556	147.1371	165.6679	174.8962
Harmonic Mean	234.7181	136.7044	158.7873	172.7419
Quadratic Mean	252.3483	166.6363	177.1586	179.0372
Cubic Mean	257.0111	174.7520	181.6798	180.9970

Data variable: Subject
Subsample selected by: Sequence × Treatment

	AB × A	**AB × B**	**BA × A**	**BA × B**
Valid Cases	6.0000	6.0000	6.0000	6.0000
Mean	6.0000	6.0000	7.0000	7.0000
Geometric Mean	4.5808	4.5808	6.1063	6.1063
Harmonic Mean	3.1297	3.1297	5.0585	5.0585
Quadratic Mean	7.0238	7.0238	7.6811	7.6811
Cubic Mean	7.7250	7.7250	8.2081	8.2081

5.1.2. Confidence Intervals

Confidence intervals for mean, median, geometric and harmonic means (t- or Z-intervals) and intervals for variance and standard deviation can be computed. Data input is in multisample format (see 5.0.1. Multisample Data Types).



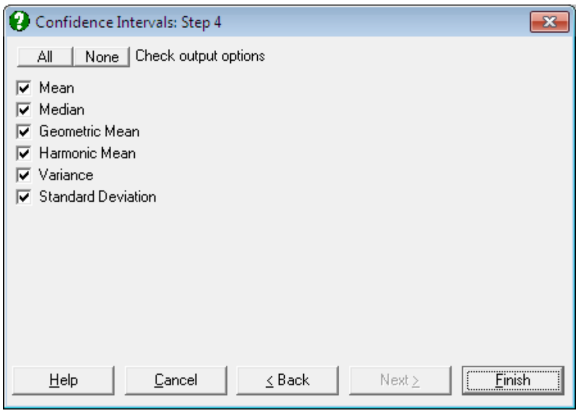
Confidence Intervals: Step 3

Select interval

● t-interval

○ Z-interval

By default, intervals for means are based on the t-distribution with a critical value of $t_{1\text{-}\alpha/2,n-1}$. It is possible to calculate intervals using the standard normal distribution with a critical value of $Z_{1-\alpha/2}$. The confidence level $1 - \alpha$ can be defined in Variable Selection Dialogue.



Confidence Intervals: Step 4

Check output options

☑ Mean

☑ Median

☑ Geometric Mean

☑ Harmonic Mean

☑ Variance

☑ Standard Deviation

Mean:

$$\text{Lower limit} = \text{Mean} - t_{1-\alpha/2,n-1}\,\text{SE}$$

$$\text{Upper limit} = \text{Mean} + t_{1-\alpha/2,n-1}\,\text{SE}$$

Median: The methods used in computing the median and its confidence limits are reported in the header. These methods can be changed using the dialogues of the Quantiles (Percentiles) procedure (see sections 5.1.3.1. Quantile Methods and 5.1.3.2. Quantile Interval Methods).

Geometric Mean: Assuming Ln(X_i) i = 1,…, n are normally distributed, the limits are defined as:

$$\text{Lower limit} = G / \text{Exp}\left(t_{1-\alpha/2,n-1}\, a_G\right)$$

$$\text{Upper limit} = G\,\text{Exp}\left(t_{1-\alpha/2,n-1}\, a_G\right)$$

where G is the geometric mean and the term a_G (which is *not* the standard error of geometric mean) is defined as:

$$a_G = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\text{Log}(G) - \text{Log}(X_i)\right)^2}$$

Harmonic Mean: Assuming 1/X_i i = 1,…, n are normally distributed, the confidence interval is:

$$\text{Lower limit} = \frac{1}{\frac{1}{H} + t_{1-\alpha/2,n-1}\, a_H}$$

$$\text{Upper limit} = \frac{1}{\frac{1}{H} - t_{1-\alpha/2,n-1}\, a_H}$$

where H is the harmonic mean and the term a_H (which is *not* the standard error of harmonic mean) is defined as:

$$a_H = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(1/H - 1/X_i\right)^2}$$

Variance: The 100(1 - α)% confidence interval for the variance is constructed using the chi-square distribution with n – 1 degrees of freedom:

$$\text{Lower limit} = \frac{s^2(n-1)}{\chi^2_{\alpha/2,n-1}}$$

$$\text{Upper limit} = \frac{s^2(n-1)}{\chi^2_{1-\alpha/2,n-1}}$$

where s^2 is the sample variance.

Standard Deviation: The lower and upper limits are the square roots of corresponding limits for variance.

Example

Open ANOVA and select Statistics 1 → Descriptive Statistics → Confidence Intervals and from the Variable Selection Dialogue select *AUC* (*C20*) as [Variable] and *Treatment* (*S19*) as [Factor] and click [Finish].

Confidence Intervals

Data variable: AUC
Subsample selected by: Treatment = A
Number of Cases: 12

	Value	Lower 95%	Upper 95%
* Mean	209.4167	169.1754	249.6580
** Median	200.5000	154.0000	290.0000
* Geometric Mean	199.8379	161.9368	246.6098
* Harmonic Mean	189.4269	153.3584	247.6786
Variance	4011.3561	2012.9935	11563.8961
Standard Deviation	63.3353	44.8664	107.5356

* t-interval
** Quantile Method: Simple Average, Interval Method: Normal Approximation

Data variable: AUC
Subsample selected by: Treatment = B
Number of Cases: 12

	Value	Lower 95%	Upper 95%
* Mean	167.1667	137.7396	196.5937
** Median	165.5000	133.0000	210.0000
* Geometric Mean	160.4173	131.3584	195.9047
* Harmonic Mean	152.6247	123.7428	199.0937
Variance	2145.0606	1076.4422	6183.7587
Standard Deviation	46.3148	32.8092	78.6369

* t-interval

** Quantile Method: Simple Average, Interval Method: Normal Approximation

Go back to the Variable Selection Dialogue omit *Treatment* (*S19*) from the [Factor] list and select the Z interval option on the next dialogue.

Confidence Intervals

Data variable: AUC
Number of Cases: 24

	Value	Lower 95%	Upper 95%
*** Mean**	188.2917	164.9290	211.6543
**** Median**	187.5000	154.0000	220.0000
*** Geometric Mean**	179.0460	156.5802	204.7352
*** Harmonic Mean**	169.0460	146.8020	199.2349
Variance	3410.0417	2059.8730	6710.0662
Standard Deviation	58.3956	45.3858	81.9150

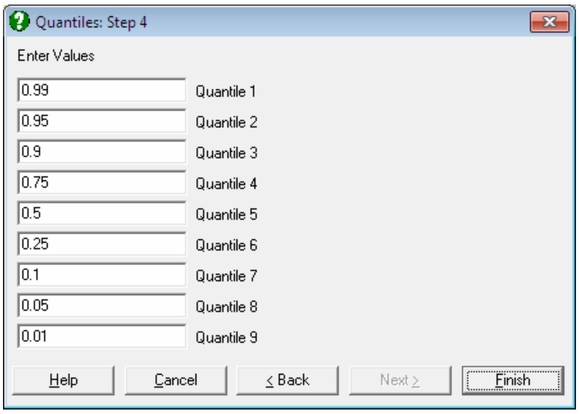
* Z-interval
** Quantile Method: Simple Average, Interval Method: Normal Approximation

5.1.3. Quantiles (Percentiles)

Quantiles and their confidence limits can be estimated for multiple samples (see 5.0.1. Multisample Data Types). A dialogue allows specifying up to nine quantiles. A quantile will be computed for any entry as long as:

$$0 < q < 1$$

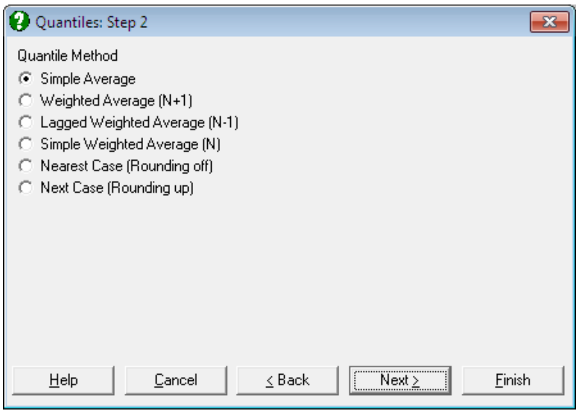
All other values will be treated as missing and they will not be shown in the output. For instance, if you wish to display only the median and quartiles in the output, enter 0.25, 0.5 and 0.75 in the first three boxes and enter 0 in the remaining. The output will have three rows only. Values entered in this dialogue will be stored by the program.


Quantiles: Step 4

Enter Values

0.99	Quantile 1
0.95	Quantile 2
0.9	Quantile 3
0.75	Quantile 4
0.5	Quantile 5
0.25	Quantile 6
0.1	Quantile 7
0.05	Quantile 8
0.01	Quantile 9

5.1.3.1. Quantile Methods


Quantiles: Step 2

Quantile Method

- ● Simple Average
- ○ Weighted Average (N+1)
- ○ Lagged Weighted Average (N-1)
- ○ Simple Weighted Average (N)
- ○ Nearest Case (Rounding off)
- ○ Next Case (Rounding up)

It is possible to compute quantiles according to six alternative definitions. The common parameters used in these definitions are as follows:

n: Valid number of cases
q: Quantile / 100 (0 < q < 1)
j: Integer part of nq
g: Fractional part of nq

Simple Average (UNISTAT default):

$nq = j + g$

if $g = 0$ then $y = (x_j + x_{j+1})/2$

if $g > 0$ then $y = x_{j+1}$

Weighted Average (N + 1):

$(n+1)q = j + g$

$y = (1-g)x_j + gx_{j+1}$

Lagged Weighted Average (N - 1):

$(n-1)q + 1 = j + g$

$y = (1-g)x_j + gx_{j+1}$

Simple Weighted Average (N):

$nq = j + g$

$y = (1-g)x_j + gx_{j+1}$

Nearest Case (Rounding-off):

$nq + 0.5 = j + g$

$y = x_j$

Next Case (Rounding-up):

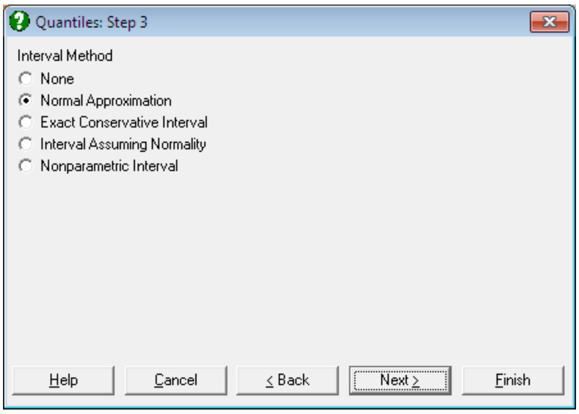
$nq = j + g$

if $g = 0$ then $y = x_j$

if $g > 0$ then $y = x_{j+1}$

5.1.3.2. Quantile Interval Methods

A further dialogue allows for selecting one of the following confidence interval methods:


Quantiles: Step 3

Interval Method

○ None
● Normal Approximation
○ Exact Conservative Interval
○ Interval Assuming Normality
○ Nonparametric Interval

None: Quantiles are displayed without their confidence limits.

Normal Approximation: Let l and u be the ranks of the lower and upper limits respectively. The approximate 100(1 - α)% confidence interval for the Q^th quantile is constructed using the normal distribution:

$$l = nq - Z_{1-\alpha/2}\sqrt{nq(1-q)}$$
$$u = 1 + nq + Z_{1-\alpha/2}\sqrt{nq(1-q)}$$

where:

$q = Q / 100 \ (0 < q < 1)$

The lower and upper limits of the confidence interval are the l^th and u^th observations of the data sorted in increasing order. For small data sets a closed interval may not exist. See Gardner & Altman (2000), p. 39.

Exact Conservative Interval: This is calculated using the binomial distribution, by finding the rank of the two observations, one corresponding to the largest cumulative probability under:

$$p_l = \alpha/2$$

and the other to the smallest over:

$p_u = 1 - \alpha/2$

using the binomial distribution with a probability equal to q. The exact conservative interval is p_u - p_l. See Gardner & Altman (2000), p. 39.

Interval Assuming Normality: Assuming the variable is normally distributed, confidence limits for the Q^{th} quantile are based on the noncentral t-distribution:

$$\text{Lower limit} = \overline{X} + t'_{\alpha/2, n-1, Z_q \sqrt{n}} SE$$

$$\text{Upper limit} = \overline{X} - t'_{\alpha/2, n-1, Z_{1-q} \sqrt{n}} SE$$

See Hahn, G. J. and Meeker, W. Q. (1991), p. 56.

Nonparametric Interval: The distribution-free confidence limits for the Q^{th} quantile should conform to the following set of constrains:

- l and u are symmetric (or nearly symmetric) around p(n + 1),
- l and u are as near to p(n + 1) as possible,
- the coverage probability is as small as possible over 1 – α, or
- B(u – 1; n, p) – B(l – 1; n, p) ≥ 1 – α,
- 0 < l < u ≤ n,
- 0 < p < 1.

where B is the cumulative binomial probability:

$$B_{k,n,p} = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

and *nearly symmetric* means that the difference between distances of l and u from q(n + 1) can only be ±1 position at the most. The left hand side of the first inequality is called the *coverage probability*. See Hahn, G. J. and Meeker, W. Q. (1991), p. 83.

5.1.3.3. Quantiles in Other Procedures

Apart from the dedicated Quantiles (Percentiles) procedure, UNISTAT also reports quantiles (or percentiles) in some other procedures. The quantile method and the quantile interval method selected in this procedure will also be valid in the following procedures, with a few exceptions.

Summary Statistics: Median and quartiles are reported (see 5.1.1. Summary Statistics).

Confidence Intervals: Median and its confidence limits are reported (see 5.1.2. Confidence Intervals).

Sample Statistics: Median and quartiles are reported. The quantile method selected in the current procedure will be valid only for the first data option Select Multiple Columns. If data has frequency counts or it is grouped (i.e. data options 2 and 3), then quantiles are computed with their own dedicated algorithms corresponding to these data types (see 5.1.4. Sample Statistics).

Histogram: Median and quartiles are drawn on the graph, if selected (see 5.3.3. Histogram).

3D Histogram: Median and quartiles are drawn on the graph, if selected (see 5.3.4. 3D Histogram).

Box-Whisker, Dot and Bar Plots: In box plot median and quartiles and in dot plot median, quartiles and user-defined percentiles can be drawn on the graph. The quantile method selected in the current procedure will be valid for all these quantiles (see 5.3.1. Box-Whisker, Dot and Bar Plots).

5.1.3.4. Quantiles Example

Open ANOVA and select Statistics 1 → Descriptive Statistics → Quantiles (Percentiles). Select *AUC* (*C20*) as [Variable] and *Treatment* (*S19*) as [Factor]. In the next two dialogues select quantile method as Simple Average and interval method Nonparametric Interval. In the last dialogue accept the quantiles given and click [Finish].

Quantiles

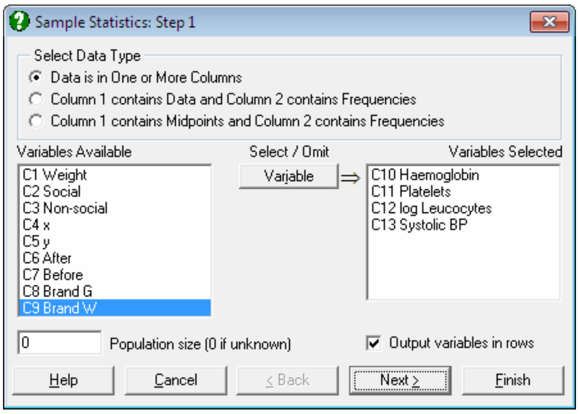
Data variable: AUC
Subsample selected by: Treatment = A
Number of Cases: 12
Quantile Method: Simple Average
Interval Method: Nonparametric Interval

Quantile	Value	Lower 95%	Upper 95%	Lower rank	Upper rank	Coverage probability
99%	294.0000	290.0000	294.0000	10	12	0.1134
95%	294.0000	290.0000	294.0000	10	12	0.4401
90%	293.0000	290.0000	294.0000	10	12	0.6067
75%	270.0000	200.0000	294.0000	6	12	0.9541
50%	200.5000	154.0000	290.0000	3	10	0.9614
25%	161.0000	97.0000	201.0000	1	7	0.9541
10%	151.0000	97.0000	154.0000	1	3	0.6067
5%	97.0000	97.0000	154.0000	1	3	0.4401
1%	97.0000	97.0000	154.0000	1	3	0.1134

Data variable: AUC
Subsample selected by: Treatment = B
Number of Cases: 12
Quantile Method: Simple Average
Interval Method: Nonparametric Interval

Quantile	Value	Lower 95%	Upper 95%	Lower rank	Upper rank	Coverage probability
99%	240.0000	210.0000	240.0000	10	12	0.1134
95%	240.0000	210.0000	240.0000	10	12	0.4401
90%	220.0000	210.0000	240.0000	10	12	0.6067
75%	200.0000	163.0000	240.0000	6	12	0.9541
50%	165.5000	133.0000	210.0000	3	10	0.9614
25%	136.5000	77.0000	168.0000	1	7	0.9541
10%	116.0000	77.0000	133.0000	1	3	0.6067
5%	77.0000	77.0000	133.0000	1	3	0.4401
1%	77.0000	77.0000	133.0000	1	3	0.1134

5.1.4. Sample Statistics



Sample Statistics: Step 1

Select Data Type

● Data is in One or More Columns
○ Column 1 contains Data and Column 2 contains Frequencies
○ Column 1 contains Midpoints and Column 2 contains Frequencies

Variables Available	Select / Omit	Variables Selected
C1 Weight C2 Social C3 Non-social C4 x C5 y C6 After C7 Before C8 Brand G C9 Brand W	Variable ⇒	C10 Haemoglobin C11 Platelets C12 log Leucocytes C13 Systolic BP

0 Population size (0 if unknown)

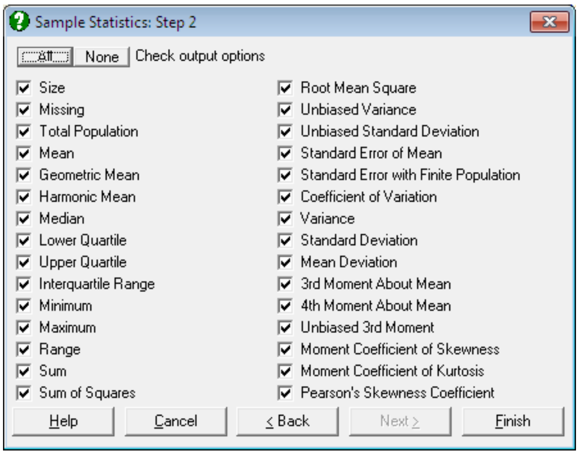
☑ Output variables in rows

The Variable Selection Dialogue for this procedure offers three types of data to analyse (see 5.0.2. One-Sample Data Types). A text box is also provided on this dialogue to enter the size of the total population from which the sample is drawn. The default value of 0 means that the total population is not known and the program assumes an infinite population. A non-zero population value affects only the standard error of mean in output.

The Output variables in rows check box allows you to transpose the output matrix. This will be useful when you wish to use the output from this procedure (such as means and standard errors) for further analysis in other procedures.

Output for different data options differ slightly. For instance, the grouped data output includes the Sheppard's correction for the second and fourth moments but it does not include minimum, maximum and range.

For ungrouped data, the method used in computing the median, lower and upper quartiles is indicated in the output. This can be one of the six methods described in the previous section 5.1.3.1. Quantile Methods.



Sample Statistics: Step 2

Check output options

- ☑ Size
- ☑ Missing
- ☑ Total Population
- ☑ Mean
- ☑ Geometric Mean
- ☑ Harmonic Mean
- ☑ Median
- ☑ Lower Quartile
- ☑ Upper Quartile
- ☑ Interquartile Range
- ☑ Minimum
- ☑ Maximum
- ☑ Range
- ☑ Sum
- ☑ Sum of Squares
- ☑ Root Mean Square
- ☑ Unbiased Variance
- ☑ Unbiased Standard Deviation
- ☑ Standard Error of Mean
- ☑ Standard Error with Finite Population
- ☑ Coefficient of Variation
- ☑ Variance
- ☑ Standard Deviation
- ☑ Mean Deviation
- ☑ 3rd Moment About Mean
- ☑ 4th Moment About Mean
- ☑ Unbiased 3rd Moment
- ☑ Moment Coefficient of Skewness
- ☑ Moment Coefficient of Kurtosis
- ☑ Pearson's Skewness Coefficient

The following statistics can be calculated for *ungrouped* (option 1) and *frequency* and *grouped* data (options 2 and 3). Let n be the number of valid observations (i.e. excluding missing values) and f_i the frequency of data point X_i given in column 2. Note that for ungrouped data $f_i = 1$, $i = 1, \ldots, n$.

Size: Number of cases (rows) in the sample, including missing values.

Missing: Number of missing cases in the sample. In frequency and grouped data a case is considered missing when either or both of value and frequency are missing.

Total Frequency:

$$N = \sum_{i=1}^{n} f_i$$

N = n for ungrouped data.

Mean: The weighted arithmetic mean is:

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$$

Geometric Mean: The weighted geometric mean is:

$$G = \sqrt[N]{\prod_{i=1}^{n} X_i^{f_i}}$$

Harmonic Mean: The weighted harmonic mean is:

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{f_i}{X_i}}$$

The following relationship should hold if $X_i \geq 0$, i = 1, …, n:

$$H \leq G \leq \overline{X}$$

Median: For ungrouped data, this is computed using the quantile method selected in step two of the Quantiles (Percentiles) procedure, as described in section 5.1.3.1. Quantile Methods.

For frequency and grouped data, both value and frequency columns are sorted in ascending order according to values. For frequency data, half of total frequency is found and the median is calculated as above. For grouped data, median is calculated by interpolation as:

$$L + \left(\frac{\frac{N}{2} - \left(\sum_{i=1}^{L} f_i \right)}{f_{\text{Median}}} \right) C$$

where:

- L is the lower class boundary of the class containing the median,
- the summation term is the sum of frequencies of all classes lower than the median class,
- C is the size of median class interval and
- N is the total frequency as defined above.

Lower Quartile: Calculations are similar to that of median, except for 25% quantile instead of 50%.

Upper Quartile: Calculations are similar to that of median, except for 75% quantile instead of 50%.

Interquartile Range: Difference between upper and lower quartiles.

Minimum: Smallest observed value in data (not available for grouped data).

Maximum: Greatest observed value in data (not available for grouped data).

Range: Difference between maximum and minimum values (not available for grouped data).

Sum: The weighted sum is:

$$\text{Sum} = \sum_{i=1}^{n} f_i X_i$$

Sum of Squares: The weighted sum of squares is:

$$\text{Ssq} = \sum_{i=1}^{n} f_i X_i^2$$

Root Mean Square (Quadratic mean):

$$\sqrt{\frac{\text{Ssq}}{N}}$$

Unbiased Variance:

$$\text{Var}_U = \frac{\text{Ssq} - \dfrac{\text{Sum}^2}{N}}{N-1}$$

Unbiased Standard Deviation:

$$\text{Std}_U = \sqrt{\text{Var}_U}$$

Standard Error of Mean:

$$\frac{\text{Std}_U}{\sqrt{N}}$$

Standard Error with Finite Population Correction: Available only when total population is known and it is greater than the total frequency.

$$\frac{\text{Std}_U}{\sqrt{N}}\sqrt{1-\frac{N}{\text{TotalPop}}}$$

Coefficient of Variation:

$$\frac{\text{Std}_U}{\overline{X}}$$

Variance:

$$\mathrm{Var_B} = \frac{\mathrm{Ssq}}{\mathrm{N}} - \overline{\mathrm{X}}^{2}$$

Standard Deviation:

$$\mathrm{Std_B} = \sqrt{\mathrm{Var_B}}$$

Sheppard's Correction for 2nd Moment (Variance): Available for only grouped data:

$$\mathrm{Var_B} - \frac{\mathrm{C}^2}{12}$$

where C is the size of uniform class interval.

Mean Deviation:

$$\frac{\sum_{i=1}^{n} f_i \left| X_i - \overline{X} \right|}{N}$$

3rd Moment About the Mean:

$$m_3 = \frac{\sum_{i=1}^{n} f_i \left(X_i - \overline{X} \right)^3}{N}$$

4th Moment About the Mean:

$$m_4 = \frac{\sum_{i=1}^{n} f_i \left(X_i - \overline{X} \right)^4}{N}$$

Unbiased 3rd Moment:

$$m_3 \frac{N^2}{(N-1)(N-2)}$$

Sheppard's Correction for the 4th Moment: Available for only grouped data:

$$m_4 - \mathrm{Var}_B \frac{C^2}{2} + \frac{7C^4}{240}$$

where C is the size of uniform class interval.

Moment Coefficient of Skewness:

$$\frac{m_3}{\mathrm{Var}_B \mathrm{Std}_B}$$

An alternative definition of skewness is given in section 5.1.1. Summary Statistics.

Moment Coefficient of Kurtosis:

$$\frac{m_4}{\mathrm{Var}_B^2}$$

An alternative definition of kurtosis is given in section 5.1.1. Summary Statistics.

Pearson's Second Coefficient of Skewness:

$$\frac{3\left(\overline{X} - \mathrm{Median}\right)}{\mathrm{Std}_B}$$

Example 1: Ungrouped data

Open PARTEST and select Statistics 1 → Descriptive Statistics → Sample Statistics. Select *Haemoglobin*, *Platelets*, *log Leucocytes*, and *Systolic BP* (*C10* to *C13*) as [Variable]s, uncheck the Output variables in rows box and click [Finish].

Sample Statistics

Quantile Method: Simple Average

	Haemoglobin	Platelets	log Leucocytes	Systolic BP
Size	10.0000	10.0000	10.0000	10.0000
Missing	0.0000	0.0000	0.0000	0.0000
Mean	-0.5300	-0.0300	-0.5900	3.1000
Geometric Mean	*	*	*	*
Harmonic Mean	*	*	*	*
Median	-0.6000	0.1000	-0.6500	2.0000
Lower Quartile	-1.5000	-1.0000	-1.6000	-2.0000
Upper Quartile	0.0000	0.6000	0.9000	8.0000
Interquartile Range	1.5000	1.6000	2.5000	10.0000
Minimum	-2.4000	-2.2000	-3.2000	-6.0000
Maximum	2.3000	1.9000	1.7000	14.0000
Range	4.7000	4.1000	4.9000	20.0000
Sum	-5.3000	-0.3000	-5.9000	31.0000
Sum of Squares	22.0700	13.3900	25.1700	437.0000
Root Mean Square	1.4856	1.1572	1.5865	6.6106
Unbiased Variance	2.1401	1.4868	2.4099	37.8778
Unbiased Standard Deviation	1.4629	1.2193	1.5524	6.1545
Standard Error of Mean	0.4626	0.3856	0.4909	1.9462
Coefficient of Variation	-2.6186	-38.5588	-2.4961	1.8834
Variance	1.9261	1.3381	2.1689	34.0900
Standard Deviation	1.3878	1.1568	1.4727	5.8387
Mean Deviation	1.1500	0.9020	1.2500	4.9200
3rd Moment About Mean	1.3179	-0.4318	-0.1544	69.6720
4th Moment About Mean	9.2971	4.2938	9.5652	2527.7857
Unbiased 3rd Moment	1.8304	-0.5998	-0.2144	96.7667
Moment Coefficient of Skewness	0.4930	-0.2790	-0.0483	0.3500
Moment Coefficient of Kurtosis	2.5060	2.3981	2.0334	2.1751
Pearson's Skewness Coefficient	0.1513	-0.3371	0.1222	0.5652

Example 2: Variables in rows

Continuing from the last example, go back to Variable Selection Dialogue, check the Output variables in rows box and click [Next]. From the Output Options Dialogue select only the last three options and click [Finish].

Sample Statistics

	Moment Coefficient of Skewness	Moment Coefficient of Kurtosis	Pearson's Skewness Coefficient
Haemoglobin	0.4930	2.5060	0.1513
Platelets	-0.2790	2.3981	-0.3371
log Leucocytes	-0.0483	2.0334	0.1222
Systolic BP	0.3500	2.1751	0.5652

Example 3: Frequency data

Open TIMESER, select Statistics 1 → Descriptive Statistics → Sample Statistics and select the second data option Column 1 contains Data and Column 2 contains Frequencies. Select *Surface Area* (*C13*) as [Column 1] and *Blemishes* (*C14*) as [Column 2] and enter 150 in the Total Population box. The following results are obtained:

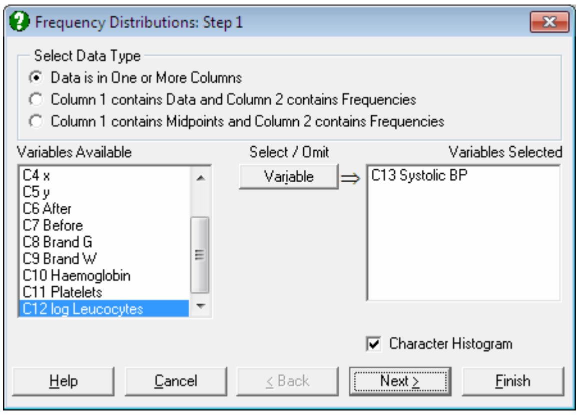
Sample Statistics

Surface Area: contains data, Blemishes contains frequencies

	Surface Area
Size	20.0000
Missing	0.0000
Total Frequency	94.0000
Total Population	150.0000
Mean	0.8462
Geometric Mean	0.8265
Harmonic Mean	0.8070
...	...
Root Mean Square	0.8653
Unbiased Variance	0.0330
Unbiased Standard Deviation	0.1817
Standard Error of Mean	0.0187
Standard Error with Finite Population	0.0115
Coefficient of Variation	0.2136
Variance	0.0327
Standard Deviation	0.1807
Mean Deviation	0.1443
3rd Moment About Mean	0.0004
4th Moment About Mean	0.0017
Unbiased 3rd Moment	0.0004
Moment Coefficient of Skewness	0.0635
Moment Coefficient of Kurtosis	1.6210
Pearson's Skewness Coefficient	0.1024

5.1.5. Frequency Distributions

The Variable Selection Dialogue will offer three types of data to analyse (see 5.0.2. One-Sample Data Types). A check box situated at bottom right on the same dialogue will enable you to select the type of output: Frequency Table or Character Histogram.



Frequency Distributions: Step 1

Select Data Type

● Data is in One or More Columns

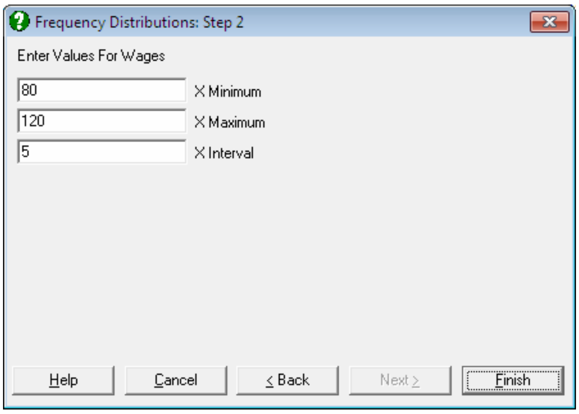
○ Column 1 contains Data and Column 2 contains Frequencies

○ Column 1 contains Midpoints and Column 2 contains Frequencies

Variables Available	Select / Omit	Variables Selected
C4 x C5 y C6 After C7 Before C8 Brand G C9 Brand W C10 Haemoglobin C11 Platelets C12 log Leucocytes	Variable ⇒	C13 Systolic BP

☑ Character Histogram

The next dialogue will prompt for the lower bound, upper bound and the class interval. The default values computed and suggested by the program will often generate a satisfactory outcome. However, the program’s suggestions can be overridden and other values used. If too small a class interval is entered the number of classes may turn out to be too large. By default, the program allows a maximum of 200 classes, though this can be increased if necessary. Also, if the specified lower limit is greater than the minimum observation or the upper limit is less than the maximum observation, then the cumulative and relative frequencies will not add up to their totals. In all three cases a message is issued, but the procedure is not aborted.


Frequency Distributions: Step 2

Enter Values For Wages

80 X Minimum

120 X Maximum

5 X Interval

Starting from the lower bound, the program scans data in steps equal to the class interval and determines observations falling within each class. Class intervals are closed from below (or include the lower limit) and are open from above (or exclude the upper limit), with the exception of the last class, which is also closed from above. If the data is ungrouped (data option 1) then the program counts the number of observations within each class. If observations are accompanied with frequency counts (data option 2) then the program adds up the frequencies of observations falling within each class for a cumulative distribution. If the data is already grouped (data option 3) then the program will reconstruct all class midpoints according to the lower bound (the first observation) and the class interval (the difference between the second and the first observations) of column 1. No checks are made to determine whether the first column does actually contain midpoints.

Frequency Table: If the Table option is selected then the output will be in the form of a table containing class numbers, midpoints, frequency counts, cumulative frequency counts, percentages and cumulative percentages.

Character Histogram: A horizontal bar character histogram is drawn. The bar lengths are scaled so that the longest bar fits into the Width parameter defined in Tools → Options → Output → Text Margins.

Example 1

Open DEMODATA, select Statistics 1 → Descriptive Statistics → Frequency Distributions and from the Variable Selection Dialogue select the data option 1 Select Multiple Columns. Select *Output 2* (*C9*) as [Variable] and accept the default values from the subsequent dialogues.

Frequency Distributions

For Output2

Class	Mid-Point	Frequency	Cumulative	Percentage	Cumulative
1	92.5000	5	5	8.6%	8.6%
2	95.0000	2	7	3.4%	12.1%
3	97.5000	2	9	3.4%	15.5%
4	100.0000	2	11	3.4%	19.0%
5	102.5000	3	14	5.2%	24.1%
6	105.0000	4	18	6.9%	31.0%
7	107.5000	7	25	12.1%	43.1%
8	110.0000	10	35	17.2%	60.3%
9	112.5000	17	52	29.3%	89.7%
10	115.0000	6	58	10.3%	100.0%

Example 2

Open DEMODATA, select Statistics 1 → Descriptive Statistics → Frequency Distributions and from the Variable Selection Dialogue select the data option 1 Select Multiple Columns. Select *Output 2* (*C9*) as [Variable], check the Character Histogram box and accept the default values from the subsequent dialogues.

Frequency Distributions

For Output2

Class	Mid-Point	Frequency	0.0000 17.0000
1	92.5000	5	******************
2	95.0000	2	********
3	97.5000	2	********
4	100.0000	2	********
5	102.5000	3	***********
6	105.0000	4	***************
7	107.5000	7	*************************
8	110.0000	10	************************************
9	112.5000	17	**
10	115.0000	6	**********************

5.1.6. Stem and Leaf Plot

A character stem and leaf plot of the selected column is drawn. The stem is the number displayed to the left of the vertical bar (|) and each leaf is a digit displayed to the right of it. Each leaf represents a separate data value. Adding the stem value to the leaf digit and multiplying by the leaf unit will give the class to which a data point belongs. For instance:

Leaf unit = 1	1 | 2 represents 12
Leaf unit = 10	1 | 2 represents 120
Leaf unit = 0.1	1 | 2 represents 1.2

Values are rounded to the nearest leaf unit. The first column in the output represents the number of cases which lie in this row or further away from the median. If the first column has brackets around it, this denotes that the median lies in this row and the number shown is the number of leaves in this row.

Example

Open DEMODATA, select Statistics 1 → Descriptive Statistics → Stem and Leaf Plot and select *Output 2* (*C9*) as [Variable].

Stem and Leaf Diagram

For Output2

```
Leaf unit = 1     1|2 represents  12

    6     9| 333344
   10     9| 5689
   15    10| 02334
   27    10| 566777888899
  (27)   11| 000011111112222333333333444
    4    11| 5555
```

5.1.7. Sequence Diagram

A vertical character plot of the selected column is drawn against the number of rows. The minimum and maximum observations are displayed on the horizontal axis. Row numbers and values are displayed on the left of the plot.

Rows containing missing data are drawn blank and a missing data marker is printed in place of their value. The resolution of this graph depends on the Width parameter defined in Tools → Options → Output → Text Margins.

Example

Open DEMODATA and select Statistics 1 → Descriptive Statistics → Sequence Diagram. Select *Dependent* (*C13*) as [Variable].

Sequence Diagram

For Dependent

```
Row     Value | 106.9000                                    121.2000
  1  121.2000 |                                                            *
  2  119.4000 |                                                     *
  3  118.0000 |                                               *
  4  116.5000 |                                         *
  5  115.4000 |                                    *
  6  114.9000 |                                  *
  7  114.8000 |                                  *
  8  114.6000 |                                 *
  9  115.0000 |                                  *
 10  116.0000 |                                       *
 11  115.6000 |                                     *
 12  116.0000 |                                       *
...       ... |
 24  107.4000 |   *
 25  107.5000 |   *
 26  107.8000 |     *
 27  107.6000 |    *
 28  107.5000 |   *
 29  107.6000 |    *
 30  107.3000 |   *
 31  107.0000 | *
 32  106.9000 | *
```

5.1.8. Scatter Diagram

Two or more columns can be selected by clicking on [Variable]. The program will draw an X-Y character plot of all possible pairs. The axes are scaled automatically and an asterisk is printed at each x-y point. If two or more points fall on the same location, their count is shown instead of an asterisk. 0 will be printed for 10 or more coinciding points.

Missing data are omitted pairwise, i.e. a point is not drawn if either or both observations are missing. The two columns need not be of equal length. Points above the length of the shorter column will be considered as missing.

The resolution of this graph depends on the Width parameter defined in Tools → Options → Output → Text Margins.

Example

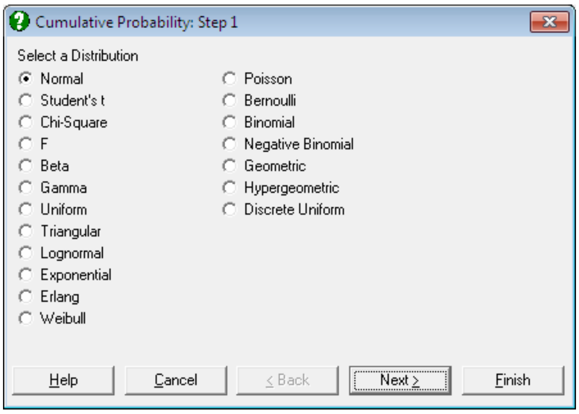
Open DEMODATA and select Statistics 1 → Descriptive Statistics → Scatter Diagram. Select *Wages* (*C2*) and *Energy* (*C3*) as [Variable]s.

Scatter Diagram

```
Energy
 124.000+
        |                                                  **
        |                                                **2    *  *
        |                                                 3*  *
        |                                               *   *
 112.000+                                      *
        |                                      *           *
        |                                     *52    *
        |                                   *  *2
        |                                    **
 100.000+                                 *
        |
        |
        |                              *
        |
 88.0000+                          *
        |   **       *     *
        |  2*  ** **  * *      *
        |  3*
        |   *
 76.0000+
         +-------------+-------------+-------------+-------------+-------------+
   80.0000       90.0000      100.0000      110.0000      120.0000      130.0000
                                 Wages
```

5.2. Distribution Functions

It is possible to calculate cumulative probabilities for given critical values and, conversely, critical values for given cumulative probabilities. This can be done either as a quick calculator-type single-entry calculations, or multiple data points can be input together with their parameter values as spreadsheet columns. Further options enable the user to produce Random Numbers with given distribution parameters and estimate and fit theoretical distributions on data. The following Distribution Functions are supported.



Continuous
Normal
Student's t
Chi-Square
F
Beta
Gamma
Uniform
Triangular
Lognormal
Exponential
Erlang
Weibull

Discrete
Poisson
Bernoulli
Binomial
Negative Binomial
Geometric
Hypergeometric
Discrete Uniform

Any one of these 19 Distribution Functions can be selected from a dialogue. There are a few options which are not implemented. These are Random Numbers and Expected Frequencies for hypergeometric distribution, and Random Numbers for negative binomial and Weibull distributions.

The routines used for Distribution Functions are highly accurate and in general all displayed digits are significant. If in some cases there appears to be a discrepancy between the values calculated by UNISTAT and the published statistical tables, the most likely reason is the rounding off errors involved in published tables due to the limited number of digits they display.

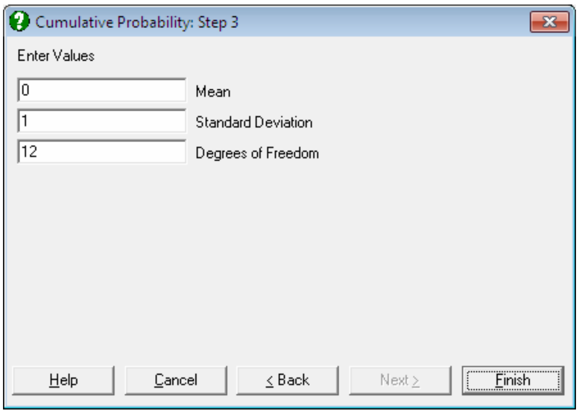
A significant portion of potential errors are trapped during the input of parameters. The user is prevented from entering illegal values for most parameters. For instance, the program will not proceed unless a number between 0 and 1 is entered for a probability value, a positive integer for a degree of freedom, a lower limit which is less than the upper limit. However, it is not possible to trap all errors at the entry stage. In such cases the value -99 will be returned for a parameter which cannot be calculated.

Density functions and mean and standard deviation formulas of all 19 distributions are given in the Appendix.

5.2.1. Cumulative Probability

This is a single-entry probability calculator, which accepts one critical value input at a time. If you have a large number of critical values and their corresponding parameter values already entered into spreadsheet columns, then use the Probabilities and Critical Values procedure.

First select the distribution you want to use. Then the program prompts for a critical value and for the necessary distribution parameters (e.g. mean and standard deviation for normal distribution, or alpha and beta coefficients for gamma distribution). Output includes the given parameters, the estimated mean and variance (which are calculated from the given parameters), cumulative probability (p), and complementary probability (1 - p).



Cumulative Probability: Step 3

Enter Values

0 Mean

1 Standard Deviation

12 Degrees of Freedom

The program stores the parameters entered for each distribution function. These numbers are suggested for entry when the same distribution is subsequently called either for cumulative probabilities or for critical values.

Example

Select Statistics 1 → Distribution Functions → Cumulative Probability and from the list of distributions, Normal. At the next dialogue enter 1.96 for the critical value and enter 0 and 1 for the mean and standard deviation at the next:

Cumulative Probability

Normal Distribution

```
                    x Value =    1.9600
                       Mean =    0.0000
         Standard Deviation =    1.0000
                       Mean =    0.0000
                   Variance =    1.0000
                  Frequency =    0.0584
     Cumulative Probability =    0.9750
  Complementary Probability =    0.0250
```

5.2.2. Critical Value

This is a single-entry probability calculator, which accepts one probability input at a time. If you have a large number of probabilities and their corresponding parameter values already entered into spreadsheet columns, then use the Probabilities and Critical Values procedure.

After selecting the distribution, enter the probability value and the necessary distribution parameters (e.g. the two degrees of freedom for the F-distribution, number of trials and probability of success for the binomial distribution). Output includes the estimated mean and variance of the distribution (which are calculated from the given parameters) and the calculated critical value.

Example

Select Statistics 1 → Distribution Functions → Critical Value and from the list of distributions select Normal. At the next dialogue enter 0.025 for the probability and enter 0 and 1 for the mean and standard deviation at the next to obtain the following results:

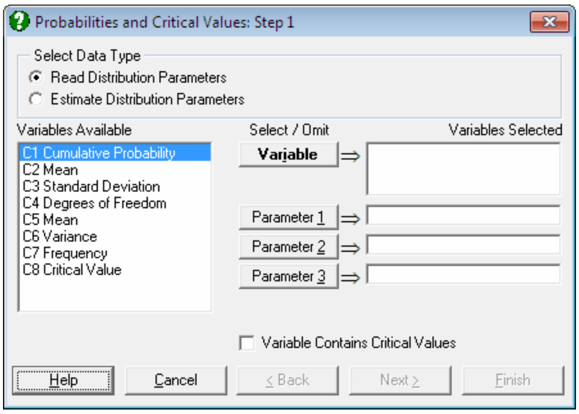
Critical Value

Normal Distribution

```
       Probability =    0.0250
              Mean =    0.0000
Standard Deviation =    1.0000
              Mean =    0.0000
          Variance =    1.0000
         Frequency =    0.0584
    Critical Value =   -1.9600
```

5.2.3. Probabilities and Critical Values

This procedure will take as input cumulative probabilities (or critical values) and their corresponding parameter values from spreadsheet columns and output the estimated critical values (or cumulative probabilities). For a few quick calculator-type calculations, you may wish to use the Cumulative Probability or Critical Value procedures instead. Two different types of analysis can be performed depending on the type of data available.



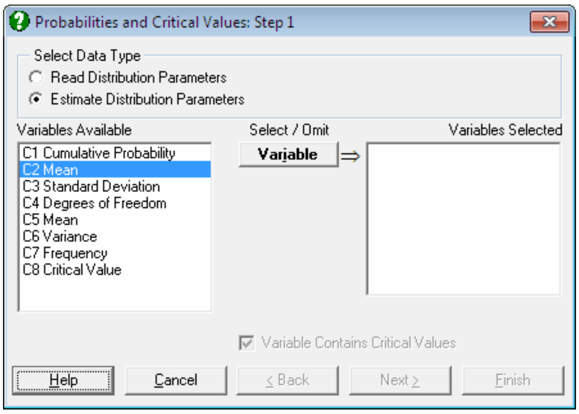
Probabilities and Critical Values: Step 1

Select Data Type
● Read Distribution Parameters
○ Estimate Distribution Parameters

Variables Available
C1 Cumulative Probability
C2 Mean
C3 Standard Deviation
C4 Degrees of Freedom
C5 Mean
C6 Variance
C7 Frequency
C8 Critical Value

□ Variable Contains Critical Values

1) **Read Distribution Parameters:** Select one or more [Variable]s containing either cumulative probabilities or critical values. If the selected column contains critical values also check the box Variable Contains Critical Values. For probability input, this box should be unchecked. Also select the parameters required by the particular distribution you will want to use by clicking [Parameter 1], [Parameter 2] or [Parameter 3]. For instance, t-distribution requires three parameters, mean, standard deviation and degrees of freedom, so at this stage you will also need to select the three columns containing these parameters. If you are not sure about how many parameters, which parameters and in which order to select, then you can consult Appendix, where all this information is given. You can also learn the parameters required by a particular distribution by running one of Cumulative Probability or Critical Value procedures first. If the number of parameter columns selected matches the number of parameters required by a particular distribution, then the usual distribution selection dialogue pops up next. Otherwise the program displays a message and does not proceed further.


Probabilities and Critical Values: Step 1

Select Data Type

○ Read Distribution Parameters

● Estimate Distribution Parameters

Variables Available

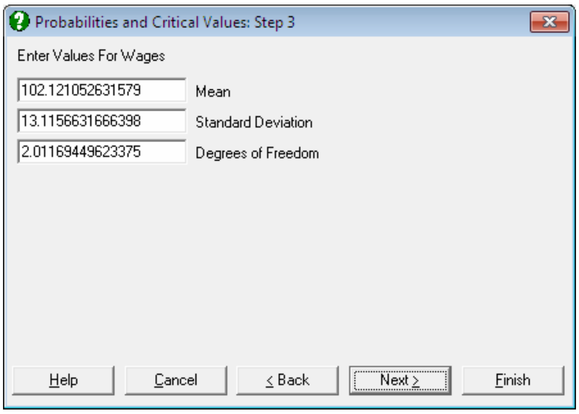
C1 Cumulative Probability
C2 Mean
C3 Standard Deviation
C4 Degrees of Freedom
C5 Mean
C6 Variance
C7 Frequency
C8 Critical Value

Select / Omit

Variables Selected

☑ Variable Contains Critical Values

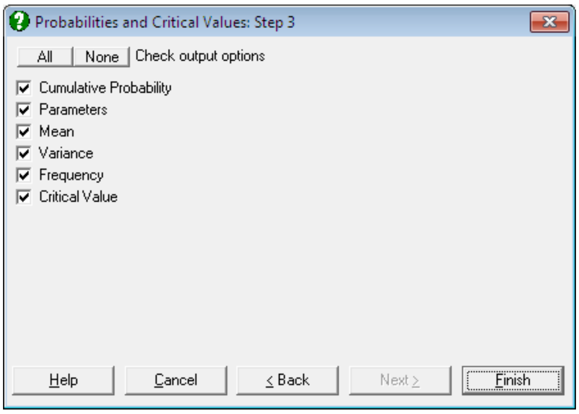
2) **Estimate Distribution Parameters:** If you do not know the parameter values then this option can be used to estimate them from data. In this case, only the [Variable]s containing critical values can be selected.


Probabilities and Critical Values: Step 3

Enter Values For Wages

102.121052631579	Mean
13.1156631666398	Standard Deviation
2.01169449623375	Degrees of Freedom

After the Distribution Functions dialogue the program will ask for the parameters of the selected distribution as in the case of single Cumulative Probability procedure. Here, however, the parameter values suggested are calculated by the program directly from the selected column, assuming that the distribution of the random variable is that given in the selected column. For instance, the suggested mean and standard deviation values for the normal distribution will be the mean and the standard deviation of the selected column. If a parameter cannot be estimated by the program then the value -99 will appear in the input field. You may override the suggested values and enter any parameter values.

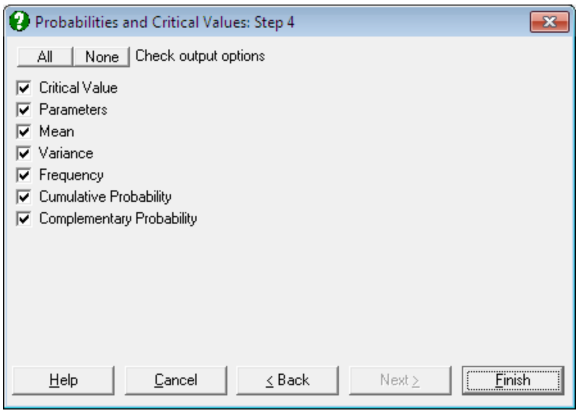


Probabilities and Critical Values: Step 3

Check output options

☑ Cumulative Probability
☑ Parameters
☑ Mean
☑ Variance
☑ Frequency
☑ Critical Value

Output includes the estimates of mean, variance and the probability density. Probabilities calculated are cumulative, i.e. they give the area enclosed under the density function from negative infinity up to the critical value given in data column. If input data are critical values, then output options will include cumulative and complementary (1 - cumulative) probabilities. If input data are cumulative probabilities, then critical values will be the last output option.



Probabilities and Critical Values: Step 4

Check output options

☑ Critical Value
☑ Parameters
☑ Mean
☑ Variance
☑ Frequency
☑ Cumulative Probability
☑ Complementary Probability

Errors are handled separately for each critical value. This means that even if all values of the selected column are illegal for the selected distribution function, a table will still be displayed with all results as missing. This helps to visualise the valid range of the random variable for given parameters.

Example 1

Consider a hypothetical example where three sets of cumulative probabilities, means and standard deviations are given and we want to find their corresponding critical values. Select File → New and then enter the following data:

	Prob	Mean	StdDev
1	0.025	0	1
2	0.05	-2	2
3	0.1	1	0.5

Select Statistics 1 → Distribution Functions → Probabilities and Critical Values, the data option Read Distribution Parameters and *C1 Prob* as [Variable], *C2 Mean* as [Parameter 1] and *C3 StdDev* as [Parameter 2]. Leave the Variable Contains Critical Values box unchecked. Select Normal from the distribution list and check all output options to obtain the following results:

Probabilities and Critical Values

Normal Distribution

Probability: Prob
Mean: Mean
Standard Deviation: StdDev

	Cumulative Probability	Mean	Standard Deviation	Mean	Variance
1	0.0250	0.0000	1.0000	0.0000	1.0000
2	0.0500	-2.0000	2.0000	-2.0000	4.0000
3	0.1000	1.0000	0.5000	1.0000	0.2500

	Frequency	Critical Value
1	0.0584	-1.9600
2	0.1031	-5.2897
3	0.1755	0.3592

If you are using UNISTAT in Stand-Alone Mode, click on the UNISTAT icon on the Output Medium Toolbar to send all output to UNISTAT spreadsheet. In Excel Add-In Mode select the output matrix as data.

Then select Statistics 1 → Distribution Functions → Probabilities and Critical Values again, but this time select the newly saved *Critical Value* column as [Variable] and leave [Parameter 1] and [Parameter 2] unchanged. Check the

Variable Contains Critical Values box, select Normal from the distribution list and check all output options to obtain the following results:

Probabilities and Critical Values

Normal Distribution

Critical Value: Critical Value
Mean: Mean
Standard Deviation: StdDev

	Critical Value	Mean	Standard Deviation	Mean	Variance
1	-1.9600	0.0000	1.0000	0.0000	1.0000
2	-5.2897	-2.0000	2.0000	-2.0000	4.0000
3	0.3592	1.0000	0.5000	1.0000	0.2500

	Frequency	Cumulative Probability	Complementary Probability
1	0.0584	0.0250	0.9750
2	0.1031	0.0500	0.9500
3	0.1755	0.1000	0.9000

Example 2

Select File → New and then enter numbers from 0 to 5 into the first 6 rows of column 1 (*C1*) and select Statistics 1 → Distribution Functions → Probabilities and Critical Values. Select the data option Estimate Distribution Parameters, *C1* as [Variable] from the variable list and Binomial from the distributions list. Accept the suggested parameters, number of trials = 5 and probability of success = .5, to obtain the following results:

Probabilities and Critical Values

Binomial Distribution

Critical Value: Success

	Critical Value	Number of Trials	Probability of Success	Mean	Variance
1	0.0000	5.0000	0.5000	2.5000	1.2500
2	1.0000	5.0000	0.5000	2.5000	1.2500
3	2.0000	5.0000	0.5000	2.5000	1.2500
4	3.0000	5.0000	0.5000	2.5000	1.2500
5	4.0000	5.0000	0.5000	2.5000	1.2500
6	5.0000	5.0000	0.5000	2.5000	1.2500

	Frequency	Cumulative Probability	Complementary Probability
1	0.0313	0.0313	0.9687
2	0.1562	0.1875	0.8125
3	0.3125	0.5000	0.5000
4	0.3125	0.8125	0.1875
5	0.1562	0.9687	0.0313
6	0.0313	1.0000	0.0000

5.2.4. Random Numbers

A random variable is generated with a specified length and with distribution parameters approximate to those specified by the user.

First select a distribution from the distributions menu. There are three Distribution Functions for which the Random Numbers procedure is not implemented: Weibull, negative binomial and hypergeometric distributions. After choosing any other distributions, the program will ask for the length of the random variable. The number suggested by the program is the maximum row number which is currently in use; i.e. the Used row number as displayed on the Status Panel of the Data Processor. Next the distribution parameters are entered. The numbers suggested by the program are the ones last entered by the user. A table containing the column of generated Random Numbers is displayed.

There are two alternative ways of comparing the generated Random Numbers with the theoretical distribution function. The first is the Expected Frequencies procedure which is explained below. This procedure will display the Expected Frequencies on the basis of the generated Random Numbers either in the form of a Table or a Character Histogram. The second alternative is the Histogram procedure. This will display a high resolution histogram of the generated Random

Numbers and optionally superimpose up to six estimated theoretical distributions simultaneously (see 5.3.3.3. Fitting Distribution Functions).

Example

Select Statistics 1 → Distribution Functions → Random Numbers, from the distributions list Student's t and enter 10 for the Variable Size. At the parameter dialogue enter:

- Mean = 0
- Standard Deviation = 1
- Degrees of Freedom = 5

When you reproduce this example you will obtain different numbers as they are randomly generated.

Random Numbers

Student's t Distribution

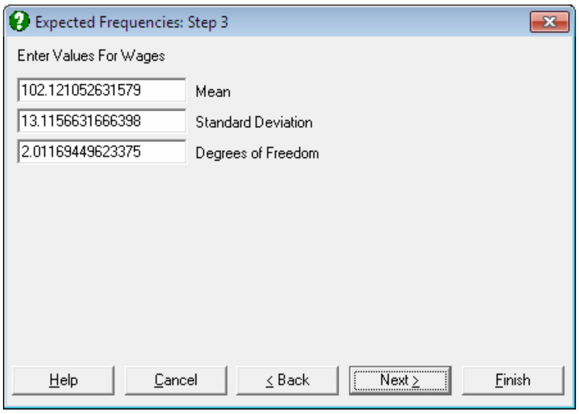
Mean: 0.0000
Standard Deviation: 1.0000
Degrees of Freedom: 5.0000

	T-Rand
1	0.0971
2	-1.0758
3	0.1968
4	-0.0896
5	-1.2651
6	0.3970
7	-1.5811
8	-2.0857
9	-3.0010
10	0.3280

5.2.5. Expected Frequencies

The distribution parameters of a random variable are estimated and the theoretical distribution with the estimated parameters is displayed.

First choose the form of data. This can be ungrouped data, data with frequency counts, or grouped data (see 5.0.2. One-Sample Data Types). Select the data column (if the data is ungrouped) or columns (if the data is with frequency counts or already grouped) containing the random variable and the distribution function. There is a check box on the same dialogue to set the form of output; Table or Character Histogram.


Expected Frequencies: Step 3

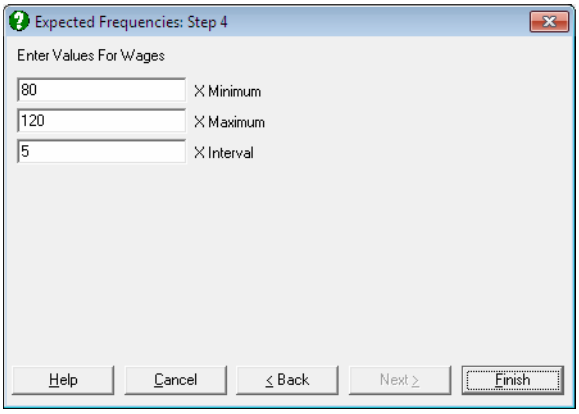
Enter Values For Wages

102.121052631579 Mean

13.1156631666398 Standard Deviation

2.01169449623375 Degrees of Freedom

The Expected Frequencies procedure is not implemented for hypergeometric distribution. When any other distribution is selected, the program will ask for the distribution parameters. The suggested numbers are those which are estimated by the program. To display the estimated distribution accept the program's suggestions. A parameter which cannot be estimated is assigned the number -99. In this case either abort the procedure or enter a number of your own choice.



Enter Values For Wages

80 X Minimum

120 X Maximum

5 X Interval

After entering the distribution parameters a dialogue is displayed which is similar to the one used in the Frequency Distributions procedure. The first three fields are for the lower bound of the first class, the upper bound of the last class and the class interval respectively.

If the form of output is Table, class midpoints, observed and expected frequencies will be displayed in a table. The second alternative is a character histogram of observed frequencies where expected frequencies are represented by (+) and are superimposed on the horizontal frequency bars.

The goodness of fit is also displayed by performing a two-sample chi-square test on the observed and expected frequency columns. Note that the degrees of freedom is adjusted by the number of parameters estimated for the distribution (see 6.3.1.2. Two Sample Chi-Square Test).

Example

Let us start by generating a column of Random Numbers using the gamma distribution. First clear all data in the spreadsheet by selecting File → New and then select Statistics 1 → Distribution Functions → Random Numbers, from the distributions list Gamma and enter 100 for the Variable Size. At the parameter dialogue enter *alpha* = 2 and *beta* = .8 to obtain a skewed distribution. If you are using UNISTAT in Stand-Alone Mode, make Data Processor active and click on the UNISTAT icon on the Input Panel. This will add the column of Random Numbers GammaRand to the Data Processor. If you are using UNISTAT in Excel Add-In Mode, then highlight the column of Random Numbers. Then select Statistics 1 → Distribution Functions → Expected Frequencies and GammaRand as [Variable] and check the Character Histogram box. The distribution list will still show the Gamma distribution. Accept the default values in the next two dialogues. You will obtain different numbers as they are generated randomly. Shapes of the histograms should look similar.

Expected Frequencies

Gamma Distribution

Observed: GammaRand
Alpha: 2.2988
Beta: 0.8697

Class	Mid-Point	Frequency	0.0000 33.0000
1	0.0000	7	*********+*
2	1.2500	32	**+
3	2.5000	33	**+******
4	3.7500	14	********************** +
5	5.0000	7	*********** +
6	6.2500	5	******+*
7	7.5000	1	**+
8	8.7500	0	+
9	10.0000	0	+
10	11.2500	1	**

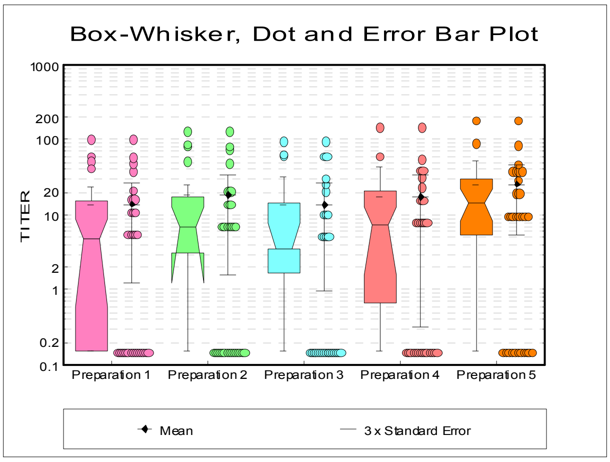
Chi-Square Statistic =	10.5291
Degrees of Freedom =	7
Right-Tail Probability =	0.1605

5.3. Descriptive Plots

With this release of UNISTAT, the Descriptive Plots menu is substantially revised and a new procedure, Bland-Altman Plot, is added. All descriptive plots can now handle categorical data. The Normal Probability Plot is no longer a univariate procedure and accordingly it is added to the Chart Gallery as an option. A new mini Chart Gallery is introduced for the three paired data procedures (3D Histogram, Bland-Altman Plot and Ladder Plot) so that each can be visualised instantly with the same data selection.

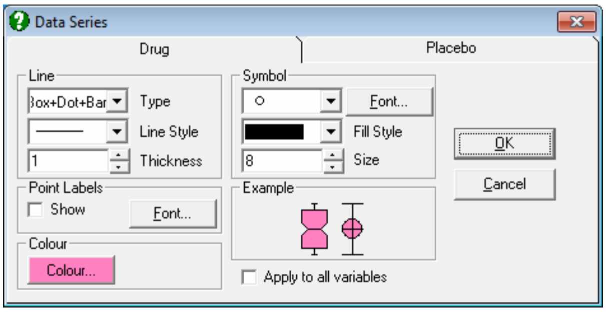
For more information on common graphics controls see 2.3. Graphics Editor.

5.3.1. Box-Whisker, Dot and Bar Plots



Multisample data can be entered in the form of multiple columns or data columns classified by factor columns. If at least one factor is selected, then a further dialogue will pop up asking for the combination of factor levels to be included. The data is plotted on the Y-axis (where the Scale Type can be one of linear, log base 10, log base e, log based to any user-defined value, logit, probit, gompit (cloglog) or loglog and the categories on the X-axis. Although an unlimited number of data series can be plotted, properties of only the first nine can be individually controlled on the Data Series dialogue that can be accessed either from the Edit → Data Series menu or by double-clicking on the graph area. The rest of the series will repeat the properties of the first nine in a circular fashion.

The Apply to all variables check box allows you to apply the current variable's settings to all selected variables.



Data Series

Drug | Placebo

Line

3ox+Dot+Bar Type

Line Style

1 Thickness

Point Labels

□ Show

Font...

Colour

Colour...

Symbol

○ Font...

Fill Style

8 Size

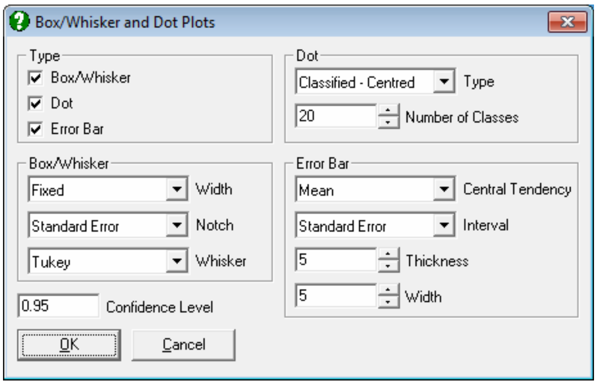
Example

□ Apply to all variables

OK

Cancel

Symbol type, symbol size, colour and Point Labels can be controlled for outlying points on Box and Whisker Plot for each data series individually.



Box/Whisker and Dot Plots

Type

☑ Box/Whisker

☑ Dot

☑ Error Bar

Box/Whisker

Fixed Width

Standard Error Notch

Tukey Whisker

0.95 Confidence Level

Dot

Classified - Centred Type

20 Number of Classes

Error Bar

Mean Central Tendency

Standard Error Interval

5 Thickness

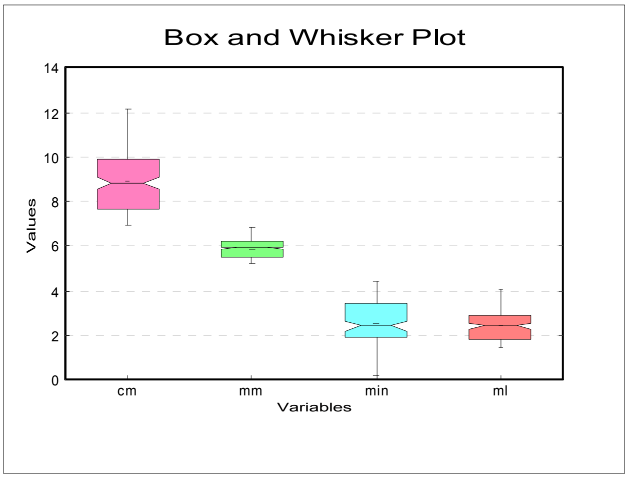
5 Width

OK

Cancel

The Edit → Width / Notch / Dots dialogue can be used to control the statistical parameters represented on the graph. The three check boxes in the Type panel allow drawing any combination of Box and Whisker Plot, Dot Plot and Error Bar Plot on the same graph. The other three frames on this dialogue are used to control the individual characteristics of each type of plot. The Confidence Level text box is included in this dialogue for the sake of convenience, although it is also available in the Variable Selection Dialogue. Changes made on this dialogue will apply to all data series.

5.3.1.1. Box and Whisker Plot



A box and whisker plot conveys the following information:

Bottom of the box: Lower quartile.

Middle of the box: Median.

Top of the box: Upper quartile.

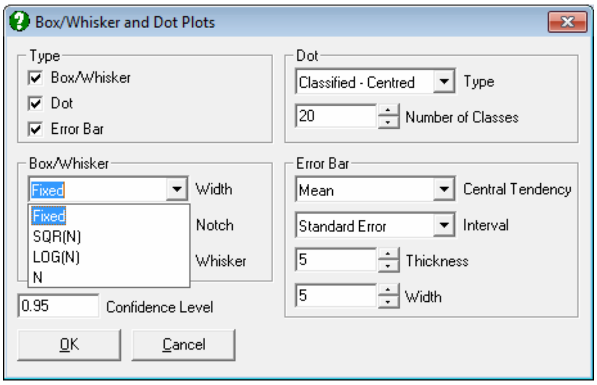
Box Width: The variable box width conveys information about the size of the sample. See below.

Notch: When there is a notch, it conveys information about the dispersion of data about the median. See below.

Lower Whisker: Lower adjacent value. Any values below this are outliers and are plotted individually. See below for alternative methods.

Upper Whisker: Upper adjacent value. Any values above this are outliers and are plotted individually. See below for alternative methods.

On the Width / Notch / Dots dialogue, the first group of controls concerns the Box and Whisker plots.



Box/Whisker and Dot Plots

Type
☑ Box/Whisker
☑ Dot
☑ Error Bar

Dot
Classified - Centred Type
20 Number of Classes

Box/Whisker
Fixed Width
Notch
Whisker

Error Bar
Mean Central Tendency
Standard Error Interval
5 Thickness
5 Width

0.95 Confidence Level

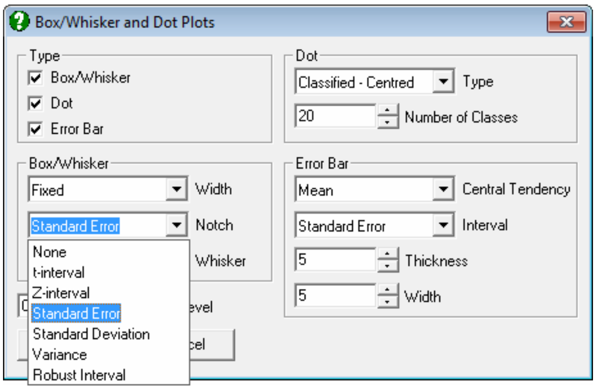
Width: The width of boxes can be used to convey information about sample sizes:

Fixed: No size information.

Sqr(n): The widths are proportional to the square root of their sample size.

Log(n): The widths are proportional to the 10 based logarithm of their sample size.

n: The widths are proportional to their sample size.



Box/Whisker and Dot Plots

Type
☑ Box/Whisker
☑ Dot
☑ Error Bar

Dot
Classified - Centred Type
20 Number of Classes

Box/Whisker
Fixed Width
Standard Error Notch
Whisker

Error Bar
Mean Central Tendency
Standard Error Interval
5 Thickness
5 Width

Notch: The extent of notches represents the following dispersion measures:

None: A notch is not drawn.

t-interval:

$$\text{Lower limit} = \text{Median} - t_{1-\alpha/2,n-1}\,\text{SE}$$
$$\text{Upper limit} = \text{Median} + t_{1-\alpha/2,n-1}\,\text{SE}$$

where $t_{1\text{-}\alpha/2,n-1}$ is the critical value from t-distribution with n - 1 degrees of freedom.

Z-interval:

$$\text{Lower limit} = \text{Median} - Z_{1-\alpha/2}\,\text{SE}$$
$$\text{Upper limit} = \text{Median} + Z_{1-\alpha/2}\,\text{SE}$$

Standard Error:

$$\text{Lower limit} = \text{Median} - \text{SE}$$
$$\text{Upper limit} = \text{Median} + \text{SE}$$

Standard Deviation: As above, but with sample standard deviation.

$$\text{Lower limit} = \text{Median} - s$$
$$\text{Upper limit} = \text{Median} + s$$

Variance: As above, but with sample variance.

$$\text{Lower limit} = \text{Median} - s^2$$
$$\text{Upper limit} = \text{Median} + s^2$$

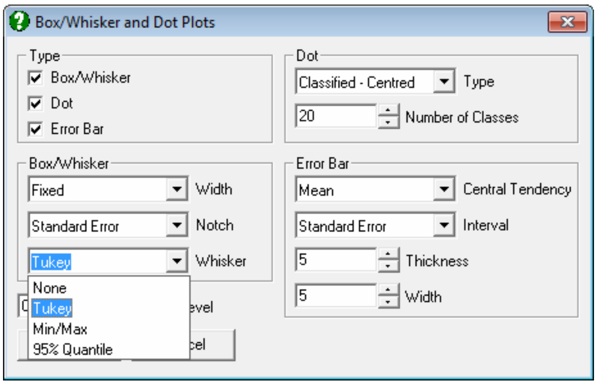
Robust Confidence Interval: The robust standard error (SE*) is defined as:

$$\text{SE}^* = \frac{1.25\,\text{IQR}}{1.35\sqrt{n}}$$

where IQR is the inter-quartile range and n is the sample size. The robust confidence interval is then defined as:

$$\text{Lower limit} = \text{Median} - Z_{1-\alpha/2}\,\text{SE}^*$$
$$\text{Upper limit} = \text{Median} + Z_{1-\alpha/2}\,\text{SE}^*$$

where $Z_{1-\alpha/2}$ is the critical value from the standard normal distribution (see McGill, R., Tukey, J. W. and Larsen, W. A. 1978).



Box/Whisker and Dot Plots

Type
☑ Box/Whisker
☑ Dot
☑ Error Bar

Dot
Classified - Centred Type
20 Number of Classes

Box/Whisker
Fixed Width
Standard Error Notch
Tukey Whisker

Error Bar
Mean Central Tendency
Standard Error Interval
5 Thickness
5 Width

Whiskers: These convey information about the dispersion of data. Any values remaining outside the extent of whiskers are called outliers.

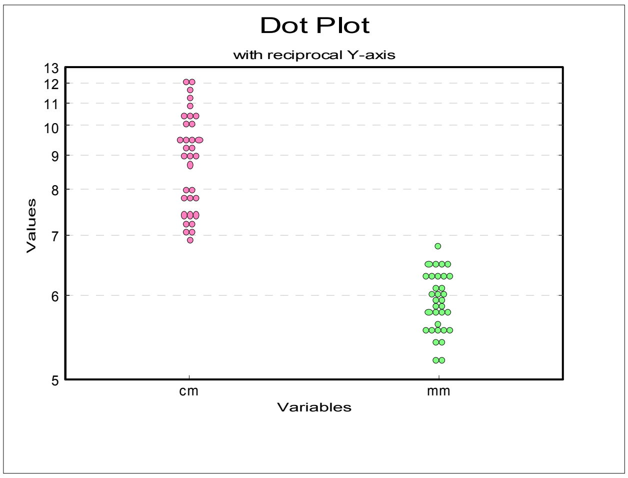
None: No whiskers and outliers are plotted.

Tukey: This is he default method. The lower whisker corresponds to the maximum of (i) lower quartile minus 1.5 times the inter-quartile range and (ii) the minimum observation and the upper whisker to the minimum of (i) upper quartile plus 1.5 times the inter-quartile range and (ii) the maximum observation.

Min / Max: Whiskers correspond to the minimum and maximum of data series.

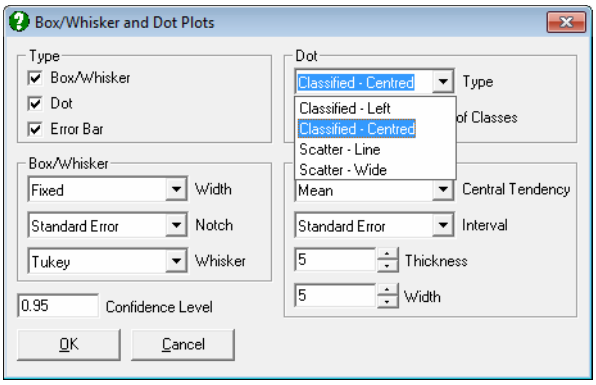
Quantiles: Whiskers correspond to the lower and upper 95% quantiles by default. The significance level can be changed by the user.

5.3.1.2. Dot Plot



The second frame contains controls for dot plots.

Type: The dots can be plotted in four different ways. The first two options will classify the observations into a specified number of classes, like in a histogram. The latter two options will plot the dots at their actual values, rather than classifying them into groups.


Box/Whisker and Dot Plots

Type
☑ Box/Whisker
☑ Dot
☑ Error Bar

Box/Whisker
Fixed Width
Standard Error Notch
Tukey Whisker

0.95 Confidence Level

Dot
Classified - Centred Type
Classified - Left
Classified - Centred
Scatter - Line
Scatter - Wide
of Classes
Mean Central Tendency
Standard Error Interval
5 Thickness
5 Width

Classified - left: Observations will be classified into groups and the dots will be left-justified.

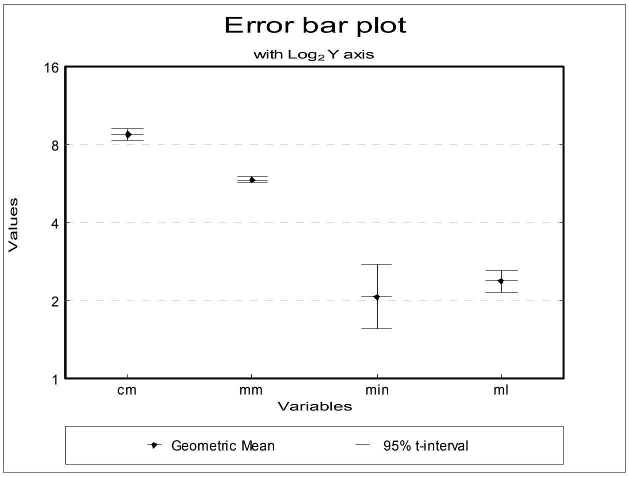
Classified - centred: Observations will be classified into groups and the dots will be centred.

Scatter - line: The actual values of observations will be plotted along a vertical line.

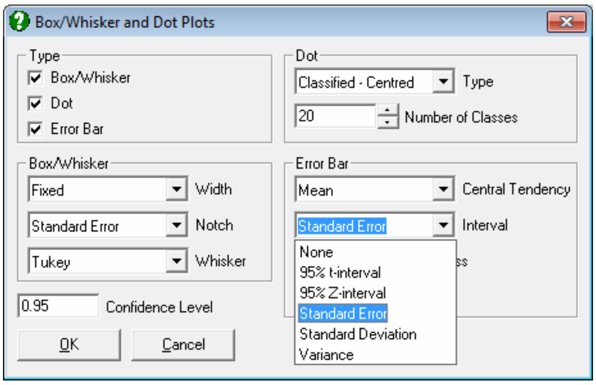
Scatter - wide: The actual values of observations will be plotted and the overlapping dots will be separated as much as possible.

Number of Classes: The classified dot plots are essentially histograms and this parameter controls the number of classes (the default is 20). The size of dots can be adjusted from the Edit → Data Series → Symbol panel to obtain the desired appearance.

5.3.1.3. Error Bar Plot



Central Tendency and Confidence Interval: The following central tendency measures and their confidence limits can be drawn.



Box/Whisker and Dot Plots

Type
☑ Box/Whisker
☑ Dot
☑ Error Bar

Dot
Classified - Centred ▾ Type
20 Number of Classes

Box/Whisker
Fixed ▾ Width
Standard Error ▾ Notch
Tukey ▾ Whisker

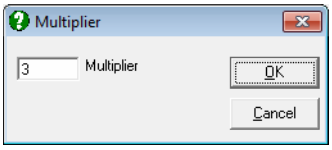
Error Bar
Mean ▾ Central Tendency
Standard Error ▾ Interval

None
95% t-interval
95% Z-interval
Standard Error
Standard Deviation
Variance

0.95 Confidence Level

- Mean
    - t-interval
    - Z-interval
    - Standard Error
    - Standard Deviation
    - Variance
- Geometric Mean
    - t-interval
    - Z-interval
- Harmonic Mean
    - t-interval
    - Z-interval
- Median
    - Quartiles
    - 95% Quantile
    - Robust Confidence Interval

When Central Tendency is Mean and one of Standard Error or Standard Deviation options is selected, a dialogue pops up asking for a multiplier.



Multiplier

3 Multiplier

Error bars for standard error will then be calculated as:

$$\text{Lower limit} = \text{Mean} - \text{k x SE}$$
$$\text{Upper limit} = \text{Mean} + \text{k x SE}$$

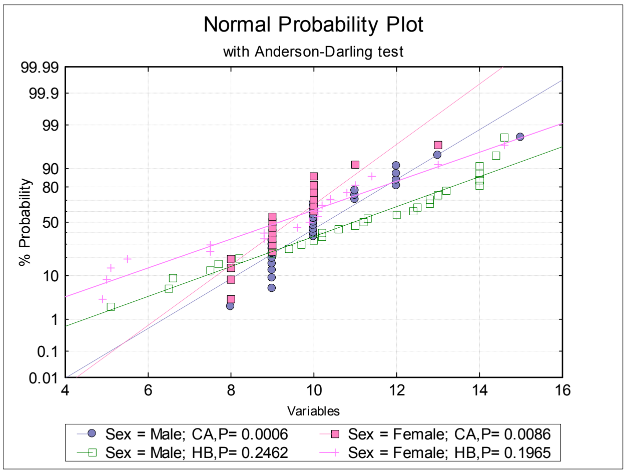
and for standard deviation:

$$\text{Lower limit} = \text{Mean} - \text{k x s}$$
$$\text{Upper limit} = \text{Mean} + \text{k x s}$$

where k is the multiplier defined by the user.

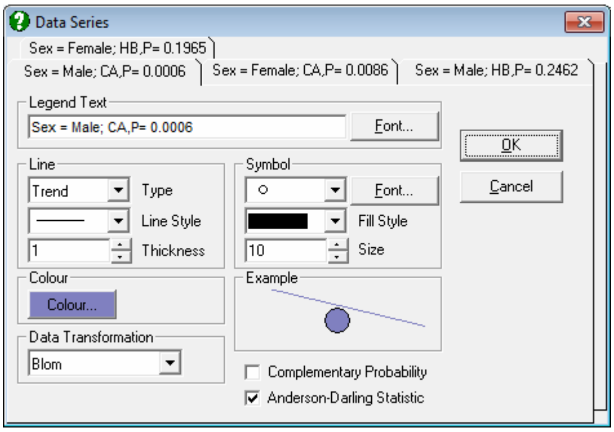
5.3.2. Normal Probability Plot

Multisample data can be entered in the form of multiple columns or data columns classified by factor columns. If at least one factor is selected, then a further dialogue will pop up asking for the combination of factor levels to be included.



If the data lies on a near-straight line, then it is said to conform to the normal distribution. By default, an Anderson-Darling Test of normality is also performed and its tail probability is reported in the legend. Smaller p-values indicate non-normality.

Edit → Data Series dialogue allows connecting data points with lines or drawing a line of best fit with or without confidence intervals. It is possible to plot probabilities or complementary probabilities.



Sex = Female; HB,P= 0.1965
Sex = Male; CA,P= 0.0006
Sex = Female; CA,P= 0.0086
Sex = Male; HB,P= 0.2462

Legend Text: Sex = Male; CA,P= 0.0006

Line
Type: Trend
Line Style: ———
Thickness: 1

Symbol: ○
Fill Style: ■
Size: 10

Colour

Data Transformation: Blom

☐ Complementary Probability
☑ Anderson-Darling Statistic

Data itself is plotted on the X-axis with all scaling options available (see Scale Type) and the corresponding Y-axis (expected normal probability) values are computed from the inverse normal cdf employing a scale transformation and plotted with a probit scale. The following approximations to the normal scores are supported (see Blom, G. 1958), where Blom transformation is the default:

Blom scores:

$$X_i = \frac{i - 3/8}{n + 1/4}, i = 1, \ldots, n.$$

Tukey scores:

$$X_i = \frac{i - 1/3}{n + 1/3}, i = 1, \ldots, n.$$

Van der Waerden scores:

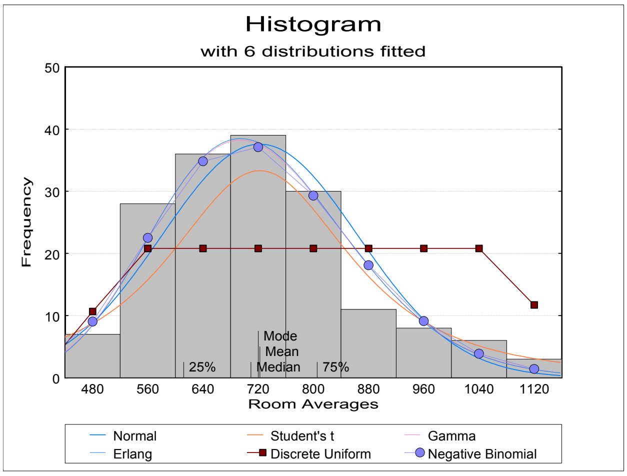
$$X_i = \frac{i}{n + 1}, i = 1, \ldots, n.$$

Note that due to these transformations a Normal Probability Plot is different from X-Y Plots with a probit axis.

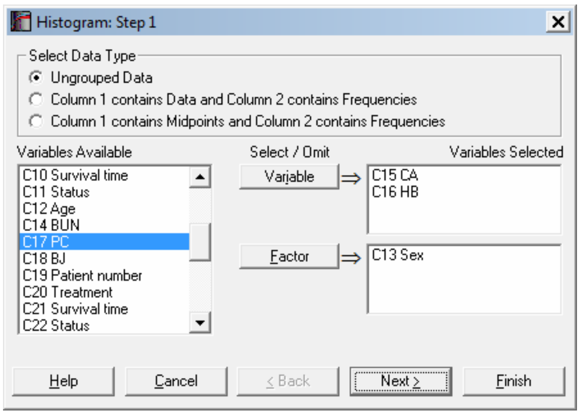
5.3.3. Histogram

Histograms can be drawn with regular or irregular class intervals and with mean, median, mode, lower and upper quartile values displayed. It is possible to fit up to six Distribution Functions simultaneously from a total of 19 continuous and discrete distributions. Frequency distributions and goodness of fit tests are displayed for fitted functions.

Multisample data can be entered in the form of multiple columns or data columns classified by factor columns. If at least one factor is selected, then a further dialogue will pop up asking for the combination of factor levels to be included. The unchecked levels will be excluded from the plot. If Run a separate analysis for each option selected is checked, a separate output will be generated for each factor level. Otherwise, one histogram will be drawn with the included factor levels.



This procedure allows choice of ungrouped data, data with frequency counts or grouped data (see 5.0.2. One-Sample Data Types). It is possible to draw frequency and cumulative histograms for string variables and histograms with irregular class widths.



Histogram: Step 1

Select Data Type

● Ungrouped Data

○ Column 1 contains Data and Column 2 contains Frequencies

○ Column 1 contains Midpoints and Column 2 contains Frequencies

Variables Available

C10 Survival time
C11 Status
C12 Age
C14 BUN
C17 PC
C18 BJ
C19 Patient number
C20 Treatment
C21 Survival time
C22 Status

Variables Selected

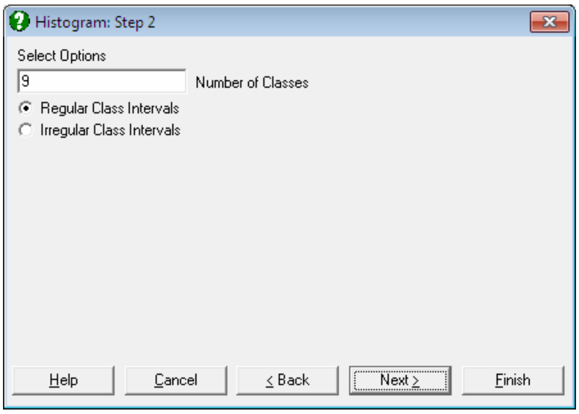
C15 CA
C16 HB

C13 Sex

You can choose to display on the X-axis either the midpoints or the lower and upper limits for each class using the Edit → Bars dialogue.

5.3.3.1. Regular and Irregular Class Intervals

A further dialogue will allow you to edit the number of classes suggested by the program and choose between regular and irregular class intervals. At this stage, the program would already have calculated the default values for the lower and upper bounds and the class interval.



Histogram: Step 2

Select Options

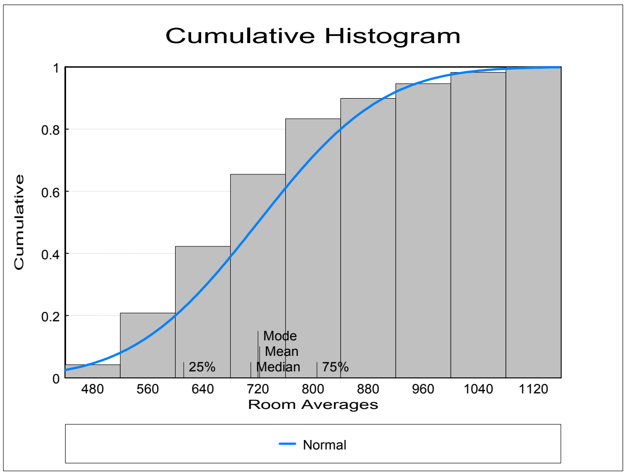
9 Number of Classes

● Regular Class Intervals

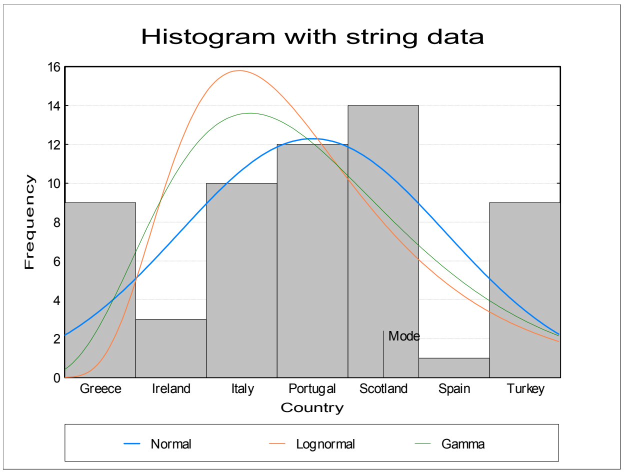
○ Irregular Class Intervals

Regular Class Intervals: If this (default) option is selected then the program will proceed with drawing the graph. The lower and upper bound and the class interval values can be edited subsequently, by opening the Edit → Axes dialogue. If the lower limit is higher than the minimum observation or the upper limit is lower than the maximum observation or more than 200 classes are generated, then a warning will be issued. In such cases the program will still proceed with plotting a histogram. If a wider class interval is entered, then the program will not rescale the Y-axis to cater for higher bars. This can be done manually.

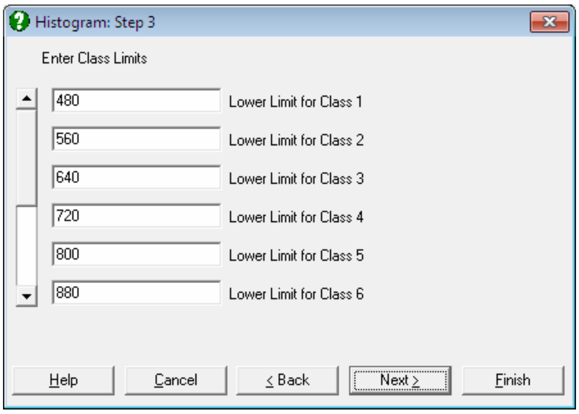
For further details of constructing regular class intervals see 5.1.5. Frequency Distributions.



If a string variable is selected, the class intervals will be regular and fixed.



Irregular Class Intervals: If this option is selected then the program will open a new dialogue to allow you to edit the suggested class intervals.



Histogram: Step 3

Enter Class Limits

480 Lower Limit for Class 1

560 Lower Limit for Class 2

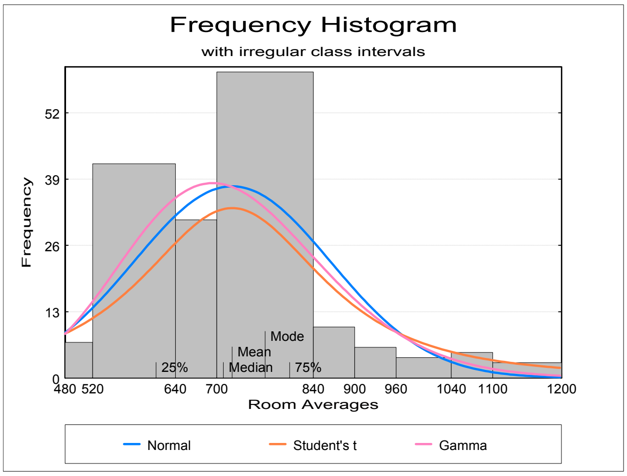
640 Lower Limit for Class 3

720 Lower Limit for Class 4

800 Lower Limit for Class 5

880 Lower Limit for Class 6

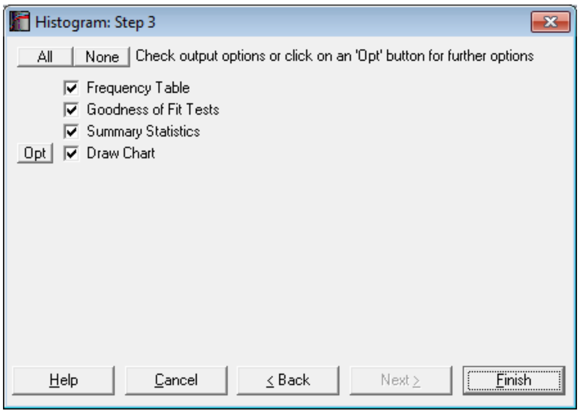
The dialogue contains a vertical scroll bar to edit up to 200 lower limits and the upper limit for the last class. Changes to the number of classes should be made before entering these values. The program will not proceed until a valid selection is made for all classes.



5.3.3.2. Histogram Output Options

The text output from this procedure includes three tables for observed and fitted frequencies, fitted distribution parameters (if any), goodness of fit tests and summary statistics. For the calculation of chi-square statistic and its degrees of freedom see 6.3.1.2. Two Sample Chi-Square Test.

As in other output options (see 2.1.5. Output Options Dialogue), when you click on the [Finish] button, the summary information and the histogram will be sent to the Output Medium with default options. If you want to edit the properties of the histogram, add or remove distribution functions, you can send it to Graphics Editor by clicking on the [Opt] button situated to the left of the Draw Chart check box.

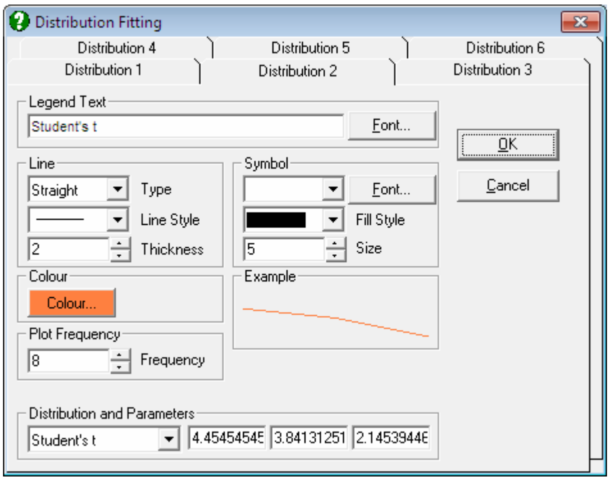

Histogram: Step 3

Check output options or click on an 'Opt' button for further options

☑ Frequency Table
☑ Goodness of Fit Tests
☑ Summary Statistics
☑ Draw Chart

5.3.3.3. Fitting Distribution Functions

When a histogram is displayed with default options, the program will already have fitted eighteen Distribution Functions (except for the negative binomial distribution) on the data. Any six of these can be displayed simultaneously by selecting the Edit → Distributions dialogue. The type, parameters and appearance of these Distribution Functions can be controlled by the user.



Legend Text: Student's t

Line — Type: Straight; Line Style: ———; Thickness: 2

Symbol — Fill Style: ■; Size: 5

Plot Frequency — Frequency: 8

Distribution and Parameters: Student's t | 4.45454545 | 3.84131251 | 2.1453944E

The Edit → Distributions dialogue features a Distribution and Parameters group at the bottom containing a drop-down list for all distributions supported. When a distribution is selected from this list, up to three more text fields are displayed immediately to the right of the list. These fields contain the estimated parameters for each distribution function (see Appendix). For instance, while for the normal distribution two fields will display the estimated mean and standard deviation, for t-distribution a third field will display the estimated degrees of freedom. A parameter which cannot be estimated is assigned the value -99. You can edit the values in each parameter field. For each distribution function you can also select the line style, thickness, colour, symbols, etc.

Any combination of continuous and discrete distribution functions can be selected for up to six distributions. The same distribution can be selected more than once. This may be useful for displaying one or more theoretical curves of the same distribution with different parameters against the fitted parameters.

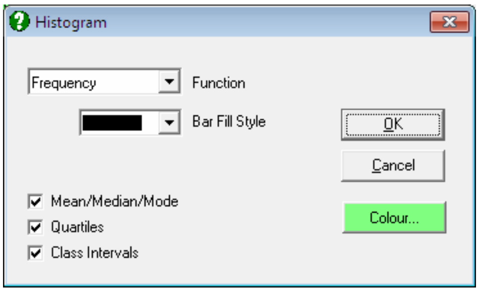
Distributions in the drop-down list are in the same order as they are in the Distribution Functions dialogue (see 5.2.1. Cumulative Probability). Hypergeometric distribution - for which the estimated frequencies procedure is not implemented - is excluded. It is also possible to plot Distribution Functions without having to fit them on a frequency histogram by means of the Plot of Distribution Functions procedure.

Colour: This controls the colour of the fitted curves.

Symbol: The usual symbol selection group can be used to display Symbols for discrete distributions. When a selection is made other than None for a discrete distribution, a symbol will be drawn on the line at each distinct value of the X-axis variable.

Plot Frequency: This control determines the resolution of fitted distribution curves. The default value of 10 means that the functions will be evaluated at every 10th pixel. This field can have a minimum value of 1, in which case the functions will be evaluated at every pixel. This will take 10 times longer to compute and it may be more difficult to distinguish various curves.

5.3.3.4. Bars



Histogram

Function: Frequency

Bar Fill Style: ■

☑ Mean/Median/Mode

☑ Quartiles

☑ Class Intervals

This dialogue provides controls for editing aspects of the histogram bars.

Function: The available options are (i) Frequency and (ii) Cumulative. Distributions can be fitted in either case.

Bar Fill Style: Bars can be filled with solid colours or with cross-hatch patterns.

Bar Colour: This controls the colour of the histogram bars.

Mean / Median / Mode: For numeric variables, this will draw a vertical line for each statistic along the X-axis. For string variables only the mode is drawn.

Quartiles: For numeric variables, a vertical line for 25% and 75% quantiles will be drawn along the X-axis.

Class Intervals: X-axis tick marks and their corresponding value labels can be drawn either in the middle of a class, or at the lower and upper boundaries. This option is available only for histograms with regular class intervals. Irregular histograms will always display class intervals. If the selected column contains String Data, tick marks will always be drawn at midpoints.

5.3.3.5. Example

Open TIMESER and select Graph → Descriptive Plots → Histogram. Select *Room Averages* (*C1*) as [Variable], accept the program's suggestion on the next two dialogues and on the Output Options Dialogue, click [Opt] situated to the left of the Draw Chart option. The histogram will be displayed in Graphics Editor. Select Edit → Distributions (or double click at the middle of the graph) and select five distributions as Normal, Student's t, Gamma, Erlang and Negative Binomial. After the graph is updated, close the Graphics Editor and click [Finish] on the Output Options Dialogue.

Histogram

Frequency Table

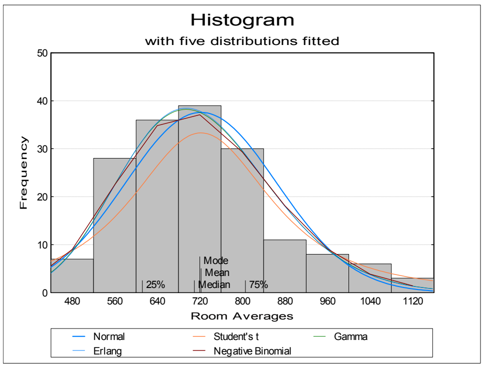
Room Averages	Observed	Normal	Student's t	Gamma	Erlang	Negative Binomial
480	7.0000	9.1004	8.8637	8.9565	8.7975	9.0346
560	28.0000	19.7499	15.9337	22.4677	22.3912	22.5061
640	36.0000	31.5472	26.3081	34.8328	34.9867	34.8212
720	39.0000	37.0937	32.6673	37.1220	37.3794	37.1007
800	30.0000	32.1072	26.8991	29.3196	29.4695	29.2992
880	11.0000	20.4574	16.4631	18.1326	18.1278	18.1035
960	8.0000	9.5939	9.1668	9.1517	9.0733	9.1170
1040	6.0000	3.3110	5.2059	3.8914	3.8164	3.8623
1120	3.0000	0.8407	3.1213	1.4293	1.3836	1.4110
Total	168.0000	163.8012	144.6289	165.3035	165.4255	165.2557

Goodness of Fit

	Parameter 1	Parameter 2	Parameter 3	Chi-Square Statistic	DoF	Right-Tail Probability
Normal	722.2976	142.6569		16.6294	6	0.0107
Student's t	722.2976	142.6569	2.0001	11.1877	5	0.0478
Gamma	25.6358	0.0355		7.5912	6	0.2696
Erlang	26.0000	0.0360		7.7877	6	0.2541
Negative Binomial	722.2976	26.5791		7.6770	6	0.2627

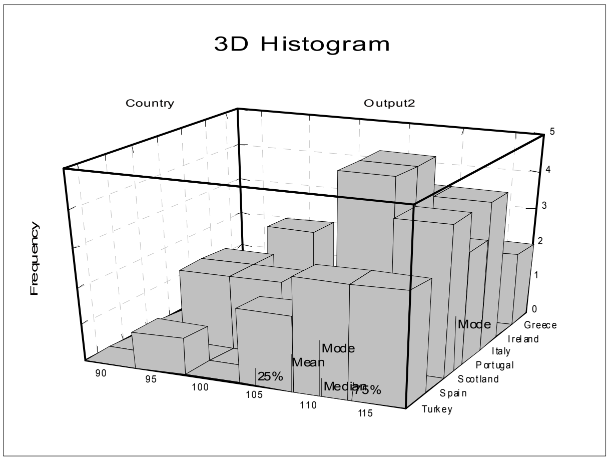
Descriptive Statistics

	Room Averages
Mean	722.2976
Median	709.5000
Mode	720.0000
Lower Quartile	612.0000
Upper Quartile	805.5000

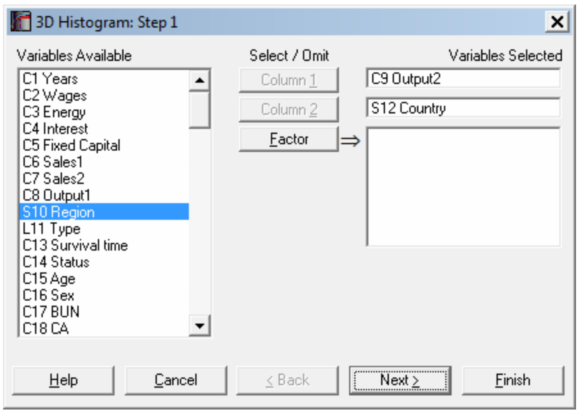


5.3.4. 3D Histogram

A 3D Histogram can be drawn by selecting two numeric or string variables as [Column 1] and [Column 2]. Optionally, one or more factors can be selected, in which case the program will display a list of all possible combinations of factor levels. The unchecked levels will be excluded from the plot. If Run a separate analysis for each option selected is checked, a separate output will be generated for each factor level. Otherwise, one histogram is drawn with the included factor levels.



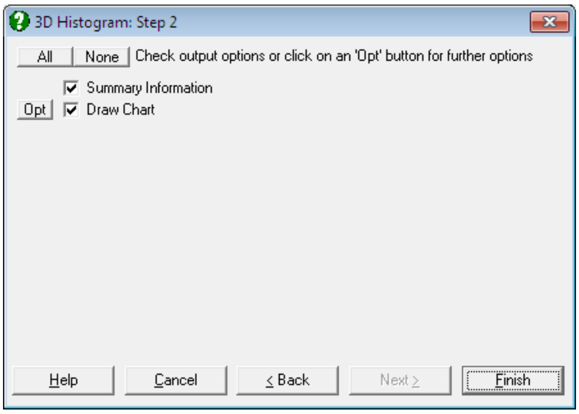
Observations falling within each class of X-variable are further classified according to classes of the Y-variable (or vice versa). Mean, median, mode and quartiles can be displayed and class intervals edited for axes displaying numeric variables.


3D Histogram: Step 1

Variables Available: C1 Years, C2 Wages, C3 Energy, C4 Interest, C5 Fixed Capital, C6 Sales1, C7 Sales2, C8 Output1, S10 Region, L11 Type, C13 Survival time, C14 Status, C15 Age, C16 Sex, C17 BUN, C18 CA

Variables Selected: Column 1: C9 Output2; Column 2: S12 Country; Factor: (empty)

The Summary Information option will produce two tables for frequencies and statistics. As in other output options, when you click on the [Finish] button, the summary information and the histogram will be sent to the Output Medium with default options. If you want to edit the properties of the histogram, then you can send it to Graphics Editor by clicking on the [Opt] button situated to the left of the Draw Chart check box.


3D Histogram: Step 2

Check output options or click on an 'Opt' button for further options

☑ Summary Information

☑ Draw Chart

The mode displayed on each axis is the mode for that particular variable, but not for the 3D distribution of frequencies. Therefore, the highest bar on the graph should not necessarily correspond to any mode values displayed.

The Edit dialogues for this procedure are similar to those of 2D Histogram procedure, except there is no possibility here to fit Distribution Functions on 3D Histograms.

Example

Open PARTESTS and select Graph → Descriptive Plots → 3D Histogram. Select *Haemoglobin* (*C10*) as [Column 1] and *Platelets* (*C11*) as [Column 2]. In the following output titles of the graph have been edited manually.

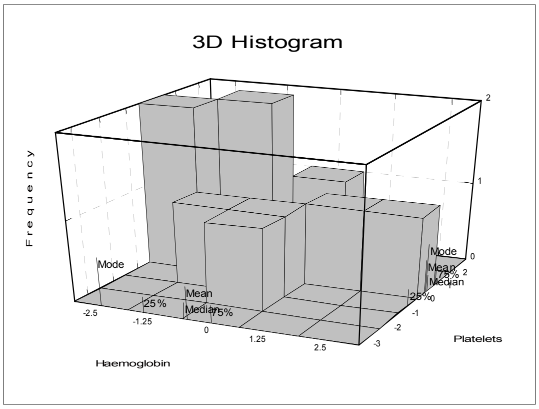
3D Histogram

Frequency Table

Platelets \ Haemoglobin	-2.5	-1.25	0	1.25	2.5	Total
-3	0	0	0	0	0	0
-2	0	0	1	0	0	1
-1	0	1	1	0	0	2
0	2	0	0	1	1	4
1	0	2	0	0	0	2
2	0	0	1	0	0	1
Total	2	3	3	1	1	10

Descriptive Statistics

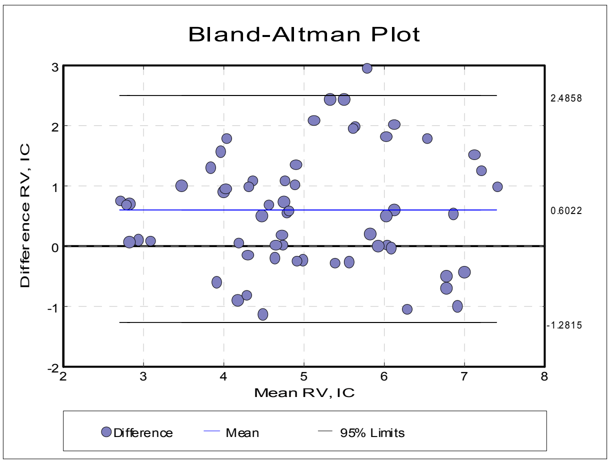
	Haemoglobin	Platelets
Mean	-0.5300	-0.0300
Median	-0.6000	0.1000
Mode	-2.5000	0.0000
Lower Quartile	-1.5000	-1.0000
Upper Quartile	0.0000	0.6000



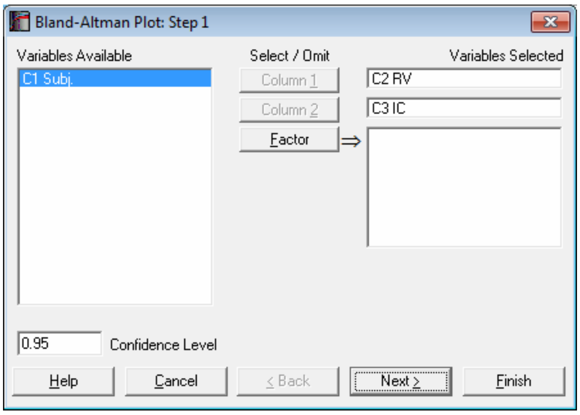
5.3.5. Bland-Altman Plot

The Bland-Altman plot (which is also known as Difference Plot or Tukey Mean Difference Plot) aims to show whether the difference between two methods is significant.

In its simplest form, the differences between observation pairs are plotted against their mean and the mean difference and its 95% confidence limit lines are drawn on the same plot.



5.3.5.1. One Measurement Per Subject



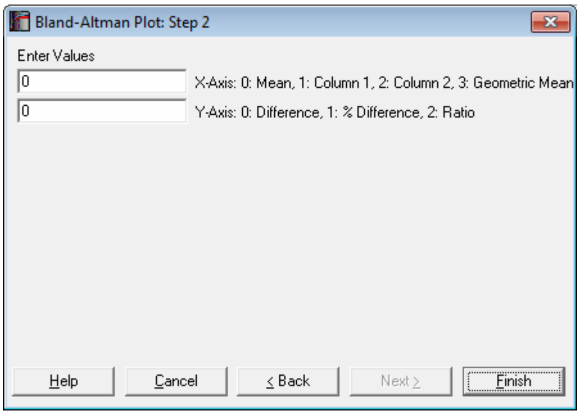
C1 Subj.

C2 RV

C3 IC

0.95 Confidence Level

Select two columns of data using the [Column 1] and [Column 2] buttons on the Variable Selection Dialogue. The next dialogue offers the following selections for axes:



Bland-Altman Plot: Step 2

Enter Values

0 X-Axis: 0: Mean, 1: Column 1, 2: Column 2, 3: Geometric Mean

0 Y-Axis: 0: Difference, 1: % Difference, 2: Ratio

X-Axis Options:

Mean: This is the recommended default option. The average of the two measurements is plotted.

Column 1: If the method represented in Column 1 is a proven benchmark method, you can plot the differences against this method, rather than the average of the two methods.

Column 2: If the method represented in Column 2 is a proven benchmark method, you can plot the differences against this method, rather than the average of the two methods.

Geometric Mean: This is mostly used when the Y-Axis represents ratios. The ratio / geometric mean options are useful when the original data was subject to a logarithmic transformation and the results are to be transformed back to the original scale.

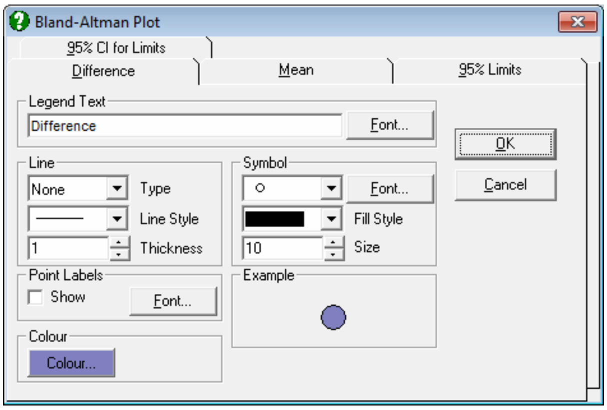
Y-Axis Options:

Difference: This is the recommended default option. The difference between the two measurements is plotted.

% Difference: 100 x *Difference* / *Mean* is plotted for each pair.

Ratio: This is mostly used when the X-Axis represents geometric mean. The ratio / geometric mean options are useful when the original data was subject to a logarithmic transformation and the results are to be transformed back to the original scale.

By default, the 95% confidence intervals are displayed for the mean and 95% confidence lines. For further information see Bland & Altman (1999). You can omit these lines by selecting None for the Line Type on the Data Series dialogue that can be accessed either from the Edit → Data Series menu or by double-clicking on the graph area.



Bland-Altman Plot

95% CI for Limits | Difference | Mean | 95% Limits

Legend Text: Difference Font...

Line
None Type
Line Style
1 Thickness

Symbol
○ Font...
Fill Style
10 Size

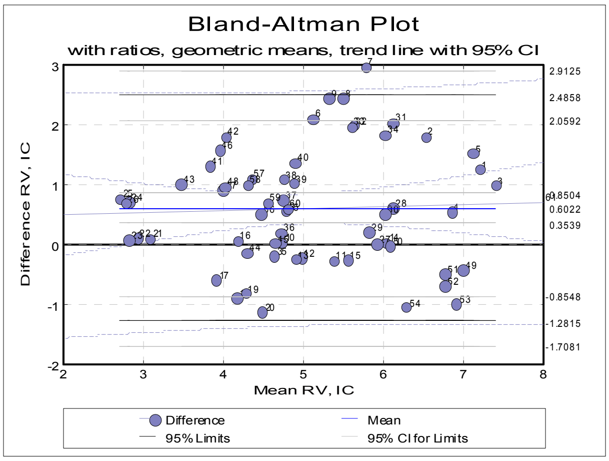
Point Labels
☐ Show Font...

Example

Colour
Colour...

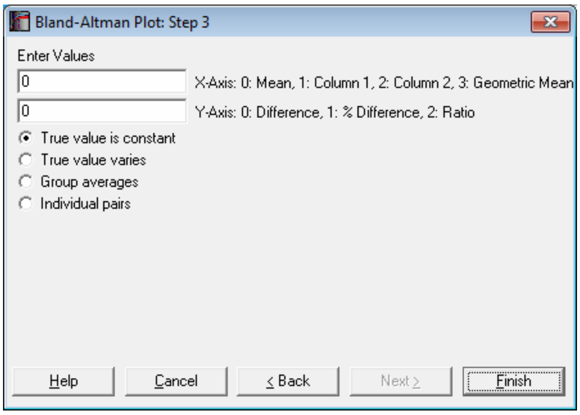
OK
Cancel

The standard X-Y Plots utilities like fitting a trend line or labelling the individual points are also available in this procedure.



5.3.5.2. Repeated Measurements Per Subject

If the data contains more than one observation per subject, it is divided into subgroups as described in section 7.3.0.1. ANOVA and GLM Data Format. In this case the program will display a list of subgroups as defined by the factor column, allowing you to choose which subgroups to include in the analysis. The unchecked levels will be excluded from the plot. The groups may be defined by more than one factor variable, in which case a list of all possible combinations of factor levels will be displayed.


Bland-Altman Plot: Step 3

Enter Values

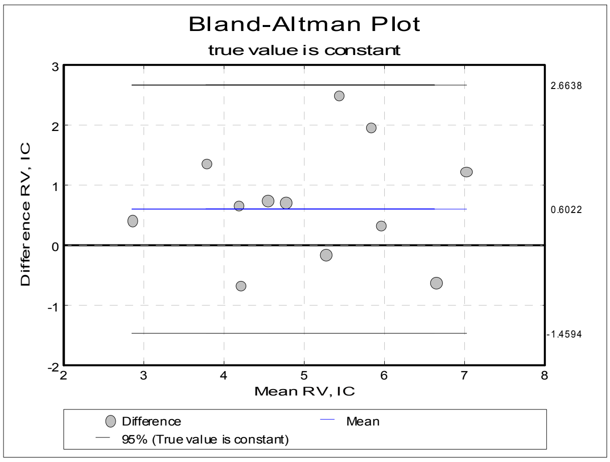
0 X-Axis: 0: Mean, 1: Column 1, 2: Column 2, 3: Geometric Mean

0 Y-Axis: 0: Difference, 1: % Difference, 2: Ratio

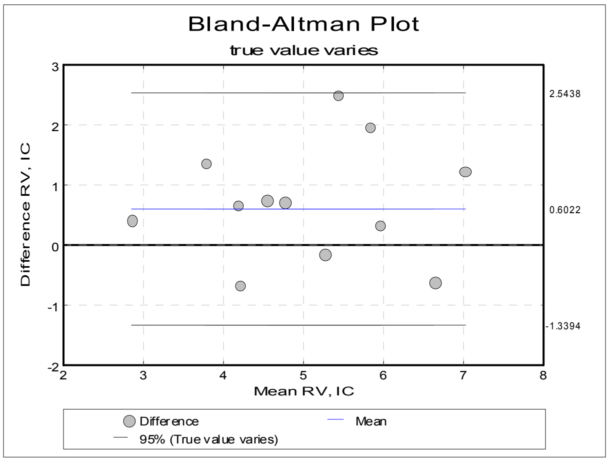
● True value is constant
○ True value varies
○ Group averages
○ Individual pairs

When there are repeated measures per subject, the confidence intervals can be computed by three alternative methods. For further information see Bland & Altman (2007). In this case one symbol is plotted for each subgroup whose area is proportional to the size of that subsample. There is a fourth option which disregards the repeated measurements and plots the individual pairs as in the previous section.

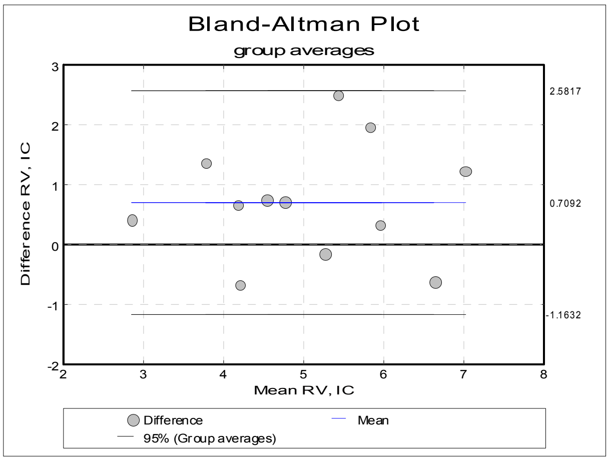
True value is constant: This is the recommended default option. Measurement errors are calculated using one-way analysis of variance for each method separately as described by Bland & Altman (2007) page 578.



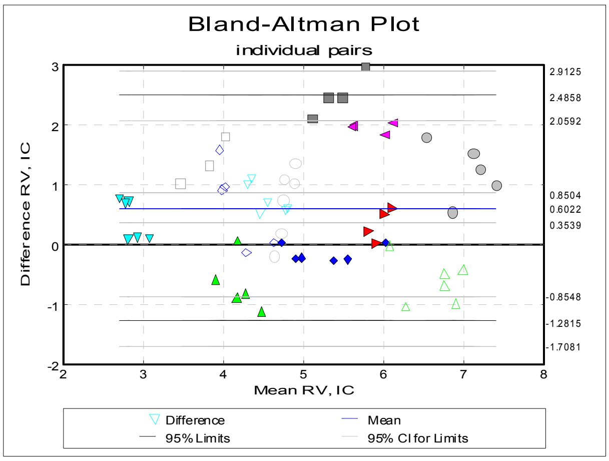
True value varies: Measurement errors are calculated using a one-way analysis of variance for differences as introduced by Bland & Altman (1999) and corrected in (2007) page 576.



Group averages: Measurements are pooled for each subject and first and then plotted as individual pairs as described by Bland & Altman (2007) page 581. Note that the mean is different from the above two options.



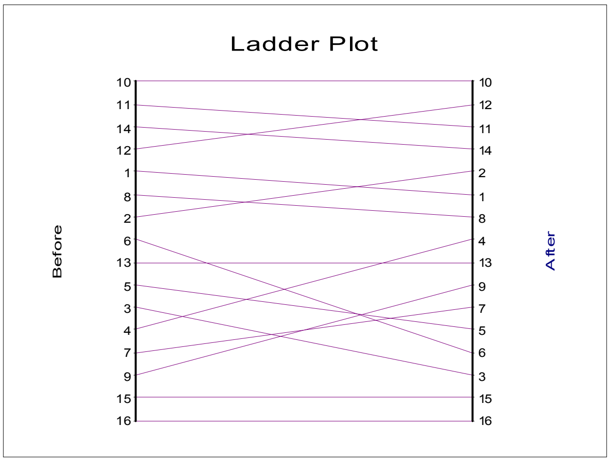
Individual pairs: The individual pairs are plotted as described in the previous section and the division into subgroups is disregarded in all computations, except that they are represented by different symbols on the graph.



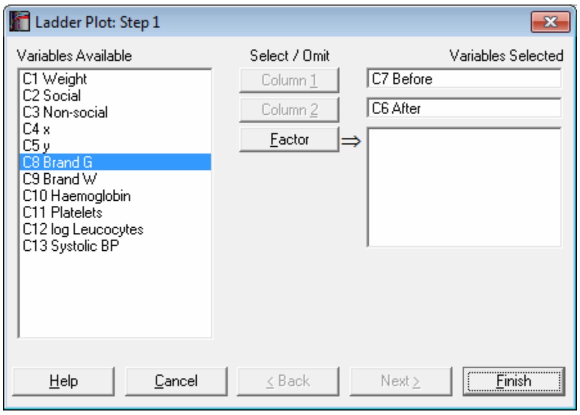
Examples

We have already reproduced above all examples in Bland & Altman (2007). The data set given in Table 1, page 572 can be found in the file BLAND-ALTMAN in UNISTAT's Examples folder. Select Graph → Descriptive Plots → Bland-Altman Plot to run the procedure.

5.3.6. Ladder Plot



The ladder plot shows the relationship between the ranks of data in two columns. This plot is used (usually) to visualise the effects of a treatment on the same set of subjects *before* and *after* a treatment. Select the two columns using [Column 1] and [Column 2] from the Variable Selection Dialogue.


Ladder Plot: Step 1

Variables Available

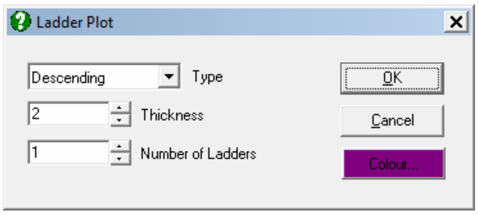
C1 Weight
C2 Social
C3 Non-social
C4 x
C5 y
C8 Brand G
C9 Brand W
C10 Haemoglobin
C11 Platelets
C12 log Leucocytes
C13 Systolic BP

Variables Selected

C7 Before
C6 After

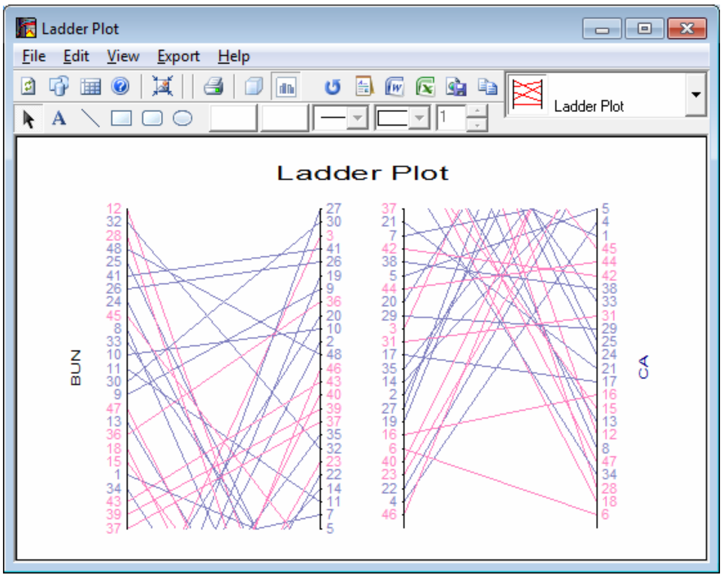
Optionally, one or more factors can be selected, in which case the program will display a list of all possible combinations of factor levels. The unchecked levels will be excluded from the plot and the remaining levels will be distinguished by different colours.

Each bar on the ladder represents a row of data. The ranks of data in each row are plotted on the left and on the right of the ladder for columns 1 and 2 respectively. The same numbers on either side are then connected by a line. Therefore, if a row has the same rank in both columns, its bar will be horizontal.



Type: Descending
Thickness: 2
Number of Ladders: 1

The plot can be split into a number of ladders laid out side by side instead of one single ladder. This is useful for plotting columns with a large number of rows. The ranking can be in Ascending or Descending order. All these options can be controlled from Edit → Ladders dialogue.



The Chart Gallery (i.e. the drop-down list of graphics options displayed on the top-right of the Graphics Editor window) provides quick access to the other two paired data plots, 3D Histogram and Bland-Altman Plot.

UNISTAT Statistical Package

Chapter 6
Statistical Tests, Correlations and Tables

6.0. Overview

A wide range of parametric and nonparametric tests, correlation coefficients and statistical tables are covered in this chapter. Most test statistics are displayed with their probability values and, where relevant, with confidence intervals. Output also includes summary statistics about the columns or groups of data used in the test. Contingency Table, Cross-Tabulation and Break-Down analyses can be performed with an unlimited number of factors.

In this chapter of the User's Guide formulas used in tests will be given where possible. The following notation:

n_1, ..., n_k, M_1, ..., M_k, s_1, ..., s_k

will stand for the number of observations, means and standard deviations of samples 1, ..., k respectively. In one sample tests n stands for the sample size and in tests with more than one variable n will stand for the sum of all valid cases in all samples. For instance, in a two sample test $n = n_1 + n_2$.

All tests can handle missing observations. The tests that do not require paired or matrix data will ignore only the individual missing values. Tests on paired data will omit a pair if either or both of the observations in a pair are missing. Tests that require data in matrix format will omit any rows containing one or more missing observations.

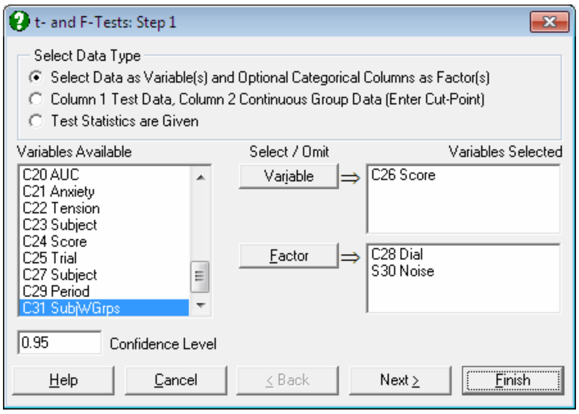
The data format for tests covered in this chapter will be one of the following types.

6.0.1. One Sample Tests

All one sample tests (t-, chi-square, Kolmogorov-Smirnov) are accessed under their two sample versions. If you wish to perform a One Sample t-Test, for example, you will need to run the t- and F-Tests procedure, where you will be able to select only one variable. If you select two or more variables, then two separate one sample tests will be performed on each variable, alongside a two sample test between them. Output Options Dialogues will allow you to choose which tests to appear in the output.

6.0.2. Two Sample Tests

For tests with two independent samples, a Variable Selection Dialogue containing three data type options will be displayed.



t- and F-Tests: Step 1

Select Data Type

● Select Data as Variable(s) and Optional Categorical Columns as Factor(s)
○ Column 1 Test Data, Column 2 Continuous Group Data (Enter Cut-Point)
○ Test Statistics are Given

Variables Available: C20 AUC, C21 Anxiety, C22 Tension, C23 Subject, C24 Score, C25 Trial, C27 Subject, C29 Period, C31 SubjWGrps

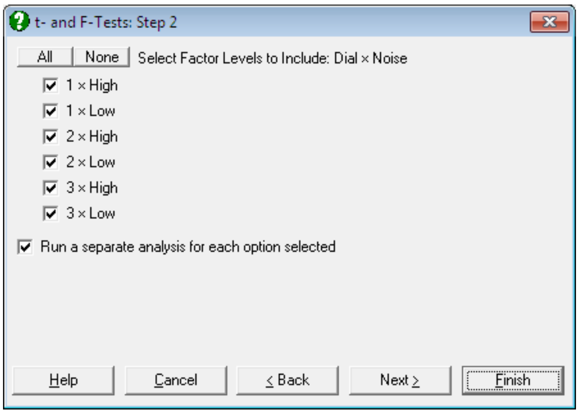
Variables Selected:
- Variable ⇒ C26 Score
- Factor ⇒ C28 Dial, S30 Noise

0.95 Confidence Level

1) **Select Data as Variable(s) and Optional Categorical Columns as Factor(s):** Data in spreadsheet columns or categorical data can be analysed under this option.

   a) **Samples are in separate columns:** If no [Factor]s and at least two [Variable]s are selected, then the program will perform the test on all possible pairs of columns, which do not need to have the same length.

   b) **At least one factor and one variable are selected:** The optional [Factor] columns can contain numeric or String Data, but should have a limited number of distinct values. In this case, a further dialogue will pop up, displaying a check list of all combinations of levels in the selected factors. There will also be a check box Run a separate analysis for each option selected, which is used to determine whether the variables or factors will have the priority in the display of output.



Select Factor Levels to Include: Dial × Noise

☑ 1 × High
☑ 1 × Low
☑ 2 × High
☑ 2 × Low
☑ 3 × High
☑ 3 × Low

☑ Run a separate analysis for each option selected

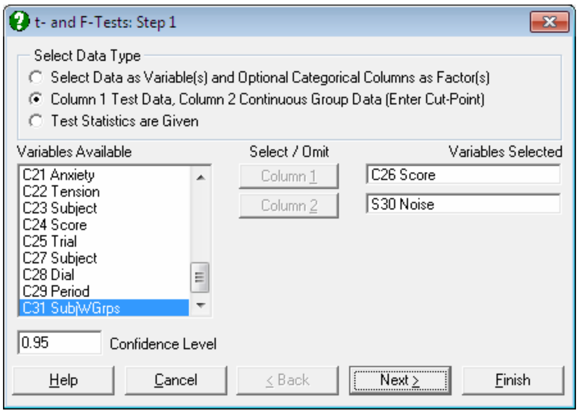
At least one factor and two variables are selected and Run a separate analysis for each option selected **is checked:** The program will perform tests on all possible pairs of variables, and the tests will be repeated for each level (or combinations of levels) checked in the second dialogue. The selected columns should have the same length.

For instance, if two variables and one factor containing three levels are selected, three check boxes will be displayed representing each level. The test will be performed between the two variables, for only those rows containing the selected level of the factor column (the inner loop). There will be as many tests as the number of levels selected (the outer loop).

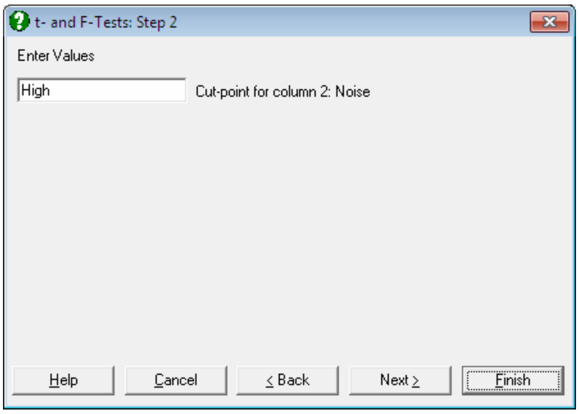
At least one factor and one variable are selected and Run a separate analysis for each option selected **is unchecked:** The program will perform tests on all possible pairs of factor levels (or combinations of levels) and the tests will be repeated for each variable. The selected columns should have the same length.

For instance, if one variable and one factor containing three levels is selected and all are checked, then the test will be performed three times (n(n - 1)/2) for each possible pair of factor levels, 1 - 2, 1 - 3 and 2 - 3. If there are two factors selected, say one having two levels and the other three, then the list will contain six check boxes, 1 x 1, 1 x 2, 1 x 3, 2 x 1, 2 x 2, 2 x 3. Suppose only the boxes 1 x 2 and 2 x 2 are checked. Then the test will be performed between those rows of the data variable containing 1 in the first factor column and 2 in the second versus those contain 2 in the first and 2 in the second. If more than one variables are selected, the same factor selections (the inner loop) will apply to all variables (the outer loop).

2) **Column 1 Test Data, Column 2 Group Data (Enter Cut-Point):** It is assumed that the first column contains a mixture of two samples and a second column contains the grouping criterion. The second column may contain continuous numeric, categorical or String Data.

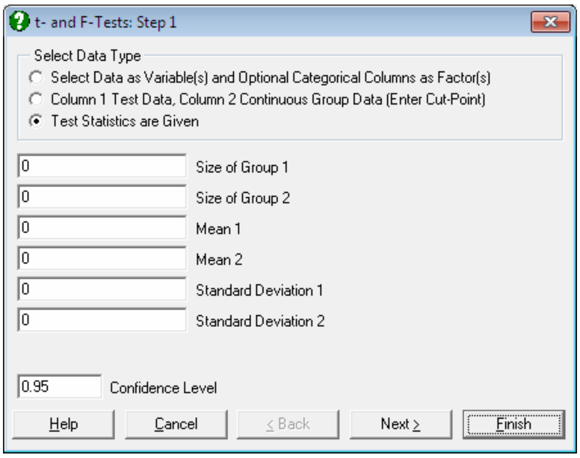


The next dialogue asks for a cut-point, i.e. the number which will separate the criterion column into two groups. All observations less than the cut-point (i.e. excluding it) will be considered as sample 1 and all those greater than or equal to the cut-point (i.e. including it) will be considered as sample 2, without actually replacing the numbers in the second column. The string variables are separated according to their lexicographic ordering.



This information will then be used to separate the corresponding values of the first column into two groups. For any grouping criterion more complex than this, use Data Processor’s Data → Recode Column procedure to convert continuous data columns into factor columns. The number of cases falling within each sample and their basic statistics (such as mean, standard deviation) can be displayed as part of output.

3) **Test Statistics are Given:** This data option is particularly useful when the raw data is not available but its parameters are known. You are expected to enter all parameters necessary for the test. For example, in t- and F-Tests, the relevant parameters are size, mean and standard deviation for the two samples.



t- and F-Tests: Step 1

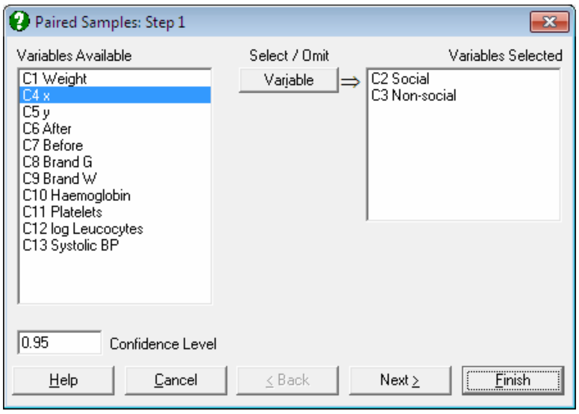
Select Data Type

○ Select Data as Variable(s) and Optional Categorical Columns as Factor(s)
○ Column 1 Test Data, Column 2 Continuous Group Data (Enter Cut-Point)
● Test Statistics are Given

0	Size of Group 1
0	Size of Group 2
0	Mean 1
0	Mean 2
0	Standard Deviation 1
0	Standard Deviation 2
0.95	Confidence Level

6.0.3. Tests with Paired Data

Any number of columns can be selected by clicking on [Variable]. The program will perform the test on all possible pairs of columns. The test is not performed on pairs with different column lengths.



Paired Samples: Step 1

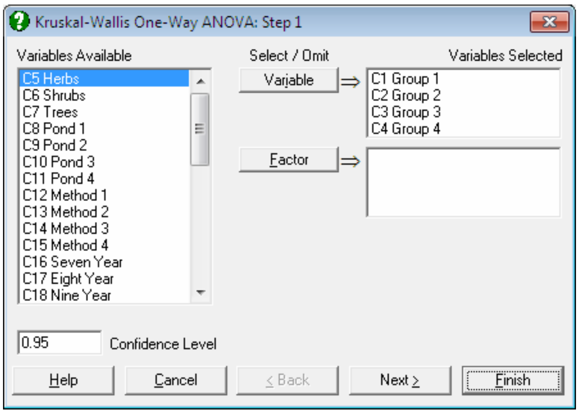
Variables Available: C1 Weight, C4 x, C5 y, C6 After, C7 Before, C8 Brand G, C9 Brand W, C10 Haemoglobin, C11 Platelets, C12 log Leucocytes, C13 Systolic BP

Variables Selected: C2 Social, C3 Non-social

0.95 Confidence Level

6.0.4. Multisample Tests

The data selection for tests with two or more independent samples is similar to the first option of two independent sample tests (see 6.0.2. Two Sample Tests). An unlimited number of variables (not necessarily of equal length) and factors can be selected. The factors can be numeric or String Data columns, but should contain a limited number of distinct values.



Kruskal-Wallis One-Way ANOVA: Step 1

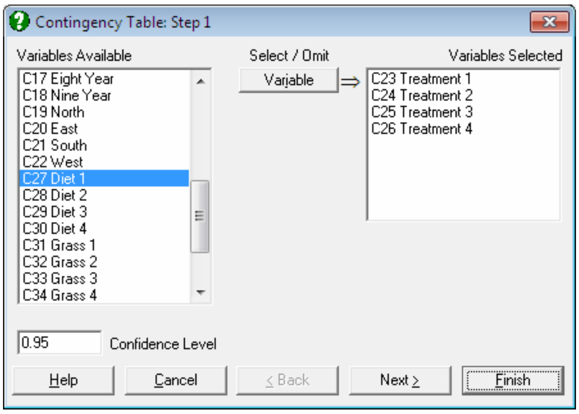
Variables Available: C5 Herbs, C6 Shrubs, C7 Trees, C8 Pond 1, C9 Pond 2, C10 Pond 3, C11 Pond 4, C12 Method 1, C13 Method 2, C14 Method 3, C15 Method 4, C16 Seven Year, C17 Eight Year, C18 Nine Year

Variables Selected: C1 Group 1, C2 Group 2, C3 Group 3, C4 Group 4

0.95 Confidence Level

6.0.5. Tests with Matrix Data

The tests in this group require two or more columns of data with equal lengths. When the Variable Selection Dialogue is displayed, highlight the desired range of columns and click on [Variable] to include them in the analysis.



Variables Selected
C23 Treatment 1
C24 Treatment 2
C25 Treatment 3
C26 Treatment 4

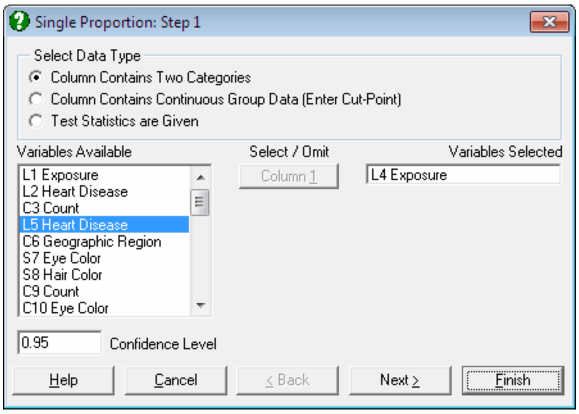
0.95 Confidence Level

6.0.6. Tests with Binary Data

The tests in this category (e.g. the Binomial Test) are accessed from the menu option Statistics 1 → Nonparametric Tests (1-2 Samples) → Binomial Proportion.

The Variable Selection Dialogue contains the following three data type options.

1) **Column Contains Two Categories:** Select one column from the Variables Available list. It can be a string, numeric, factor or continuous data column but it should contain only two distinct values (levels). The program will not proceed with the test if this requirement is not met.

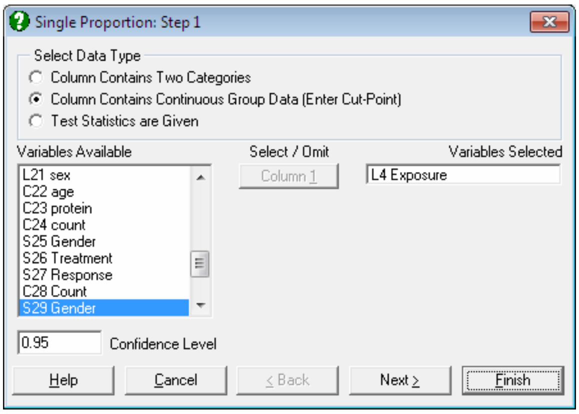


● Column Contains Two Categories
○ Column Contains Continuous Group Data (Enter Cut-Point)
○ Test Statistics are Given

Variables Selected: L4 Exposure

0.95 Confidence Level

2) **Column Contains Continuous Group Data:** One column containing string or numeric factor or continuous data can be selected. It will be split into two groups, i.e. those below and above a certain value.

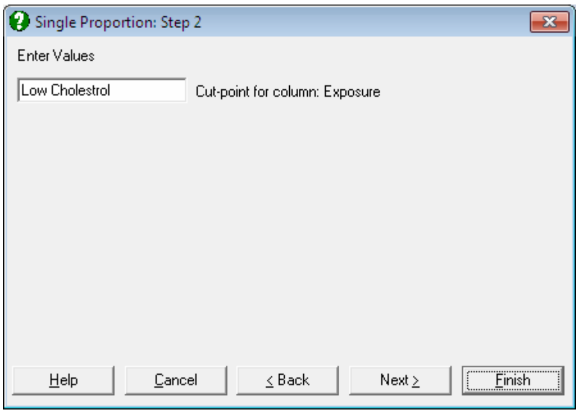


○ Column Contains Two Categories
● Column Contains Continuous Group Data (Enter Cut-Point)
○ Test Statistics are Given

Variables Selected: L4 Exposure

0.95 Confidence Level

At the next dialogue, you will be asked to enter a cut-point. All observations less than the cut-point (i.e. excluding it) will be considered as being in group 1 and all those greater than or equal to the cut-point (i.e. including it) will be considered as being in group 2. The string variables are separated according to their lexicographic ordering. The output includes the number of cases falling within each group.

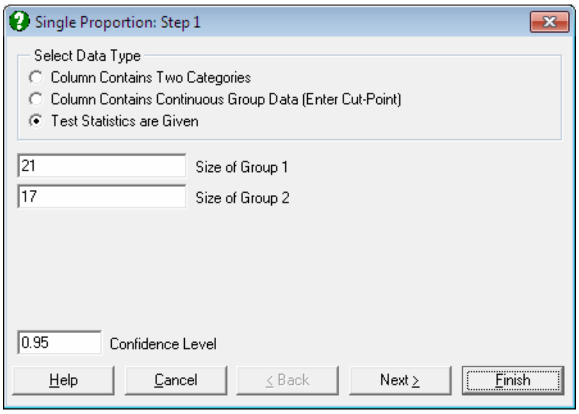


Single Proportion: Step 2

Enter Values

Low Cholestrol Cut-point for column: Exposure

3) **Frequencies are Given:** Enter size of groups 1 and 2.



Single Proportion: Step 1

Select Data Type

○ Column Contains Two Categories
○ Column Contains Continuous Group Data (Enter Cut-Point)
● Test Statistics are Given

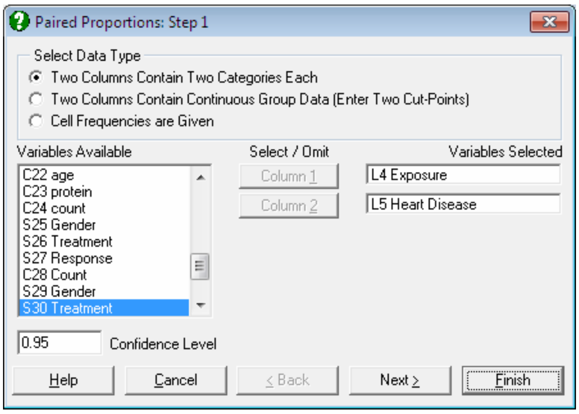
21 Size of Group 1

17 Size of Group 2

0.95 Confidence Level

6.0.7. 2 x 2 Tables

Tests for 2 x 2 tables are grouped under the menu items Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Proportions and Paired Proportions. All these tests can also be performed in Contingency Table and Cross-Tabulation procedures (see 6.6.2.3. 2 x 2 Table Statistics).



Paired Proportions: Step 1

Select Data Type

● Two Columns Contain Two Categories Each
○ Two Columns Contain Continuous Group Data (Enter Two Cut-Points)
○ Cell Frequencies are Given

Variables Available: C22 age, C23 protein, C24 count, S25 Gender, S26 Treatment, S27 Response, C28 Count, S29 Gender, S30 Treatment

Variables Selected: L4 Exposure, L5 Heart Disease

0.95 Confidence Level

The Variable Selection Dialogue for these procedures contains the following three data type options.

1) **Two Columns Contain Two Categories Each:** Choose two columns from the Variables Available list which contain only two distinct values. The program will not proceed with the test if this requirement is not met. The selected columns may contain string or numeric values.

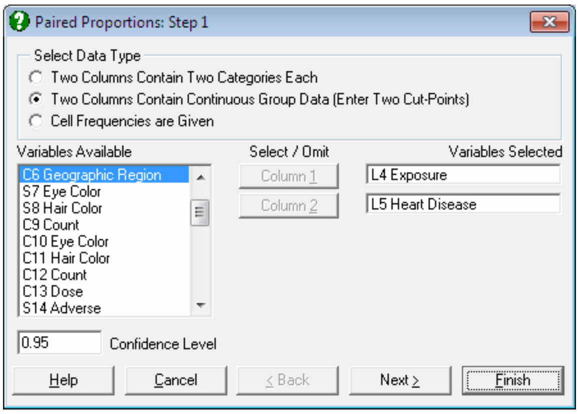
A 2 x 2 table containing four cells (i.e. (1,1), (1,2), (2,1), (2,2)) is formed for each test.

i) For Paired Proportions, the two selected columns should have the same length. The number of cases falling within each cell of the table is counted by the program. The cell (1,1) contains the number of pairs where each member is the smaller of the two values in its own column and the cell (2,2) contains the number of pairs with larger values. In other words, Paired Proportions, forms a 2 x 2 table in a way similar to that in Cross-Tabulation.

ii) For the Unpaired Proportions procedure, the selected columns may have unequal lengths. The program will determine the number of cases in each cell separately for each column, such that the cells (1,1) and (2,1) contain the number of small and large values in the first column respectively, and (1,2) and (2,2) that of small and large values in the second column. Note that this is an entirely different way of forming a 2 x 2 table compared with Paired Proportions and Cross-Tabulation.

2) **Two Columns Contain Continuous Group Data (Enter Two Cut-Points):** As in the previous data option, two columns containing string or

numeric values can be selected. However, unlike the previous option, these columns may contain any number of distinct values.



Paired Proportions: Step 1

Select Data Type

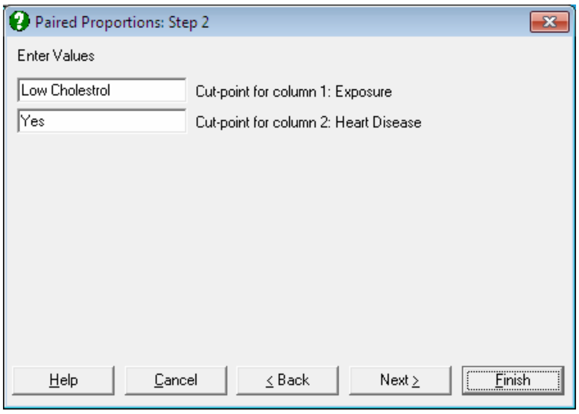
○ Two Columns Contain Two Categories Each
● Two Columns Contain Continuous Group Data (Enter Two Cut-Points)
○ Cell Frequencies are Given

Variables Available: C6 Geographic Region, S7 Eye Color, S8 Hair Color, C9 Count, C10 Eye Color, C11 Hair Color, C12 Count, C13 Dose, S14 Adverse

Variables Selected: L4 Exposure, L5 Heart Disease

0.95 Confidence Level

You will be asked to enter a cut-point for each column. These cut-points will separate each column into two groups. For each group, cases less than the cut-point (i.e. excluding it) will be considered as being in group 1 and all those greater than or equal to the cut-point (i.e. including it) will be considered as being in group 2. The string variables are separated according to their lexicographic ordering.



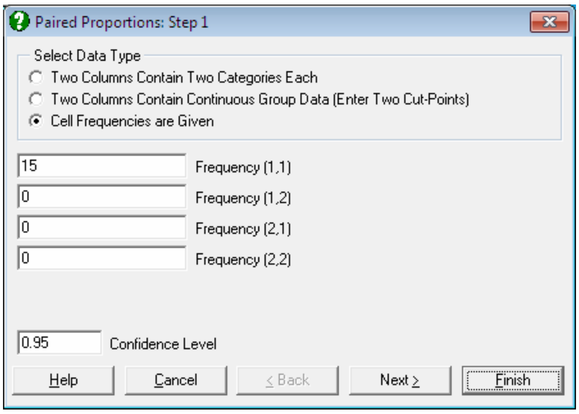
Paired Proportions: Step 2

Enter Values

Low Cholestrol — Cut-point for column 1: Exposure
Yes — Cut-point for column 2: Heart Disease

The method of counting frequencies for the Unpaired Proportions procedure is different from that in Paired Proportions and Cross-Tabulation, as explained above in (1).

3) **Cell Frequencies are Given:** Enter the four elements of the table in row order; i.e. the first column of the first row, the second column of the first row, the first column of the second row and the second column of the second row.


Paired Proportions: Step 1

Select Data Type
○ Two Columns Contain Two Categories Each
○ Two Columns Contain Continuous Group Data (Enter Two Cut-Points)
● Cell Frequencies are Given

15	Frequency (1,1)
0	Frequency (1,2)
0	Frequency (2,1)
0	Frequency (2,2)
0.95	Confidence Level

i) For Paired Proportions, the cells of the table are entered in the order of (1,1), (1,2), (2,1), (2,2).

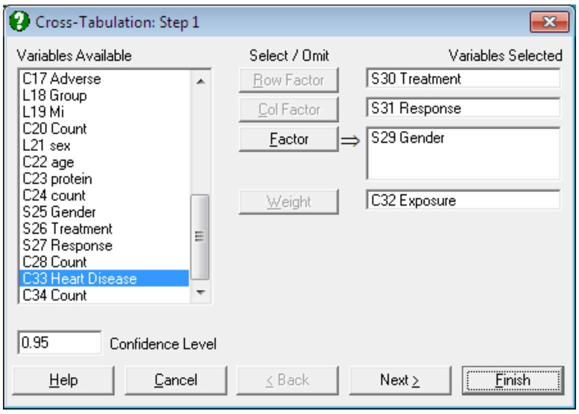
When the four frequency values for a 2 x 2 table are already available in the spreadsheet, you do not have to type them again into the Cell Frequencies are Given dialogue. Instead, you can use the Contingency Table procedure (see 6.6.2.3. 2 x 2 Table Statistics).

ii) For the Unpaired Proportions procedure, the cells contain the following frequencies:

(1,1) contains the number of small values in the first column
(1,2) contains the number of small values in the second column
(2,1) contains the number of large values in the first column
(2,2) contains the number of large values in the second column.

6.0.8. R x C Tables

The Cross-Tabulation procedure requires selection of a [Row Factor] and a [Column Factor]. A table with R rows (the number of levels in the row factor) and C columns (the number of levels in the column factor) will be formed. The cells of the table will contain the number of pairs that correspond to each unique combination of levels of the two factors.



Cross-Tabulation: Step 1

Row Factor: S30 Treatment

Col Factor: S31 Response

Factor: S29 Gender

Weight: C32 Exposure

0.95 Confidence Level

A column containing weights (or frequencies) and an unlimited number of [Factor]s can be selected for Stratified Analysis.

6.1. Parametric Tests

t-tests, F-tests and Hotelling's T-Squared Test are grouped under the Statistics 1 → Parametric Tests option. Test results include tail probabilities and confidence intervals. Although the default value for the confidence level (1 - α) is 0.95, this can be edited to any other value between 0 and 1 on the Variable Selection Dialogue.

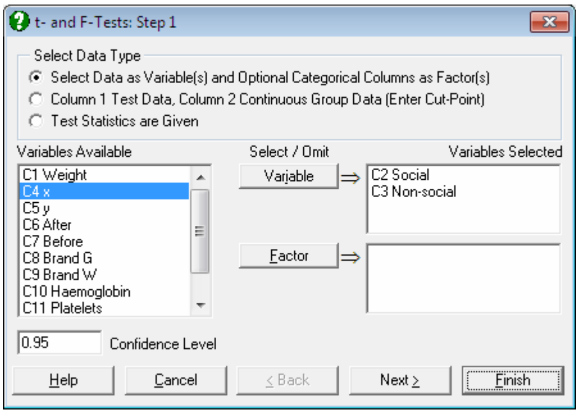
6.1.1. t- and F-Tests

All t- and F-Tests can be accessed under this menu item and the results presented in a single page of output.

If you wish to perform a One Sample t-Test, you can select only one variable. If you select two or more variables, then for each pair, two separate one sample t-tests will be performed on each variable, alongside the two sample tests between them. A paired t-test will be performed only when the two selected variables have the same size. Output Options Dialogue will allow you to choose which tests to appear in the output.

The t-test is used to determine whether the difference between two means is significant. The null hypothesis tested in all four types of t-test is that "the difference between two population means is zero". When the alternative hypothesis is "the difference is not equal to zero", the two-tailed probability should be compared against the given significance level α (usually 0.05). If the calculated probability is greater than α, then the null hypothesis cannot be rejected. Otherwise, we can conclude that the two means are significantly different. In this case, the confidence interval for the difference will not enclose 0. When the alternative hypothesis is a difference in one direction (i.e. one mean is greater or less than the other), then the one-tailed probability is compared against α. UNISTAT reports both one and two-tailed probabilities, where the former is the half of the latter.

The data for this test can be in one of the three types supported for Two Sample Tests.


t- and F-Tests: Step 1

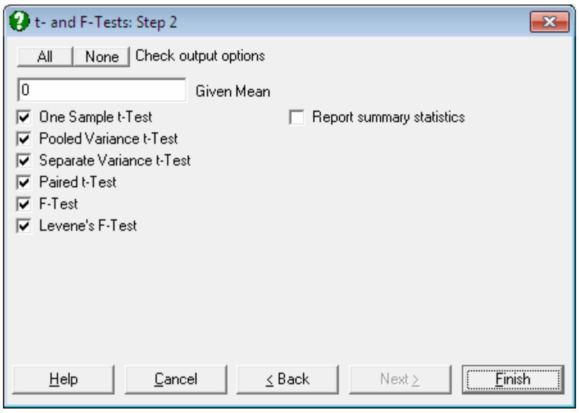
Select Data Type
● Select Data as Variable(s) and Optional Categorical Columns as Factor(s)
○ Column 1 Test Data, Column 2 Continuous Group Data (Enter Cut-Point)
○ Test Statistics are Given

Variables Available: C1 Weight, C4 x, C5 y, C6 After, C7 Before, C8 Brand G, C9 Brand W, C10 Haemoglobin, C11 Platelets

Variables Selected: C2 Social, C3 Non-social

0.95 Confidence Level

After the variable selection is complete, you will be able to select which tests you wish to have displayed in the output. The output consists of the t-value, its degrees of freedom, one and two-tailed probabilities and the confidence interval for the specified confidence level. When the Report summary statistics box is checked, summary information (number of valid cases, missing observations or pairs, mean and standard deviation) about the selected variables is also displayed.


t- and F-Tests: Step 2

Check output options

0 Given Mean

☑ One Sample t-Test
☑ Pooled Variance t-Test
☑ Separate Variance t-Test
☑ Paired t-Test
☑ F-Test
☑ Levene's F-Test

□ Report summary statistics

6.1.1.1. One Sample t-Test

If only one variable is selected, the program will perform only a one-sample t-test against the given mean. By default, the given mean is 0, testing whether the mean of the sample is different from zero. If two or more variables have been selected, then the program will perform two separate one-sample t-tests on each pair of variables, using the same given mean specified in the Output Options Dialogue. Missing cases are omitted by case and the degrees of freedom is adjusted accordingly.

The null hypothesis that “the population mean is equal to the given mean” (a scalar) is tested. The population variance is assumed to be unknown. The t-statistic is computed as:

$$t = \frac{(\overline{X} - M)}{s}\sqrt{n}$$

$$df = n - 1$$

where M is the given mean.

Example

Example 4.2 on p. 101 from Armitage & Berry (2002). The null hypothesis “the population mean is not significantly different from 24” is tested at 95% and 99% levels.

Open PARTESTS, select Statistics 1 → Parametric Tests → t- and F-Tests, select the first column *Weight* (*C1*) as [Variable] and click [Next]. Type 24 into the Given Mean box, select all output options (including the Report summary statistics box) and click [Next] to obtain the following results:

t- and F-Tests

For Weight

	Valid Cases	Missing	Mean	Standard Deviation	Difference	Standard Error
Mean(Weight) – 24	20	0	21.0000	5.9116	-3.0000	1.3219

	t-Statistic	Degrees of Freedom	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
Mean(Weight) – 24	-2.2695	19.0000	0.0175	0.0351	-5.7667	-0.2333

Since the two-tailed probability is less than 5% we reject the null hypothesis and conclude that the population mean is significantly different from 24 at a 95% level. The same result can also be obtained from the reported confidence interval for the difference between means (-5.7667 to -0.2333), since it does not include zero.

If, however, the confidence level is increased to 99%, the null hypothesis should not be rejected, as the two-tailed probability is greater than 1%. This can also be observed from the confidence interval by repeating the test with a 99% confidence level. Select the test again and edit the Confidence Level box to 0.99 in Variable Selection Dialogue. This time, the confidence interval includes zero.

	Lower 99%	Upper 99%
Mean(Weight) – 24	-6.7818	0.7818

6.1.1.2. Pooled Variance t-Test

The null hypothesis "two population means are equal" is tested. It is assumed that the two populations are independent and their standard deviations are the same. The assumption of equal variances can be tested by using the F-test or Levene's F-Test, which is part of the standard output of this procedure. If the two-tailed probability for the F-value is greater than the specified α (such as 0.01 or 0.05), then the hypothesis of equal variances is not rejected and the t-test can use the pooled-variance estimate (equal variances). Otherwise the t-test should be based on separate variance estimates (unequal variances).

The t-statistic for equal population variances is calculated as follows:

$$t = \frac{M_1 - M_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$df = n_1 + n_2 - 2$$

where:

$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

is the pooled estimate of the population variance.

Missing values are omitted by case and the degrees of freedom is adjusted accordingly.

Example 1

Table 87 on p. 231 from Cohen, L. & M. Holliday (1983). The null hypothesis "empathy scores of social and non-social work students have the same mean" is tested at a 95% confidence level.

Open PARTESTS and select Statistics 1 → Parametric Tests → t- and F-Tests. Select the data option 1 and *Social* and *Non-social* (*C2* and *C3*) as [Variable]s. Enter 0 for the Given Mean and select all output options. The following results are obtained:

t- and F-Tests

For Social and Non-social

	Valid Cases	Missing	Mean	Standard Deviation
Mean(Social) – 0	10	0	75.5000	4.5031
Mean(Non-social) – 0	10	0	63.1000	5.9712
Pooled Variance				5.2884
Separate Variance				
Paired	10	0		5.2957

	Difference	Standard Error	t-Statistic	Degrees of Freedom
Mean(Social) – 0	75.5000	1.4240	53.0196	9.0000
Mean(Non-social) – 0	63.1000	1.8883	33.4169	9.0000
Pooled Variance	12.4000	2.3650	5.2431	18.0000
Separate Variance	12.4000	2.3650	5.2431	16.7351
Paired	12.4000	1.6746	7.4045	9.0000

	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
Mean(Social) – 0	0.0000	0.0000	72.2787	78.7213
Mean(Non-social) – 0	0.0000	0.0000	58.8284	67.3716
Pooled Variance	0.0000	0.0001	7.4313	17.3687
Separate Variance	0.0000	0.0001	7.4042	17.3958
Paired	0.0000	0.0000	8.6117	16.1883

	Variance 1	Variance 2	F-Statistic	d.f. Numerator	d.f. Denominator
F-Test	20.2778	35.6556	1.7584	9	9
Levene's F Test			0.6515	1	18

	Right-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
F-Test	0.2066	0.4132	0.4368	7.0791
Levene's F Test		0.4301		

This result shows that there is a significant difference at the 0.1% level, between the empathy scores of social work students and non-social work students.

Example 2

Example on p. 30, Gardner & Altman (2000). Blood pressure level data for diabetics and non diabetics are not available but all necessary parameters to perform a t-test are given.

Size of Group 1	100
Size of Group 2	100
Mean 1	146.4
Mean 2	140.4
Standard Deviation 1	18.5
Standard Deviation 2	16.8

Select Statistics 1 → Parametric Tests → t- and F-Tests, select the data option 3 Test Statistics are Given and enter the above data. Leave the default value of the confidence level unchanged at 0.95. Check all output options except the Report summary statistics box. The following results are obtained:

t- and F-Tests

Test Statistics are Given

	Difference	Standard Error	t-Statistic	Degrees of Freedom
Mean(Group 1) – 0	146.4000	1.8500	79.1351	99.0000
Mean(Group 2) – 0	140.4000	1.6800	83.5714	99.0000
Pooled Variance	6.0000	2.4990	2.4010	198.0000
Separate Variance	6.0000	2.4990	2.4010	196.1884

	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
Mean(Group 1) – 0	0.0000	0.0000	142.7292	150.0708
Mean(Group 2) – 0	0.0000	0.0000	137.0665	143.7335
Pooled Variance	0.0086	0.0173	1.0720	10.9280
Separate Variance	0.0086	0.0173	1.0717	10.9283

	F-Statistic	d.f. Numerator	d.f. Denominator
F-Test	1.2126	99	99

	Right-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
F-Test	0.1696	0.3391	0.8159	1.8022

Note that paired t-test and Levene's F-Test cannot be computed when the data option Test Statistics are Given is selected.

Next, click on the [Last Procedure Dialogue] button to re-display the Variable Selection Dialogue. Edit the value of the confidence level to 0.99 and click [Finish]. All results will be as above except for the confidence intervals. The interval for pooled variance t-test will be:

	Lower 99%	Upper 99%
Pooled Variance	-0.4996	12.4996

And finally, edit the confidence level to 0.9 and repeat the procedure to obtain:

	Lower 90%	Upper 90%
Pooled Variance	1.8702	10.1298

6.1.1.3. Separate Variance t-Test

The null hypothesis "the means of two populations are equal" is tested. It is assumed that their standard deviations may be different. The resulting t-statistic is based on a number of degrees of freedom which is reduced by a factor depending on the extent of the differences in variances.

$$t = \frac{M_1 - M_2}{\sqrt{a_1 + a_2}}$$

where:

$$df = \frac{(a_1 + a_2)^2}{\frac{a_1^2}{n_1 - 1} + \frac{a_2^2}{n_2 - 1}}$$

and where:

$$a_1 = \frac{s_1^2}{n_1},\ a_2 = \frac{s_2^2}{n_2}$$

The reported degrees of freedom (Satterthwaite's approximation) may not be an integer but the nearest integer is used to calculate the tail probabilities.

Missing values are omitted by case and the degrees of freedom is adjusted accordingly.

Example

Table 89 on p. 233 from Cohen, L. & M. Holliday (1983). The raw data on social perceptiveness scores of nursery school and non-nursery school children are not available, but all parameters necessary to perform a t-test are given.

Size of Group 1	71
Size of Group 2	64
Mean 1	19.5
Mean 2	15.3
Standard Deviation 1	3.4
Standard Deviation 2	4.6

Select Statistics 1 → Parametric Tests → t- and F-Tests. Select the data option 3 Test Statistics are Given and enter the above values. Check only the Separate Variance t-test output option to obtain the following results:

t- and F-Tests

Test Statistics are Given

	Difference	**Standard Error**	**t-Statistic**	**Degrees of Freedom**
Separate Variance	4.2000	0.7025	5.9790	115.1866

	1-Tail Probability	**2-Tail Probability**	**Lower 95%**	**Upper 95%**
Separate Variance	0.0000	0.0000	2.8086	5.5914

This result shows that there is a significant difference at the 0.1% level, of the social perceptiveness of nursery school and non-nursery school children.

6.1.1.4. Paired t-Test

This test will be available only when the following conditions are met:

- The data option 1 is selected
- At least two variables are selected
- The selected pairs have the same length.

Two or more columns can be selected by clicking on [Variable]. The test will be performed between all possible pairs, as long as the two columns have the same size. For each test, any pair of cases with one or more missing values is omitted and the degrees of freedom adjusted. It is also possible to perform t-tests between subgroups of two variables defined by one or more factor columns. In this case, the Run a separate analysis for each option selected box must be unchecked.

A paired t-test is performed between two variables, such as values of a sample before and after a certain treatment. The null hypothesis tested is “the difference between pairs is zero” against the alternative hypothesis that “the difference between pairs is not equal to zero”.

The t-statistic is calculated as follows:

$$t = \frac{M_D}{s_D}$$

$$df = n - 1$$

where M_D and s_D are the mean and standard error of D and:

$$D_i = X_{1i} - X_{2i},\ i = 1,....,n$$

Example 1

Example 8.3.4 on pp. 454-54, Larson, H. J. (1982). The null hypothesis “reaction times before consumption of beverage *x* and after consumption *y* on individuals are equal” is tested.

Open PARTESTS and select Statistics 1 → Parametric Tests → t- and F-Tests. Select *x* and *y* (*C4* and *C5*) as [Variable]s and check only the One-sample t-test and Paired t-test boxes to obtain the following results:

t- and F-Tests

For x and y

	Difference	Standard Error	t-Statistic	Degrees of Freedom
Mean(x) – 0	602.4000	29.3342	20.5358	9.0000
Mean(y) – 0	803.7000	19.6413	40.9190	9.0000
Paired	-201.3000	15.1056	-13.3262	9.0000

	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
Mean(x) – 0	0.0000	0.0000	536.0415	668.7585
Mean(y) – 0	0.0000	0.0000	759.2684	848.1316
Paired	0.0000	0.0000	-235.4712	-167.1288

This result shows that there is a significant difference at the 5% level, of the reaction time of individuals before and after consumption of beverage.

Example 2

Example on p. 31, Gardner & Altman (2000). Data on testing the difference between the systolic blood pressure levels for 16 middle aged men before and after a standard exercise are given. The difference between the two columns should be in the order of *After - Before*.

Open PARTESTS and select Statistics 1 → Parametric Tests → t- and F-Tests and select *Before* and *After* (*C6* and *C7*) as [Variable]s and check all output options to obtain the following results:

t- and F-Tests

For After and Before

	Valid Cases	Missing	Mean	Standard Deviation
Mean(After) – 0	16	0	147.7500	12.3477
Mean(Before) – 0	16	0	141.1250	13.6229
Pooled Variance				13.0010
Separate Variance				
Paired	16	0		5.9652

	Difference	Standard Error	t-Statistic	Degrees of Freedom
Mean(After) – 0	147.7500	3.0869	47.8630	15.0000
Mean(Before) – 0	141.1250	3.4057	41.4376	15.0000
Pooled Variance	6.6250	4.5965	1.4413	30.0000
Separate Variance	6.6250	4.5965	1.4413	29.7148
Paired	6.6250	1.4913	4.4425	15.0000

	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
Mean(After) – 0	0.0000	0.0000	141.1704	154.3296
Mean(Before) – 0	0.0000	0.0000	133.8659	148.3841
Pooled Variance	0.0799	0.1599	-2.7624	16.0124
Separate Variance	0.0800	0.1600	-2.7662	16.0162
Paired	0.0002	0.0005	3.4464	9.8036

	Variance 1	Variance 2	F-Statistic	d.f. Numerator	d.f. Denominator
F-Test	152.4667	185.5833	1.2172	15	15
Levene's F Test			0.1228	1	30

	Right-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
F-Test	0.3542	0.7084	0.4253	3.4838
Levene's F Test		0.7284		

6.1.1.5. F-Test

The F-Test is used to compare variances or standard deviations of two samples. The null hypothesis tested is “the two populations have equal variances”. Columns selected for this test need not be equal in size. Output displays the F-value, degrees of freedom, the right and two-tailed probabilities from the F-distribution and the confidence interval for the specified confidence level. When the alternative hypothesis is “the two population variances are not equal”, use the two-tailed probability. When $s_1 > s_2$ the F-value is calculated as follows:

$$F = \frac{s_1^2}{s_2^2}$$

$$\text{df numerator} = n_1 - 1$$

$$\text{df denominator} = n_2 - 1$$

If $s_1 > s_2$ then the F-value is inverted and the two degrees of freedom are interchanged. In other words, the F-value is always the larger variance divided by the smaller variance.

Example 1

Example 5.1 on p. 151 from Armitage & Berry (2002). The null hypothesis “the two population variances are not significantly different” is tested at 95% level. The raw data are not available, but it is sufficient to know the number of cases in each group and their variances to perform an F-test .

Size of Group 1	10
Size of Group 2	20
Variance 1	1.232
Variance 2	0.304

Select Statistics 1 → Parametric Tests → t- and F-Tests, the data option 3 Test Statistics are Given. As this dialogue asks for standard deviations rather than variances, enter Sqr(1.232) and Sqr(0.304) for the two standard deviations. The mean values are not used for F-test. In the Output Options Dialogue, check only the F-test and Levene's F-Test boxes. The following results are obtained:

t- and F-Tests

Test Statistics are Given

	F-Statistic	d.f. Numerator	d.f. Denominator
F-Test	4.0526	9	19

	Right-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
F-Test	0.0049	0.0099	1.4071	14.9272

Since the null hypothesis suggests a two-tailed test (equal vs. not equal) then the two-tailed probability should be compared with α. This result shows there is no significant difference between the two population variances at 5% level for the two-tailed test.

When the data option 3 Test Statistics are Given is selected the Levene's test cannot be computed.

Example 2

Example 8.3.2 on pp. 450-51, Larson, H. J. (1982). The null hypothesis "the population variances for hours of services given by 60 watt light bulbs of brand G and brand W are the same" is tested.

Open PARTESTS and select Statistics 1 → Parametric Tests → t- and F-Tests. Select *Brand G* and *Brand W* (*C8* and *C9*) as [Variable]s and check only the F-test, Levene's F-Test and Report summary statistics boxes to obtain the following results:

t- and F-Tests

For Brand G and Brand W

	Variance 1	Variance 2	F-Statistic	d.f. Numerator	d.f. Denominator
F-Test	2222.2143	653.8778	3.3985	7	9
Levene's F Test			1.4112	1	16

	Right-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
F-Test	0.0459	0.0917	0.8097	16.3918
Levene's F Test		0.2522		

Since the null hypothesis suggests a two-tailed test (equal vs. not equal) then we should look at the 2-tail probability. This result shows that there is no significant difference between the two population variances at 5% level.

6.1.1.6. Levene's F-Test

Levene's F-Test has the advantage of being less sensitive to deviations from normality and is considered to be more powerful than the classical F-test. The alternative hypothesis for Levene's test is "the two population variances are not equal" and the probability reported is comparable to the two-tailed probability for the F-test. The test statistic, which has an F-distribution, is computed as follows:

$$F = \frac{(n-2)\sum_{i=1}^{2} n_i (\overline{Z}_i - \overline{Z})^2}{\sum_{i=1}^{2}\sum_{j=1}^{n_i} (Z_{ij} - \overline{Z}_i)^2}$$

$$\text{df numerator} = 1$$

$$\text{df denominator} = n_1 + n_2 \text{ - } 2$$

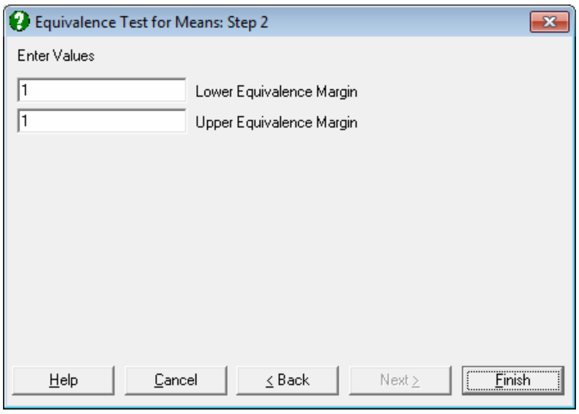
where:

$$Z_{ij} = \left| X_{ij} - \sum_{j=1}^{n_i} X_{ij} \right| \quad \overline{Z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij} \quad \overline{Z} = \frac{1}{n}\sum_{i=1}^{2} n_i \overline{Z}_i$$

Missing values are omitted by case. If the data option 3 Test Statistics are Given is selected then the Levene's test will not be available.

6.1.2. Equivalence Test for Means

t-tests are used to decide whether two means are significantly different from each other. If you wish to find out whether two means cannot be said to be different within predefined boundaries (lower and upper equivalence bounds), use this test. The null hypothesis tested is that the two means are not equivalent, i.e. difference between them is less than the lower equivalence bound or greater than the upper equivalence bound. If the alternative hypothesis is true, namely that the difference is between the two equivalence bounds, then the two means are said to be equivalent.



Equivalence Test for Means: Step 2

Enter Values

1 Lower Equivalence Margin

1 Upper Equivalence Margin

When the lower and upper equivalence bounds are 0, this test is equivalent to the standard t-test, except that here the confidence limits are reported at 1 - 2α level, rather than the usual 1 - α.

This is the parametric version of equivalence test for binomial proportions (see 6.4.3.5. Equivalence Test for Binomial Proportion).

Example

Open PARTESTS and select Statistics 1 → Parametric Tests → Equivalence Test for Means and select *Before* and *After* (*C6* and *C7*) as [Variable]s and check all output options to obtain the following results:

Equivalence Test for Means

For Before and After

Lower Equivalence Margin = 1.0000

Lower Equivalence	Difference	Standard Error	t-Statistic	Degrees of Freedom
Pooled Variance	-6.6250	4.5965	-1.2237	30.0000
Separate Variance	-6.6250	4.5965	-1.2237	29.7148

Lower Equivalence	1-Tail Probability	2-Tail Probability	Lower 90%	Upper 90%
Pooled Variance	0.1153		-14.4265	
Separate Variance	0.1153		-14.4289	

Upper Equivalence Margin = 1.0000

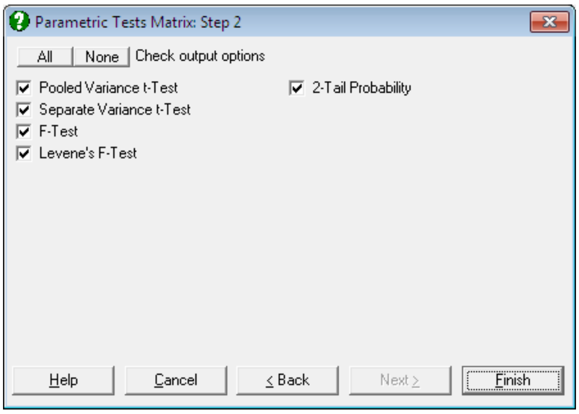
Upper Equivalence	Difference	Standard Error	t-Statistic	Degrees of Freedom
Pooled Variance	-6.6250	4.5965	-1.2237	30.0000
Separate Variance	-6.6250	4.5965	-1.2237	29.7148

Upper Equivalence	1-Tail Probability	2-Tail Probability	Lower 90%	Upper 90%
Pooled Variance	0.1153			1.1765
Separate Variance	0.1153			1.1789

Overall	1-Tail Probability	Lower 90%	Upper 90%
Pooled Variance	0.1153	-14.4265	1.1765
Separate Variance	0.1153	-14.4289	1.1789

6.1.3. Parametric Tests Matrix

This procedure will compute the four most commonly used Parametric Tests (pooled and separate variance t-tests and classical and Levene's F-Tests) for all pairs of selected variables.



Parametric Tests Matrix: Step 2

All | None | Check output options

☑ Pooled Variance t-Test
☑ Separate Variance t-Test
☑ F-Test
☑ Levene's F-Test

☑ 2-Tail Probability

The Variable Selection Dialogue is of the multisample type (see 6.0.4. Multisample Tests), allowing selection of multiple variables and factors. It is possible to perform tests between variables in separate columns, as well as between the groups defined by levels of factor columns.

The output is in the form of a matrix in each cell of which the test statistic, its degrees of freedom and the tail probability are displayed for up to four tests. The tests to be performed can be selected from the Output Options Dialogue.

You can choose to display one- or two-tailed probabilities for all test statistics.

Example

Open DEMODATA and select Statistics 1 → Parametric Tests → 6.1.3. Parametric Tests Matrix. Select *Wages*, *Energy*, *Interest* and *Fixed capital* (*C2* to *C5*) as [Variable]s and check all three options at the Output Options Dialogue to obtain the following results:

Parametric Tests Matrix

	Wages			Energy		
	Test	DoF	2-Tail P	Test	DoF	2-Tail P
Wages t-Pl				0.3079	112.0000	0.7588
t-Sp				0.3079	110.5460	0.7588
F-				1.2591	56.0000	0.1957
F-Lv				1.9049	56.0000	0.1703
Energy t-Pl	0.3079	112.0000	0.7588			
t-Sp	0.3079	110.5460	0.7588			
F-	1.2591	56.0000	0.1957			
F-Lv	1.9049	56.0000	0.1703			
Interest t-Pl	7.4684	113.0000	0.0000	6.6811	113.0000	0.0000
t-Sp	7.4634	112.0421	0.0000	6.6701	108.4311	0.0000
F-	1.1621	56.0000	0.2868	1.4632	56.0000	0.0776
F-Lv	0.8365	56.0000	0.3624	5.2207	56.0000	0.0242
Fixed Capital t-Pl	7.4734	113.0000	0.0000	6.6840	113.0000	0.0000
t-Sp	7.4683	111.9722	0.0000	6.6728	108.2920	0.0000
F-	1.1699	56.0000	0.2783	1.4731	56.0000	0.0740
F-Lv	0.7881	56.0000	0.3766	5.1776	56.0000	0.0248

	Interest			Fixed Capital		
	Test	DoF	2-Tail P	Test	DoF	2-Tail P
Wages t-Pl	7.4684	113.0000	0.0000	7.4734	113.0000	0.0000
t-Sp	7.4634	112.0421	0.0000	7.4683	111.9722	0.0000
F-	1.1621	56.0000	0.2868	1.1699	56.0000	0.2783
F-Lv	0.8365	56.0000	0.3624	0.7881	56.0000	0.3766
Energy t-Pl	6.6811	113.0000	0.0000	6.6840	113.0000	0.0000
t-Sp	6.6701	108.4311	0.0000	6.6728	108.2920	0.0000
F-	1.4632	56.0000	0.0776	1.4731	56.0000	0.0740
F-Lv	5.2207	56.0000	0.0242	5.1776	56.0000	0.0248
Interest t-Pl				-0.0070	114.0000	0.9945
t-Sp				-0.0070	113.9987	0.9945
F-				1.0068	57.0000	0.4899
F-Lv				0.0015	57.0000	0.9689
Fixed Capital t-Pl	-0.0070	114.0000	0.9945			
t-Sp	-0.0070	113.9987	0.9945			
F-	1.0068	57.0000	0.4899			
F-Lv	0.0015	57.0000	0.9689			

6.1.4. Hotelling’s T-Squared Test

This is the multidimensional equivalent of One Sample t-Test. The null hypothesis “the population mean vector is equal to the given mean vector” is tested. Hotelling’s T-Squared statistic is computed as follows:

$$T^2 = (\bar{x} - \mu)' S^{-1} (\bar{x} - \mu)$$

where:

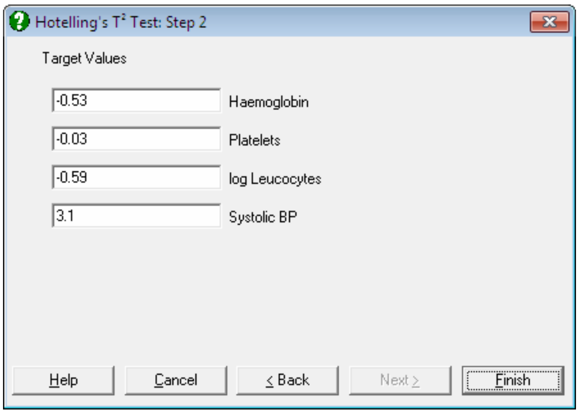
- $\bar{x}$ is the sample mean vector.
- μ is the expected mean vector (target level).
- S is the sample covariance matrix.

The test statistic (which is F-distributed) is found as:

$$F_{p,n-p} = \frac{(n-p)T^2}{p(n-1)}$$

df numerator = p
df denominator = n - p

where p is the number of variables and n is the number of valid cases.


Hotelling's T² Test: Step 2

Target Values

-0.53 Haemoglobin
-0.03 Platelets
-0.59 log Leucocytes
3.1 Systolic BP

Select two or more columns by clicking on [Variable]. The next dialogue prompts for the given target levels, where the mean value of each variable is offered by default. Any rows containing at least one missing value are omitted. The output includes the sample covariance matrix, observed means, target levels, Hotelling's T-Squared statistic and its tail probability.

Also see the related quality control procedure 9.3.4. Hotelling's T-Squared Analysis.

Example

Example 13.3 on p. 474 Armitage & Berry (2002). Measurements are made on babies when they were 25 and 50 days old. The null hypothesis "there is no significant difference between measurements on 25 and 50 days" is tested.

Open PARTESTS and select Statistics 1 → Parametric Tests → Hotelling's T-Squared Test. Select *Haemoglobin*, *Platelets*, *log Leucocytes* and *Systolic BP* (*C10* to *C13*) as variables and all target levels as zero. The following results are obtained.

Hotelling's T-Squared Test

	Target Values	**Mean**	**Difference**
Haemoglobin	0.0000	-0.5300	-0.5300
Platelets	0.0000	-0.0300	-0.0300
Log Leucocytes	0.0000	-0.5900	-0.5900
Systolic BP	0.0000	3.1000	3.1000

Hotelling's T-Squared Statistic =	7.4391
F(4,6) =	1.2398
Right-Tail Probability =	0.3869

The result is not significant at 10% level. Thus do not reject the null hypothesis.

6.2. Correlations

Correlation Coefficients measure the degree of association between two sets of data. They take on values ranging from -1 to +1 (inclusive), meaning complete negative and positive correlations respectively. A zero value means that the two data sets have no association. In this case they are said to be uncorrelated.

Two of the Correlation Coefficients in this group, Spearman and Kendall correlations are nonparametric. This means that it is the relative positions of data points in the sample that matters, rather than their nominal values. These routines involve highly demanding sorting and ranking phases, which may be time consuming with large data files.

Confidence intervals are reported for all Correlation Coefficients. Assuming the two samples have a joint bivariate normal distribution, the confidence interval for their correlation coefficient is computed after applying the Fisher’s z transformation:

$$z = \frac{1}{2}\mathrm{Ln}\left(\frac{1+r}{1-r}\right)$$

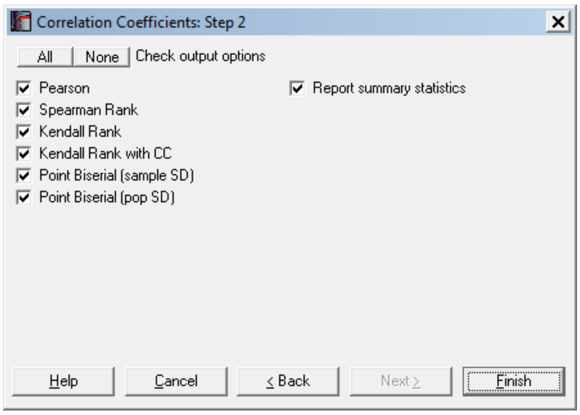
$$z_1 = z - \frac{Z_{1-\alpha/2}}{\sqrt{n-3}}, \quad z_2 = z + \frac{Z_{1-\alpha/2}}{\sqrt{n-3}}$$

$$\mathrm{LL} = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \quad \mathrm{UL} = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$$

where $Z_{1-\alpha/2}$ is the critical value from the standard normal distribution.

6.2.1. Correlation Coefficients

Four Correlation Coefficients (Pearson product moment, Spearman rank, Kendall rank and point biserial) can be accessed under this menu item and the results presented in a single page of output.



Correlation Coefficients: Step 2

Check output options

☑ Pearson
☑ Spearman Rank
☑ Kendall Rank
☑ Kendall Rank with CC
☑ Point Biserial (sample SD)
☑ Point Biserial (pop SD)

☑ Report summary statistics

Two or more columns can be selected by clicking on [Variable]. Correlations will be computed between all possible pairs, as long as the two columns have the same size. For each test, any pair of cases with one or more missing values is omitted and the degrees of freedom adjusted. Output Options Dialogue will allow you to choose which tests to appear in the output.

If a factor column is selected, then it is assumed that the data is not paired and only the point serial correlation is computed.

6.2.1.1. Pearson Product Moment Correlation

The aim of this correlation coefficient is to establish the degree of linear relationship between two variables. The coefficient is defined as the covariance of the two samples divided by the product of their standard deviations.

$$r = \frac{\mathrm{Cov}(XY)}{S_X S_Y}$$

The probability value is based on Student's t-distribution, where the t-statistic is calculated as:

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

$$df = n - 2$$

This correlation coefficient is a relatively poor measure of association since it does not take into consideration the individual distributions of the two variables. The effect of outliers may be considerable. This makes it difficult to conclude that one linear correlation is significantly better than another. The nonparametric Correlation Coefficients Spearman's rho and Kendall's tau are more robust measures.

Pairs with one or more missing values are omitted and the degrees of freedom is adjusted. The output includes the correlation coefficient, its confidence interval, t-statistic, degrees of freedom and one- and two-tailed probabilities.

Example

Table 8.5 on p. 89, Gardner & Altman (2000). The null hypothesis "basal metabolic rate and total energy expenditure are not correlated" is tested at 95% confidence level.

Open CORRCOEF, select Statistics 1 → Correlation Coefficients, select *Basal* and *Energy* (*C1* and *C2*) as [Variable]s, select all output options (including the Report summary statistics box) and click [Next] to obtain the following results:

Correlation Coefficients

For Basal and Energy

	Valid Cases	Missing	Mean	Standard Deviation
Basal	13	0	5.6515	0.4650
Energy	13	0	8.0662	1.2381
Paired	13	0		

	Correlation Coefficient	Degrees of Freedom	* Test Statistic	1-Tail Probability	2-Tail Probability
Pearson	0.7283	11	3.5249	0.0024	0.0048
Spearman Rank	0.6190	11	2.6139	0.0120	0.0241
Kendall Rank	0.4258		2.0171	0.0218	0.0437
Kendall Rank with CC	0.4387		2.0782	0.0188	0.0377
Point Biserial (sample SD)	-0.7866	24	-6.2419	0.0000	0.0000
Point Biserial (pop SD)	-0.8022	24	-6.5828	0.0000	0.0000

	Lower 95%	Upper 95%
Pearson	0.2961	0.9129
Spearman Rank	0.1032	0.8724
Kendall Rank	-0.1635	0.7912
Kendall Rank with CC	-0.1481	0.7970
Point Biserial (sample SD)	-0.8998	-0.5743
Point Biserial (pop SD)	-0.9076	-0.6019

* Z-statistic for Kendall rank, t-statistic otherwise

This result shows that there is a significant correlation between the two variables.

6.2.1.2. Spearman's Rank Correlation

Correlation between relative rankings of the two variables is measured rather than their nominal values. In this way each variable is transformed into a uniformly distributed variable and the effect of outliers is minimised. Spearman's correlation coefficient (also called rho) is calculated as follows:

$$\varrho = \frac{T_x + T_y - R}{2\sqrt{T_x T_y}}$$

where R is the sum of squared differences between the ranks of corresponding cases of the two variables and:

$$T_x = \frac{n^3 - n - K_x}{12}$$

$$T_y = \frac{n^3 - n - K_y}{12}$$

where K_x and K_y are the sum of k^3 - k where k is the number of ties at a given rank within each variable. The tail probability of rho is determined by comparing the following t-statistic with the Student's t distribution:

$$t = \varrho\sqrt{\frac{n-2}{1-\varrho^2}}$$

$$df = n - 2$$

Pairs with at least one missing value are omitted and the degrees of freedom is adjusted. The output includes the correlation coefficient, its confidence interval, t-statistic, degrees of freedom and one- and two-tailed probabilities.

Example

Example 19.13 on p. 401 from Zar, J. H. (2010). The null hypothesis "there is no correlation between the ranks of values in the two variables" is tested.

Open CORRCOEF, select Statistics 1 → Correlation Coefficients. Select *X* and *Y* (*C3* and *C4*) as [Variable]s and select only the Spearman Rank output option to obtain the following results:

Correlation Coefficients

For X and Y

	Correlation Coefficient	Degrees of Freedom	* Test Statistic	1-Tail Probability	2-Tail Probability
Spearman Rank	0.8511	10	5.1261	0.0002	0.0004

	Lower 95%	Upper 95%
Spearman Rank	0.5418	0.9574

* Z-statistic for Kendall rank, t-statistic otherwise

This result shows that there is a significant rank correlation and the null hypothesis should be rejected. Note that the denominator evaluates to 240, not 242 as in the book.

6.2.1.3. Kendall's Rank Correlation

Like Spearman's rho this is also a rank correlation coefficient (also called tau) and as such it has the same advantage over Pearson Product Moment Correlation.

Additionally, it provides a more robust nonparametric measure by comparing the relative ordering of ranks rather than their numeric difference as in the case of Spearman's rho. Kendall's tau is calculated as:

$$\tau = \frac{R}{\sqrt{\frac{(n^2 - n - K_x)(n^2 - n - K_y)}{4}}}$$

where R is the number of times a case is greater than other cases in both variables summed over all cases, and K_x and K_y are the sum of k^2 - k where k is the number of ties at a given rank within each variable. For tau with continuity correction R is augmented by one.

The tail probability of tau is determined from the normal distribution with a standard deviation:

$$S = \frac{J(2n+5) - P_x - P_y}{18} + \frac{Q_x Q_y}{9J(n-2)} + \frac{K_x K_y}{2J}$$

where:

- P_x = sum of $(k^2 - k)(k - 2)$ for X
- P_y = sum of $(k^2 - k)(k - 2)$ for Y
- Q_x = sum of $(k^2 - k)(2k + 5)$ for X
- Q_y = sum of $(k^2 - k)(2k + 5)$ for Y
- $J = n^2 - n$.

Pairs with at least one missing value are omitted and the degrees of freedom is adjusted. The output includes the correlation coefficient, its confidence interval, t-statistic, degrees of freedom and one- and two-tailed probabilities.

Example

Table 56 on p. 160 from Cohen, L. & M. Holliday (1983). Ten trainees on a management course have been rated on a personality measure *Introversion* and on an *Attitude to Change* scale. The null hypothesis "there is no correlation between these two rankings" is tested.

Open CORRCOEF and select Statistics 1 → Correlation Coefficients. Select *Introversion* and *Attitude* (*C5* and *C6*) as variables and select only the Kendall Rank output option to obtain the following results:

Correlation Coefficients

For Introversion and Attitude

	Correlation Coefficient	Degrees of Freedom	* Test Statistic	1-Tail Probability	2-Tail Probability
Kendall Rank	0.6286		2.4545	0.0071	0.0141
Kendall Rank with CC	0.6519		2.5455	0.0055	0.0109

	Lower 95%	Upper 95%
Kendall Rank	-0.0017	0.9014
Kendall Rank with CC	0.0377	0.9086

* Z-statistic for Kendall rank, t-statistic otherwise

This result shows that there is a significant rank correlation at the 1% level, between the *Introversion / extraversion* rating and the *Attitude to Change* rating.

6.2.1.4. Point Biserial Correlation

This is an alternative to the linear (Pearson's) correlation coefficient when the first variable is continuous and the second variable is binary. Let n_p and n_q be the respective numbers of Ps and Qs and n the total number of valid cases. There are two alternative ways of calculating the coefficient:

Using sample standard deviation:

$$r = \frac{M_p - M_q}{SD_{samp}} \sqrt{\frac{n_p n_q}{n^2}}$$

where SD_{samp} is the sample standard deviation of the two samples combined:

$$SD_{samp} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

Using population standard deviation:

$$r = \frac{M_p - M_q}{SD_{pop}} \sqrt{\frac{n_p n_q}{n^2}} = \frac{M_p - M_q}{SD_{samp}} \sqrt{\frac{n_p n_q}{n(n-1)}}$$

where SD_{pop} is the population standard deviation of the two samples combined:

$$SD_{pop} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Before this release of UNISTAT, the version with sample standard deviation was used.

In both cases, the following t-value is compared with the t-distribution:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$df = n - 2$$

The data for this test can be in one of the three types supported for Two Sample Tests. If the last data option Test Statistics are Given is selected the program will prompt for sizes, means and standard deviations of the two samples. Missing values are omitted by case and the degrees of freedom is adjusted accordingly.

Example 1: Point biserial correlation using sample standard deviation

Table 57 on p. 164 from Cohen, L. & M. Holliday (1983). Examination scores of on and off campus social work students is given in one column of the table and their residence pattern in a second column.

Open CORRCOEF and select Statistics 1 → Correlation Coefficients. Select *Score* (*C7*) as [Variable] and *Off Campus* (*C8*) as [Factor], and select the Point Biserial and Report Summary Statistics output options to obtain the following results:

Correlation Coefficients

Data variable: Score
Subsample selected by: Off Campus = 0,1

	Valid Cases	Missing	Mean	Standard Deviation
0	6	0	82.3333	5.1251
1	4	0	65.0000	4.0825

	Correlation Coefficient	Degrees of Freedom	* Test Statistic	1-Tail Probability	2-Tail Probability
Point Biserial	0.8480	8	4.5260	0.0010	0.0019

	Lower 95%	Upper 95%
Point Biserial	0.4686	0.9633

* Z-statistic for Kendall rank, t-statistic otherwise

This result shows that there is a significant correlation at the 0.1% level between examinations scores and residence.

Example 2: Point biserial correlation using population standard deviation

Example 19.16 on p. 410 from Zar, J. H. (2010). The null hypothesis that there is no correlation between blood-clotting time and drug is tested.

Open CORRCOEF and select Statistics 1 → Correlation Coefficients. Select *X1* (*C12*) as [Factor] and *Y1* (*C13*) as [Variable], and select only the Point Biserial output option to obtain the following results:

Correlation Coefficients

Data variable: Y1
Subsample selected by: X1 = 0,1

	Correlation Coefficient	Degrees of Freedom	* Test Statistic	1-Tail Probability	2-Tail Probability
Point Biserial	-0.5983	11	-2.4765	0.0154	0.0308

	Lower 95%	Upper 95%
Point Biserial	-0.8643	-0.0706

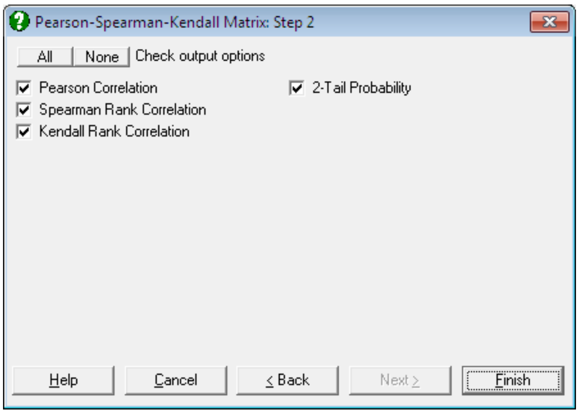
* Z-statistic for Kendall rank, t-statistic otherwise

Since P < 0.05, we reject the null hypothesis.

6.2.2. Pearson-Spearman-Kendall Correlations Matrix

This procedure computes the three most commonly used Correlation Coefficients (i.e. Pearson, Spearman, Kendall) for all combinations of column pairs selected from the Variables Available list. The output is in the form of a matrix in each cell of which up to three Correlation Coefficients, the number of cases and the probability values are reported. It is possible to display one- or two-tailed probabilities.

Columns to be analysed are selected from the Variables Available list by clicking on [Variable]. Selected columns do not need to have equal length. The program will consider each column pair separately and calculate the coefficients for only those pairs with equal number of cases. Matrix cells for columns with unequal lengths will be left blank. Also, the program will handle missing values for each pair of columns separately and omit the missing values pairwise.


Pearson-Spearman-Kendall Matrix: Step 2

Check output options

☑ Pearson Correlation
☑ Spearman Rank Correlation
☑ Kendall Rank Correlation
☑ 2-Tail Probability

An Output Options Dialogue will allow you to select which correlations to be displayed in the output.

Computing the rank Correlation Coefficients can be time consuming. When k columns are chosen from the data matrix, this procedure will compute 3k(k - 1)/2 Correlation Coefficients together with their tail probability values.

Example

Open CORRCOEF and select Statistics 1 → Correlation Coefficients → Pearson-Spearman-Kendall Correlations Matrix. Select *Introversion*, *Attitude* and *Score* (*C5* to *C7*) as [Variable]s ad click [Finish].

Pearson-Spearman-Kendall Matrix

	Introversion			Attitude		
	Corr	No	2-Tail P	Corr	No	2-Tail P
Introversion Prsn				0.8185	10	0.0038
Spmn				0.7976	10	0.0057
Kndl				0.6286	10	0.0141
Attitude Prsn	0.8185	10	0.0038			
Spmn	0.7976	10	0.0057			
Kndl	0.6286	10	0.0141			
Score Prsn	0.0000	10	1.0000	0.1178	10	0.7459
Spmn	0.0851	10	0.8152	0.2031	10	0.5736
Kndl	0.0449	10	0.8575	0.1413	10	0.5840

	Score		
	Corr	No	2-Tail P
Introversion Prsn	0.0000	10	1.0000
Spmn	0.0851	10	0.8152
Kndl	0.0449	10	0.8575
Attitude Prsn	0.1178	10	0.7459
Spmn	0.2031	10	0.5736
Kndl	0.1413	10	0.5840
Score Prsn			
Spmn			
Kndl			

6.2.3. Partial Correlation Matrix

Partial correlation is used to obtain the linear correlation between two variables after the effects of some other variables are filtered out. The latter are referred to as control variables or covariates. The number of covariates included gives the order of partial correlation. Here only the formula for a first order coefficient will be given as higher levels quickly get complicated:

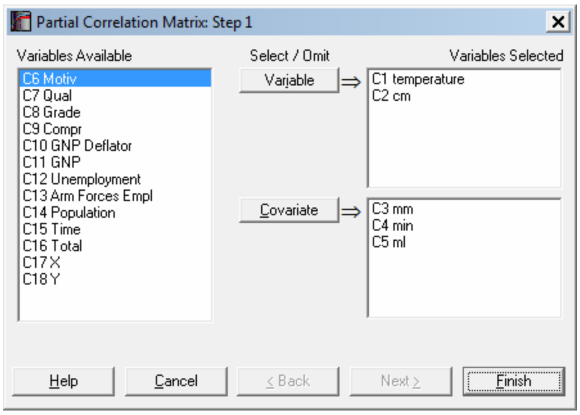
$$r_{ij.k} = \frac{r_{ij} - r_{ki} r_{kj}}{(1 - r_{ki}^2)(1 - r_{kj}^2)}$$

where i and j are the indices for variables to be correlated and k is for the covariate. All correlation coefficients on the right hand side of the equation are zero order Pearson Product Moment Correlation.

Like the Pearson-Spearman-Kendall Correlations Matrix procedure, Partial Correlation Matrix can compute more than one coefficient at a time and display the results in the form of a matrix. Each cell of the output matrix displays the correlation coefficient, its degrees of freedom and probability from the t-distribution. The degrees of freedom is calculated as:

$$df = n - m - 2$$

where n is the number of valid cases and m is the number of covariates.


Partial Correlation Matrix: Step 1

Variables Available: C6 Motiv, C7 Qual, C8 Grade, C9 Compr, C10 GNP Deflator, C11 GNP, C12 Unemployment, C13 Arm Forces Empl, C14 Population, C15 Time, C16 Total, C17 X, C18 Y

Variables Selected (Variable): C1 temperature, C2 cm

Covariate: C3 mm, C4 min, C5 ml

As of this version of UNISTAT, we have introduced a new regression based algorithm in calculating the partial correlations. Now it is possible to select an unlimited number of columns to be correlated or to be used as covariates. Missing values are handled for each correlation separately, i.e. any cases containing a missing value in the two variables correlated or in any selected covariates are omitted. Also, we now display the degrees of freedom rather than the number of cases in the output.

Select the columns to be correlated from the Variables Available list by clicking on [Variable] and covariates by clicking on [Covariate]. If no covariates are selected then the program will compute the zero order (Pearson) correlations. Coefficients will be displayed for only those pairs of variables and covariates with equal size.

Example

Example 20.2 on p. 439 from Zar, J. H. (2010). In this particular example, the partial correlation for each pair is computed using all the rest of the variables as covariates. Here it will be sufficient to generate one of the partial correlation coefficients.

Open REGRESS, select Statistics 1 → Correlation Coefficients → Partial Correlation and select *temperature, cm* (*C1-C2*) as [Variable]s and *mm, min, ml* (*C3-C5*) as [Covariate]s to obtain the following results:

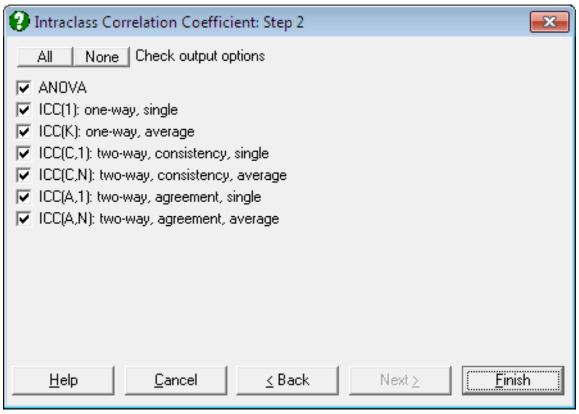
Partial Correlation Matrix

3 Order Correlations
Controlling for: mm, min, ml

	temperature			**cm**		
	Corr	**DoF**	**2-Tail P**	**Corr**	**DoF**	**2-Tail P**
Temperature				0.1943	28	0.3036
cm	0.1943	28	0.3036			

6.2.4. Intraclass Correlation Coefficients

The intraclass correlation coefficient is a multivariate generalisation of the more commonly used Correlation Coefficients on paired data. It can be considered as a correlation coefficient for k categories (columns) with n cases (rows). Its most common application is, like the Kappa tests (see 6.5.10. Kappa Test for Inter-Observer Variation), the test of agreement between k raters on n subjects.



☑ ANOVA
☑ ICC(1): one-way, single
☑ ICC(K): one-way, average
☑ ICC(C,1): two-way, consistency, single
☑ ICC(C,N): two-way, consistency, average
☑ ICC(A,1): two-way, agreement, single
☑ ICC(A,N): two-way, agreement, average

UNISTAT supports six categories of intraclass correlation coefficient, each representing a combination of the following properties:

One-way / Two-way: The degree of agreement when, raters are assigned to subjects randomly / all raters rate all subjects, respectively.

Consistency / Agreement: The degree of, consistency among / absolute agreement for, subjects respectively.

Single / Average: The agreement, among subjects / for the average of n independent subjects, respectively.

The output options include the ANOVA table, six correlation coefficients, their significance tests and confidence intervals.

ANOVA: A one-way repeated measures ANOVA table is displayed.

ICC(1): Intraclass correlation coefficient for the case of one-way, single measurement.

ICC(K): Intraclass correlation coefficient for one-way, average measurement.

ICC(C,1): Intraclass correlation coefficient for two-way, consistency, single measurement.

ICC(C,N): Intraclass correlation coefficient for two-way, consistency, average measurement.

ICC(A,1): Intraclass correlation coefficient for two-way, absolute agreement, single measurement.

ICC(A,N): Intraclass correlation coefficient for two-way, absolute agreement, average measurement.

See Shrout, P. E., and Fleiss, J. L. (1979) and McGraw, K. O. & Wong, S. P. (1996) for details.

Example

The data is taken from Shrout P.E. and Fleiss J.L. (1979).

Open CORRCOEF, select Statistics 1 → Nonparametric Tests (Multisample) → REGRESS, select Statistics 1 → Correlation Coefficients → Partial Correlation and select *temperature, cm* (*C1-C2*) as [Variable]s and *mm, min, ml* (*C3-C5*) as [Covariate]s to obtain the following results:

Partial Correlation Matrix

3 Order Correlations
Controlling for: mm, min, ml

	temperature			cm		
	Corr	DoF	2-Tail P	Corr	DoF	2-Tail P
Temperature				0.1943	28	0.3036
cm	0.1943	28	0.3036			

6.2.4. Intraclass Correlation Coefficients and select *Rater 1* to *Rater 4* (*C9* to *C12*) as [Variable]s to obtain the following results:

Intraclass Correlation Coefficients

Variables Selected: Rater 1, Rater 2, Rater 3, Rater 4
Number of Columns: 4
Number of Rows: 6

ANOVA

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Between Raters	97.458	3	32.486	31.866	0.1293
Between Subjects	56.208	5	11.242	11.027	0.2246
Within Subjects	112.750	18	6.264	6.144	0.3086
Error	15.292	15	1.019		
Total	168.958	23	7.346		

Intraclass Correlation Coefficients

	Value	F-Statistic	d.f. Numerator	d.f. Denominator
ICC(1): one-way, single	0.1657	1.7947	5	18
ICC(K): one-way, average	0.4428	1.7947	5	18
ICC(C,1): two-way, consistency, single	0.7148	11.0272	5	15
ICC(C,N): two-way, consistency, average	0.9093	11.0272	5	15
ICC(A,1): two-way, agreement, single	0.2898	11.0272	5	5
ICC(A,N): two-way, agreement, average	0.6201	11.0272	5	5

	Right-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
ICC(1): one-way, single	0.1648	0.3295	-0.1329	0.7226
ICC(K): one-way, average	0.1648	0.3295	-0.8844	0.9124
ICC(C,1): two-way, consistency, single	0.0001	0.0003	0.3425	0.9459
ICC(C,N): two-way, consistency, average	0.0001	0.0003	0.6757	0.9859
ICC(A,1): two-way, agreement, single	0.0099	0.0198	0.0188	0.7611
ICC(A,N): two-way, agreement, average	0.0099	0.0198	0.0711	0.9272

6.3. Goodness of Fit Tests

The Goodness of Fit Tests are used to determine whether two samples have similar distributions. The main difference between chi-square tests and Kolmogorov-Smirnov Tests is that the former are used with frequency data while the latter with continuous data.

If you want to test whether a sample is normally distributed (when the population mean and standard deviation are not known and are to be estimated from the data), you can use the Normality Tests procedure. The four tests supported are Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling.

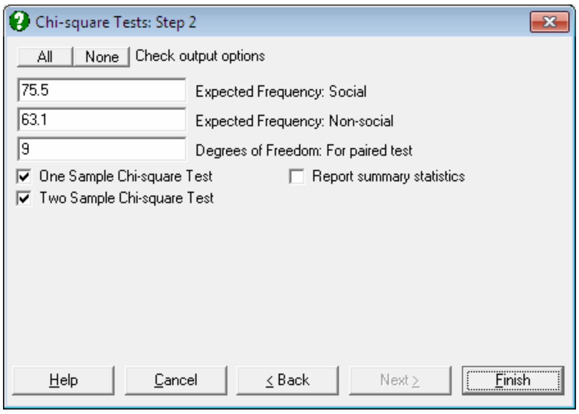
The Chi-Square Tests expect data in the form of frequency counts. In Stand-Alone Mode, raw data can easily be transformed into the frequency counts format using the Data Processor's **Freq()** function (see 3.4.2.5. Statistical Functions).

In case of one sample tests, the second sample is assumed to have a known distribution function such as uniform or normal. In two sample tests the second sample may be just another set of observed data, in which case the test will only determine whether the two samples have similar distributions, or it may contain expected values from a theoretical distribution function, in which case the test will determine whether the first sample has a distribution consistent with a particular theoretical distribution function.

6.3.1. Chi-Square Tests

One sample and two sample chi-squared tests can be accesses under one menu item and the results will be presented in a single page of output.

If you wish to perform a one sample chi-squared test, you can select only one variable. If you select two or more variables, then two separate one sample tests will be performed on each variable, alongside a two sample test between them. A two sample chi-squared test will be performed only when the two selected [Variable]s have the same length. The Output Options Dialogue will allow you to choose which tests to appear in the output.



Check output options

75.5 Expected Frequency: Social

63.1 Expected Frequency: Non-social

9 Degrees of Freedom: For paired test

☑ One Sample Chi-square Test

☑ Two Sample Chi-square Test

□ Report summary statistics

The default expected frequency value suggested by the program is the mean of observed frequencies. These can be changed to any other values. When more than two variables are selected, however, the program does not stop and display the Output Options Dialogue and proceeds with the default expected frequencies.

When the Report summary statistics box is checked, summary information about the selected variables (number of valid cases, missing observations or pairs, mean and standard deviation) is also displayed.

6.3.1.1. One Sample Chi-Square Test

The null hypothesis “observed frequencies are all equal to the given (expected) frequency” is tested. The chi-square statistic is computed as:

$$C = \sum \frac{(fo_i - fe_i)^2}{fe_i}$$

$$df = n - 1$$

where fo_i is the i^{th} observed frequency and fe is the expected frequency.

Example

Example 10.3.1 on p. 529, Larson, H. J. (1982). A die is rolled 200 times and the number of times each number occurs is recorded in a table. The null hypothesis “all six numbers are equally likely” is tested.

Open GOODFIT and select Statistics 1 → Goodness of Fit Tests → Chi-Square Tests. Select *Frequency* (*C1*) as [Variable], accept the program’s suggestion of 33.33

as the expected value, check the Report summary statistics box and click [Finish].

Chi-square Tests

	Valid Cases	Missing	Mean	Standard Deviation
Frequency	6	0	33.3333	1.6330

	Expected Frequency	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Frequency	33.3333	2.8600	5	0.7216

This result shows there is no significant difference between the observed frequency and the expected frequency at 5% level. Hence we accept that the likelihood of six numbers are not significantly different.

6.3.1.2. Two Sample Chi-Square Test

This test computes the goodness of fit for two columns containing frequency data. In general, observed frequencies (which are assumed to be in column 1) are compared with expected or theoretical frequencies (which are assumed to be in column 2). Normally, the sums of the two columns are expected to be the same. If this is not the case the program will normalise the values of the second column such that their sum is equal to the first column's sum. The chi-square statistic is computed as:

$$C = \sum \frac{(fo_i - fe_i)^2}{fe_i}$$

$$df = n - 1$$

where fo_i and fe_i are the i^th observed and expected frequencies respectively.

More than two variables can be selected by clicking on [Variable]. The test will be performed on all possible pairs with equal length. Any pair of cases with at least one missing value is omitted and the degrees of freedom is adjusted.

When a given set of frequencies is compared with a theoretical distribution, allowance should be made in the degrees of freedom for the estimated parameters of the distribution. For instance, if the theoretical distribution (column 2) is normal, the degrees of freedom for the test should be n - 3, to reflect the effect of the estimated distribution parameters, mean and standard deviation. For a

Poisson distribution the degrees of freedom is n - 2, as the mean of the distribution should be estimated. To find out about degrees of freedom for other distributions see Appendix.

If only two variables are selected, then the program will prompt for the degrees of freedom, displaying a default value of n - 1. If more than two variables are selected, the program uses this value for all pairs of variables and does not prompt for user input.

The output includes the chi-square statistic, degrees of freedom and the right tail probability.

Example

Example 11.1 on p. 395 from Armitage & Berry (1994). The first column of data contains the observed frequencies of bacterial counts and the second expected frequencies from Poisson distribution.

Open GOODFIT and select Statistics 1 → Goodness of Fit Tests → Chi-Square Tests. Select *Observed* (*C2*) as the first variable and *Expected* (*C3*) as the second. Enter the degrees of freedom as 6 (instead of the suggested 7 since the Poisson distribution uses 1 parameter), to obtain the following results:

Chi-square Tests

	Valid Cases	Missing	Mean	Standard Deviation
Observed	8	0	50.0000	39.2538
Expected	8	0	49.9875	36.6127
Observed – Expected	8	0		

	Expected Frequency	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Observed	50.0000	215.7200	7	0.0000
Expected	49.9875	187.7151	7	0.0000
Observed – Expected		6.0150	6	0.4215

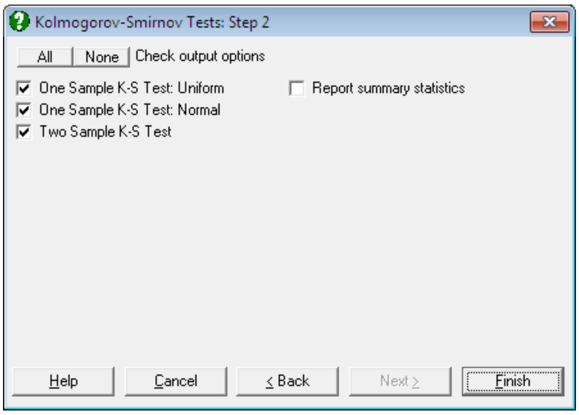
This result shows there is no significant difference between the observed and the expected frequencies, though they are both significantly different from zero.

6.3.2. Kolmogorov-Smirnov Tests

One sample and two sample Kolmogorov-Smirnov Tests can be accesses under one menu item and the results are presented in a single page of output.

If you wish to perform a one sample Kolmogorov-Smirnov test, you can select only one variable. If you select two or more variables, then two separate one sample tests will be performed on each variable, alongside a two sample test between them. Output Options Dialogue will allow you to choose which tests to appear in the output.

When the Report summary statistics box is checked, summary information about the selected variables (number of valid and missing cases, mean and standard deviation) is also displayed.



☑ One Sample K-S Test: Uniform
☑ One Sample K-S Test: Normal
☑ Two Sample K-S Test
□ Report summary statistics

6.3.2.1. One Sample Kolmogorov-Smirnov Test: Uniform Distribution

The null hypothesis “the cumulative distribution of the observed set of data is uniform” is tested. It is assumed that the underlying distribution is continuous. The program computes cumulative proportions from the sample and finds the absolute value of their maximum difference from the cumulative uniform distribution (i.e. the Kolmogorov-Smirnov statistic).

$$K = |\, cpo_i - cpe_i \,|_{max}$$

where cpo_i and cpe_i are the i^{th} observed and uniform cumulative proportions respectively.

The output includes the test statistic and its two-tailed probability (which is computed from the Smirnov formula).

Example

Example 3.4 on pp. 72-74 from Sprent, P. (1993), where the null hypothesis "the population is uniformly distributed" is tested at a 95% confidence level.

Open GOODFIT and select Statistics 1 → Goodness of Fit Tests → Kolmogorov-Smirnov Tests. Select *Distance* (*C4*) as [Variable] and check One Sample K-S Test: Uniform and Report summary statistics boxes to obtain the following results:

Kolmogorov-Smirnov Tests

For Distance

	Valid Cases	Missing	Mean	Standard Deviation
Distance (Uniform)	20	0	2.7350	1.5246

	Abs(Maximum difference)	Test Statistic	2-Tail Probability	Lilliefors Probability
Distance (Uniform)	0.2217	0.9915	0.2793	

Since the probability is larger than 5%, do not reject the null hypothesis.

6.3.2.2. One Sample Kolmogorov-Smirnov Test: Normal Distribution

This is similar to Kolmogorov-Smirnov test for uniform distribution except that the observed cumulative proportions are compared with the normal cumulative proportions.

The two-tailed probability value computed from Smirnov formula is reported. An alternative probability definition by Lilliefors (1967), adopting the correction introduced by Dallal and Wilkinson (1986), is also reported. The probability values from Smirnov and Lilliefors formulas can be quite different. The Smirnov probability should only be used when the population mean and standard deviation are known and the Lilliefors probability should be used when these entities are to be estimated from data.

One sample Kolmogorov-Smirnov test for normality with this latter probability value is also known as Lilliefors test. Lilliefors probability values are also reported as part of Normality Tests output for Kolmogorov-Smirnov test.

Example

Example 3.6 on pp. 77-79 from Sprent, P. (1993), where the null hypothesis "the population death age is normally distributed" is tested.

Open GOODFIT and select Statistics 1 → Goodness of Fit Tests → Kolmogorov-Smirnov Tests. Select *Age* (*C7*) as [Variable], check only the One Sample K-S Test: Normal box to obtain the following results:

Kolmogorov-Smirnov Tests

For Age

	Abs(Maximum difference)	Test Statistic	2-Tail Probability	Lilliefors Probability
Age (Normal)	0.1921	1.0522	0.2182	0.0062

The null hypothesis is rejected at the 99% confidence level. UNISTAT reports the exact probability while Sprent compares the test statistic with the published critical value for 1%.

6.3.2.3. Two Sample Kolmogorov-Smirnov Test

Cumulative distributions computed from the two given data sets are compared. The test statistic is:

$$K = D\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where:

$$D = | cp1_i - cp2_i |_{max}$$

and $cp1_i$ and $cp2_i$ are the i^th^ cumulative proportions of the first and second samples respectively.

Data in one of the three types supported for two sample tests can be entered (see 6.0.2. Two Sample Tests). The output includes the number of cases in two samples, the maximum difference, the test statistic, its two-tailed probability from the Smirnov distribution.

If the last data option is selected, the program will prompt for the number of cases in each sample and the maximum absolute cumulative difference. It is possible to use this procedure to calculate the cumulative Smirnov distribution probabilities without having to perform the test itself.

Example

Open GOODFIT and select Statistics 1 → Goodness of Fit Tests → Kolmogorov-Smirnov Tests. Select *Distance* (*C4*) and *Age* (*C7*) as [Variable]s and check all boxes on the output options dialogue.

Kolmogorov-Smirnov Tests

For Distance and Age

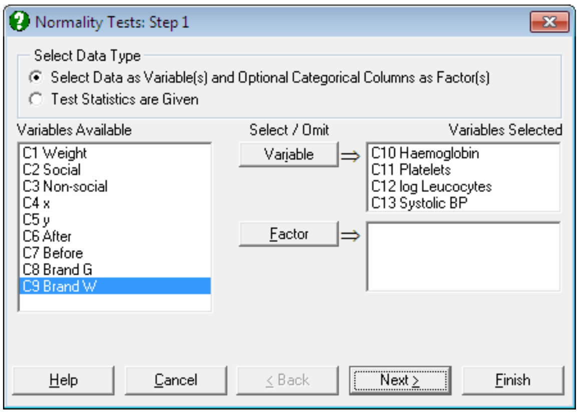
	Valid Cases	**Missing**	**Mean**	**Standard Deviation**
Distance (Uniform)	20	0	2.7350	1.5246
Age (Uniform)	30	0	61.4333	25.0430
Distance (Normal)	20	0	2.7350	1.5246
Age (Normal)	30	0	61.4333	25.0430
Distance – Age				

	Abs(Maximum difference)	**Test Statistic**	**2-Tail Probability**	*** Lilliefors Probability**
Distance (Uniform)	0.2217	0.9915	0.2793	
Age (Uniform)	0.3182	1.7428	0.0046	
Distance (Normal)	0.1112	0.4975	0.9655	0.2000
Age (Normal)	0.1921	1.0522	0.2182	0.0062
Distance - Age	1.0000	3.4641	0.0000	

* Lilliefors probability = 0.2 means 0.2 or greater.

6.3.3. Normality Tests

Four commonly used tests of normality can be performed; Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling. The test statistics are displayed with their probability values and optionally, with basic sample statistics (number of cases, mean and standard deviation). The latter three tests (Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling) are also known as EDF tests since they are based on the Empirical Distribution Function (EDF). These tests are based on the assumption that population mean and standard deviation are not known and are to be estimated from the data. Without this assumption the probability values may be quite different. See Stephens, M. A. (1974), (1986).



Normality Tests: Step 1

Select Data Type

● Select Data as Variable(s) and Optional Categorical Columns as Factor(s)

○ Test Statistics are Given

Variables Available

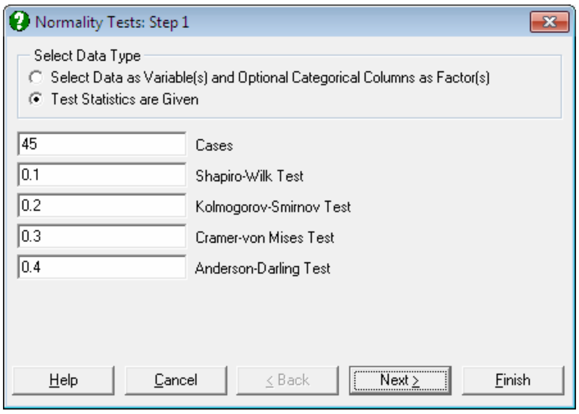
C1 Weight
C2 Social
C3 Non-social
C4 x
C5 y
C6 After
C7 Before
C8 Brand G
C9 Brand W

Variables Selected

C10 Haemoglobin
C11 Platelets
C12 log Leucocytes
C13 Systolic BP

Multisample data can be entered either in the form of multiple columns (not necessarily of equal length) or data columns classified by one or more factor columns (see 6.0.4. Multisample Tests). If at least one factor column is selected, then a further dialogue will pop up asking for the combination of factor levels to be included.

It is also possible to use this procedure as a probability calculator when the data is not available but the number of cases and the test statistics are known.



Normality Tests: Step 1

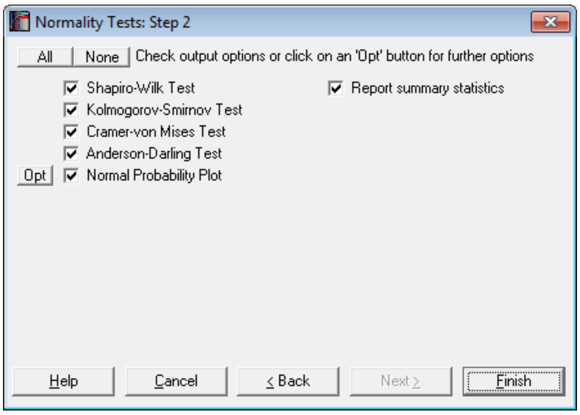
Select Data Type

○ Select Data as Variable(s) and Optional Categorical Columns as Factor(s)

● Test Statistics are Given

45	Cases
0.1	Shapiro-Wilk Test
0.2	Kolmogorov-Smirnov Test
0.3	Cramer-von Mises Test
0.4	Anderson-Darling Test

On the Output Options Dialogue, you can select only the desired test statistics and their probabilities to be displayed in the output. As of this version of UNISTAT, it is also possible to display a Normal Probability Plot for the variables selected. The Anderson-Darling Test probability is also reported on the graph. For further information see 5.3.2. Normal Probability Plot



Normality Tests: Step 2

Check output options or click on an 'Opt' button for further options

☑ Shapiro-Wilk Test

☑ Kolmogorov-Smirnov Test

☑ Cramer-von Mises Test

☑ Anderson-Darling Test

Opt ☑ Normal Probability Plot

☑ Report summary statistics

When the Report summary statistics box is checked, summary information (number of valid cases, missing observations or pairs, mean and standard deviation) about the selected variables is also displayed.

6.3.3.1. Shapiro-Wilk Test

The test statistic and its probability value are computed according to Royston (1995), which works accurately for samples with 3 to 5000 observations:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_i\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

where a$_i$ are some coefficients dependent on the sample size.

Earlier versions of UNISTAT also featured the classic Shapiro-Wilk (1965) normality test for samples with 50 or less observations and an overall test of normality by Shapiro & Wilk (1968), when all sample sizes are between 7 and 20 inclusive. If you wish to replace Royston (1995) method with the classic Shapiro-Wilk (1965) and its accompanying overall normality test, enter the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] section:

```
OverallNormality=1
```

For a description of the classical test see 10.1.3.1. Normality Tests for Bioassays.

Example

The two samples given here are taken from Shapiro, S. S. and M. B. Wilk. (1965), p. 606.

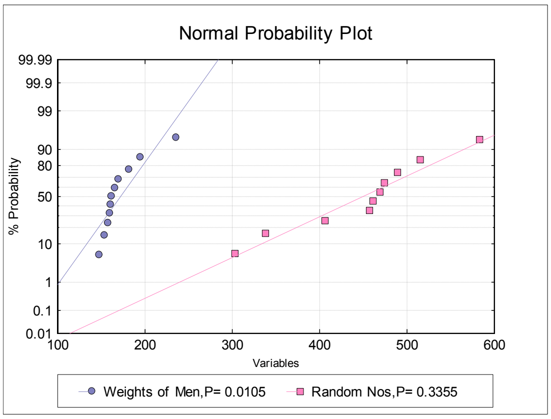
Open GOODFIT, select Statistics 1 → Goodness of Fit Tests → Normality Tests and *Weights of Men* (*C5*) and *Random Nos* (*C6*) as [Variable]s to obtain the following results:

Normality Tests

Smaller probabilities indicate non-normality.

	Valid Cases	Missing	Mean	Standard Deviation	Shapiro-Wilk	Prob
Weights of Men	11	0	172.0000	24.9520	0.7888	0.0067
Random Nos	10	0	449.5000	82.0762	0.9427	0.5831

	Kolmogorov-Smirnov	Prob	Cramer-von Mises	Prob	Anderson-Darling	Prob
Weights of Men	0.2592	0.0374	0.1639	0.0125	0.9468	0.0105
Random Nos	0.2364	0.1163	0.0718	0.2377	0.3775	0.3355



6.3.3.2. Kolmogorov-Smirnov Test

The difference between cumulative proportions of the sample and the corresponding cumulative proportions from the normal distribution are computed and the absolute value of their maximum difference is reported (see 6.3.2. Kolmogorov-Smirnov Tests):

$$K = | cpo_i - cpe_i |_{max}$$

where cpo_i and cpe_i are the ith observed and normal cumulative proportions respectively.

The probability value for this test is computed according to Lilliefors (1967), with the correction introduced by Dallal and Wilkinson (1986). The maximum probability value that can be computed is 0.2. Therefore, when the probability value is reported as 0.2, this should be interpreted as probability ≥ 0.2.

6.3.3.3. Cramer-von Mises Test

For a one sample case, where population mean and standard deviation are not known, but are estimated from the data, the Cramer-von Mises test statistic is defined as:

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(Z_i - \frac{2i-1}{2n} \right)^2$$

6.3.3.4. Anderson-Darling Test

For a one sample case, where population mean and standard deviation are not known, but are estimated from the data, the Anderson-Darling test statistic is defined as:

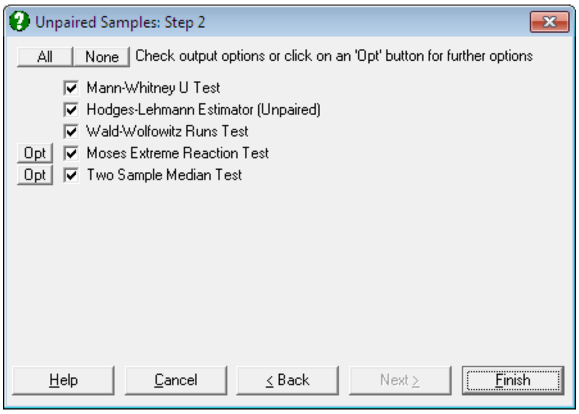
$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\mathrm{Ln}(Z_i) + (2n+1-2i)\mathrm{Ln}(1-Z_i)\right]$$

6.4. Nonparametric Tests with One or Two Samples

One and two sample nonparametric tests and tests on proportions are grouped in this section.

6.4.1. Unpaired Samples

Data in one of the three types supported for Two Sample Tests can be used for these tests. Missing values are omitted by case.



Check output options or click on an 'Opt' button for further options

☑ Mann-Whitney U Test
☑ Hodges-Lehmann Estimator (Unpaired)
☑ Wald-Wolfowitz Runs Test
☑ Moses Extreme Reaction Test
☑ Two Sample Median Test

Moses Extreme Reaction Test and Two Sample Median Test have a further dialogue each, which can be accessed by clicking on their [Opt] buttons situated to the left of the check boxes. If [Finish] is clicked before [Opt], then the program will use the default values suggested by the program, without displaying their further dialogues.

6.4.1.1. Mann-Whitney U Test

This test is used to determine whether two independent random samples have been drawn from the same population. The null hypothesis tested is that "the population relative frequency distributions are identical" against the alternative hypothesis that "they are different" (two-tailed test).

The output includes the number of cases, rank sums, mean ranks, and U scores for the two samples as well as the test statistic, correction for ties and the asymptotic (normal and t-) and exact two-tailed probability values, with and without continuity correction.

The test statistic U for sample 1 is obtained by summing the number of times cases in sample 1 are smaller than cases in sample 2. U for sample 2 is found similarly. The smaller U value is chosen as the test statistic. A small or large U value indicates that the two samples are not similarly distributed. U values can also be calculated as:

$$U_1 = n_1 n_2 + n_1(n_1+1)/2 - R_1$$

$$U_2 = n_1 n_2 + n_2(n_2+1)/2 - R_2$$

where R_1 and R_2 are the sum of ranks for groups 1 and 2 respectively.

The program will compute and display a Z statistic which is corrected for ties and with no continuity correction as:

$$Z = \frac{U - E_{MW}}{SD_{MW}}$$

where the mean of the Mann-Whitney distribution is given as:

$$E_{MW} = \frac{n_1 n_2}{2}$$

and its standard deviation as:

$$SD_{MW} = \sqrt{\frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{N_1} \left(w_{i1} - E_{MW} / n_1\right)^2}$$

where:

$$n = n_1 + n_2$$

and w_{i1} is the rank of the ith case belonging to group 1, supposing that group 1 has the smaller U.

The Z statistic with continuity correction is:

$$Z = \frac{U - E_{MW} - \mathrm{Sign}\left(U - E_{MW}\right)/2}{SD_{MW}}$$

One- and two-tailed probabilities from normal and t-distributions (with $n-1$ degrees of freedom) are displayed for Z-statistic without and with continuity correction.

The following alternative definition of the standard deviation (given by Armitage & Berry (2002) p. 276 and Gardner & Altman (2000) p. 40) is not used here as it does not take ties into consideration:

$$SD_{MW} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

An exact p-value is also computed which is accurate for data sets with or without ties. By default, it is reported for $n \leq 150$, though this limit can be changed by the user. To do this, the following line should be entered and edited in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
WMWMaxExactSize=150
```

This limit can be extended if there are no ties in data. However, if ties exist, the exact p-value for $n > 150$ may take a long time to compute.

It is also possible to save the complete exact one-tailed cumulative probability distribution of the test statistic in its rank sum form by including the following line in the [Options] section of *Unistat65.ini*:

```
WMWSaveDist=1
```

By default, the distribution will be saved to the following file:

Documents\Unistat65\WMWExactDist.txt

This file name can be changed by entering and editing the following line in the [Options] section of *Unistat65.ini*:

```
WMWSaveDistFile=..\Documents\Unistat65\WMWExactDist.txt
```

Example 1

Example 10.3 on p. 279 from Armitage & Berry (2002). An estimate of the median difference is required. A comparison of 32 inpatients and 32 outpatients is made.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Samples. Select *Inpatients* (*C16*) and *Outpatients* (*C17*) as [Variable]s and check only the Mann-Whitney U Test output option to obtain the following results:

Unpaired Samples

Mann-Whitney U Test

	Cases	Rank Sum	Mean Rank	U
Inpatients	32	858.0000	26.8125	694.0000
Outpatients	32	1222.0000	38.1875	330.0000
Total	64	2080.0000	32.5000	

Correction for Ties =	410.5000

	U	Test Statistic	1-Tail Probability	2-Tail Probability
Asymptotic Normal	330.0000	-2.4670	0.0068	0.0136
Asymptotic Normal with CC		-2.4603	0.0069	0.0139
Asymptotic t		-2.4670	0.0082	0.0164
Asymptotic t with CC		-2.4603	0.0083	0.0166
Exact			0.0065	0.0131

It is concluded that the medians of the two samples are significantly different. A t-test cannot detect a significant difference between the two sample means. This example shows the power of Mann-Whitney U Test when the assumption of normality fails.

Example 2

Example 8.11 on p. 164 from Zar, J. H. (2010). The null hypothesis "there is no difference between the heights of male and female students" is tested.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Samples. Select *Males* (*C18*) and *Females* (*C19*) as [Variable]s and check only the Mann-Whitney U Test output option to obtain the following results:

Unpaired Samples

Mann-Whitney U Test

	Cases	Rank Sum	Mean Rank	U
Males	7	60.0000	8.5714	3.0000
Females	5	18.0000	3.6000	32.0000
Total	12	78.0000	6.5000	

Correction for Ties =	0.0000

	U	Test Statistic	1-Tail Probability	2-Tail Probability
Asymptotic Normal	3.0000	-2.3548	0.0093	0.0185
Asymptotic Normal with CC		-2.2736	0.0115	0.0230
Asymptotic t		-2.3548	0.0191	0.0382
Asymptotic t with CC		-2.2736	0.0220	0.0440
Exact			0.0088	0.0177

Zar reports the exact two-tailed probability as 0.018 and since this is less than 0.05, we reject the null hypothesis.

If the WMWSaveDist=1 line is included in the [Options] section of *Documents\Unistat65\Unistat65.ini* file, the exact one-tailed cumulative distribution of the rank sum is saved to the WMWExactDist.txt file as follows:

Rank Sum	One Tail Probability
28	1.26262626262626E-03
29	2.52525252525253E-03
30	5.05050505050505E-03
31	8.83838383838384E-03
32	1.51515151515152E-02
33	0.023989898989899
34	3.66161616161616E-02
35	0.053030303030303
36	7.44949494949495E-02
37	0.101010101010101
38	0.133838383838384
39	0.171717171717172
40	0.215909090909091
41	0.265151515151515
42	0.319444444444444
43	0.377525252525252
44	0.438131313131313
45	0.5
46	0.561868686868687
47	0.622474747474747
48	0.680555555555555
49	0.734848484848485
50	0.784090909090909
51	0.828282828282828
52	0.866161616161616
53	0.898989898989899
54	0.92550505050505
55	0.946969696969697
56	0.963383838383838
57	0.976010101010101
58	0.984848484848485
59	0.991161616161616
60	0.994949494949495
61	0.997474747474747
62	0.998737373737374
63	1

6.4.1.2. Hodges-Lehmann Estimator (Unpaired)

If the product of the two sample sizes does not exceed 2 x 10^9 then an estimate of the difference between the two sample medians and its confidence interval are computed.

First, all n_1 x n_2 differences between each pair of numbers from the two samples are sorted in increasing order. Then, the median (the Hodges-Lehmann estimator or the shift parameter) is found.

The output includes a table where the minimum, maximum, mean and standard deviation of the rank sum are displayed. The mean of the rank sum is different from the mean of the Mann-Whitney statistic, whereas their standard deviations are the same.

The limits of the asymptotic confidence interval are the Kth smallest and the Kth largest difference:

$$K = E_{MW} - Z_{1-\alpha/2} SD_{MW}$$

where K is rounded up to the nearest integer and the mean and standard deviation of the Mann-Whitney statistic are as given in the previous section.

The exact confidence interval is also displayed, which is based on the exact distribution of the Mann-Whitney statistic. To determine the lower bound of the exact interval (the K_lth smallest difference), find k_l such that:

$$\Pr(U \geq k_1) \leq \alpha/2$$

round k_l up to the nearest integer and calculate:

$$K_l = n_1 n_2 - k_l + 1$$

The upper limit is determined likewise, for:

$$\Pr(U \leq k_u) \leq \alpha/2 \,.$$

For the paired case of this test see 6.4.2.2. Hodges-Lehmann Estimator (Paired).

Example 1

Example 10.4 on p. 283 from Armitage & Berry (2002). Gain in weight of rats receiving diets with high and low protein content are measured. The null hypothesis “there is no difference in median weights” is tested at 95% level.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Samples. Select *High* (*C1*) and *Low* (*C2*) as [Variable]s and check the Hodges-Lehmann Estimator (Unpaired) output option to obtain the following results:

Unpaired Samples

Hodges-Lehmann Estimator (Unpaired)

For High and Low

	Minimum	Maximum	Mean	Standard Deviation
Rank Sum	78.0000	162.0000	120.0000	11.8270

	K	Difference Between Medians	Lower 95%	Upper 95%
Asymptotic	19	18.5000	-3.0000	40.0000
Exact			-3.0000	40.0000

6.4.1.3. Wald-Wolfowitz Runs Test

The null hypothesis "two independent samples have been drawn from the same population" is tested against the alternative hypothesis "they differ in respect of their medians, variability or skewness". It is assumed that the variable under consideration has a continuous distribution.

All cases from the two samples are sorted together. If the two distributions are similar, then cases belonging to two samples must be scattered randomly. Then the program counts the number of runs (i.e. the number of groups of cases which belong to the same sample). If there are ties between cases belonging to two samples then the minimum and the maximum possible number of runs are reported separately. Two sets of results using the normal approximation are reported.

Asymptotic without Continuity Correction: In this case the Z-statistic is defined as:

$$Z = \frac{R - M}{s}$$

where:

$$M = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$s^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

Asymptotic with Continuity Correction: The Z-statistic with continuity correction is defined as:

$$Z = \frac{R - M - \text{Sign}(R - M)/2}{s}$$

In some applications, the test statistic with continuity correction is reported for $n_1 + n_2 < 50$ and without continuity correction otherwise. The same normal approximation is also used for the Runs Test.

Exact: The exact one- and two-tailed probabilities are reported. Their use is recommended for $n \le 30$.

Data in one of the three types supported for Two Sample Tests can be used for this test. Missing values are omitted by case.

Example

Table 100 on p. 251 from Cohen, L. & M. Holliday (1983). Aggression scores in 20 nursery school children following violent (Condition 1) and neutral (Condition 0) cartoons are given.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Samples. Select *Score* (*C13*) as [Variable] and *Condition* (*C14*) as [Factor]. From the next dialogue uncheck the Run a separate analysis for each option selected box and select only the Wald-Wolfowitz Runs Test option:

Unpaired Samples

Wald-Wolfowitz Runs Test

Data variable: Score
Subsample selected by: Condition

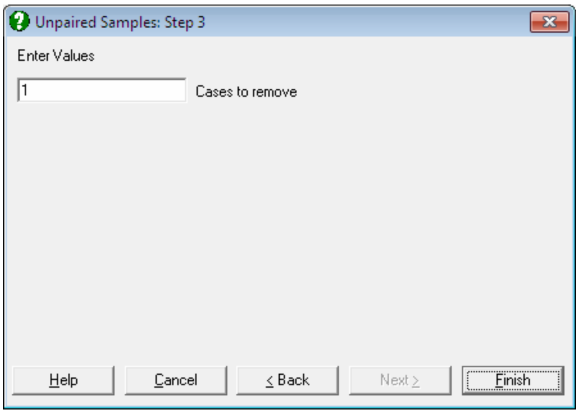
Condition	Cases	Mean	Standard Deviation	Standard Error
0	10	24.2000	19.5209	6.1731
1	10	46.2000	14.1327	4.4692
Total	20	35.2000	17.0411	3.8105

	Number of Runs	Z-Statistic	1-Tail Probability	2-Tail Probability
Asymptotic	8	-1.3784	0.0840	0.1681
Asymptotic with CC		-1.1487	0.1253	0.2507
Exact			0.1276	

This result is not significant at the 10% level. Hence do not reject the null hypothesis “watching violent cartoons does not cause a significant change in the aggression of nursery school children”.

6.4.1.4. Moses Extreme Reaction Test

This test is used to determine the difference in range between two samples. Cases from the two samples are ranked together. Ranks corresponding to the smallest and largest group 1 cases are determined. The span is the difference between these two ranks plus one.


Unpaired Samples: Step 3

Enter Values

1 Cases to remove

The program will prompt for the number of cases to be trimmed from either side of the span. The suggested number is either 1 or the integer closest to 5% of the number of cases in group 1, whichever is larger. However, this number can be changed by the user. The output includes the number of cases in two groups as well as the span and the one-tailed probability.

The exact one-tailed probability is computed for n ≤ 150. This limit can be changed by entering the following line with the appropriate number in the [Options] section of *Documents**Unistat65**Unistat65.ini* file:

WMWMaxExactSize=150

Example

Open DEMODATA and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Samples. Select *Wages* (*C2*) and *Energy* (*C3*) as [Variable]s and click on the [Opt] button next to the Moses Extreme Reaction Test option. Accept the default value of 3 from the next dialogue.

Unpaired Samples

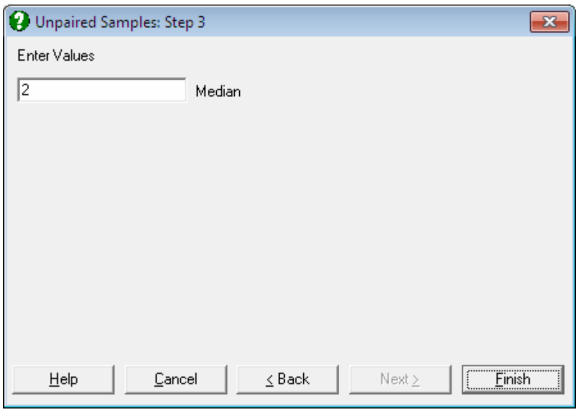
Moses Extreme Reaction Test

	Cases
Wages	57
Energy	57
Total	114

	span	1-Tail Probability
whole of group 1	108	0.0567
3 case(s) removed from ends	92	0.0302

6.4.1.5. Two Sample Median Test

This test is used to determine whether two samples are drawn from populations with similar medians. The median for the two combined samples is calculated, the two samples are dichotomised and a 2 x 2 table is formed. It is possible to edit the computed median and to enter any values. The output includes the generated 2 x 2 table, chi square test statistics without and with a continuity correction and the exact probabilities.


Unpaired Samples: Step 3

Enter Values

Median: 2

Asymptotic without Continuity Correction: The following chi-square statistic with one degree of freedom is compared with the chi-square distribution:

$$C = \frac{n\left(g_1(n_2 - g_2) - g_2(n_1 - g_1)\right)^2}{(g_1 + g_2)(n_1 + n_2 - g_1 - g_2)n_1 n_2}$$

Asymptotic with Continuity Correction: In this case the numerator is slightly different:

$$C = \frac{n\left(\left| g_1(n_2 - g_2) - g_2(n_1 - g_1)\right| - n/2\right)^2}{(g_1 + g_2)(n_1 + n_2 - g_1 - g_2)n_1 n_2}$$

where g_1 and g_2 are the number of cases greater than the median in samples 1 and 2 respectively.

Exact: Two-tailed and table probabilities are reported using Fisher's exact probability formula (see 6.4.5.2. Fisher's Exact Test).

Example

Example 8.18 on p. 156 from Zar, J. H. (1999). The null hypothesis "the medians of the two sampled populations are equal" is tested.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Samples and select *Assistant A* (*C20*) and *Assistant B* (*C21*) as [Variable]s. Note that these are the rank data in descending order. Select the Two Sample Median Test output option to obtain the following results:

Unpaired Samples

Two Sample Median Test

	> Median	<=Median	Total
Assistant A	6	5	11
Assistant B	6	8	14
Total	12	13	25

	Median	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Asymptotic	12.5000	0.3372	1	0.5615
Asymptotic with CC		0.0315	1	0.8592

	2-Tail Probability	Table Probability
Fisher's Exact	0.6951	0.2668

Since P > 0.05, do not reject the null hypothesis. In the 5th edition of *Biostatistical Analysis* (2010) Example 8.15 on p. 173, Zar employs a different method where observations at the median are omitted. With this approach the total number of valid cases is 23 and the chi-squared statistic with continuity correction is 0.473.

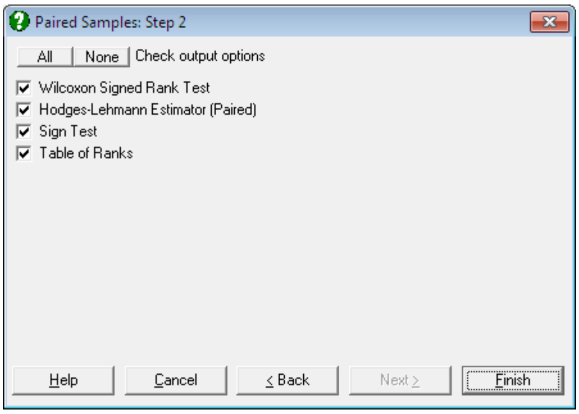
6.4.2. Paired Samples

Like the nonparametric tests on unpaired samples of the previous section, the tests in this section are also used to assess the significance of the difference between population distributions of two samples. In this case the two samples are assumed to consist of matched pairs.

In general, a test is run on paired data by selecting two numeric data columns as [Variable]s. When three or more columns are selected the test will be performed on all possible pairs with equal length (see 6.0.3. Tests with Paired Data). The missing values are omitted pairwise.

Despite the section title Paired Samples, it is also possible to select a single column. When only one column is selected, the test is performed against a hypothetical second variable consisting of zeros.

An Output Options Dialogue offering four options is displayed next.


Paired Samples: Step 2

All | None | Check output options

☑ Wilcoxon Signed Rank Test
☑ Hodges-Lehmann Estimator (Paired)
☑ Sign Test
☑ Table of Ranks

6.4.2.1. Wilcoxon Signed Rank Test

This test is used to assess the significance of the difference between population distributions of the two samples consisting of matched pairs. The absolute values of the difference between the pairs are ranked and the rank sums of negative and positive differences are computed. The signed ranks can be displayed using the Table of Ranks option below.

A very small or a very large sum indicates that the two samples do not have similar distributions. The smaller of the two values is selected as the test statistic.

Missing values are omitted pairwise. The output includes a table displaying the number of positive and negative differences, rank sums and mean ranks. One- and two-tailed asymptotic probabilities are reported without and with continuity correction, as well as the one- and two-tailed exact probabilities.

Asymptotic without Continuity Correction: The Z-statistic is defined as:

$$Z = \frac{R - E_{WSR}}{SD_{WSR}}$$

Asymptotic with Continuity Correction: In this case the Z-statistic is:

$$Z = \frac{R - E_{WSR} - \mathrm{Sign}\left(R - E_{WSR}\right)/2}{SD_{WSR}}$$

where the mean of the Wilcoxon Signed Rank distribution is given as:

$$E_{WSR} = \frac{n(n+1)}{4}$$

and its standard deviation:

$$SD_{WSR} = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{T}{48}}$$

R is the smaller sum of the like-signed ranks (the test statistic) and T is the sum of t^3 - t where t is the number of ties at a given rank.

If n > 20 then the Z statistic provides a good approximation for the distribution of the test statistic. The exact p-value is reported for n ≤ 150 and it is accurate for data sets with or without ties. To change this limit, the following line should be entered and edited in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
WMWMaxExactSize=150
```

It is also possible to save the complete exact one-tailed cumulative probability distribution of the test statistic by including the following line in the [Options] section of *Unistat65.ini*:

```
WMWSaveDist=1
```

By default, the distribution will be saved to the following file:

..*Documents**Unistat65**WMWExactDist.txt*

This file name can be changed by entering and editing the following line in the [Options] section of *Documents**Unistat65**Unistat65.ini* file:

```
WMWSaveDistFile=..\Documents\Unistat65\WMWExactDist.txt
```

Example 1

Example 10.2 on p. 275 from Armitage & Berry (2002). The null hypothesis "there is no difference between the effects of the drug and the placebo" is tested.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Samples. Select *Drug* (*C5*) and *Placebo* (*C6*) as [Variable]s and check only the Wilcoxon Signed Rank Test output option to obtain the following results:

Paired Samples

Wilcoxon Signed Rank Test

For Drug and Placebo

	Cases	Rank Sum	Mean Rank
Negative Differences	6	38.0000	6.3333
Positive Differences	4	17.0000	4.2500
Total	10	55.0000	5.5000

Correction for Ties = 1.5000

	W	Z-Statistic	1-Tail Probability	2-Tail Probability
Asymptotic	17.0000	-1.0273	0.1521	0.3043
Asymptotic with CC		-1.0787	0.1404	0.2807
Exact			0.1611	0.3223

Since the two-tailed probability is far greater than 5% the test result is not significant. Therefore, do not reject the null hypothesis.

If the WMWSaveDist=1 line is included in the [Options] section of *Documents**Unistat65**Unistat65.ini* file, the exact one-tailed cumulative distribution of the rank sum is saved to the WMWExactDist.txt file as follows (shortened):

Rank Sum	One Tail Probability
0	0.0009765625
2.5	0.0048828125
5	0.01171875
6.5	0.013671875
7.5	0.021484375
8	0.0224609375
9	0.0302734375
9.5	0.0322265625
10	0.0390625
10.5	0.04296875
11.5	0.056640625
12	0.064453125
12.5	0.068359375
13	0.076171875
…	…
…	…
42	0.931640625
42.5	0.935546875
43	0.943359375
43.5	0.95703125
44.5	0.9609375
45	0.9677734375
45.5	0.9697265625
46	0.9775390625
47	0.978515625
47.5	0.986328125
48.5	0.98828125
50	0.9951171875
52.5	0.9990234375
55	1

Example 2

Example 9.4, p. 185 from Zar, J. H. (2010). The null hypothesis “deer hindleg length is the same as foreleg length” is tested.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Samples. Select *Hindleg* (*C7*) and *Foreleg* (*C8*) as [Variable]s and check only the Wilcoxon Signed Rank Test output option to obtain the following results:

Paired Samples

Wilcoxon Signed Rank Test

For Hindleg and Foreleg

	Cases	Rank Sum	Mean Rank
Negative Differences	2	4.0000	2.0000
Positive Differences	8	51.0000	6.3750
Total	10	55.0000	5.5000

Correction for Ties = 0.7500

	W	Z-Statistic	1-Tail Probability	2-Tail Probability
Asymptotic	4.0000	-2.3536	0.0093	0.0186
Asymptotic with CC		-2.4047	0.0081	0.0162
Exact			0.0059	0.0117

Since the probability is less than 5%, reject the null hypothesis.

6.4.2.2. Hodges-Lehmann Estimator (Paired)

This statistic will estimate the median difference. First, the difference between each pair is computed for n cases. Then the averages of all combinations of differences (also known as Walsh averages) are computed. The $n(n + 1)/2$ averages are sorted in increasing order and their median (the Hodges-Lehmann estimator or the shift parameter) is found.

The output includes a table where the minimum, maximum, mean and standard deviation of the test statistic are displayed.

The limits of the asymptotic confidence interval are the K*th smallest and the K*th largest difference:

$$K^* = E_{WSR} - Z_{1-\alpha/2} SD_{WSR}$$

where K* is rounded up to the nearest integer and the mean and standard deviation of the signed rank statistic are as given in the previous section.

The exact confidence interval is also displayed, which is based on the exact distribution of the test statistic. To determine the lower bound of the exact interval (the K^*_l smallest difference), find K^*_l such that:

$$\Pr(R \le K_l^*) \le \alpha/2$$

round K^*_l up to the nearest integer. The upper limit is determined likewise, for:

$$\Pr(R \ge K_u^*) \le \alpha/2$$

For the unpaired case of this test see 6.4.1.2. Hodges-Lehmann Estimator (Unpaired).

Example 1

Example 10.2 on p. 276 from Armitage & Berry (2002). An estimate of the median difference is required.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Samples. Select *Drug* (*C5*) and *Placebo* (*C6*) as [Variable]s and check only the Hodges-Lehmann Estimator (Paired) output option to obtain the following results:

Paired Samples

Hodges-Lehmann Estimator (Paired)

For Drug and Placebo

	Minimum	Maximum	Mean	Standard Deviation
Rank Sum	0.0000	55.0000	27.5000	9.7340

	K*	Median of Walsh Differences	Lower 95%	Upper 95%
Asymptotic	9	-1.0000	-4.5000	1.0000
Exact			-4.5000	1.0000

Example 2

Table 5.3 on p. 42, Gardner & Altman (2000). Beta endorphin concentrations in subjects before and after running in a half marathon are measured. We would like to estimate the sample median for the pairwise averages between differences and the 95% confidence intervals.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Samples. Select *After* (*C9*) and *Before* (*C10*) as [Variable]s and check only the Hodges-Lehmann Estimator (Paired) output option to obtain the following results:

Paired Samples

Hodges-Lehmann Estimator (Paired)

For After and Before

	Minimum	Maximum	Mean	Standard Deviation
Rank Sum	0.0000	66.0000	33.0000	11.2472

	K*	Median of Walsh Differences	Lower 95%	Upper 95%
Asymptotic	11	18.8250	11.9000	25.1000
Exact			11.9000	25.1000

6.4.2.3. Sign Test

This is a weaker version of Wilcoxon Signed Rank Test. The negative and positive differences are counted and the ties are ignored. Since the probability that either sum exceeds the other is 0.5, it is equivalent to a Binomial Test with p = 0.5.

The exact probability is calculated from the binomial distribution. The asymptotic probability is based on a normal approximation:

$$Z = \frac{\mathrm{Max}(n_p, n_n) - 0.5(n_p + n_n) - 0.5}{0.5\sqrt{n_p + n_n}}$$

where n_p and n_n are the numbers of positive and negative differences respectively. In both cases a two-tailed probability is reported. The output consists of the number of negative and positive differences, number of ties, test statistic and the exact binomial and asymptotic two-tailed probabilities.

Example 1

Example 10.1 on p. 274 from Armitage & Berry (2002). The null hypothesis "there is no difference between the effects of the drug and the placebo" is tested.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Samples. Select *Drug* (*C5*) and *Placebo* (*C6*) as [Variable]s and check only the Sign Test output option to obtain the following results:

Paired Samples

Sign Test

For Drug and Placebo

	Cases
Negative Differences	6
Positive Differences	4
Ties	0
Total	10

	Value	Z-Statistic	1-Tail Probability	2-Tail Probability
Asymptotic	6.0000	0.3162	0.3759	0.7518
Exact			0.3770	0.7539

Example 2

Example 24.10, p. 538 from Zar, J. H. (2010). The null hypothesis is "deer hindleg length is the same as foreleg length".

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Samples. Select *Hindleg* (*C7*) and *Foreleg* (*C8*) as [Variable]s and check only the Sign Test output option to obtain the following results:

Paired Samples

Sign Test

For Hindleg and Foreleg

	Cases
Negative Differences	2
Positive Differences	8
Ties	0
Total	10

	Value	Z-Statistic	1-Tail Probability	2-Tail Probability
Asymptotic	8.0000	1.5811	0.0569	0.1138
Exact			0.0547	0.1094

Since the two-tailed exact binomial probability is greater than 5%, do not reject the null hypothesis.

6.4.2.4. Table of Ranks

When this output option is selected, a table will be formed displaying the two columns, their differences, the signed rank of their absolute value and the differences ordered in ascending order. The signed ranks are the intermediate values used in Wilcoxon Signed Rank Test.

The last column, ordered difference, can be used to run a Walsh Test. This test is used to determine whether two samples have been drawn from symmetrically distributed populations. It is assumed that the distributions are continuous. The test can be performed meaningfully only on small data sets with n ≤ 15.

First the signed difference for each matched pair is computed and then differences are ranked in increasing size. The null hypothesis is that "the average of differences is equal to zero" against the alternative hypothesis that "the population mean is other than zero". Output displays the two data columns, their differences and the ranked difference. For probability values, tables for the Walsh Test must be consulted.

Example

Table 99 on p. 248 from Cohen, L. & M. Holliday (1983). With and without practice errors in a manual dexterity selection test are given for 11 candidates.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Samples. Select *Without* (*C11*) and *With* (*C12*) as [Variable]s and check only the Table of Ranks output option to obtain the following results:

Paired Samples

Table of Ranks

For Without and With

Row	Without	With	Difference	Signed Rank	Ordered Difference
1	11.0000	6.0000	5.0000	7.5000	-1.0000
2	4.0000	2.0000	2.0000	3.0000	0.0000
3	5.0000	4.0000	1.0000	1.5000	1.0000
4	9.0000	3.0000	6.0000	9.5000	2.0000
5	5.0000	5.0000	0.0000	0.0000	3.0000
6	13.0000	7.0000	6.0000	9.5000	4.0000
7	5.0000	6.0000	-1.0000	-1.5000	4.0000
8	7.0000	3.0000	4.0000	5.5000	5.0000
9	8.0000	4.0000	4.0000	5.5000	5.0000
10	10.0000	7.0000	3.0000	4.0000	6.0000
11	12.0000	7.0000	5.0000	7.5000	6.0000

Consult tables for critical values of the Walsh Test with n = 11. We see from the table that a two-tailed test with n =11 is significant at the 5.6% level if:

$$\max[d_7, \tfrac{1}{2}(d_5+d_{11})] < 0 \text{ or } \min[d_5, \tfrac{1}{2}(d_1+d_7)] > 0$$

In this example:

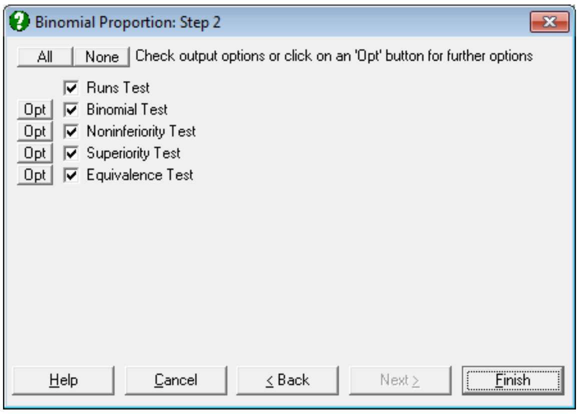
$$\max[4, \tfrac{1}{2}(3+6)] < 0 \text{ or } \min[3, \tfrac{1}{2}(-1+4)] > 0$$
$$\max[4, 4\tfrac{1}{2}] < 0 \text{ or } \min[3, 2\tfrac{1}{2}] > 0$$

Since 3 > 0, this result is significant at the 5.6% level. Hence reject the null hypothesis that “manual dexterity does not change with practice”.

6.4.3. Binomial Proportion

One of the three data types supported for binary data can be selected (see 6.0.6. Tests with Binary Data).

You can set the value of some user-defined parameters (such as expected proportion) using the [Opt] buttons in the Output Options Dialogue. If [Finish] is clicked instead, the default value suggested by the program will be used.


Binomial Proportion: Step 2

Check output options or click on an 'Opt' button for further options

☑ Runs Test
☑ Binomial Test
☑ Noninferiority Test
☑ Superiority Test
☑ Equivalence Test

6.4.3.1. Runs Test

This test is used to determine the randomness of cases belonging to two outcomes within a sample. The number of runs R (i.e. the number of groups of cases which belong to the same group) in the raw data is counted. If the last data option Test Statistics are Given is selected then an [Opt] button will also be available for Runs Test, allowing entry of a number of runs value.

Two sets of results are reported using the normal approximation.

Asymptotic without Continuity Correction: In this case the Z-statistic is defined as:

$$Z = \frac{R - M}{s}$$

where:

$$M = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

$$s^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

Asymptotic with Continuity Correction: The Z-statistic with continuity correction is defined as:

$$Z = \frac{R - M - \mathrm{Sign}(R - M)/2}{s}$$

In some applications, the test statistic with continuity correction is reported for $n_1 + n_2 < 50$ and without continuity correction otherwise.

The output includes the number of cases in each group as well as the number of runs. The same normal approximation is also used for the Wald-Wolfowitz Runs Test.

Example

Example 25.8, p. 598 from Zar, J. H. (2010). The null hypothesis "the sequence is in a random order" is tested.

Open NONPAR12 and select Statistics 1 → Nonparametric Tests (1-2 Samples) → Binomial Proportion, the data option 1 Column Contains Two Categories. Then select *Species* (*S15*) as [Column 1] and check only the Runs Test output option to obtain the following results:

Binomial Proportion

Data option: Column Contains Two Categories

	Cases
Species = A	9
Species = B	13
Total	22

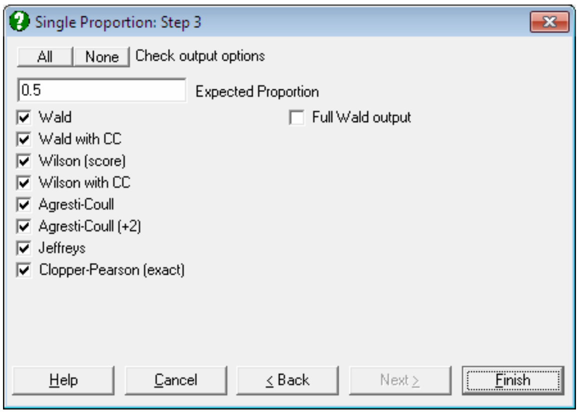
Runs Test

	Number of Runs	**Z-Statistic**	**1-Tail Probability**	**2-Tail Probability**
Asymptotic	8	-1.4197	0.0779	0.1557
Asymptotic with CC		-1.6460	0.0499	0.0998

This result is not significant at the 5% level (i.e. p > 0.05) and therefore do not reject the null hypothesis "the sequence is in a random order". Note that the number of runs is given wrongly as 9 in the book.

6.4.3.2. Binomial Test

This test compares the observed ratio of two groups (e.g. successes and failures) in a sample with a given expected ratio. There are many different methods to estimate the confidence intervals and tail probabilities for a Binomial Proportion. For details see Newcombe, R. G. (1998).



Expected Proportion: 0.5

☑ Wald
☑ Wald with CC
☑ Wilson (score)
☑ Wilson with CC
☑ Agresti-Coull
☑ Agresti-Coull (+2)
☑ Jeffreys
☑ Clopper-Pearson (exact)

☐ Full Wald output

It is also possible to perform this test for each binary factor in a 2 x 2 table using the Contingency Table and Cross-Tabulation procedures (see 6.6.2.3. 2 x 2 Table Statistics).

The further options dialogue is accessed by clicking on the [Opt] button situated to the left of the Binomial Test check box. By default, the program suggests an expected proportion of 0.5, however, this can be changed to any value between 0 and 1. The output includes a summary table for the number of cases in each group as well as the observed and expected ratios. You can choose to display any of the following eight more commonly used methods.

Wald: This is the standard asymptotic method without continuity correction. Confidence limits with normal approximation to binomial distribution are:

$$\mathrm{LL}, \mathrm{UL} = \hat{\mathrm{p}} \pm \mathrm{Z}_{\alpha/2}(\mathrm{SE}_{\hat{\mathrm{p}}})$$

where:

$$\hat{p} = n_1 / n$$

is the observed proportion and:

$$SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$$

is the sample standard error. The standard error used in confidence interval calculations is the sample standard error, which is based on the observed proportion.

On the next line, the standard error based on the null hypothesis (H_0: observed proportion is equal to the expected proportion) and the corresponding one- and two-tailed normal probabilities are reported:

$$SE_{p_0} = \sqrt{p_0(1-p_0)/n}$$

$$Z_{p_0} = (\hat{p} - p_0)/SE_{p_0}$$

where p_0 is the expected proportion.

If the Full Wald Output box is checked, then the missing parts of the Wald output, i.e. one- and two-tailed probabilities based on the sample standard error:

$$Z_{\hat{p}} = (\hat{p} - p_0)/SE_{\hat{p}}$$

and the confidence limits under H_0:

$$LL, UL = \hat{p} \pm Z_{\alpha/2}(SE_{p_0})$$

are also displayed. The user should take care with the interpretation of this extended output.

Wald with Continuity Correction: A continuity correction term of 1/(2n) is included:

$$LL, UL = \hat{p} \pm \left(Z_{\alpha/2}(SE_{\hat{p}}) + 1/(2n)\right)$$

In this case, the Z-statistic based on the expected proportion (the null hypothesis H_0: proportion is equal to the expected proportion) is:

$$Z_{p_0} = \left((\hat{p} - p_0) - \mathrm{Sign}(\hat{p} - p_0)/(2n)\right)/SE_{p_0}$$

If the Full Wald Output box is checked, then the missing parts of the Wald output, i.e. one- and two-tailed probabilities based on the sample standard error and the confidence limits under H$_0$ are also displayed. The user should take care with the interpretation of this extended output.

Wilson (score): The confidence limits without continuity correction are:

$$\mathrm{LL,UL} = \frac{2n\hat{p} + Z_{\alpha/2}^2 \pm Z_{\alpha/2}\sqrt{Z_{\alpha/2}^2 + \hat{p}(1-\hat{p})4n}}{2\left(n + Z_{\alpha/2}^2\right)}$$

Earlier versions of UNISTAT report these confidence limits for the Asymptotic without Continuity Correction case.

Wilson with Continuity Correction:

$$\mathrm{LL,UL} = \frac{2n\hat{p} + Z_{\alpha/2}^2 \pm 1 \pm Z_{\alpha/2}\sqrt{Z_{\alpha/2}^2 \pm 2 - 1/n + 4\hat{p}\left(n(1-\hat{p}) \pm 1\right)}}{2\left(n + Z_{\alpha/2}^2\right)}$$

Earlier versions of UNISTAT report these confidence limits for the Asymptotic with Continuity Correction case.

Agresti-Coull: This is similar to Wald interval but $Z_{\alpha/2}^2/2$ (i.e. half of the square of normal critical value) is added to numbers of successes and failures:

$$\mathrm{LL,UL} = \tilde{p} \pm Z_{\alpha/2}(\mathrm{SE}_{\tilde{p}})$$

where:

$$\mathrm{SE}_{\tilde{p}} = \sqrt{\tilde{p}\left(1-\tilde{p}\right)/\tilde{n}}$$

$$\tilde{n}_1 = n_1 + Z_{\alpha/2}^2/2$$

$$\tilde{n} = n + Z_{\alpha/2}^2$$

$$\tilde{p} = \tilde{n}_1/\tilde{n}$$

If the Full Wald Output box is checked, then the following Z-statistic and its one- and two-tailed probabilities are also displayed:

$$Z_{\tilde{p}} = \left(\tilde{p} - p_0\right)/\mathrm{SE}_{\tilde{p}}$$

Agresti-Coull (+2): This similar to the Agresti-Coull interval except that 2 (a crude approximation to $Z^2_{\alpha/2}/2$) is added to the numbers of successes and failures:

$$\tilde{n}_1 = n_1 + 2$$

$$\tilde{n} = n + 4$$

$$\tilde{p} = \tilde{n}_1 / \tilde{n}$$

If the Full Wald Output box is checked, then the following Z-statistic and its one- and two-tailed probabilities are also displayed:

$$Z_{\tilde{p}} = \left(\tilde{p} - p_0\right) / SE_{\tilde{p}}$$

Jeffreys: The confidence limits are defined as the following critical values from the inverse beta distribution:

$$LL = \beta_{\alpha/2, n_1+0.5, n-n_1+0.5}$$

$$UL = \beta_{1-\alpha/2, n_1+0.5, n-n_1+0.5}$$

Clopper-Pearson (exact): The exact one- and two-tailed binomial probabilities are reported. The exact confidence interval is calculated as:

$$LL = \frac{n_1}{n_1 + (n - n_1 + 1)F_{\alpha, 2n-2n_1+2, 2n_1}}$$

$$UL = \frac{n_1 + 1}{n_1 + 1 + (n - n_1)/F_{\alpha, 2n_1+2, 2n-2n_1}}$$

Example 1

Table I on p. 861 from Newcombe, R. G. (1998) where examples with five confidence intervals supported by UNISTAT are given. The following group sizes are given for the second column of the results table.

Size of Group 1	15
Size of Group 2	133
Expected Proportion	0.5

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Binomial Proportion and select the data option 3 Cell Frequencies are Given. Enter the above group sizes and check only the Binomial Test output option to obtain the following results:

Binomial Proportion

Data option: Test Statistics are Given

	Cases
Group 1	15
Group 2	133
Total	148

Binomial Test

Expected Proportion =	0.5000
Observed Proportion =	0.1014

	Proportion used in SE	Standard Error	Z-Statistic	1-Tail Probability	2-Tail Probability
Wald	0.1014	0.0248			
H0	0.5000	0.0411	-9.6995	0.0000	0.0000
Wald with CC	0.1014	0.0248			
H0	0.5000	0.0411	-9.6173	0.0000	0.0000
Wilson (score)					
Wilson with CC					
Agresti-Coull	0.1114	0.0255			
Agresti-Coull (+2)	0.1118	0.0256			
Jeffreys					
Clopper-Pearson (exact)				0.0000	0.0000

	Lower 95%	Upper 95%
Wald	0.0527	0.1500
H0		
Wald with CC	0.0494	0.1534
H0		
Wilson (score)	0.0624	0.1605
Wilson with CC	0.0598	0.1644
Agresti-Coull	0.0614	0.1615
Agresti-Coull (+2)	0.0617	0.1619
Jeffreys	0.0604	0.1576
Clopper-Pearson (exact)	0.0578	0.1617

Example 2

Example 4.6 on p. 115 from Armitage & Berry (2002). Patients' preference for two analgesic drugs, X and Y is recorded. The null hypothesis "the ratio of preferences is not different from 50%" is tested.

Size of Group 1	65
Size of Group 2	35
Expected Proportion	0.5

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Binomial Proportion and select the data option 3 Cell Frequencies are Given. Enter values in the above table and check only the Binomial Test output option to obtain the following results:

Binomial Proportion

Data option: Test Statistics are Given

	Cases
Group 1	65
Group 2	35
Total	100

Binomial Test

Expected Proportion =	0.5000
Observed Proportion =	0.6500

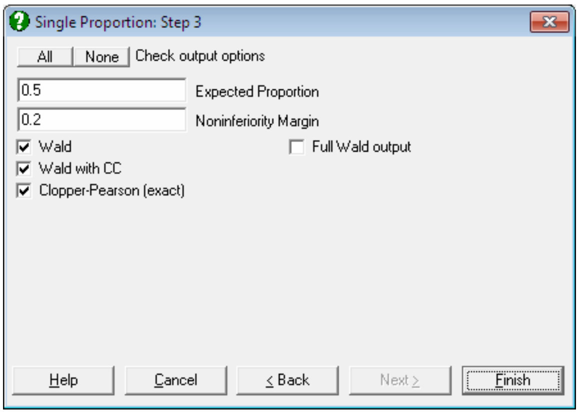
	Proportion used in SE	**Standard Error**	**Z-Statistic**	**1-Tail Probability**	**2-Tail Probability**
Wald	0.6500	0.0477			
H0	0.5000	0.0500	3.0000	0.0013	0.0027
Wald with CC	0.6500	0.0477			
H0	0.5000	0.0500	2.9000	0.0019	0.0037
Wilson (score)					
Wilson with CC					
Agresti-Coull	0.6445	0.0470			
Agresti-Coull (+2)	0.6442	0.0469			
Jeffreys					
Clopper-Pearson (exact)				0.0018	0.0035

	Lower 95%	Upper 95%
Wald H0	0.5565	0.7435
Wald with CC H0	0.5515	0.7485
Wilson (score)	0.5525	0.7364
Wilson with CC	0.5474	0.7409
Agresti-Coull	0.5524	0.7365
Agresti-Coull (+2)	0.5522	0.7362
Jeffreys	0.5533	0.7382
Clopper-Pearson (exact)	0.5482	0.7427

This result is significant at the 1% level. Hence reject the null hypothesis "the patients have no significant preference for a particular analgesic drug".

6.4.3.3. Noninferiority Test

The null hypothesis tested is that "the expected proportion is worse than the expected proportion by a given margin δ". The alternative hypothesis is "the observed proportion is not inferior."



Check output options

0.5 Expected Proportion

0.2 Noninferiority Margin

☑ Wald
☐ Full Wald output
☑ Wald with CC
☑ Clopper-Pearson (exact)

The noninferiority test similar to Binomial Test with the exception that the expected proportion is reduced by the noninferiority margin δ. Also, the Z-statistic is based on the observed proportion (unlike the Binomial Test where it is based on the expected proportion H_0). The confidence limits are reported at 1 - 2α level, rather than the usual 1 - α.

Wald: By default, the Z-statistic and the confidence interval are both based on the sample standard error:

$$Z_{\hat{p}} = \left(\hat{p} - p_0^*\right) / SE_{\hat{p}}$$

$$LL, UL = \hat{p} \pm Z_{\alpha/2}(SE_{\hat{p}})$$

where:

$$\hat{p} = n_1 / n$$

is the observed proportion and:

$$SE_{\hat{p}} = \sqrt{\hat{p}\left(1 - \hat{p}\right) / n}$$

and:

$$p_0^* = p_0 - \delta$$

If the Full Wald Output box is checked, then on the next line, the Z-statistic and confidence interval based on the noninferiority limit are also reported:

$$Z_{p_0^*} = \left(\hat{p} - p_0^*\right) / SE_{p_0^*}$$

$$LL, UL = p_0^* \pm Z_{\alpha/2}(SE_{p_0^*})$$

where:

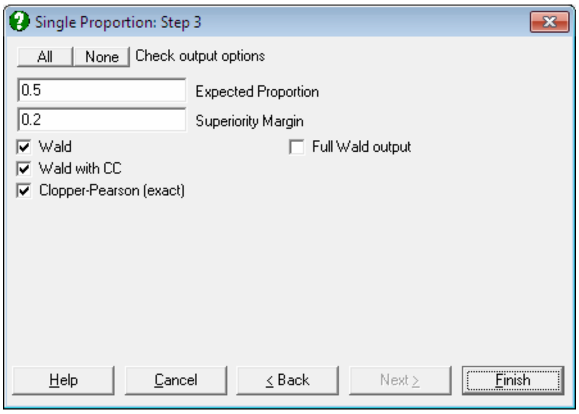
$$SE_{p_0^*} = \sqrt{p_0^*\left(1 - p_0^*\right) / n}$$

Wald with Continuity Correction: A continuity correction term of 1/(2n) is included as in the Binomial Test.

Clopper-Pearson (exact): The exact one- and two-tailed binomial probabilities and the exact confidence interval are reported.

6.4.3.4. Superiority Test

This is identical to Noninferiority Test except that the given margin δ is added to the expected proportion, rather than subtracted.



Single Proportion: Step 3

Check output options

0.5 Expected Proportion

0.2 Superiority Margin

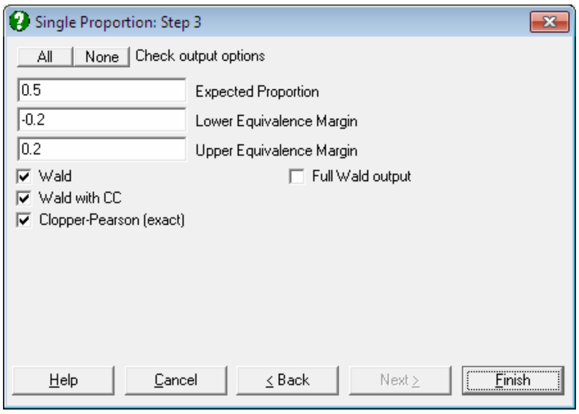
☑ Wald □ Full Wald output

☑ Wald with CC

☑ Clopper-Pearson (exact)

6.4.3.5. Equivalence Test for Binomial Proportion

This is, in effect, a combined Noninferiority Test and Superiority Test. An overall test table displays the larger one-tail probability comparing the two tests and their corresponding lower and upper confidence limits.



Single Proportion: Step 3

Check output options

0.5 Expected Proportion

-0.2 Lower Equivalence Margin

0.2 Upper Equivalence Margin

☑ Wald □ Full Wald output

☑ Wald with CC

☑ Clopper-Pearson (exact)

This is the nonparametric version of equivalence test for means (see 6.1.2. Equivalence Test for Means).

Example 1

Size of Group 1	228
Size of Group 2	534

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Binomial Proportion and select the data option 3 Cell Frequencies are Given. Enter the above group sizes (and enter 1 for the Number of Runs to proceed) and click on the [Opt] button for the Equivalence Test output option. Enter the following and click [Finish]:

Expected Proportion	0.28
Lower Equivalence Margin	-0.1
Upper Equivalence Margin	0.1

Binomial Proportion

Data option: Test Statistics are Given

	Cases
Group 1	228
Group 2	534
Total	762

Equivalence Test

Expected Proportion =	0.2800
Observed Proportion =	0.2992
Lower Equivalence Margin =	-0.1000

Lower Equivalence	**Proportion used in SE**	**Standard Error**	**Z-Statistic**	**1-Tail Probability**	**2-Tail Probability**
Wald	0.2992	0.0166	7.1865	0.0000	
H0	0.1800				
Wald with CC	0.2992	0.0166	7.1469	0.0000	
H0	0.1800				
Clopper-Pearson (exact)				0.0000	

Lower Equivalence	**Lower 90%**	**Upper 90%**
Wald	0.2719	
H0		
Wald with CC	0.2713	
H0		
Clopper-Pearson (exact)	0.2719	

Upper Equivalence Margin =	0.1000

Upper Equivalence	Proportion used in SE	Standard Error	Z-Statistic	1-Tail Probability	2-Tail Probability
Wald	0.2992	0.0166	-4.8701	0.0000	
H0	0.3800				
Wald with CC	0.2992	0.0166	-4.8305	0.0000	
H0	0.3800				
Clopper-Pearson (exact)				0.0000	

Upper Equivalence	Lower 90%	Upper 90%
Wald		0.3265
H0		
Wald with CC		0.3272
H0		
Clopper-Pearson (exact)		0.3277

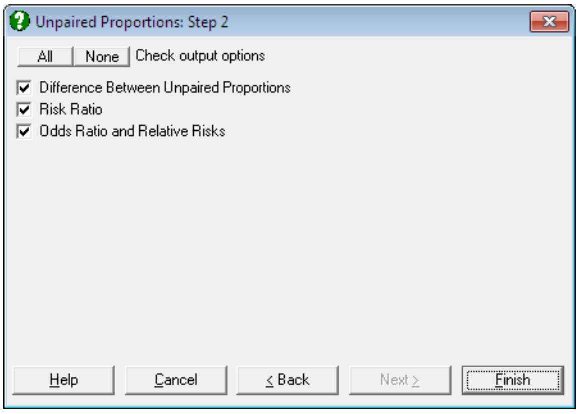
Overall	1-Tail Probability	Lower 90%	Upper 90%
Wald	0.0000	0.2719	0.3265
Wald with CC	0.0000	0.2713	0.3272
Clopper-Pearson (exact)	0.0000	0.2719	0.3277

6.4.4. Unpaired Proportions

A 2 x 2 table is formed to perform the procedures listed under this topic. This can be done from raw data consisting of two columns (not necessarily of equal length) or by directly entering the four cell frequencies in the following order:

(1,1) contains the number of category 1s in the first sample
(1,2) contains the number of category 1s in the second sample
(2,1) contains the number of category 2s in the first sample
(2,2) contains the number of category 2s in the second sample

This data structure is explained in detail at the beginning of this chapter (see 6.0.7. 2 x 2 Tables). The user should take care to distinguish this table from a table formed on the same pair of columns by the Paired Proportions procedure. Here, the total table frequency is the sum of valid cases in sample 1 and sample 2, whereas in Paired Proportions (as in 2 x 2 cross-tabulation) the total frequency is the number of valid pairs.



Check output options

☑ Difference Between Unpaired Proportions
☑ Risk Ratio
☑ Odds Ratio and Relative Risks

When the four frequency values for a 2 x 2 table are already available in the spreadsheet, you do not have to type them again into the Cell Frequencies are Given dialogue. All statistics available under Binomial Proportion, Unpaired Proportions and Paired Proportions procedures are also available in Contingency Table and Cross-Tabulation procedures (see 6.6.2.3. 2 x 2 Table Statistics).

6.4.4.1. Difference Between Unpaired Proportions

Let:

$$p_1 = \frac{n_{11}}{n_{11} + n_{21}}$$

and:

$$p_2 = \frac{n_{12}}{n_{12} + n_{22}}$$

The confidence interval for the difference between two proportions uses the separate sample variance, which is defined as:

$$\mathrm{Var}_{\mathrm{separate}} = \frac{p_1(1-p_1)}{n_{11} + n_{21}} + \frac{p_2(1-p_2)}{n_{12} + n_{22}}$$

and the confidence interval is:

$$\mathrm{LL}, \mathrm{UL} = (p_1 - p_2) \pm Z_{\alpha/2}\sqrt{\mathrm{Var}_{\mathrm{separate}}}$$

The pooled variance is used to test the difference between two proportions. The test statistic is based on the following normal approximation:

$$Z = \frac{p_1 - p_2}{\sqrt{\mathrm{Var}_{\mathrm{pooled}}}}$$

where:

$$\mathrm{Var}_{\mathrm{pooled}} = p(1-p)\left(\frac{1}{n_{11} + n_{21}} + \frac{1}{n_{21} + n_{22}}\right)$$

and:

$$p = \frac{n_{11} + n_{21}}{n}$$

The pooled confidence limits are computed as:

$$\mathrm{LL}, \mathrm{UL} = \left(p_1 \text{ - } p_2\right) \pm Z_{\alpha/2}\sqrt{\mathrm{Var}_{\mathrm{pooled}}}$$

Example 1

Example 4.10 on p. 125 from Armitage & Berry (2002). The effect of two treatments on the mortality rates of two random groups is assessed. The data is given in the form of a 2 x 2 table.

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Proportions and select the data option 3 Cell Frequencies are Given. Enter 41 in (Sample 1 = 1), 64 in (Sample 2 = 1), 216 in (Sample 1 = 2) and 180 in (Sample 2 = 2). Select only the Difference Between Unpaired Proportions output option to obtain the following results:

Unpaired Proportions

Data option: Cell Frequencies are Given

	Sample 1	Sample 2	Total
1	41	64	105
2	216	180	396
Total	257	244	501

Difference Between Unpaired Proportions

Proportion 1 =	0.1595
Proportion 2 =	0.2623

	Difference	Standard Error	Z-Statistic	1-Tail Probability	2-Tail Probability
Pooled Variance	-0.1028	0.0364	-2.8247	0.0024	0.0047
Separate Variance		0.0363	-2.8341	0.0023	0.0046

	Lower 95%	Upper 95%
Pooled Variance	-0.1741	-0.0315
Separate Variance	-0.1738	-0.0317

In this example, Armitage and Berry report the Z-statistic and separate variance confidence interval only.

Example 2

Example on p. 32, Gardner & Altman (1989). Number of patients responding to treatment in two groups is given in the form of a 2 x 2 table.

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Proportions and select the data option 3 Cell Frequencies are Given. Enter 61 in

(Sample 1 = 1), 45 in (Sample 2 = 1), 19 in (Sample 1 = 2) and 35 in (Sample 2 = 2) to obtain the following results:

Unpaired Proportions

Data option: Cell Frequencies are Given

	Sample 1	Sample 2	Total
1	61	45	106
2	19	35	54
Total	80	80	160

Difference Between Unpaired Proportions

Proportion 1 =	0.7625
Proportion 2 =	0.5625

	Difference	Standard Error	Z-Statistic	1-Tail Probability	2-Tail Probability
Pooled Variance	0.2000	0.0748	2.6750	0.0037	0.0075
Separate Variance		0.0731	2.7369	0.0031	0.0062

	Lower 95%	Upper 95%
Pooled Variance	0.0535	0.3465
Separate Variance	0.0568	0.3432

In this example, Gardner and Altman report the separate variance confidence interval only.

6.4.4.2. Risk Ratio

Risk Ratio is defined as (see Gardner & Altman 2000, p. 58.):

$$R = \frac{n_{11} / (n_{11} + n_{12})}{n_{21} / (n_{21} + n_{22})}$$

where the logarithm of R has a standard error of:

$$SE_{Ln(R)} = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21}} - \frac{1}{n_{21} + n_{22}}}$$

and the confidence limits are reported as:

$$LL, UL = Exp\left(Ln(R) \pm Z_{\alpha/2} SE_{Ln(R)}\right)$$

Example

Example on p. 59, Gardner & Altman (2000). Prevelance of Helicobacter pylori infection in preschool children according to mother's histoty of ulcer is given as a 2 x 2 table.

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Proportions and select the data option 3 Cell Frequencies are Given. Enter 6 in (Sample 1 = 1), 112 in (Sample 2 = 1), 16 in (Sample 1 = 2) and 729 in (Sample 2 = 2). Select only the *Risk Ratio* output option.

Unpaired Proportions

Data option: Cell Frequencies are Given

	Sample 1	Sample 2	Total
1	6	112	118
2	16	729	745
Total	22	841	863

Risk Ratio

	Value	Lower 95%	Upper 95%
Risk Ratio	2.0479	1.0131	4.1397

6.4.4.3. Odds Ratio and Relative Risks

The last part of output reports the odds ratio:

$$\mathrm{OR} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

where the logarithm of OR has a standard error of:

$$\mathrm{SE}_{\mathrm{Ln(OR)}} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

and the confidence intervals, which are also known as logit limits, are:

$$\mathrm{LL, UL} = \mathrm{Exp}\left(\mathrm{Ln(OR)} \pm Z_{\alpha/2} \mathrm{SE}_{\mathrm{Ln(OR)}}\right)$$

For the odds ratio for paired cases see 6.4.5.4. Odds Ratio (Paired).

The relative risk for column 1 (cohort 1) is given as:

$$RR_1 = \frac{n_{11}\left(n_{21} + n_{22}\right)}{n_{21}\left(n_{11} + n_{12}\right)}$$

and its confidence interval:

$$LL, UL = RR_1 Exp\left(-Z_{\alpha/2} SE_{Ln(RR_1)}\right), RR_1 Exp\left(Z_{\alpha/2} SE_{Ln(RR_1)}\right)$$

where:

$$SE_{Ln(RR_1)} = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21}} - \frac{1}{n_{21} + n_{22}}}$$

The relative risk for column 2 (cohort 2) is found by interchanging the indices.

Example

Example 4.11 on p. 128 from Armitage & Berry (2002). Association of bronchial carcinoma and asbestos exposure is investigated. The data is given in the form of a 2 x 2 table.

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Unpaired Proportions and select the data option 3 Cell Frequencies are Given. Enter 148 in (Sample 1 = 1), 372 in (Sample 2 = 1), 75 in (Sample 1 = 2) and 343 in (Sample 2 = 2). Select only the Odds Ratio and Relative Risks output option to obtain the following results:

Unpaired Proportions

Data option: Cell Frequencies are Given

	Sample 1	Sample 2	Total
1	148	372	520
2	75	343	418
Total	223	715	938

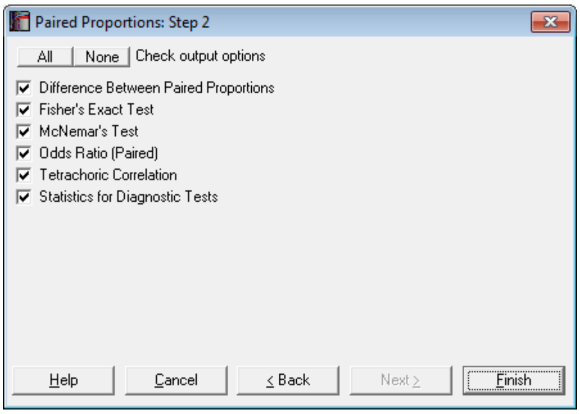
Odds Ratio and Relative Risks

	Value	Lower 95%	Upper 95%
Odds Ratio	1.8195	1.3290	2.4911
Exact		1.3152	2.5283
Relative Risk (Cohort 1)	1.5863	1.2401	2.0290
Relative Risk (Cohort 2)	0.8718	0.8126	0.9353

6.4.5. Paired Proportions

A 2 x 2 table is formed to perform procedures in this section. The data can be in the form of two binary factors or two continuous variables split into two groups by two cutpoints. It is also possible to enter directly the four cell frequencies as explained at the beginning of this chapter (see 6.0.7. 2 x 2 Tables). The user should take care to distinguish this table from a table formed on the same pair of columns by the Unpaired Proportions procedure. Here, the total table frequency is the number valid pairs (as in 2 x 2 cross-tabulation), whereas in Unpaired Proportions the total frequency is the sum of valid cases in sample 1 and sample 2.

The two columns usually contain measurements on the same sample before and after a certain treatment.


Paired Proportions: Step 2

All | None | Check output options

☑ Difference Between Paired Proportions
☑ Fisher's Exact Test
☑ McNemar's Test
☑ Odds Ratio (Paired)
☑ Tetrachoric Correlation
☑ Statistics for Diagnostic Tests

When the four frequency values for a 2 x 2 table are already available in the spreadsheet, you do not have to type them again into the Cell Frequencies are Given dialogue. All statistics available under Binomial Proportion, Unpaired Proportions and Paired Proportions procedures are also available in Contingency Table and Cross-Tabulation procedures (see 6.6.2.3. 2 x 2 Table Statistics).

6.4.5.1. Difference Between Paired Proportions

As in the previous section, the null hypothesis "the two proportions are equal" is tested (see 6.4.4.1. Difference Between Unpaired Proportions). That is:

$$p_1 - p_2 = 0$$

where:

$$p_1 = \frac{n_{11} + n_{12}}{n} \quad , \quad p_2 = \frac{n_{11} + n_{21}}{n}$$

See Gardner & Altman (2000) p. 52, the traditional method. The test statistic based on normal approximation is defined as:

$$Z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

with a standard error of:

$$SE = \frac{\sqrt{n_{12} + n_{21} - (n_{12} - n_{21})^2 / n}}{n}$$

The asymptotic confidence limits are computed as:

$$LL, UL = \left(p_1 - p_2\right) \pm Z_{\alpha/2} SE$$

The exact probability is determined using the binomial distribution and the exact confidence limits are based on the definitions introduced for the Binomial Test:

$$LL = (2\pi_L - 1)(n_{12} + n_{21}) / n$$

$$UL = (2\pi_U - 1)(n_{12} + n_{21}) / n$$

Example 1

Example 4.9 on p. 123 from Armitage & Berry (2002). Distribution of sputum according to results of culture on two media are given in the form of a 2 x 2 table.

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Proportions and select the data option 3 Cell Frequencies are Given. Enter 20 in (1,1), 12 in (1,2), 2 in (2,1), 16 in (2,2) and check only the Difference Between Paired Proportions output option to obtain the following results:

Paired Proportions

Data option: Cell Frequencies are Given

	1	2	Total
1	20	2	22
2	12	16	28
Total	32	18	50

Difference Between Paired Proportions

Proportion 1 =	0.0400
Proportion 2 =	0.2400

	Difference	2-Tail Probability	Lower 95%	Upper 95%
Asymptotic	-0.2000	0.0075	-0.3358	-0.0642
Exact Binomial		0.0129	0.0402	0.2700

Example 2

Example on p. 32, Gardner & Altman (1989). Inadequacy of monitoring in hospital of deaths and survivors among asthma patients is given in the form of a 2 x 2 table.

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Proportions and select the data option 3 Cell Frequencies are Given. Enter 10 in (1,1), 3 in (1,2), 13 in (2,1), 9 in (2,2) and check only the Difference Between Paired Proportions output option to obtain the following results:

Paired Proportions

Data option: Cell Frequencies are Given

	1	2	Total
1	10	13	23
2	3	9	12
Total	13	22	35

Difference Between Paired Proportions

Proportion 1 =	0.3714
Proportion 2 =	0.0857

	Difference	Standard Error	Z-Statistic	1-Tail Probability	2-Tail Probability
Asymptotic	0.2857	0.1036	2.5000	0.0062	0.0124
Exact Binomial				0.0106	0.0213

	Lower 95%	Upper 95%
Asymptotic	0.0827	0.4887
Exact Binomial	0.0398	0.4201

6.4.5.2. Fisher's Exact Test

Under the assumption of independence of column and row factors, the probability of the observed 2 x 2 table (the table probability) follows the hypergeometric distribution:

$$P_T = \frac{\left(r_1!r_2!c_1!c_2!\right)/n!}{n_{11}!n_{12}!n_{21}!n_{22}!}$$

where n_{ij} are the cell frequencies and:

$$r_i = n_{i1} + n_{i2}$$

$$c_j = n_{1j} + n_{2j}$$

$$n = n_{11} + n_{21} + n_{12} + n_{22}$$

The following four probabilities are reported.

Right-Tail Probability: Sum of all possible table probabilities with the same observed row and column totals where n_{11} is greater than or equal to the observed n_{11}. Use this to test positive association between the two factors.

Left-Tail Probability: Sum of all possible table probabilities with the same observed row and column totals where n_{11} is less than or equal to the observed n_{11}. Use this to test negative association between the two factors.

2-Tail Probability: Sum of all table probabilities with the same observed row and column totals. Use this to test association between the two factors.

Table Probability: The probability of the observed table P_T as defined above.

You can perform Fisher's Exact Test for R x C Tables (i.e. tables larger than 2 x 2), using the Cross-Tabulation procedure (see 6.6.2.2.2. Fisher's Exact Test).

Example

Example 24.22 on p. 568 from Zar, J. H. (2010). Data is given in the form of 2 x 2 contingency tables.

Paired Proportions

Data option: Cell Frequencies are Given

	1	2	Total
1	12	7	19
2	2	9	11
Total	14	16	30

Fisher's Exact Test

	Left-Tail Probability	Right-Tail Probability	Two-Tail Probability	Table Probability
Fisher's Exact	0.99787	0.02119	0.02589	0.01906

Zar reports the right-tail and two-tailed probabilities.

6.4.5.3. McNemar Test

McNemar is a chi-square statistic used to test whether the first row and first column totals are equal.

Asymptotic without Continuity Correction: The chi-square statistic is:

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

$$df = 1$$

Asymptotic with Continuity Correction:

$$\chi^2 = \frac{\left(|n_{12} - n_{21}| - 1\right)^2}{n_{12} + n_{21}}$$

$$df = 1$$

Exact Binomial: The exact probability is calculated from the binomial function:

$$P = \sum_{i=0}^{n_{12}} \binom{n_{12}+n_{21}}{i} 0.5^{n_{12}+n_{21}}$$

and the exact confidence limits are given by Liddell (1983) as:

$$\pi_L = \frac{n_{12}}{(n_{21}+1)F_{1-\alpha/2,2(n_{21}+1),2n_{12}}}$$

$$\pi_U = \frac{(n_{12}+1)F_{1-\alpha/2,2(n_{12}+1),2n_{21}}}{n_{21}}$$

Example

Example 24.17 on p. 571 from Zar, J. H. (2010). Data is recorded in the form of a 2 x 2 contingency table. The null hypothesis “the proportion of persons experiencing relief is the same with both locations” is tested.

		Lotion 1	
		Relief	**No relief**
Lotion 2	**Relief**	12	5
	No relief	11	22

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Proportions and select the data option 3 Cell Frequencies are Given. Enter 12 in (1,1), 11 in (1,2), 5 in (2,1), 22 in (2,2). Next, select only the McNemar Test option.

Paired Proportions

Data option: Cell Frequencies are Given

	1	**2**	**Total**
1	12	5	17
2	11	22	33
Total	23	27	50

McNemar's Test

	Chi-Square Statistic	**Degrees of Freedom**	**Right-Tail Probability**
Asymptotic	2.2500	1	0.1336
Asymptotic with CC	1.5625	1	0.2113

	2-Tail Probability	**Lower 95%**	**Upper 95%**
Exact Binomial	1.0000	0.7047	8.0769

Since p > 0.05 do not reject the null hypothesis.

6.4.5.4. Odds Ratio (Paired)

The odds ratio for paired cases is computed as:

$$\mathrm{OR} = \frac{n_{12}}{n_{21}}$$

The exact confidence limits are based on the definitions introduced for the Binomial Test:

$$\mathrm{LL} = \pi_L / (1 - \pi_L)$$

$$\mathrm{UL} = \pi_U / (1 - \pi_U)$$

For the odds ratio for unpaired cases see 6.4.4.3. Odds Ratio and Relative Risks.

Example

Example on p. 66, Gardner Altman (2000). Inadequacy of monitoring in hospital of deaths and survivors among asthma patients is given in the form of a 2 x 2 table.

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Proportions and select the data option 3 Cell Frequencies are Given. Enter 10 in (1,1), 3 in (1,2), 13 in (2,1), 9 in (2,2) and check only the Odds Ratio (Paired) output option:

Paired Proportions

Data option: Cell Frequencies are Given

	1	2	Total
1	10	13	23
2	3	9	12
Total	13	22	35

Odds Ratio (Paired)

	Value	Lower 95%	Upper 95%
Odds Ratio (Paired)	4.3333	1.1908	23.7074

6.4.5.5. Tetrachoric Correlation

The Tetrachoric Correlation coefficient is computed as follows:

$$r_{Cos} = \mathrm{Cos}\left(\frac{\pi}{1+\sqrt{R}}\right)$$

where:

$$R = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

is the tetrachoric ratio.

Example

Table 58 on p. 167 from Cohen, L. & M. Holliday (1983). The raw data is not available on individual success ratings, but a 2 x 2 contingency table is given on satisfactory / unsatisfactory ratings on a basic computing course.

Frequency (1,1)	40
Frequency (1,2)	10
Frequency (2,1)	20
Frequency (2,2)	30

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Proportions → Tetrachoric Correlation, select the data option Cell Frequencies are Given and enter the values as given in the above table to obtain the following results:

Paired Proportions

Data option: Cell Frequencies are Given

	1	**2**	**Total**
1	40	10	50
2	20	30	50
Total	60	40	100

Tetrachoric Correlation

	Ratio	**Tetrachoric Correlation**
	6.0000	0.6132

6.4.5.6. Statistics for Diagnostic Tests

Define the 2 x 2 table entries as follows:

	Positive Actual	**Negative Actual**	**Total**
Positive Estimate	TP	FP	TP + FP
Negative Estimate	FN	TN	FN + TN
Total	TP + FN	FP + TN	TOTAL

where:

TP: True Positive: Correct acceptance,
TN: True Negative: Correct rejection,
FP: False Positive: False alarm (Type I error),
FN: False Negative: Missed detection (Type II error).

There are two important points we need to emphasize here to ensure that the 2 x 2 table generated by the Paired Proportions procedure conforms to these definitions. First, you need to select the factor representing the *Actual* state as *Column 1* from the Variable Selection Dialogue so that it appears at the top of the table. Secondly, as the Paired Proportions procedure sorts the factor levels in ascending order, the smaller values of the two factors selected are assumed to represent the positive outcome. If the larger values represent the positive outcome in your data, you will need to recode your factor columns first, so that the smaller values represent the positive outcome. Otherwise the statistics below will not be computed correctly.

Many of the statistics displayed here are proportions and their confidence intervals are computed employing the Wald (asymptotic) and Clopper-Pearson (exact) methods for binomial proportions (see 6.4.3.2. Binomial Test). Confidence intervals for likelihood ratios are computed as in Simel D., Samsa G., Matchar D. (1991).

Sensitivity: True positive rate or the probability of diagnosing a case as positive when it is actually positive.

TP / (TP + FN)

Specificity: True negative rate or the probability of diagnosing a case as negative when it is actually negative.

TN / (TN + FP)

Accuracy: The rate of correctly classified or the probability of true positive results, including true positive and true negative.

Sensitivity * Prevalence + Specificity * (1 - Prevalence)

(TP + TN) / TOTAL

Prevalence: The actual positive rate.

(TP + FN) / TOTAL

Apparent Prevalence: The estimated positive rate.

(TP + FP) / TOTAL

Youden's Index: Confidence intervals are calculated as in Bangdiwala S.I., Haedo A.S., Natal M.L. (2008).

Sensitivity + Specificity
TP / (TP + FN) + TN / (FP + TN)

Positive Predictive Value: PPV

TP / (TP + FP)

Negative Predictive Value: NPV

TN / (FN + TN)

Positive Likelihood Ratio: LR+

Sensitivity / (1 - Specificity)
(TP / (TP + FN)) / (1 - (TN / (FP + TN)))

Negative Likelihood Ratio: LR-

(1 – Sensitivity) / Specificity
(1 - (TP / (TP + FN))) / (TN / (FP + TN))

Diagnostic Odds Ratio: Confidence intervals are calculated as in Scott I.A., Greenburg P.B., Poole P.J. (2008).

Positive Likelihood Ratio / Negative Likelihood Ratio
(TP * TN) / (FP * FN)

Weighted Positive Likelihood Ratio: WLR+. LR+ is weighted by prevalence.

(Prevalence * Sensitivity) / ((1-Prevalence)(1-Specificity))
TP / FP

Weighted Negative Likelihood Ratio: WLR-. LR- is weighted by prevalence.

(Prevalence (1-Sensitivity)) / ((1-Prevalence) Specificity)
FN / TN

Example

Table 19.47, Case (b) on p. 694 from Armitage & Berry (2002). The data is available as a 2 x 2 contingency table:

Frequency (1,1)	90
Frequency (1,2)	90
Frequency (2,1)	10
Frequency (2,2)	810

Select Statistics 1 → Nonparametric Tests (1-2 Samples) → Paired Proportions → Statistics for Diagnostic Tests, select the data option Cell Frequencies are Given and enter the values as given in the above table to obtain the following results:

Paired Proportions

Data option: Cell Frequencies are Given

	1	**2**	**Total**
1	90	90	180
2	10	810	820
Total	100	900	1000

Statistics for Diagnostic Tests

Smaller factor level represents the positive outcome.
Confidence Intervals: Row 1: Asymptotic Normal, Row 2: Exact Binomial

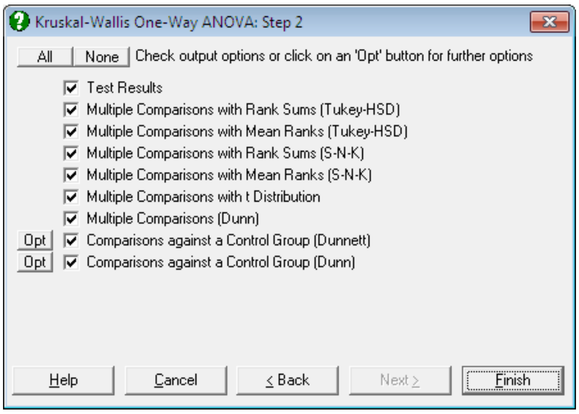
	Value	Standard Error	Lower 95%	Upper 95%
Sensitivity	0.9000	0.0300	0.8412	0.9588
			0.8238	0.9510
Specificity	0.9000	0.0100	0.8804	0.9196
			0.8785	0.9188
Accuracy	0.9000	0.0095	0.8814	0.9186
			0.8797	0.9179
Prevalence	0.1000	0.0095	0.0814	0.1186
			0.0821	0.1203
Apparent Prevalence	0.1800	0.0121	0.1562	0.2038
			0.1567	0.2052
Youden's Index	0.8000			
			0.7023	0.8698
Positive Predictive Value	0.5000	0.0373	0.4270	0.5730
			0.4247	0.5753
Negative Predictive Value	0.9878	0.0038	0.9803	0.9953
			0.9777	0.9941
Positive Likelihood Ratio	9.0000		7.3201	11.0654
Negative Likelihood Ratio	0.1111		0.0617	0.2001
Diagnostic Odds Ratio	81.0000		40.6821	161.2749
Weighted Positive Likelihood Ratio	1.0000		0.8133	1.2295
Weighted Negative Likelihood Ratio	0.0123		0.0067	0.0229

6.5. Multisample Nonparametric Tests

Nonparametric tests that are performed on two or more samples are collected under this section. A large number of tests which are not included here can be accessed from the Contingency Table or Cross-Tabulation procedures (see 6.6.2.2. R x C Table Statistics), where they are computed as part of table statistics.

6.5.1. Kruskal-Wallis One-Way ANOVA

Data entry is in multisample format (see 6.0.4. Multisample Tests). Each sample can be entered in a separate column (not necessarily of equal length), or they can be stacked in one or more columns and subsamples defined by an unlimited number of factor columns. Missing values are omitted by case.



Kruskal-Wallis One-Way ANOVA: Step 2

Check output options or click on an 'Opt' button for further options

☑ Test Results
☑ Multiple Comparisons with Rank Sums (Tukey-HSD)
☑ Multiple Comparisons with Mean Ranks (Tukey-HSD)
☑ Multiple Comparisons with Rank Sums (S-N-K)
☑ Multiple Comparisons with Mean Ranks (S-N-K)
☑ Multiple Comparisons with t Distribution
☑ Multiple Comparisons (Dunn)
Opt ☑ Comparisons against a Control Group (Dunnett)
Opt ☑ Comparisons against a Control Group (Dunn)

6.5.1.1. Kruskal-Wallis ANOVA Test Results

This test is used to evaluate the degree of association between samples. It is assumed that the samples have similar distributions and that they are independent. All cases in all samples are ranked together and then the rank sum of each sample is found. The test statistic is calculated as follows:

$$H = \frac{12R}{N(N+1)} - 3(N+1)$$

$$df = M - 1$$

where N is the total number of cases in all samples, M is the number of variables and R is the total of the squared sum of ranks for each sample divided by the respective sample size.

The test statistic corrected for ties is:

$$H' = \frac{H}{1 - \frac{K}{N^3 - N}}$$

where K is sum of k³ - k and k is the number of tied cases for a particular rank.

The one-tail probability is reported from the chi-square distribution.

6.5.1.2. Kruskal-Wallis ANOVA Multiple Comparisons

Eight nonparametric Multiple Comparisons can be performed as part of this procedure. The last two are comparisons against a control group (which require further inputs) and the rest are comparisons between all possible pairs.

Multiple comparisons with rank sums (Tukey-HSD)

Nonparametric Multiple Comparisons are performed in a way similar to the Tukey-HSD test using rank sums. The standard error is computed as:

$$SE = \sqrt{\frac{N(NM)(NM+1)}{12}}$$

This test requires equal group sizes.

Multiple comparisons with mean ranks (Tukey-HSD)

Nonparametric Multiple Comparisons are performed in a way similar to the Tukey-HSD test using mean ranks. In this case the standard error is computed as:

$$SE = \sqrt{\frac{M(NM+1)}{12}}$$

This test requires equal group sizes.

Multiple comparisons with rank sums (S-N-K)

Nonparametric Multiple Comparisons are performed in a way similar to the Student-Newman-Keuls test using mean ranks. In this case the standard error is computed as:

$$SE = \sqrt{\frac{N(NM)(NM+1)}{12}}$$

This test requires equal group sizes.

Multiple comparisons with mean ranks (S-N-K)

Nonparametric Multiple Comparisons can also be performed in a way similar to the Student-Newman-Keuls test using rank sums. The standard error is computed as follows:

$$SE = \sqrt{\frac{M(NM+1)}{12}}$$

This test requires equal group sizes.

Multiple comparisons with t-distribution

If group sizes are not equal and all possible pairs are to be compared then this option can be selected. Nonparametric Multiple Comparisons are performed in a way similar to the Tukey-HSD test using mean ranks. In this case the standard error is computed as:

$$SE = \sqrt{\left(\frac{N(N+1)}{12}\frac{N-1-H}{(N-M)}\right)\left(\frac{1}{n_B}+\frac{1}{n_A}\right)}$$

Multiple comparisons (Dunn)

If group sizes are not equal and all possible pairs are to be compared, then this option can be selected.

The standard error, which has a correction term for tied ranks, is computed as follows:

$$SE = \sqrt{\left(\frac{N(N+1)}{12}-\frac{K}{12(N-1)}\right)\left(\frac{1}{n_B}+\frac{1}{n_A}\right)}$$

where N is the total number of cases, K is the sum of k^3 - k and k is the number of tied cases for a particular rank (as in Kruskal-Wallis One-Way ANOVA). In comparisons group mean ranks are used.

Comparisons against a control group (Dunnett)

If each group of data is to be tested against a control group and all groups are of the same size then select this option. If the group sizes are not equal then the next option (Dunn's test) should be used.

The standard error is computed as follows:

$$\mathrm{SE} = \sqrt{\frac{n(np)(np+1)}{6}}$$

The only other difference between the Dunnett test introduced here and the Dunnett test *per se* is that here the group rank sums are used while the latter uses group mean ranks.

Comparisons against a control group (Dunn)

If each group of data is to be tested against a control group and all groups are not of the same size then select this option. If the group sizes are equal then the previous option (Dunnett test) may also be employed.

The standard error, which has a correction term for tied ranks, is computed as in the Dunn's test above.

6.5.1.3. Kruskal-Wallis ANOVA Examples

Example 1

Example 10.6 on p. 287 from Armitage & Berry (2002). Counts of adult worms in four groups of rats are given. The null hypothesis "there is no significant difference between the rats" is tested.

Open NONPARM1 and select Statistics 1 → Nonparametric Tests (Multisample) → Kruskal-Wallis ANOVA. Select *Group 1*, *Group 2*, *Group 3* and *Group 4* (*C1* to *C4*) as [Variable]s and then select only Test Results to obtain the following results:

Kruskal-Wallis One-Way ANOVA

	Cases	Rank Sum	Mean Rank
Group 1	5	42.0000	8.4000
Group 2	5	53.0000	10.6000
Group 3	5	36.0000	7.2000
Group 4	5	79.0000	15.8000
Total	20	210.0000	10.5000

Correction for Ties =	0.0008
Chi-Square Statistic =	6.2047
Degrees of Freedom =	3
Right-Tail Probability =	0.10207

This result is not significant at the 10% level. Hence do not reject the null hypothesis.

Example 2

Examples 10.10 on p. 216 and 11.7 on p. 241 from Zar, J. H. (2010). A researcher wants to test the null hypothesis "the abundance of the flies is the same in all three vegetation layers" at a 95% significance level. If they were found to be different, then the researcher would also like to know which ones.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Kruskal-Wallis ANOVA and include *Herbs* (*C5*), *Shrubs* (*C6*) and *Trees* (*C7*) in the analysis by clicking [Variable]. Check only the Test Results and the Multiple Comparisons with Rank Sums (Tukey-HSD) boxes to obtain the following results:

Kruskal-Wallis One-Way ANOVA

	Cases	Rank Sum	Mean Rank
Herbs	5	64.0000	12.8000
Shrubs	5	30.0000	6.0000
Trees	5	26.0000	5.2000
Total	15	120.0000	8.0000

Correction for Ties =	0.0000
Chi-Square Statistic =	8.7200
Degrees of Freedom =	2
Right-Tail Probability =	0.0128

Since the right tail probability is less than 5%, the null hypothesis is rejected. Next the researcher would like to find which vegetation layers have different abundance of the flies.

Multiple Comparisons with Rank Sums (Tukey-HSD)

Method: 95% Tukey-HSD interval.
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Rank Sum	Trees	Shrubs	Herbs	
Trees	5	26.0000			**	|
Shrubs	5	30.0000			**	|
Herbs	5	64.0000	**	**		|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
Herbs - Trees	38.0000	10.0000	3.8000	3.3145	0.0197
Shrubs - Trees	4.0000	10.0000	0.4000	3.3145	0.9569
Herbs - Shrubs	34.0000	10.0000	3.4000	3.3145	0.0428

Comparison	Lower 95%	Upper 95%	Result
Herbs - Trees	4.8551	71.1449	**
Shrubs - Trees	-29.1449	37.1449	
Herbs - Shrubs	0.8551	67.1449	**

Homogeneous Subsets:

Group 1:	Trees Shrubs
Group 2:	Herbs

The overall conclusion is that fly abundance is the same for *Trees* and *Shrubs* but it is different for *Herbs*.

Example 3

Examples 10.11 on p. 217 and 11.8 on p. 242 from Zar, J. H. (2010). The null hypothesis that "pH is the same in all four ponds" is tested at a 95% significance level. If they were found to be different, then we would also like to know which ones. The data has unequal column lengths.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Kruskal-Wallis ANOVA and include *Pond 1*, *Pond 2*, *Pond 3* and *Pond 4* (*C8* to *C11*) in the analysis by clicking [Variable]. Check only the Test Results and the Multiple Comparisons (Dunn) boxes to obtain the following results:

Kruskal-Wallis One-Way ANOVA

	Cases	Rank Sum	Mean Rank
Pond 1	8	55.0000	6.8750
Pond 2	8	132.5000	16.5625
Pond 3	7	145.0000	20.7143
Pond 4	8	163.5000	20.4375
Total	31	496.0000	16.0000

Correction for Ties =	0.0056
Chi-Square Statistic =	11.9435
Degrees of Freedom =	3
Right-Tail Probability =	0.0076

Since the right tail probability is less than 5%, the null hypothesis is rejected. Next we would like to find which ponds have a different pH.

Multiple Comparisons (Dunn)

Method: 95% Dunn interval.
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean Rank	Pond 1	Pond 2	Pond 4	Pond 3	
Pond 1	8	6.8750			**	**	|
Pond 2	8	16.5625					| |
Pond 4	8	20.4375	**				|
Pond 3	7	20.7143	**				|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
Pond 3 - Pond 1	13.8393	4.6923	2.9493	2.6383	0.0191
Pond 4 - Pond 1	13.5625	4.5332	2.9918	2.6383	0.0166
Pond 2 - Pond 1	9.6875	4.5332	2.1370	2.6383	0.1956
Pond 3 - Pond 2	4.1518	4.6923	0.8848	2.6383	1.0000
Pond 4 - Pond 2	3.8750	4.5332	0.8548	2.6383	1.0000
Pond 3 - Pond 4	0.2768	4.6923	0.0590	2.6383	1.0000

Comparison	Lower 95%	Upper 95%	Result
Pond 3 - Pond 1	1.4597	26.2188	**
Pond 4 - Pond 1	1.6027	25.5223	**
Pond 2 - Pond 1	-2.2723	21.6473	
Pond 3 - Pond 2	-8.2278	16.5313	
Pond 4 - Pond 2	-8.0848	15.8348	
Pond 3 - Pond 4	-12.1028	12.6563	

Homogeneous Subsets:	
Group 1:	Pond 1 Pond 2
Group 2:	Pond 2 Pond 4 Pond 3

The overall conclusion is that water pH is the same in *Pond 2*, *Pond 4* and *Pond 3* but is different in *Pond 1*.

Example 4

Example 1, p. 291, Conover, W. J. (1999). The null hypothesis that "the four methods (i.e. columns) are equivalent" is tested at a 95% confidence level.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Kruskal-Wallis ANOVA and include *Method 1*, *Method 2*, *Method 3*, *Method 4* (*C12* to *C15*) in the analysis by clicking [Variable]. Check only the Test Results and the Multiple Comparisons with t-Distribution boxes to obtain the following results:

Kruskal-Wallis One-Way ANOVA

	Cases	Rank Sum	Mean Rank
Method 1	9	196.5000	21.8333
Method 2	10	153.0000	15.3000
Method 3	7	207.0000	29.5714
Method 4	8	38.5000	4.8125
Total	34	595.0000	17.5000

Correction for Ties =	0.0064
Chi-Square Statistic =	25.6288
Degrees of Freedom =	3
Right-Tail Probability =	0.0000

Since the right tail probability is less than 5%, the null hypothesis is rejected. Therefore, we can now ask the question which methods are different.

Multiple Comparisons with t Distribution

Method: 95% t interval.
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	Method 4	Method 2	Method 1	Method 3	
Method 4	8	4.8125		**	**	**	|
Method 2	10	15.3000	**		**	**	|
Method 1	9	21.8333	**	**		**	|
Method 3	7	29.5714	**	**	**		|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
Method 3 - Method 4	24.7589	2.5465	9.7227	2.0423	0.0000
Method 1 - Method 4	17.0208	2.3908	7.1192	2.0423	0.0000
Method 2 - Method 4	10.4875	2.3339	4.4935	2.0423	0.0001
Method 3 - Method 2	14.2714	2.4248	5.8857	2.0423	0.0000
Method 1 - Method 2	6.5333	2.2607	2.8899	2.0423	0.0071
Method 3 - Method 1	7.7381	2.4796	3.1207	2.0423	0.0040

Comparison	Lower 95%	Upper 95%	Result
Method 3 - Method 4	19.5583	29.9596	**
Method 1 - Method 4	12.1381	21.9036	**
Method 2 - Method 4	5.7210	15.2540	**
Method 3 - Method 2	9.3194	19.2234	**
Method 1 - Method 2	1.9163	11.1504	**
Method 3 - Method 1	2.6741	12.8021	**

Homogeneous Subsets:

Group 1:	Method 4
Group 2:	Method 2
Group 3:	Method 1
Group 4:	Method 3

The overall conclusion is that all methods are different.

6.5.2. Jonckheere’s Trend

This test is used when there are three or more conditions and it is possible to predict the ordering of results. A predicted trend across the scores of different groups is evaluated without ranking the data. Data should be arranged in predicted order, from lowest to highest.

Data entry is in multisample format (see 6.0.4. Multisample Tests). Each sample should be entered in a separate column (not necessarily of equal length). Missing values are omitted by case.

The test statistic displayed is corrected for ties. The one-tail probability is reported using the normal distribution.

Example

Table 112 on p. 275 from Cohen, L. & M. Holliday (1983). Correctly spelt words by pupils of three different ages are given. The null hypothesis “spelling ability is the same for all age groups” is tested.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Jonckheere’s Trend and include *Seven Year*, *Eight Year*, *Nine Year* (*C16* to *C18*) in the analysis by clicking [Variable] to obtain the following results:

Jonckheere’s Trend

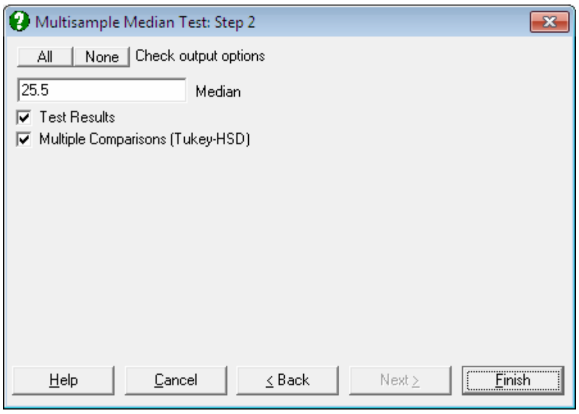
	cases
Seven Year	10
Eight Year	10
Nine Year	10
Total	30

```
                 P =    198.0000
                 Q =    50.0000
               P-Q =    148.0000
      variance P-Q =    2635.5501
       Z-Statistic =    2.8829
1-Tail Probability =    0.00197
```

This result is significant at the 1% level. Hence reject the null hypothesis.

6.5.3. Multisample Median Test

Data entry is in multisample format (see 6.0.4. Multisample Tests). Each sample can be entered in a separate column (not necessarily of equal length), or they can be stacked in one or more columns and subsamples defined by an unlimited number of factor columns. Missing values are omitted by case.



Check output options

25.5 Median

☑ Test Results

☑ Multiple Comparisons (Tukey-HSD)

A further dialogue allows you to override the median computed from data and enter any values. The two output options can also be selected from the same dialogue.

6.5.3.1. Multisample Median Test Results

Like the Two Sample Median Test, this is also used to determine whether the samples have been drawn from populations with the same median. But here, the number of samples is not limited to two. The number of cases less than or equal to and greater than the overall median is found for each sample. Then the chi-square statistic is calculated from the obtained frequencies.

The one-tail probability is reported using the chi-square distribution with M - 1 degrees of freedom.

6.5.3.2. Median Multiple Comparisons

A Tukey-HSD type comparison is made between all possible pairs of groups. The value used in comparisons is the number of cases greater than the overall median for each sample (i.e. the first column of the table displayed in test results).

If the total number of cases N is an odd number the standard error is computed as:

$$SE = \sqrt{\frac{n(N+1)}{4N}}$$

If N is an even number, then the standard error is:

$$SE = \sqrt{\frac{nN}{4(N-1)}}$$

where n is the harmonic mean of group sizes.

6.5.3.3. Multisample Median Test Example

Examples 10.12 on p. 201 and 11.11 on p. 227 from Zar, J. H. (1999). The null hypothesis "median elm tree height is the same on all four sides of a building" is tested at a 95% significance level. If they are found to be different, then we would also like to know which ones.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Multisample Median Test and include *North*, *East*, *South* and *West* (*C19* to *C22*) in the analysis by clicking [Variable]. Check the Test Results and the Multiple Comparisons boxes to obtain the following results:

Multisample Median Test

	Cases	> Median	<= Median
North	12	4	8
East	12	3	9
South	12	10	2
West	12	6	6
Total	48	23	25

Median =	7.9000
Chi-Square Statistic =	9.6000
Degrees of Freedom =	3
Right-Tail Probability =	0.0223

Since the right tail probability is less than 5% the null hypothesis is rejected. In the 5th edition of *Biostatistical Analysis* (2010), Examples 10.12 on p. 219 and 11.9 on p. 245, Zar employs a different Method where observations at the median are omitted. With this approach the total number of valid cases is 46 and the chi-squared statistic is 11.182.

Next we ask the question which groups are significantly different at 95%.

Multiple Comparisons (Tukey-HSD)

Method: 95% Tukey-HSD interval.
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	East	North	West	South	
East	12	3.0000				**	I
North	12	4.0000					I I
West	12	6.0000					I I
South	12	10.0000	**				I

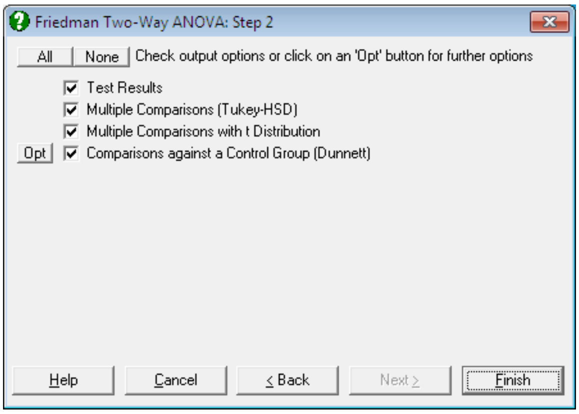
Comparison	Difference	Standard Error	q Stat	Table q	Probability
South - East	7.0000	1.7504	3.9991	3.6332	0.0242
West - East	3.0000	1.7504	1.7139	3.6332	0.6192
North - East	1.0000	1.7504	0.5713	3.6332	0.9777
South - North	6.0000	1.7504	3.4278	3.6332	0.0726
West - North	2.0000	1.7504	1.1426	3.6332	0.8507
South - West	4.0000	1.7504	2.2852	3.6332	0.3696

Comparison	Lower 95%	Upper 95%	Result
South - East	0.6406	13.3594	**
West - East	-3.3594	9.3594	
North - East	-5.3594	7.3594	
South - North	-0.3594	12.3594	
West - North	-4.3594	8.3594	
South - West	-2.3594	10.3594	

Homogeneous Subsets:
Group 1: East North West
Group 2: North West South

The overall conclusion is that although elm trees on the *East* and *South* definitely have different heights at 95%, nothing can be said about the ones on the *North* and *West*.

6.5.4. Friedman Two-Way ANOVA



Check output options or click on an 'Opt' button for further options

☑ Test Results
☑ Multiple Comparisons (Tukey-HSD)
☑ Multiple Comparisons with t Distribution
☑ Comparisons against a Control Group (Dunnett)

Data entry is in matrix format (see 6.0.5. Tests with Matrix Data). Columns selected for this test must have equal number of rows and rows containing at least one missing value are omitted.

6.5.4.1. Friedman ANOVA Test Results

This test is used to determine whether the M samples have been drawn from the same population. Cases are ranked and the mean rank is calculated for each sample. The test statistic is calculated as follows:

$$H = \frac{12R}{NM(M+1)} - 3N(M+1)$$

where N is the number of rows, M is the number of columns and R is the sum of squares of column rank totals.

The test statistic corrected for ties is:

$$H' = \frac{H}{1 - \frac{K}{N(M^3 - M)}}$$

where K is sum of k^3-k and k is the number of tied cases for a particular rank. H' has a chi-square distribution with M - 1 degrees of freedom.

Although the chi-square statistic is commonly used, here we also report an alternative definition of the Friedman statistic based on the F distribution, with M - 1 and (N - 1)(M - 1) degrees of freedom, which is said to produce more accurate results (see Conover, W. J. 1999, p. 301). This version of the Friedman statistic is computed as follows: First the ranks (R_{ij}, i = 1, ..., N, j = 1, ..., M) in each row and then their column totals are found (R_j, j = 1, ..., M). The test statistic is defined as:

$$H = \frac{(N-1)[B - NM(M+1)^2]/4}{A - B}$$

where:

$$A = \sum\sum R_{ij}^2$$

$$B = \frac{1}{N}\sum R_j^2$$

The test statistic displayed is corrected for ties. The output includes rank sum and mean rank for each variable and correction for ties. The one-tail probability is reported using both chi-square distribution with M - 1 degrees of freedom and F-distribution with N - 1 and (N - 1)(M - 1) degrees of freedom.

6.5.4.2. Friedman ANOVA Multiple Comparisons

If the null hypothesis is rejected as result of the Friedman's test, then a multiple comparison can be run to find out which column effects are different.

Multiple comparisons with rank sums (Tukey-HSD)

Nonparametric Multiple Comparisons are performed in a way similar to the Tukey-HSD test using rank sums. The standard error is computed as:

$$SE = \sqrt{\frac{NM(M+1)}{12}}$$

Multiple comparisons with t-distribution

Comparisons are made using rank sums and the t-distribution. The standard error is computed as:

$$SE = \sqrt{\frac{2N(A-B)}{(N-1)(M-1)}}$$

Comparisons against a control group (Dunnett)

If each group of data is to be tested against a control group then select this option. The standard error is computed as follows:

$$SE = \sqrt{\frac{NM(M+1)}{6}}$$

The only difference between the Dunnett test introduced here and the Dunnett test *per se* is that here the group rank sums are used while the latter uses group mean ranks.

6.5.4.3. Friedman ANOVA Examples

Example 1

Example 10.5 on p. 286 from Armitage & Berry (2002). Clotting times (min) of plasma from eight subjects, treated by four methods are given. The null hypothesis "there is no difference between the four treatments" is tested.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Friedman Two-Way ANOVA and select *Treatment 1* to *Treatment 4* (*C23* to *C26*) as [Variable]s. Select Test Results as the only output option to obtain the following results:

Friedman Two-Way ANOVA

	Cases	Rank Sum	Mean Rank
Treatment 1	8	11.0000	1.3750
Treatment 2	8	16.0000	2.0000
Treatment 3	8	23.5000	2.9375
Treatment 4	8	29.5000	3.6875
Total	32	80.0000	2.5000

```
   Number of Columns =    4
      Number of Rows =    8
   Correction for Ties =  0.0125
  Chi-Square Statistic =  15.1519
   Degrees of Freedom =   3
 Right-Tail Probability = 0.00169
              F(3,21) =   11.9871
 Right-Tail Probability = 0.0001
```

Both tests are significant at the 1% level. Hence reject the null hypothesis.

Example 2

Example 12.5 on p. 278 from Zar, J. H. (2010). A researcher wants to test the null hypothesis "time for effectiveness is the same for all three anesthetics", or in other words that all means are the same against the alternative hypothesis that they are not all equal.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Friedman Two-Way ANOVA and include *Treatment A* to *Treatment C* (*C40* to *C42*) in the analysis by clicking [Variable]. Check the Test Results output option only to obtain the following results:

Friedman Two-Way ANOVA

	Cases	Rank Sum	Mean Rank
Treatment A	5	6.0000	1.2000
Treatment B	5	15.0000	3.0000
Treatment C	5	9.0000	1.8000
Total	15	30.0000	2.0000

```
   Number of Columns =    3
      Number of Rows =    5
   Correction for Ties =  0.0000
  Chi-Square Statistic =  8.4000
   Degrees of Freedom =   2
 Right-Tail Probability = 0.0150
               F(2,8) =   21.0000
 Right-Tail Probability = 0.0007
```

As the right tail probability is less than 5% reject the null hypothesis.

Example 3

Example 1 on p. 371, Conover, W. J. (1999). A researcher wants to test the null hypothesis "the treatments in blocks (i.e. columns) have identical effects" at a 95% confidence level.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Friedman Two-Way ANOVA and include *Grass 1* to *Grass 4* (*C31* to *C34*) in the analysis by clicking [Variable]. Select only the Test Results and Multiple comparisons with t-distribution output options to obtain the following results:

Friedman Two-Way ANOVA

	Cases	Rank Sum	Mean Rank
Grass 1	12	38.0000	3.1667
Grass 2	12	23.5000	1.9583
Grass 3	12	24.5000	2.0417
Grass 4	12	34.0000	2.8333
Total	48	120.0000	2.5000

```
   Number of Columns =    4
      Number of Rows =    12
    Correction for Ties = 0.0583
  Chi-Square Statistic =  8.0973
  Degrees of Freedom =    3
Right-Tail Probability =  0.0440
             F(3,33) =    3.1922
Right-Tail Probability =  0.0362
```

Since the right tail probability is less than 5%, reject the null hypothesis. Therefore, we can proceed with Multiple Comparisons to find out which treatments are different.

Multiple Comparisons with t Distribution

Method: 95% t interval.
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Rank Sum	Grass 2	Grass 3	Grass 4	Grass 1	
Grass 2	12	23.5000				**	|
Grass 3	12	24.5000				**	|
Grass 4	12	34.0000					| |
Grass 1	12	38.0000	**	**			|

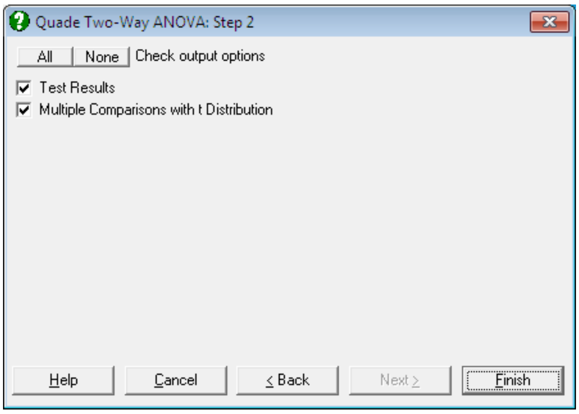
Comparison	Difference	Standard Error	q Stat	Table q	Probability
Grass 1 - Grass 2	14.5000	5.6434	2.5694	2.0345	0.0149
Grass 4 - Grass 2	10.5000	5.6434	1.8606	2.0345	0.0717
Grass 3 - Grass 2	1.0000	5.6434	0.1772	2.0345	0.8604
Grass 1 - Grass 3	13.5000	5.6434	2.3922	2.0345	0.0226
Grass 4 - Grass 3	9.5000	5.6434	1.6834	2.0345	0.1017
Grass 1 - Grass 4	4.0000	5.6434	0.7088	2.0345	0.4834

Comparison	Lower 95%	Upper 95%	Result
Grass 1 - Grass 2	3.0183	25.9817	**
Grass 4 - Grass 2	-0.9817	21.9817	
Grass 3 - Grass 2	-10.4817	12.4817	
Grass 1 - Grass 3	2.0183	24.9817	**
Grass 4 - Grass 3	-1.9817	20.9817	
Grass 1 - Grass 4	-7.4817	15.4817	

Homogeneous Subsets:	
Group 1:	Grass 2 Grass 3 Grass 4
Group 2:	Grass 4 Grass 1

The overall conclusion is that *Grass 1/2* and *Grass 1/3* are different.

6.5.5. Quade Two-Way ANOVA



☑ Test Results
☑ Multiple Comparisons with t Distribution

Data entry is in matrix format (see 6.0.5. Tests with Matrix Data). Columns selected for this test must have equal number of rows and rows containing at least one missing value are omitted.

6.5.5.1. Quade ANOVA Test Results

In the analysis of several related variables a more powerful alternative to the Friedman Two-Way ANOVA is Quade Two-Way ANOVA by ranks. First the ranks (R_{ij}, i = 1, ..., N, j = 1, ..., M) and the range of data in each row are found and range values are also ranked (Q_i, i = 1, ..., N). The test statistic is:

$$H = \frac{(N-1)B}{A-B}$$

where:

$$A = \sum\sum S_{ij}^2$$

$$B = \frac{1}{N}\sum S_j^2$$

$$S_j = \sum_i S_{ij}$$

$$S_{ij} = Q_i\left[R_{ij} - \frac{M+1}{2}\right]$$

The test statistic displayed is corrected for ties. The output includes the rank sum and mean rank for each variable and the correction for ties. The one-tail probability is reported using the F-distribution with M - 1 and (N - 1)(M - 1) degrees of freedom.

6.5.5.2. Quade ANOVA Multiple Comparisons

If the null hypothesis is rejected as result of the Quade's test, then a multiple comparison can be run to find out which column effects are different. Nonparametric Multiple Comparisons are performed in a way similar to the Tukey-HSD test using rank sums and the t-distribution. The standard error is:

$$SE = \sqrt{\frac{2N(A-B)}{(N-1)(M-1)}}$$

6.5.5.3. Quade ANOVA Example

Example 2 on p. 375, Conover, W. J. (1999). A researcher wants to test the null hypothesis that "the treatments in blocks (i.e. columns) have identical effects" at a 95% confidence level.

Open NONPARM1, select Statistics 1 → Nonparametric Tests (Multisample) → Quade Two-Way ANOVA and include *Brand A*, *Brand B*, *Brand C*, *Brand D* and *Brand E*, (*C35* to *C39*) in the analysis by clicking [Variable] to obtain the following results:

Quade Two-Way ANOVA

	Cases	Rank Sum	Mean Rank
Brand A	7	21.5000	3.0714
Brand B	7	15.0000	2.1429
Brand C	7	15.0000	2.1429
Brand D	7	26.0000	3.7143
Brand E	7	27.5000	3.9286
Total	35	105.0000	3.0000

Number of Columns =	5
Number of Rows =	7
F(4,24) =	3.8293
Right-Tail Probability =	0.0152

Since the right tail probability is less than 5%, reject the null hypothesis. Therefore, proceed with the Multiple Comparisons to find out which treatments are different.

Multiple comparisons with t-distribution

Method: 95% t interval.
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Rank Sum	Brand B	Brand C	Brand A	Brand D	Brand E	
Brand B	7	-38.0000				**	**	|
Brand C	7	-14.0000					**	| |
Brand A	7	-9.5000					**	| |
Brand D	7	23.5000	**					| |
Brand E	7	38.0000	**	**	**			|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
Brand E - Brand B	76.0000	22.0586	3.4454	2.0639	0.0021
Brand D - Brand B	61.5000	22.0586	2.7880	2.0639	0.0102
Brand A - Brand B	28.5000	22.0586	1.2920	2.0639	0.2087
Brand C - Brand B	24.0000	22.0586	1.0880	2.0639	0.2874
Brand E - Brand C	52.0000	22.0586	2.3574	2.0639	0.0269
Brand D - Brand C	37.5000	22.0586	1.7000	2.0639	0.1021
Brand A - Brand C	4.5000	22.0586	0.2040	2.0639	0.8401
Brand E - Brand A	47.5000	22.0586	2.1534	2.0639	0.0416
Brand D - Brand A	33.0000	22.0586	1.4960	2.0639	0.1477
Brand E - Brand D	14.5000	22.0586	0.6573	2.0639	0.5172

Comparison	Lower 95%	Upper 95%	Result
Brand E - Brand B	30.4732	121.5268	**
Brand D - Brand B	15.9732	107.0268	**
Brand A - Brand B	-17.0268	74.0268	
Brand C - Brand B	-21.5268	69.5268	
Brand E - Brand C	6.4732	97.5268	**
Brand D - Brand C	-8.0268	83.0268	
Brand A - Brand C	-41.0268	50.0268	
Brand E - Brand A	1.9732	93.0268	**
Brand D - Brand A	-12.5268	78.5268	
Brand E - Brand D	-31.0268	60.0268	

Homogeneous Subsets:

Group 1:	Brand B Brand C Brand A
Group 2:	Brand C Brand A Brand D
Group 3:	Brand D Brand E

The overall conclusion is that *Brand B* & *E*, *B* & *D*, *C* & *E*, and *A* & *E* are different.

6.5.6. Kendall's Concordance Coefficient

This test is particularly useful for evaluating various readings on the same set of variables. Each variable is ranked and its mean rank is found. The test statistic is calculated as follows:

$$H = \frac{12R}{M^2(N^3 - N) - MK}$$

where R is the sum of squared differences from the mean rank and K is the sum of k^3 - k and k is the number of tied cases for a particular rank.

Data entry is in matrix format (see 6.0.5. Tests with Matrix Data). Columns selected for this test must have equal number of rows and rows containing at least one missing value are omitted.

The test statistic displayed is corrected for ties. For each variable the rank sum and mean rank are displayed. The correction factor is also reported. One-tail probability is obtained using the chi-square distribution with M - 1 degrees of freedom.

WARNING! *For this procedure UNISTAT expects the data to be entered as variables in rows and cases in columns. If the data is not already in this form, use Data Processor's* Data → Transpose Matrix *facility to obtain the correct format.*

Example

Example 20.5 on p. 450, Zar, J. H. (2010). The data table given in the book needs to be transposed to run the test in UNISTAT. The null hypothesis "there is no association among the three variables" is tested at a 95% confidence level.

Open NONPARM2, select Statistics 1 → Nonparametric Tests (Multisample) → Kendall's Concordance and include *C1* to *C12* in the analysis by clicking [Variable] to obtain the following results:

Kendall’s Concordance Coefficient

	Cases	Rank Sum	Mean Rank
C1	3	14.5000	4.8333
C2	3	21.0000	7.0000
C3	3	30.5000	10.1667
C4	3	6.0000	2.0000
C5	3	11.0000	3.6667
C6	3	3.5000	1.1667
C7	3	20.0000	6.6667
C8	3	11.5000	3.8333
C9	3	29.0000	9.6667
C10	3	28.5000	9.5000
C11	3	22.5000	7.5000
C12	3	36.0000	12.0000
Total	36	234.0000	6.5000

Number of Columns =	12
Number of Rows =	3
Coefficient =	0.9241
Correction for Ties =	15.0000
Chi-Square Statistic =	30.4965
Degrees of Freedom =	11
Right-Tail Probability =	0.0013

Since the right tail probability is less than 5% reject the null hypothesis.

6.5.7. Page’s L Trend

This is an extension of Friedman Two-Way ANOVA by ranks. Page’s L Trend is useful when trends between several variables are examined. It is defined as:

$$L = \sum ct_c$$

where t_c is the column rank totals and c is the column number.

Data entry is in matrix format (see 6.0.5. Tests with Matrix Data). Columns selected for this test must have equal number of rows and rows containing at least one missing value are omitted.

The output includes the rank sum and mean rank for each variable. For the probability level, tables for Page’s L must be consulted.

Example

Table 85 on p. 227 from Cohen, L. & M. Holliday (1983). Cognitive gain scores in predicted order under four learning conditions are given. The null hypothesis “cognitive gains in the sample are not related to the predicted order of experimental treatments” is tested.

Open NONPARM2, select Statistics 1 → Nonparametric Tests (Multisample) → Page’s L Trend and select *Condition 1*, *Condition 2*, *Condition 3* and *Condition 4* (*C13* to *C16*) as [Variable]s to obtain the following results:

Page’s L Trend

	Cases	Rank Sum	Mean Rank
Condition 1	6	8.5000	1.4167
Condition 2	6	12.0000	2.0000
Condition 3	6	17.0000	2.8333
Condition 4	6	22.5000	3.7500
Total	24	60.0000	2.5000

```
   Number of Columns =    4
      Number of Rows =    6
     Page’s L Trend =    173.5000
```

For the probability value consult critical values table for Page’s L test

From tables, the critical value for 6 subjects and 4 conditions at 0.1% level is 172. Therefore, this result is significant at 0.001 (or 0.1%) level and thus reject the null hypothesis.

6.5.8. Cochran's Q

Data for this test must be in dichotomised form, i.e. in terms of 0 and 1. The hypothesis tested is whether three or more matched sets of frequencies or proportions differ significantly among themselves.

$$Q = \frac{(M-1)(MV)}{MT-H}$$

where V is the sum of C^2 - T^2, H is the sum of R^2 when Cs are column totals, Rs are row totals and T is sum total. Q has a chi-square distribution with M - 1 degrees of freedom. Missing values are not supported.

Example 1

Table 84 on p. 224 from Cohen, L. & M. Holliday (1983). Frequency of correct solutions to each of four problems by ten subjects are given. The null hypothesis "there is no significant difference in difficulty of the four problems" is tested.

Open NONPARM2, select Statistics 1 → Nonparametric Tests (Multisample) → Cochran's Q and select *Problem 1*, *Problem 2*, *Problem 3* and *Problem 4* (*C17* to *C20*) as [Variable]s to obtain the following results:

Cochran's Q

	Cases	0	1
Problem 1	10	4	6
Problem 2	10	3	7
Problem 3	10	3	7
Problem 4	10	1	9
Total	40	11	29

```
  Number of Columns =   4
     Number of Rows =   10
Chi-Square Statistic =  1.9655
 Degrees of Freedom =   3
Right-Tail Probability = 0.57959
```

This result is not significant at the 10% level. Hence do not reject the null hypothesis.

Example 2

Example 12.6 on p. 282, Zar, J. H. (2010). A researcher wants to test the null hypothesis "the proportion of humans attacked by mosquitoes is the same for all 5 clothing types" at a 95% confidence level.

Open NONPARM2, select Statistics 1 → Nonparametric Tests (Multisample) → Cochran's Q and include *Light loose*, *Light dark*, *Dark long*, *Dark short* and *None* (*C21* to *C25*) in the analysis by clicking [Variable] to obtain the following results:

Cochran's Q

	Cases	0	1
Light loose	8	6	2
Light dark	8	4	4
Dark long	8	4	4
Dark short	8	1	7
None	8	3	5
Total	40	18	22

Number of Columns =	5
Number of Rows =	8
Chi-Square Statistic =	6.9474
Degrees of Freedom =	4
Right-Tail Probability =	0.1387

Since the right tail probability is greater than 5%, do not reject the null hypothesis.

6.5.9. Kappa Test for Inter-Category Variation

Cohen's Kappa is a measure of agreement between two or more raters classifying a sample of items into one of k mutually exclusive and exhaustive unordered categories. Three versions of Kappa test are supported. The first two can be accessed under the menu option for multisample nonparametric tests and the third appears among the statistics for square cross-tabulations (see 6.6.2.2. R x C Table Statistics). You can also use the REGRESS, select Statistics 1 → Correlation Coefficients → Partial Correlation and select *temperature, cm* (*C1-C2*) as [Variable]s and *mm, min, ml* (*C3-C5*) as [Covariate]s to obtain the following results:

Partial Correlation Matrix

3 Order Correlations
Controlling for: mm, min, ml

	temperature			cm		
	Corr	DoF	2-Tail P	Corr	DoF	2-Tail P
Temperature				0.1943	28	0.3036
cm	0.1943	28	0.3036			

6.2.4. Intraclass Correlation Coefficients procedure to compare raters under up to six different sets of assumptions.

The first version is a generalised implementation of Kappa test for measuring agreement among many raters (or the inter-category agreement) introduced by Fleiss, J. L., (1971). The algorithm takes into account the correction in computing large sample variances by Fleiss, J. L., Nee, J. C. M. and Landis, J. R. (1979).

The data for this test is assumed to be in the form of a table where the rows represent subjects and the columns represent categories of classification. A fixed number of observers, say n, are assumed to rate each case. A *rating* means assigning the value 1 to a category for a subject. Therefore, each subject (row) should have n ratings and row totals should be equal to n. If this condition is not met the program will display a message and abort the procedure. Any rows containing one or more missing values are omitted.

If the data has not been formed into a frequency table, you can use the Cross-Tabulation procedure to generate the frequency counts and the Kappa test simultaneously (see 6.6.2.2. R x C Table Statistics).

Example

The data is taken from Fleiss, J L (1971).

Open NONPARM2, select Statistics 1 → Nonparametric Tests (Multisample) → Kappa Test Inter-Category Variation and select *Category1* to *Category5* (*C26* to *C30*) as [Variable]s, to obtain the following results:

Kappa Test (Inter-Category Variation)

	Total	Pexp	Pobs	Kappa	Z-Stat	Prob
Category1	26	0.1444	0.3538	0.2448	5.3335	0.0000
Category2	26	0.1444	0.3538	0.2448	5.3335	0.0000
Category3	30	0.1667	0.6000	0.5200	11.1723	0.0000
Category4	55	0.3056	0.6327	0.4711	10.1355	0.0000
Category5	43	0.2389	0.6698	0.5661	12.1506	0.0000
Total	180	0.2199	0.5556	0.4302	17.9253	0.0000

Standard Deviation =	0.0244
95% Confidence Interval =	0.3825 <> 0.4780

6.5.10. Kappa Test for Inter-Observer Variation

This version will calculate a test statistic to measure the degree of agreement between two raters. The present implementation is the original form of Kappa test as introduced by Cohen, J. A. (1960) and Cohen, J. A. (1968), and also takes into account the correction in the calculation of large sample variances by Fleiss, J. L., Cohen, J., Everitt, B. S., (1969). A weighted Kappa is computed if weights are given and an unweighted Kappa otherwise.

The data for this test must be in the form of a square matrix. Columns of the matrix can be selected from the Variables Available list by clicking on [Variable]. If the number of rows is not equal to the number of columns the test cannot be performed. Missing values are not allowed.

If a weighted analysis is to be performed weights should also be arranged in the form of a square matrix with the same dimensions. Columns of the weights matrix can be selected from the Variables Available list by clicking on [Weight].

Example

The data is taken from Fleiss, J. L., Cohen, J., Everitt, B. S., (1969).

Open NONPARM2, select Statistics 1 → Nonparametric Tests (Multisample) → Kappa Test Inter-Observer Variation and select *Diagnosis 1*, *Diagnosis 2* and *Diagnosis 3* (*C31* to *C33*) as [Variable]s to obtain the following results:

Kappa Test (Inter-Observer Variation)

Variables Selected: Diagnosis1, Diagnosis2, Diagnosis3

Expected Proportion = 0.7000
Observed Proportion = 0.4750

	Value	Standard Error	Z-Statistic	1-Tail Probability
Kappa	0.4286	0.0537	7.9792	0.0000

	2-Tail Probability	Lower 95%	Upper 95%
Kappa	0.0000	0.3233	0.5338

Keeping the *Diagnosis* columns as [Variable]s, select *Weight 1*, *Weight 2* and *Weight 3* (*C34* to *C36*) as [Weight]s:

Kappa Test (Inter-Observer Variation)

Variables Selected: Diagnosis1, Diagnosis2, Diagnosis3
Weights: Weight1, Weight2, Weight3

Expected Proportion = 0.5672
Observed Proportion = 0.7867

	Value	Standard Error	Z-Statistic	1-Tail Probability
Weighted Kappa	0.5071	0.0570	8.8967	0.0000

	2-Tail Probability	Lower 95%	Upper 95%
Weighted Kappa	0.0000	0.3954	0.6188

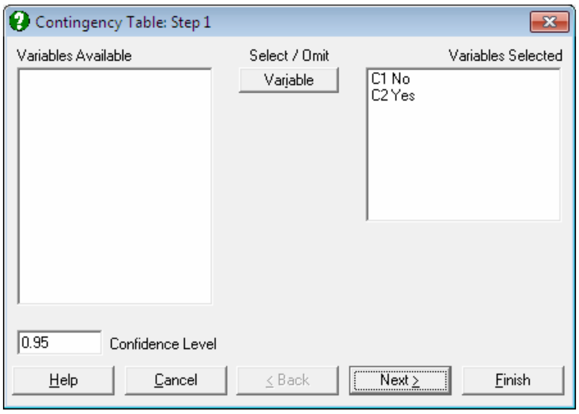
6.6. Tables

Contingency Table, Cross-Tabulation and Break-Down analyses are brought together under this section.

Contingency Table and Cross-Tabulation are provided as separate procedures to enable the user to analyse contingency tables using two distinct data types. The main difference between the two is that while the Contingency Table procedure accepts frequency count data already formed into a table, Cross-Tabulation will *generate* a frequency counts table from categorical variables. These can be numeric or string variables containing a limited number of distinct values. The form of output and the statistics reported are similar for the two procedures, with the exception of Stratified Analysis which is only available for multi-way cross-tabulations. Both procedures report an extensive set of statistics, including those covered under Binomial Proportion, Unpaired Proportions and Paired Proportions procedures. Information common to both procedures will be discussed in section 6.6.2. Cross-Tabulation.

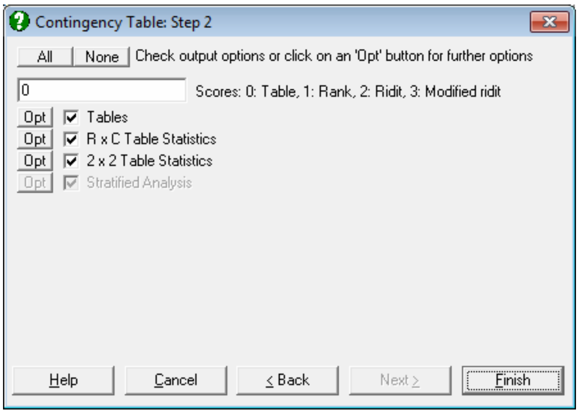
6.6.1. Contingency Table

This procedure assumes that the data is in the form of frequency counts and entered in a table format. Any number of data columns can be selected as the columns of a Contingency Table, provided that their lengths are equal. If this condition is not met, then the program will not proceed.



After selecting the variables, an Output Options Dialogue asking for the score method to be used (see 6.6.2.0. Scores) and containing four check boxes (with an

[Opt] button to their left) will be displayed. Each one of these options has further options. The full output from the Contingency Table procedure is large and computations are demanding. For a full discussion of these options see section 6.6.2. Cross-Tabulation.


Contingency Table: Step 2
All
None
Check output options or click on an 'Opt' button for further options
0
Scores: 0: Table, 1: Rank, 2: Ridit, 3: Modified ridit
Opt ☑ Tables
Opt ☑ R x C Table Statistics
Opt ☑ 2 x 2 Table Statistics
Opt ☑ Stratified Analysis
Help
Cancel
< Back
Next >
Finish

Example 1

Example 23.1 on p. 490 from Zar, J. H. (2010). The null hypothesis "Human hair colour is independent of sex in the population sampled" is tested.

Open TABLES, select Statistics 1 → Tables → Contingency Table and select *Black*, *Brown*, *Blond* and *Red* (*C1* to *C4*) as [Variable]s. From the Tables options select only the Frequency and from R x C Table Statistics select only the Chi-square Tests output options. Go back to Step 2 Output Options Dialogue and click [Finish] to obtain the following results:

Contingency Table

2 Rows x 4 Columns

Frequency

	Black	Brown	Blond	Red	Row Sum
R1	32.0000	43.0000	16.0000	9.0000	100.0000
R2	55.0000	65.0000	64.0000	16.0000	200.0000
Column Sum	87.0000	108.0000	80.0000	25.0000	300.0000

Chi-square Tests

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	8.9872	3	0.0295
Likelihood-Ratio	9.5121	3	0.0232
+ Yates Correction			
# Linear-by-linear	2.6155	1	0.1058
~ McNemar-Bowker			

+ Reported for 2 x 2 tables.
# Table scores
~ Reported for 3 x 3 or larger square tables.
Cells with expected count < 5 = 0 (0.00%)
Minimum expected count = 8.3333

Phi =	0.1731
Contingency Coefficient =	0.1705
Cramer's V =	0.1731

In this example Zar only reports Pearson's chi-square statistic and its tail probability.

Example 2

Example 23.4 on p. 499 from Zar, J. H. (2010). The null hypothesis "the ability of snails to resist the current is no different between the two species" is tested. Open TABLES, select Statistics 1 → Tables → Contingency Table, select *Resisted* and *Yielded* (*C10* and *C11*) as [Variable]s. Leave output option selections as in the previous example.

Contingency Table

2 Rows x 2 Columns

Frequency

	Resisted	Yielded	Row Sum
R1	12.0000	7.0000	19.0000
R2	2.0000	9.0000	11.0000
Column Sum	14.0000	16.0000	30.0000

Chi-square Tests

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	5.6622	1	0.0173
Likelihood-Ratio	6.0162	1	0.0142
+ Yates Correction	3.9993	1	0.0455
# Linear-by-linear	5.4734	1	0.0193
~ McNemar-Bowker			

# Table scores
~ Reported for 3 x 3 or larger square tables.
Cells with expected count < 5 = 0 (0.00%)
Minimum expected count = 5.1333

Phi =	0.4344
Contingency Coefficient =	0.3985
Cramer's V =	0.4344

Zar reports the chi-square statistic with Yates correction and its tail probability.

Example 3

Example 8.4 on p. 231 from Armitage & Berry (2002). The effects of PAS and streptomycin in the treatment of pulmonary tuberculosis are given. The null hypothesis “the probabilities for rows (columns) to fall into different columns (rows) are the same” is tested.

Open TABLES, select Statistics 1 → Tables → Contingency Table and include *Negative smear*, *NegSmr-PosCult* and *NegSmr-NegCult* (*C7* to *C9*) as [Variable]s. Include only Frequency, Expected and Chi-square Statistics output options to obtain the following results:

Contingency Table

3 Rows x 3 Columns

Frequency

	Negative smear	NegSmr-PosCult	NegSmr-NegCult	Row Sum
R1	56.0000	30.0000	13.0000	99.0000
R2	46.0000	18.0000	20.0000	84.0000
R3	37.0000	18.0000	35.0000	90.0000
Column Sum	139.0000	66.0000	68.0000	273.0000

Expected

	Negative smear	NegSmr-PosCult	NegSmr-NegCult
R1	50.4066	23.9341	24.6593
R2	42.7692	20.3077	20.9231
R3	45.8242	21.7582	22.4176

Chi-square Tests

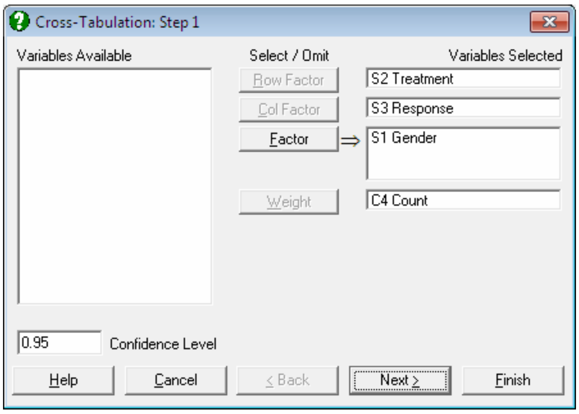
	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	17.6284	4	0.0015
Likelihood-Ratio	17.7770	4	0.0014
+ Yates Correction			
# Linear-by-linear	11.4263	1	0.0007
McNemar-Bowker	14.9937	3	0.0018

+ Reported for 2 x 2 tables.
# Table scores
Cells with expected count < 5 = 0 (0.00%)
Minimum expected count = 20.3077

Phi =	0.2541
Contingency Coefficient =	0.2463
Cramer's V =	0.1797

6.6.2. Cross-Tabulation

Multi-way cross-tabulations with or without weights can be performed.


Cross-Tabulation: Step 1

Variables Available

Select / Omit

Variables Selected

Row Factor: S2 Treatment

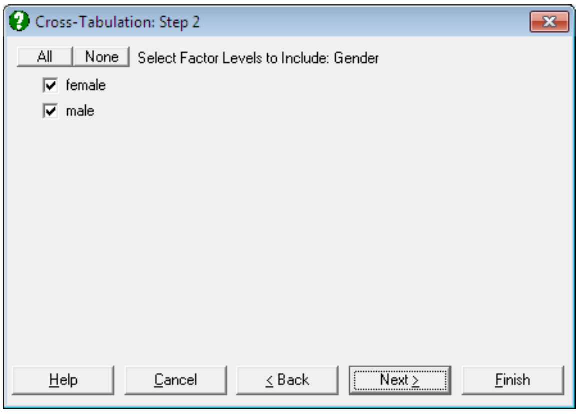
Col Factor: S3 Response

Factor: S1 Gender

Weight: C4 Count

0.95 Confidence Level

At least one factor column should be assigned as the [Row Factor] and another as the [Column Factor], whose categories are to be displayed on the rows and columns of the generated table respectively. Optionally, an unlimited number of further factor columns can be selected to run multi-way analyses. In this case, a further set of output options for Stratified Analysis will become available.

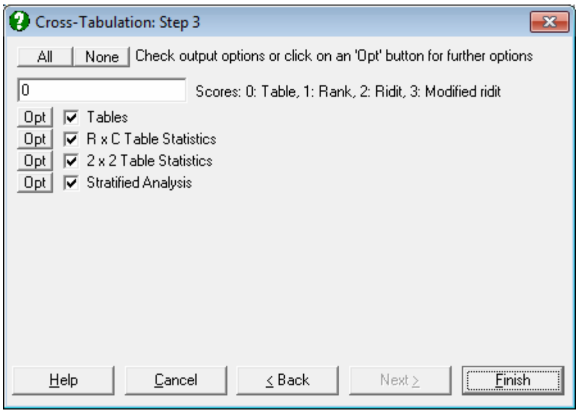

Cross-Tabulation: Step 2

Select Factor Levels to Include: Gender

☑ female

☑ male

The program will sort each column separately, determine the number of categories in each and then count the frequency of occurrence of every combination of categories. For instance, in a two-way analysis, if the variable [Row Factor] takes on values 0, 1, and 2 and the variable [Column Factor] takes on four different values such as 3, 4, 5 and 6, the program will produce a 3 by 4 table (number of rows by number of columns). When further [Factor] variables are included in the analysis, a separate table and its associated statistics will be generated for each combination of factor levels selected.


Cross-Tabulation: Step 3

All | None | Check output options or click on an 'Opt' button for further options

0 Scores: 0: Table, 1: Rank, 2: Ridit, 3: Modified ridit

Opt ☑ Tables
Opt ☑ R x C Table Statistics
Opt ☑ 2 x 2 Table Statistics
Opt ☑ Stratified Analysis

After processing the data, the program will prompt for output options. Not all output options will be available for all data sets. For instance, 2 x 2 Table Statistics option will be available when row and column factors have two levels each, and the Stratified Analysis option will only be available for Cross-Tabulation procedure when a factor column is selected.

6.6.2.0. Scores

Some of the statistics computed for this procedure depend on the ordering of rows and columns of the contingency table. These statistics are Linear-by-linear (also known as Mantel-Haenszel chi-square test), Pearson correlation, Eta and Cochran-Armitage trend test. First define the notation used in this section:

n - sum total of all frequencies
r - number of rows in contingency table
c - number of columns in contingency table
r_i - sum of frequencies in the i^{th} row
c_j - sum of frequencies in the j^{th} column

One of the following four methods can be used:

Table scores: This is the default score method. The column and row index numbers are used as factor levels.

Rank scores: These are the rank values as used in Spearman rank correlation coefficient and they are computed as:

$$R_i = \sum_{k=1,i-1} r_k + (c_i + 1), i = 1, \ldots, r$$

$$R_j = \sum_{l=1,j-1} c_l + (r_j + 1), j = 1, \ldots, c$$

Ridit scores: These are the rank scores divided by the total sample size:

$$Rr_i = R_i/n, i = 1, \ldots, r$$

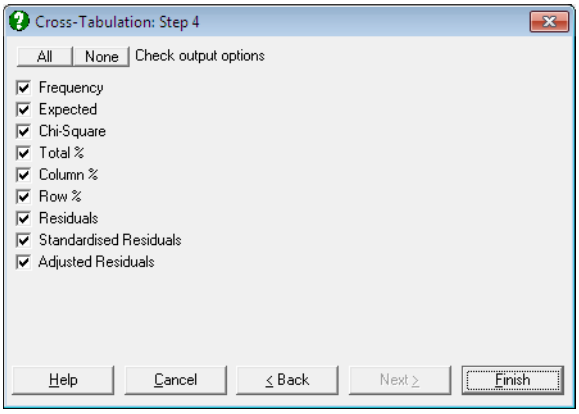
$$Rr_j = R_j/n, j = 1, \ldots, c$$

Modified ridit scores: These are the rank scores divided by the total sample size + 1:

$$Rr_i = R_i/(n+1), i = 1, \ldots, r$$

$$Rr_j = R_j/(n+1), j = 1, \ldots, c$$

6.6.2.1. Tables



Cross-Tabulation: Step 4

Check output options

☑ Frequency
☑ Expected
☑ Chi-Square
☑ Total %
☑ Column %
☑ Row %
☑ Residuals
☑ Standardised Residuals
☑ Adjusted Residuals

Each output option is displayed as a separate table, which can be sent to Data Processor for further analysis.

Frequency: For Contingency Table, this is the data as it exists in the spreadsheet. It may contain non-integer and / or missing data. For Cross-Tabulation, it consists of the computed frequency figures for categories represented by the row and column factors. Therefore, it always contains integers and no missing data. Define:

$$\text{observed cell frequency} = n_{ij}$$

This table also displays the row sum and column sum values on the table margins.

Expected Frequency: For each cell, the expected frequency is calculated as:

$$E_{ij} = \frac{r_i c_j}{n}$$

Chi-square: The contribution of each cell to the overall chi-square value is calculated as:

$$\chi^2_{ij} = \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Total Percent: Percentage of the cell frequency out of sum total of all table frequencies:

$$T_{ij} = \frac{100 n_{ij}}{n}$$

This table also displays the row sum (i.e. percentage of the row total out of sum total) and column sum (i.e. percentage of the column total out of sum total) percentages on the table margins.

Row Percent: Percentage of the cell frequency out of sum of frequencies in the same row.

$$R_i = \frac{100 n_{ij}}{r_i}$$

Column Percent: Percentage of the cell frequency out of sum of frequencies in the same column.

$$C_j = \frac{100 n_{ij}}{c_j}$$

Residual: Difference between observed and expected frequencies.

$$R_{ij} = n_{ij} - E_{ij}$$

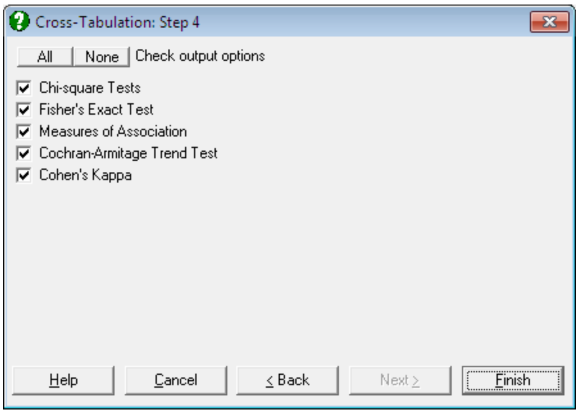
Standardised Residual:

$$SR_{ij} = \frac{R_{ij}}{\sqrt{E_{ij}}}$$

Adjusted Residual:

$$AR_{ij} = \frac{R_{ij}}{\sqrt{E_{ij}(1 - r_i / n)(1 - c_j / n)}}$$

6.6.2.2. R x C Table Statistics



Check output options

☑ Chi-square Tests
☑ Fisher's Exact Test
☑ Measures of Association
☑ Cochran-Armitage Trend Test
☑ Cohen's Kappa

6.6.2.2.1. Chi-square Tests

Several chi-square significance tests and chi-square related coefficients are reported. The program also reports the number of cells with an expected frequency less than 5, its percentage out of total frequency and the minimum expected frequency. The aim of this information is to enable the user to judge whether the asymptotic probabilities reported here are reliable enough, or the exact probabilities like Fisher's should be used.

Pearson Chi-square: This provides a measure of association between row and column factors.

$$\chi_P^2 = \sum_i^r \sum_j^c \chi_{ij}^2$$

$$df = (r-1)(c-1)$$

For rows and columns containing 0s only, the degrees of freedom is adjusted accordingly.

Likelihood Ratio: This involves the ratio of observed and expected frequencies.

$$\chi_{LR}^2 = -2\sum_i \sum_j n_{ij} Ln\left(\frac{E_{ij}}{n_{ij}}\right)$$

$$df = (r-1)(c-1)$$

Yates Correction: Also known as continuity correction, this is reported for 2 x 2 tables only.

$$\chi_Y^2 = \frac{Max\left(0, |n_{11} - n_{22}| - n/2\right)^2}{r_1 r_2 c_1 c_2}$$

$$df = 1$$

Linear-by-linear: This is also known as Mantel-Haenszel chi-square test. It tests the linear association between row and column factors:

$$\chi_L^2 = (n-1)R^2$$

$$df = 1$$

where R-squared is the Pearson correlation between the row factor and the column factor. The value of this test depends on the score method selected (see Scores).

McNemar-Bowker: This is a test of symmetry for square tables. Since it is identical to McNemar Test for 2 x 2 tables, it is reported only for 3 x 3 or larger square tables.

$$\chi_B^2 = \sum_{i=1}^{r-1} \sum_{j=i+1}^{c} \frac{\left(n_{ij} - n_{ji}\right)^2}{n_{ij} + n_{ji}}$$

$$df = r\left(r-1\right)/2$$

Phi Coefficient: For tables other than 2 x 2:

$$\varphi = \sqrt{\frac{\chi_P^2}{n}}, 0 \le \varphi \le \left(\mathrm{Min}(r,c) - 1\right)$$

and for 2 x 2 tables:

$$\varphi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}}, -1 \le \varphi \le 1$$

with a magnitude equal to Pearson correlation coefficient R (see Linear-by-linear Test above). Note that phi coefficient for 2 x 2 tables can be negative.

Contingency Coefficient:

$$CC = \sqrt{\frac{\chi_P^2}{\chi_P^2 + n}}, 0 \le CC \le \sqrt{(m-1)/m}$$

where:

$$m = \left(\mathrm{Min}(r,c) - 1\right)$$

Cramer’s V Coefficient: For tables other than 2 x 2:

$$V = \sqrt{\frac{\chi_P^2}{n\left(\mathrm{Min}(r,c) - 1\right)}}, 0 \le V \le 1$$

and for 2 x 2 tables:

$$V = \varphi, -1 \le V \le 1$$

Therefore, Cramer’s V coefficient for 2 x 2 tables can be negative.

6.6.2.2.2. Fisher’s Exact Test

Fisher’s Exact Test for R x C Tables is also known as the Freeman-Halton test. The network algorithm developed by Mehta and Patel (1983) is used to calculate the two-tailed and table probabilities. This procedure is computationally demanding and it may not be feasible for large tables. For 2 x 2 tables the results are identical to that of Fisher's Exact Test which is available under the Paired Proportions procedure (see 6.6.2.3. 2 x 2 Table Statistics below. The latter procedure also reports the left- and right- tail probabilities in addition to two-tailed and table probabilities.

6.6.2.2.3. Measures of Association

Tests in this section are used to assess the association between row and column factors of the table. All test statistics (with the exception of eta) are reported with their standard errors and asymptotic confidence intervals using the standard normal distribution:

$$\text{Stat} \pm Z_{\alpha/2}\text{StdErr}$$

The first group of tests (Somer's delta, Goodman-Kruskal's gamma, Kendall's tau b and tau c) will rank observations as concordant or discordant. Each observation is checked pairwise with all other observations to see whether its relative ordering is the same in the sequence. The ordering is called concordant if it is the same and discordant if it is different. These tests are computationally demanding. If the table is large, they may take a long time to compute.

The following entities are computed:

- C = number of concordant pairs
- D = number of discordant pairs
- T_x = number of ties in columns
- T_y = number of ties in rows.

Goodman-Kruskall's Gamma: This coefficient does not make an adjustment for ties. It is defined as:

$$\Gamma = \frac{C-D}{C+D}, -1 \le \Gamma \le 1$$

The value of gamma can be taken as the probability of correctly guessing the order of a pair of cases on one variable once the ordering on the other variable is known.

Kendall's tau b: This statistic reflects the effect of ties both in columns and rows:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

Kendall's tau c: This is also called Stuart's tau c.

$$\tau_c = \frac{2m(C - D)}{T^2(m - 1)}$$

where T is sum total of all ranks and m is the minimum of number of columns or number of rows.

$$m = \mathrm{Min}(r, c)$$

Pearson Correlation: The value of this test depends on the score method selected (see Scores).

$$R = \frac{ss_{RC}}{\sqrt{ss_R ss_C}}$$

where:

$$ss_R = \sum_i \sum_j n_{ij}\left(r_i - \bar{r}\right)$$

$$ss_C = \sum_i \sum_j n_{ij}\left(c_i - \bar{c}\right)$$

$$ss_{RC} = \sum_i \sum_j n_{ij}\left(r_i - \bar{r}\right)\left(c_i - \bar{c}\right)$$

Spearman Rank Correlation: The only difference from Pearson correlation is that ranks of factor levels (rank scores) are always used.

$$R_S = \frac{ss_{RC}}{\sqrt{ss_R ss_C}}$$

Eta: The value of this statistic depends on the score method selected (see Scores). The asymmetric eta for the column factor is defined as:

$$\eta_Y = \sqrt{1 - \frac{S_{YW}}{S(Y)}}$$

where Yj are the levels of column factor and:

$$S_{YW} = \sum_i \sum_j n_{ij} Y_j^2 - \sum_{i=1}^{r} \frac{1}{r_i} \left(\sum_{j=1}^{c} Y_j n_{ij} \right)^2$$

Eta for the row factor can be obtained by reversing the indices.

The third group of tests of association will calculate a row statistic (given the column factor), a column statistic (given the row factor) and a symmetric statistic. Here we will only define the statistic for the row factor. The column statistic can be obtained by reversing the indices.

Somer's Delta: Somer's delta also uses concordant / discordant pairs to test asymmetric association between the row and column factors.

$$\delta_R = \frac{C - D}{D_R}$$

where:

$$D_R = n^2 \sum_i^r r_i^2$$

and the symmetric delta is:

$$\delta = \frac{C - D}{\frac{1}{2}(D_R + D_C)}$$

Goodman-Kruskall's Lambda:

$$\lambda_R = \frac{\sum_i r_{i\,max} - r_{max}}{n - r_{max}}$$

where:

$$r_{i\max} = \mathrm{Max}_i(n_{ij})$$

$$r_{\max} = \mathrm{Max}_i(r_i)$$

and the symmetric lambda is the average of the two asymmetric lambdas:

$$\lambda = \frac{\sum_i r_{i\max} \sum_j c_{j\max} - r_{\max} - c_{\max}}{2n - r_{\max} - c_{\max}}$$

Uncertainty Coefficient: This measures the proportion of uncertainty (entropy) in the column variable Y that is explained by the row variable X:

$$U_R = \frac{H(X) + H(Y) - H(XY)}{H(Y)}$$

where:

$$H(X) = -\sum_i \left(\frac{r_i}{n}\right) \mathrm{Ln}\left(\frac{r_i}{n}\right)$$

$$H(Y) = -\sum_j \left(\frac{c_j}{n}\right) \mathrm{Ln}\left(\frac{c_j}{n}\right)$$

$$H(XY) = -\sum_i \sum_j \left(\frac{n_{ij}}{n}\right) \mathrm{Ln}\left(\frac{n_{ij}}{n}\right)$$

and the symmetric uncertainty coefficient is:

$$U = \frac{2\left[H(X) + H(Y) - H(XY)\right]}{H(X) + H(Y)}$$

6.6.2.2.4. Cochran-Armitage Trend Test

This test is available only for 2 x c or r x 2 tables. It provides a test statistic for the trend of binary proportions in levels of a factor with two or more levels (the response variable). The latter is assumed to contain binary responses (e.g. yes / no) and its levels are internally recoded as 0 and 1. For a table with two columns and r rows, the trend is defined as:

$$T = \frac{\sum_{i=1}^{r} n_{i1}\left(X_i - \overline{X}\right)}{\sqrt{p_1\left(1-p_1\right)s^2}}$$

where:

$$p_1 = \frac{c_1}{n}$$

and:

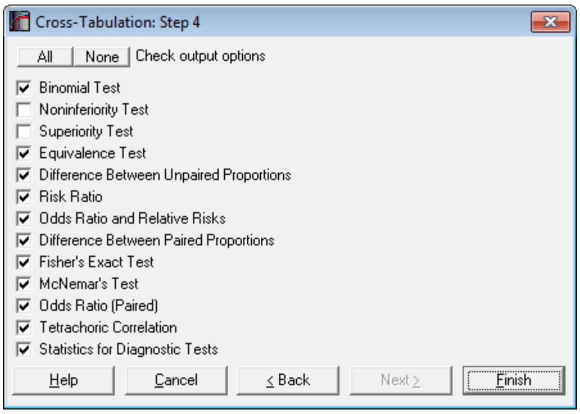
$$s^2 = \sum_{i=1}^{r} r_i\left(X_i - \overline{X}\right)^2$$

Here, X_i is the ith level of the row variable and $\overline{X}$ is the weighted average of row variable levels. Therefore, Cochran-Armitage trend test depends on the score method selected (see Scores).

6.6.2.2.5. Cohen's Kappa

Kappa is computed for square tables. Only an unweighted Kappa test is performed. For details see 6.5.10. Kappa Test for Inter-Observer Variation.

6.6.2.3. 2 x 2 Table Statistics



Check output options

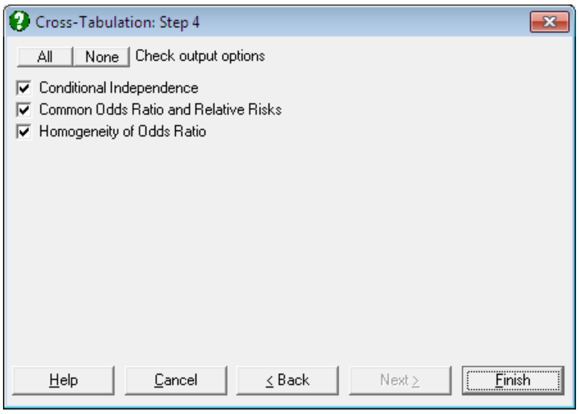
☑ Binomial Test
☐ Noninferiority Test
☐ Superiority Test
☑ Equivalence Test
☑ Difference Between Unpaired Proportions
☑ Risk Ratio
☑ Odds Ratio and Relative Risks
☑ Difference Between Paired Proportions
☑ Fisher's Exact Test
☑ McNemar's Test
☑ Odds Ratio (Paired)
☑ Tetrachoric Correlation
☑ Statistics for Diagnostic Tests

When a 2 x 2 table is formed, all statistics available under Binomial Proportion, Unpaired Proportions and Paired Proportions sections will also be available in this procedure. The user should take care to distinguish between the table formed by the Cross-Tabulation procedure and one formed on the same pair of columns by the Unpaired Proportions procedure. Here, the total table frequency is the number valid pairs (as in Paired Proportions), whereas in Unpaired Proportions the total frequency is the sum of valid cases in sample 1 and sample 2.

- Binomial Test**:** This test is performed on the row factor and the column factor separately, i.e. on the row sums and the column sums of the table. Therefore, one needs to take this into consideration when the Contingency Table option is selected. By default, the program uses an expected proportion of 0.5. You can change this to any value between 0 and 1 from Binomial Proportion → Binomial Test procedure.
- Noninferiority Test**:** This test is also performed on the row factor and the column factor separately. Expected proportion (default 0.5) and the noninferiority margin (default 0.2) can be changed from Binomial Proportion → Noninferiority Test procedure.
- Superiority Test**:** This test is also performed on the row factor and the column factor separately. Expected proportion (default 0.5) and the superiority margin (default 0.2) can be changed from Binomial Proportion → Superiority Test procedure.
- Equivalence Test for Binomial Proportion**:** This test is also performed on the row factor and the column factor separately. Expected proportion (default 0.5), the lower equivalence margin (default -0.2) and the upper equivalence margin (default 0.2) can be changed from Binomial Proportion → Equivalence Test for Binomial Proportion procedure.
- Difference Between Unpaired Proportions
- Risk Ratio
- Odds Ratio and Relative Risks
- Difference Between Paired Proportions
- Fisher's Exact Test
- McNemar Test
- Odds Ratio (Paired)
- Tetrachoric Correlation
- Statistics for Diagnostic Tests

This facility is especially useful when the four frequency values for a 2 x 2 table are already available in the spreadsheet. In this case they will not have to be typed again into the Cell Frequencies are Given dialogue of the nonparametric tests.

6.6.2.4. Stratified Analysis


Cross-Tabulation: Step 4

Check output options

☑ Conditional Independence
☑ Common Odds Ratio and Relative Risks
☑ Homogeneity of Odds Ratio

The tests in this group are available only for multi-way 2 x 2 cross-tabulations, i.e. when one or more factor variables are selected apart from the row and column factors, which have only two levels each. When a factor variable is selected, all tables and tests described above will be displayed once for each level (stratum) of the factor variable. The tests described in this section are reported only once at the end of the output.

6.6.2.4.1. Conditional Independence

These two chi-square statistics will test the independence of the row and column factors across strata.

Cochran:

$$\chi_C^2 = \frac{E_{11}}{\sqrt{\sum_{k=1}^{k} n_k p_{1k}(1-p_{1k})}}$$

$$df = 1$$

where:

$$E_{11} = \sum_k n_{11k} E_{11k}$$

$$p_{1k} = \frac{c_{1k}}{n_k}$$

and n_k is the total frequency of the kth stratum.

Mantel-Haenszel: This is similar to Cochran test but introduces a continuity correction:

$$\chi^2_{MH} = \frac{\left(\left|E_{11}\right| - 1/2\right)Sign\left(E_{11}\right)}{\sqrt{\sum_{k=1}^{k} \frac{r_{1k} r_{2k}}{n_k - 1} p_{1k}\left(1 - p_{1k}\right)}}$$

$$df = 1$$

6.6.2.4.2. Common Odds Ratio and Relative Risks

The common odds ratio for multi-way 2 x 2 tables is estimated by Mantel-Haenszel and logit methods. The odds ratio for the kth stratum is defined as:

$$OR_k = \frac{n_{11k} n_{22k}}{n_{12k} n_{21k}}$$

See Odds Ratio and Relative Risks.

Mantel-Haenszel estimate: For multi-way tables the common odds ratio is defined as:

$$OR_{MH} = \frac{\sum_k \frac{n_{11k} n_{22k}}{n_k}}{\sum_k \frac{n_{12k} n_{21k}}{n_k}}$$

and cohort 1 is:

$$RR_{MH} = \frac{\sum_k \frac{n_{11k} c_{2k}}{n_k}}{\sum_k \frac{n_{21k} c_{1k}}{n_k}}$$

Logit estimate:

$$OR_L = \mathrm{Exp}\left(\frac{\sum\limits_k w_k \mathrm{Ln}(OR_k)}{\sum\limits_k w_k}\right)$$

where:

$$w_k = \frac{1}{\mathrm{Var}(\mathrm{Ln}(OR_k))}$$

and cohort 1 is:

$$RR_L = \mathrm{Exp}\left(\frac{\sum\limits_k w_k \mathrm{Ln}(RR_k)}{\sum\limits_k w_k}\right)$$

6.6.2.4.3. Homogeneity of Odds Ratio

The null hypothesis “odds ratios across strata are equal” is tested.

Breslow-Day: This test is based on Mantel-Haenszel estimate of the common odds ratio:

$$\chi^2_{BD} = \sum_k \frac{(n_{11k} - \alpha_{11k})^2}{\beta_{11k}}$$

$$df = K - 1$$

where:

$$\frac{\alpha_{11k}(n_k - r_{1k} - c_{1k} + \alpha_{11}k)}{(r_{1k} - \alpha_{11k})(c_{1k} - \alpha_{11k})} = OR_{MH}$$

$$\beta_{11k} = \left(\frac{1}{\alpha_{11k}} + \frac{1}{r_{1k} - \alpha_{11k}} + \frac{1}{c_{1k} - \alpha_{11k}} + \frac{1}{n_k - r_{1k} - c_{1k} - \alpha_{11k}}\right)^{-1}$$

Tarone: This is a modification of Breslow-Day test:

$$\chi_T^2 = \chi_{BD}^2 - \frac{\sum_k \left(n_{11k} - \alpha_{11k}\right)^2}{\sum_k \beta_{11k}}$$

$$df = K - 1$$

Example

Open TABLES and select Statistics 1 → Tables → Cross-Tabulation. Select *Gender* (*S12*) as [Factor], *Treatment* (*S13*) as [Row Factor], *Response* (*S14*) as [Column Factor] and *Count* (*C15*) as [Weight]. At the next dialogue check all output options.

Output for the second stratum Gender = female is removed to save space.

Cross-Tabulation

Treatment (2 Rows) x Response (2 Columns)

Subsample selected by: Gender = female

Frequency

Treatment \ Response	Better	Same	Row Sum
Active	16.0000	11.0000	27.0000
Placebo	5.0000	20.0000	25.0000
Column Sum	21.0000	31.0000	52.0000

Expected

Treatment \ Response	Better	Same
Active	10.9038	16.0962
Placebo	10.0962	14.9038

Chi-Square

Treatment \ Response	Better	Same
Active	2.3818	1.6135
Placebo	2.5723	1.7426

Total %

Treatment \ Response	Better	Same	Row Sum
Active	30.77%	21.15%	51.92%
Placebo	9.62%	38.46%	48.08%
Column Sum	40.38%	59.62%	100.00%

Column %

Treatment \ Response	Better	Same
Active	76.19%	35.48%
Placebo	23.81%	64.52%

Row %

Treatment \ Response	Better	Same
Active	59.26%	40.74%
Placebo	20.00%	80.00%

Residuals

Treatment \ Response	Better	Same
Active	5.0962	-5.0962
Placebo	-5.0962	5.0962

Standardised Residuals

Treatment \ Response	Better	Same
Active	1.5433	-1.2702
Placebo	-1.6039	1.3201

Adjusted Residuals

Treatment \ Response	Better	Same
Active	2.8827	-2.8827
Placebo	-2.8827	2.8827

Chi-square Tests

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	8.3102	1	0.0039
Likelihood-Ratio	8.6334	1	0.0033
Yates Correction	6.7595	1	0.0093
# Linear-by-linear	8.1504	1	0.0043
~ McNemar-Bowker			

# Table scores
~ Reported for 3 x 3 or larger square tables.
Cells with expected count < 5 = 0 (0.00%)
Minimum expected count = 10.0962

Phi =	0.3998
Contingency Coefficient =	0.3712
Cramer's V =	0.3998

Fisher's Exact Test

	2-Tail Probability	**Table Probability**
Fisher's Exact	0.0052	0.0036

For an extended output see Fisher's exact test for 2 x 2 tables.

Measures of Association

	Value	**Standard Error**	**Lower 95%**	**Upper 95%**
Goodman-Kruskal's gamma	0.7067	0.1590	0.3951	1.0183
Kendall's tau b	0.3998	0.1247	0.1554	0.6441
Kendall's tau c	0.3920	0.1237	0.1495	0.6346
# Pearson Correlation	0.3998	0.1247	0.1554	0.6441
Spearman Rank Correlation	0.3998	0.1247	0.1554	0.6441
# Eta (col)	0.3998			
# Eta (row)	0.3998			
Somer's delta (col)	0.3926	0.1239	0.1498	0.6354
Somer's delta (row)	0.4071	0.1266	0.1590	0.6552
Somer's delta symmetric	0.3997	0.1246	0.1554	0.6440
Lambda (col)	0.2381	0.2160	-0.1852	0.6614
Lambda (row)	0.3600	0.1782	0.0108	0.7092
Lambda symmetric	0.3043	0.1729	-0.0346	0.6433
Uncertainty coefficient (col)	0.1231	0.0793	-0.0323	0.2784
Uncertainty coefficient (row)	0.1199	0.0775	-0.0320	0.2718
Uncertainty coefficient symmetric	0.1215	0.0783	-0.0321	0.2750

# Table scores

## *Cochran-Armitage Trend Test*

	Z-Statistic	**1-Tail Probability**	**2-Tail Probability**
# Cochran-Armitage Trend Test	-1.2251	0.1103	0.2205

# Table scores
Binary variable is recoded to 0s and 1s.
Reported for 2 x K tables.

Cohen's Kappa

Expected Proportion =	0.4963
Observed Proportion =	0.6923

	Value	Standard Error	Z-Statistic	1-Tail Probability
Kappa	0.3891	0.1239	3.1404	0.0008

	2-Tail Probability	Lower 95%	Upper 95%
Kappa	0.0017	0.1463	0.6320

Binomial Test

For Treatment
Subsample selected by: Gender = female

Expected Proportion =	0.5000
Observed Proportion =	0.5192

	Proportion used in SE	Standard Error	Z-Statistic	1-Tail Probability
Wald	0.5192	0.0693		
H0	0.5000	0.0693	0.2774	0.3908
Wald with CC	0.5192	0.0693		
H0	0.5000	0.0693	0.1387	0.4449
Wilson (score)				
Wilson with CC				
Agresti-Coull	0.5179	0.0669		
Agresti-Coull (+2)	0.5179	0.0668		
Jeffreys				
Clopper-Pearson (exact)				0.4449

	2-Tail Probability	Lower 95%	Upper 95%
Wald		0.3834	0.6550
H0	0.7815		
Wald with CC		0.3738	0.6646
H0	0.8897		
Wilson (score)		0.3869	0.6490
Wilson with CC		0.3778	0.6578
Agresti-Coull		0.3869	0.6490
Agresti-Coull (+2)		0.3870	0.6487
Jeffreys		0.3855	0.6509
Clopper-Pearson (exact)	0.8899	0.3763	0.6599

Binomial Test

For Response
Subsample selected by: Gender = female

Expected Proportion =	0.5000
Observed Proportion =	0.4038

	Proportion used in SE	Standard Error	Z-Statistic	1-Tail Probability
Wald	0.4038	0.0680		
H0	0.5000	0.0693	-1.3868	0.0828
Wald with CC	0.4038	0.0680		
H0	0.5000	0.0693	-1.2481	0.1060
Wilson (score)				
Wilson with CC				
Agresti-Coull	0.4105	0.0658		
Agresti-Coull (+2)	0.4107	0.0657		
Jeffreys				
Clopper-Pearson (exact)				0.1058

	2-Tail Probability	Lower 95%	Upper 95%
Wald		0.2705	0.5372
H0	0.1655		
Wald with CC		0.2609	0.5468
H0	0.2120		
Wilson (score)		0.2816	0.5393
Wilson with CC		0.2731	0.5487
Agresti-Coull		0.2814	0.5395
Agresti-Coull (+2)		0.2819	0.5396
Jeffreys		0.2786	0.5394
Clopper-Pearson (exact)	0.2116	0.2701	0.5490

Difference Between Unpaired Proportions

Proportion 1 =	0.7619
Proportion 2 =	0.3548

	Difference	Standard Error	Z-Statistic	1-Tail Probability
Pooled Variance	0.4071	0.1412	2.8827	0.0020
Separate Variance		0.1266	3.2158	0.0007

	2-Tail Probability	Lower 95%	Upper 95%
Pooled Variance	0.0039	0.1303	0.6838
Separate Variance	0.0013	0.1590	0.6552

Risk Ratio

	Value	Lower 95%	Upper 95%
Risk Ratio	2.1472	1.2620	3.6533

Odds Ratio and Relative Risks

	Value	Lower 95%	Upper 95%
Odds Ratio	5.8182	1.6755	20.2034
Exact		1.4609	25.2654
Relative Risk (Cohort 1)	2.9630	1.2740	6.8913
Relative Risk (Cohort 2)	0.5093	0.3103	0.8357

Difference Between Paired Proportions

Proportion 1 =	0.2115
Proportion 2 =	0.0962

	Difference	Standard Error	Z-Statistic	1-Tail Probability
Asymptotic	-0.1154	0.0752	-1.5000	0.0668
Exact Binomial				0.1051

	2-Tail Probability	Lower 95%	Upper 95%
Asymptotic	0.1336	-0.2629	0.0321
Exact Binomial	0.2101	-0.2399	0.0533

Fisher's Exact Test

	Left-Tail	Right-Tail	Two-Tail	Table
Probability	0.9994	0.0042	0.0052	0.0036

McNemar's Test

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Asymptotic	2.2500	1	0.1336
Asymptotic with CC	1.5625	1	0.2113

	2-Tail Probability	Lower 95%	Upper 95%
Exact Binomial	1.0000	0.7047	8.0769

Odds Ratio (Paired)

	Value	Lower 95%	Upper 95%
Odds Ratio (Paired)	0.4545	0.7047	8.0769

Tetrachoric Correlation

	Ratio	Tetrachoric Correlation
	5.8182	0.6052

Statistics for Diagnostic Tests

	Value	Standard Error	Lower 95%	Upper 95%
Sensitivity	0.6316	0.1107	0.4147	0.8485
			0.3836	0.8371
Specificity	0.5429	0.0842	0.3778	0.7079
			0.3665	0.7117
Accuracy	0.5741	0.0673	0.4422	0.7060
			0.4321	0.7077
Prevalence	0.3519	0.0650	0.2245	0.4792
			0.2268	0.4938
Apparent Prevalence	0.5185	0.0680	0.3853	0.6518
			0.3784	0.6566
Youden's Index	0.1744			
			-0.2500	0.5488
Positive Predictive Value	0.4286	0.0935	0.2453	0.6119
			0.2446	0.6282
Negative Predictive Value	0.7308	0.0870	0.5603	0.9013
			0.5221	0.8843
Positive Likelihood Ratio	1.3816		0.8394	2.2739
Negative Likelihood Ratio	0.6787		0.3593	1.2818
Diagnostic Odds Ratio	2.0357		0.6478	6.3975
Weighted Positive Likelihood Ratio	0.7500		0.4394	1.2801
Weighted Negative Likelihood Ratio	0.3684		0.1902	0.7136

Smaller factor level represents the positive outcome.
Confidence Intervals: Row 1: Asymptotic Normal, Row 2: Exact Binomial

Treatment (2 Rows) x Response (2 Columns)

Subsample selected by: Gender = male

. . .

Conditional Independence

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Cochran	8.4650	1	0.0036
Mantel-Haenszel	7.1983	1	0.0073

Common Odds Ratio and Relative Risks

	Value	Standard Error	Lower 95%	Upper 95%
Common Odds Ratio	3.3132	0.4232	1.4456	7.5934
Cohort 1	2.1636	0.2867	1.2336	3.7948
Cohort 2	0.6420	0.1586	0.4705	0.8761
Logit(Odds Ratio)	3.2941	0.4300	1.4182	7.6515
Cohort 1	2.1059	0.2890	1.1951	3.7108
Cohort 2	0.6613	0.1580	0.4852	0.9013
Ln(Odds Ratio)	1.1979	0.4232	0.3685	2.0273

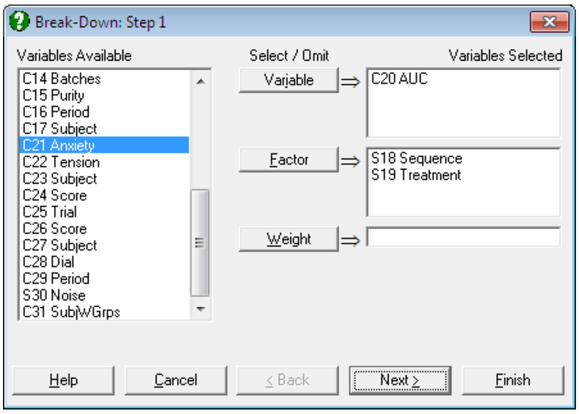
Homogeneity of Odds Ratio

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Breslow-Day	1.4929	1	0.2218
Tarone	1.4905	1	0.2221

6.6.3. Break-Down

This procedure produces a table featuring the number of cases, means and standard deviations for an unlimited number of continuous variables, broken down by an unlimited number of categorical variables. Each row of the list corresponds to a particular combination of levels of categorical variables (or factors) which can contain numeric or String Data. Factors are selected by clicking on [Factor] (see 6.0.8. R x C Tables). It is compulsory to select at least one factor column and a data variable (i.e. a continuous numeric data column) by clicking on [Variable]. A table will be displayed for each data variable, keeping the factor selections unchanged. Optionally, you can also select a column containing weights by clicking on [Weight]. In this case, all statistics displayed will be weighted by this column.

A similar analysis can be performed in the Regression and ANOVA module using the Table of Means procedure.

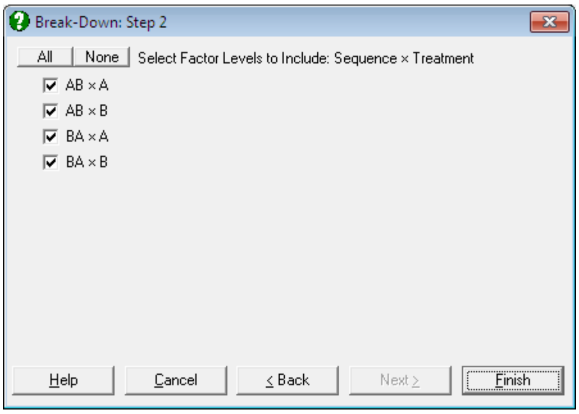


Break-Down: Step 1

Variables Available: C14 Batches, C15 Purity, C16 Period, C17 Subject, C21 Anxiety, C22 Tension, C23 Subject, C24 Score, C25 Trial, C26 Score, C27 Subject, C28 Dial, C29 Period, S30 Noise, C31 SubjWGrps

Variables Selected:
- Variable ⇒ C20 AUC
- Factor ⇒ S18 Sequence, S19 Treatment
- Weight ⇒

The program will sort each factor separately, determine the number of categories in each and then compute the number of cases, mean and standard deviation of the continuous data variable for every combination of categories.



Select Factor Levels to Include: Sequence × Treatment

☑ AB × A
☑ AB × B
☑ BA × A
☑ BA × B

Example

Open ANOVA and select Statistics 1 → Tables → Break-Down. Select *AUC* (*C20*) as [Variable] and *Sequence* (*S18*) and *Treatment* (*S19*) as [Factor]s.

Break-Down

Data variable: AUC

Sequence × Treatment	Cases	Mean	Standard Deviation	Standard Error
AB × A	6	247.0000	56.6110	23.1113
AB × B	6	157.3333	60.1387	24.5515
BA × A	6	171.8333	47.2246	19.2794
BA × B	6	177.0000	29.5025	12.0444

6.7. Sample Size and Power Estimation

Before testing a hypothesis it is desirable to know how large a sample size should be selected to achieve a desired precision. This depends on a number of factors specific to the nature of the test, such as sample variance, confidence level (α probability or Type I error) or minimum detectable difference.

It is also important to know how likely it is not to reject a null hypothesis when in fact it is false (β probability or Type II error), or in other words, to know what the power of the test is (1 - β), i.e. what is the probability of rejecting the null hypothesis when it is in fact false.

This section brings together seven broad classes of commonly used hypothesis tests and provides methods of estimating the sample size, power of the test and other parameters. The types of tests supported here are:

1) One Sample
2) Two Samples
3) Variance
4) Correlation
5) Two Correlations
6) Two Proportions
7) ANOVA

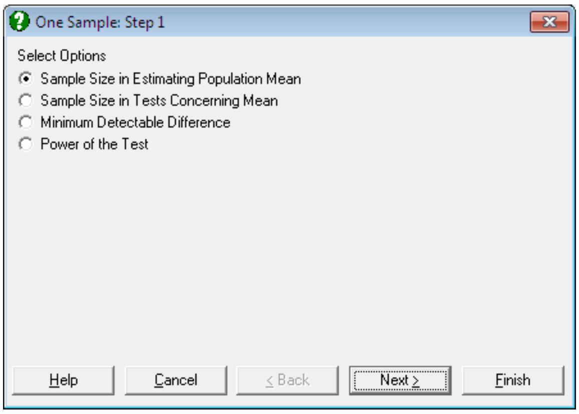
An eighth option is also provided to compute power of the test from the phi statistic and vice versa, which are used in estimating the sample size and power of the test in ANOVA and two sample tests. Therefore, UNISTAT does not require use of OC Curves published by Pearson and Hartley (1951), pp. 112-130.

Although some topics seem to have been excluded from this list, these are often special cases of the methods already provided. For instance, sample size and power of the test in Regression Analysis can be estimated using the Correlation option above. Many different types of ANOVA can also be accommodated simply by entering the relevant statistics in place of the existing parameters. In such cases, you are recommended to consult a statistics book to establish which of the existing procedures can be used as a substitute (see Zar, J. H. 2010).

In procedures where a selection of one or two-tailed estimation is available, the default is always set for two-tailed, corresponding to the null hypothesis that “the entities tested are equal” against the alternative hypothesis that “they are not equal”. Where the alternative hypothesis states a relationship of one entity being greater or less than the other, the one-tailed option should be selected.

In some procedures, the parameter to be estimated may occur on both sides of an equation and therefore it cannot be calculated directly. In such cases, an iterational algorithm is employed to determine the correct level of the parameter and usually convergence is achieved within a few iterations. In such procedures you are provided with two further input fields to control the two convergence parameters, tolerance and the maximum number of iterations. The default values of these two parameters are 0.001 and 100 respectively and they produce satisfactory results in most cases. If the convergence cannot be achieved within these values, then the program will report this in the output. Then you may edit the default values to obtain convergence.

6.7.1. One Sample


One Sample: Step 1

Select Options

● Sample Size in Estimating Population Mean
○ Sample Size in Tests Concerning Mean
○ Minimum Detectable Difference
○ Power of the Test

The first procedure provided here concerns estimation of population mean and as such it is inherently different from the last three, which are derived from the same equation, each time solving for a different parameter.

6.7.1.1. Sample Size in Estimating the Population Mean

The sample size necessary to achieve the desired level of precision in estimating a population mean is given by the following formula:

$$n = \frac{s^2 t^2_{\alpha(2),(n-1)}}{d^2}$$

Here:

- s^2 is the sample variance with ν degrees of freedom,
- d is the half-width of the desired confidence interval,
- 1 - α is the confidence level,
- 1 - β is the assurance that the confidence interval will not be greater,
- $t_{\alpha(2),(n-1)}$ is the two-tailed critical value for α from t-distribution with (n - 1) degrees of freedom.

Since critical value of t on the right hand side of this equation depends on the sample size n, the left hand side cannot be computed directly. Therefore, an iterational convergence algorithm is employed.

The user is expected to enter:

- Denominator Degrees of Freedom
- Sample Variance
- Half-Width of Confidence Interval
- Assurance for Confidence Interval
- Confidence Level

and the program will output the estimated sample size.

Example

Example 7.7 on p. 115 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → One Sample and the Sample Size in Estimating Population Mean option. Enter the following data at the next dialogue to obtain the Estimated Sample Size:

Sample Size and Power Estimation: One Sample

Sample Size in Estimating Population Mean

```
Denominator Degrees of Freedom =     2.0000
               Sample Variance =     0.4008
 Half-Width of Confidence Interval =     0.2500
Assurance for Confidence Interval =     0.9000
              Confidence Level =     0.9500
         Estimated Sample Size =     27.0740
```

6.7.1.2. Sample Size in Tests Concerning the Mean

The purpose of the experiment is to test if the sample is taken from a population with a specified mean “H_0: $\mu = \mu_0$” or it is from a population with a different mean “H_0: $\mu \neq \mu_0$”.

Sample size in tests about the mean is determined from:

$$n = \frac{s^2}{\delta^2}(t_{\alpha(2),\nu} + t_{\beta(1),\nu})^2$$

Here:

- s^2 is the sample variance with ν degrees of freedom,
- δ is the minimum detectable difference between the two population means at the desired confidence interval,
- α is the probability of committing a Type I error and 1- α is the confidence level, which can be one or two-tailed.
- β is the probability of committing a Type II error and 1 - β is the power of the test,
- $t_{\alpha(2),\nu}$ is the (one or) two-tailed critical value for α from t-distribution with ν degrees of freedom,
- $t_{\beta(1),\nu}$ is the one-tailed critical value for β from t-distribution with ν degrees of freedom.

Since critical values of t on the right hand side of this equation are dependent on the sample size n, an iterational algorithm is employed.

The user is expected to enter:

- Sample Variance
- Minimum detectable difference
- Power of the test
- Confidence Level
- 1 or 2 tailed test

and the program will output the estimated sample size.

Example

Example 7.8 on p. 116 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → One Sample and the Sample Size in Tests

Concerning Mean option. Enter the following data at the next dialogue to obtain the Estimated Sample Size:

Sample Size and Power Estimation: One Sample

Sample Size in Tests Concerning Mean

```
              Sample Variance =    1.5682
Minimum Detectable Difference =    1.0000
            Power of the Test =    0.9000
             Confidence Level =    0.9500
            1 or 2 Tailed Test =   2.0000
        Estimated Sample Size =    18.4991
```

6.7.1.3. Minimum Detectable Difference for One Sample

Rearranging the equation in section 6.7.1.2. Sample Size in Tests Concerning the Mean we obtain:

$$\delta = (t_{\alpha(2),\nu} + t_{\beta(1),\nu})\sqrt{\frac{s^2}{n}}$$

Here, δ can be computed directly, without an iterational algorithm, as all parameters on the right hand side can be computed using the given parameters.

The user is expected to enter:

- Sample Size
- Sample Variance
- Power of the test
- Confidence Level
- 1 or 2 tailed test

and the program will output the minimum detectable difference.

Example

Example 7.9 on p. 117 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → One Sample and the Minimum Detectable Difference option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: One Sample

Minimum Detectable Difference

```
                 Sample Size =    25.0000
             Sample Variance =    1.5682
           Power of the Test =    0.9000
            Confidence Level =    0.9500
            1 or 2 Tailed Test =  2.0000
Minimum Detectable Difference =   0.8470
```

6.7.1.4. Power of the Test for One Sample

Rearranging the equation in section 6.7.1.2. Sample Size in Tests Concerning the Mean we obtain:

$$t_{\beta(1),\nu} = \frac{\delta}{\sqrt{\frac{s^2}{n}}} - t_{\alpha(2),\nu}$$

Here, $t_{\beta(1),\nu}$ can be computed directly, as all parameters on the right hand side can be computed using the given parameters. Then β is obtained from t-distribution with ν degrees of freedom.

The user is expected to enter:

- Sample Size
- Sample Variance
- Minimum Detectable Difference
- Confidence Level
- 1 or 2 tailed test

and the program will output the t-statistic and its p-value.

Example

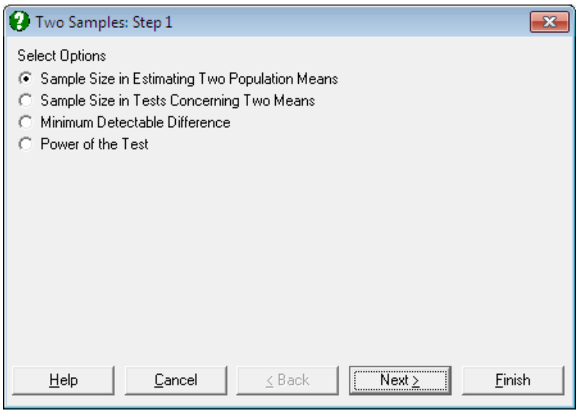
Example 7.10 on p. 118 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → One Sample and the Power of the Test option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: One Sample

Power of the Test

```
                       Sample Size =    12.0000
                   Sample Variance =    1.5682
    Minimum Detectable Difference =    1.0000
                  Confidence Level =    0.9500
                1 or 2 Tailed Test =    2.0000
                Power of the Test:
                       t-Statistic =    0.9480
               2-Tail Probability =    0.8204
```

6.7.2. Two Samples



Two Samples: Step 1

Select Options

- ● Sample Size in Estimating Two Population Means
- ○ Sample Size in Tests Concerning Two Means
- ○ Minimum Detectable Difference
- ○ Power of the Test

The first procedure provided here concerns estimation of the difference between two population means and as such it is different from the last three. The latter are derived from the same equation, each time solving for a different parameter.

6.7.2.1. Sample Size in Estimating Two Population Means

The sample size, which is assumed to be the same for both samples, is computed from:

$$n = \frac{2s_p^2 t_{\alpha(2),2(n-1)}^2}{d^2}$$

Here:

- s_p^2 is the pooled sample variance with ν degrees of freedom,
- d is the half-width of the desired confidence interval,
- 1 - α is the confidence level,
- 1 - β is the assurance that the confidence interval will not be greater,
- $t_{\alpha(2),2(n-1)}$ is the two-tailed critical value for α from t-distribution with 2(n - 1) degrees of freedom,

If the two sample sizes are not equal, and one of them is given, then the other sample size is estimated according to the following formula, which is derived from their harmonic mean:

$$n_2 = \frac{nn_1}{2n_1 - n}$$

Since critical values of t and F on the right hand side of this equation depend on the sample size n, the left hand side cannot be computed directly. Instead, an iterational algorithm is employed and usually the convergence is achieved within a few iterations.

The user is expected to enter:

- Denominator Degrees of Freedom
- Pooled Variance
- Half-Width of Confidence Interval
- Assurance for Confidence Interval
- Confidence Level
- n_1 is given when $n_1 \neq n_2$ (optional)

and the program will output the estimated sample size.

In case when the two samples are assumed to be of equal size, the value of the n_1 field should be zero.

Example

Example 8.3 on p. 147 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Two Samples and the Sample Size in Estimating Two Population Means option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Two Samples

Sample Size in Estimating Two Population Means

```
 Denominator Degrees of Freedom =    11.0000
               Pooled Variance =    0.5193
 Half-Width of Confidence Interval =    0.5000
 Assurance for Confidence Interval =    0.9000
              Confidence Level =    0.9500
            N1 Given (Optional) =    0.0000
          Estimated Sample Size =    17.2934
```

6.7.2.2. Sample Size in Tests Concerning Two Means

The purpose of the experiment is to test if the two samples are taken from one population "H_0: $\mu_1 = \mu_2$" or from two distinct populations "H_0: $\mu_1 \neq \mu_2$".

The sample size is computed from:

$$n = \frac{2s_p^2}{\delta^2}(t_{\alpha(2),\nu} + t_{\beta(1),\nu})^2$$

Here:

- s_p^2 is the pooled sample variance with ν degrees of freedom,
- δ is the minimum detectable difference between the two population means at the desired confidence interval,
- α is the probability of committing a Type I error and 1- α is the confidence level, which can be one or two-tailed.
- β is the probability of committing a Type II error and 1 - β is the power of the test, which is the probability of detecting a real difference,
- $t_{\alpha(2),\nu}$ is the (one or) two-tailed critical value for α from t-distribution with ν degrees of freedom,
- $t_{\beta(1),\nu}$ is the one-tailed critical value for β from t-distribution with ν degrees of freedom.

Since critical values on the right hand side of this equation are dependent on the sample size n, an iterational algorithm is employed.

The user is expected to enter:

- Pooled Sample Variance
- Minimum detectable difference
- Power of the test
- Confidence Level
- 1 or 2 tailed test
- n_1 is given when $n_1 \neq n_2$ (optional)

and the program will output the estimated sample size.

If the two samples are assumed to have the same size, enter 0 into the N1 Given (optional) field. If the size of one of the samples is fixed, then enter the value into this field. In this case the program will calculate the second sample size from the following formula, which is derived from their harmonic mean:

$$n_2 = \frac{nn_1}{2n_1 - n}$$

Example

Example 8.4 on p. 149 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Two Samples and the Sample Size in Tests Concerning Two Means option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Two Samples

Sample Size in Tests Concerning Two Means

```
               Pooled Variance =    0.5200
Minimum Detectable Difference =    0.5000
            Power of the Test =    0.9000
             Confidence Level =    0.9500
            1 or 2 Tailed Test =    2.0000
           N1 Given (optional) =    30.0000
         Estimated Sample Size =    44.7241
              N2 when N1 Given =    87.8328
```

6.7.2.3. Minimum Detectable Difference for Two Samples

Rearranging the equation in section 6.7.1.2. Sample Size in Tests Concerning the Mean we obtain:

$$\delta = (t_{\alpha(2),\nu} + t_{\beta(1),\nu})\sqrt{\frac{2s_p^2}{n}}$$

Here, δ can be computed directly, without an iterational algorithm, as all parameters on the right hand side can be computed using the given parameters.

The user is expected to enter:

- Sample Size
- Pooled Sample Variance
- Power of the test
- Confidence Level

- 1 or 2 tailed test

and the program will output the minimum detectable difference for this sample size. For different sample sizes enter their harmonic mean:

$$n = \frac{2n_1 n_2}{n_1 + n_2}$$

Example

Example 8.5 on p. 150 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Two Samples and the Minimum Detectable Difference option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Two Samples

Minimum Detectable Difference

```
                  Sample Size =    20.0000
              Pooled Variance =    0.5193
            Power of the Test =    0.9000
             Confidence Level =    0.9500
            1 or 2 Tailed Test =   2.0000
Minimum Detectable Difference =    0.7585
```

6.7.2.4. Power of the Test for Two Samples

Rearranging the equation in section 6.7.1.2. Sample Size in Tests Concerning the Mean we obtain:

$$t_{\beta(1),\nu} = \frac{\delta}{\sqrt{\frac{2s_p^2}{n}}} - t_{\alpha(2),\nu}$$

Here, $t_{\beta(1),\nu}$ can be computed directly and then β is obtained from the t-distribution with ν degrees of freedom.

An alternative method of estimating power of the test is also provided which is based on the noncentral F-distribution. The phi statistic is calculated from:

$$\varphi = \sqrt{\frac{n\delta^2}{4s_p^2}}$$

and its p-value is automatically calculated by the program (Pearson and Hartley (1951), pp. 112-130).

The user is expected to enter:

- Sample Size
- Pooled Sample Variance
- Minimum Detectable Difference
- Confidence Level
- 1 or 2 tailed test

and the program will output the computed t-statistic and its p-value, as well as the phi statistic and its and its p-value.

If the two samples have different sizes then their harmonic mean should be entered:

$$n = \frac{2n_1 n_2}{n_1 + n_2}$$

Example

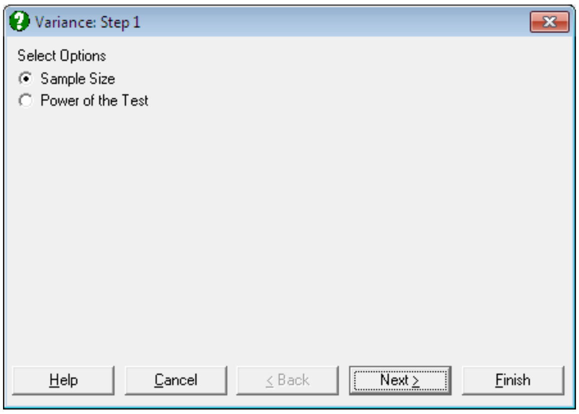
Example 8.6 on p. 151 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Two Samples and the Power of the Test option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Two Samples

Power of the Test

```
                     Sample Size =    15.0000
                Pooled Variance =    0.5193
Minimum Detectable Difference =    1.0000
               Confidence Level =    0.9500
               1 or 2 Tailed Test =    2.0000
              Power of the Test:
                        t-Statistic =    1.7519
              2-Tail Probability =    0.9546
                                 Phi =    2.6872
                       Probability =    0.9561
```

6.7.3. Variance



Select Options

● Sample Size

○ Power of the Test

6.7.3.1. Sample Size for Variance

The relationship between the estimates of sample and population variances is given as follows:

$$\frac{\chi^2_{1-\beta,\nu}}{\chi^2_{\alpha,\nu}} = \frac{\sigma_0^2}{s^2}$$

Here:

- σ_0^2 is the estimate of the population variance,
- s^2 is the sample variance with ν degrees of freedom,
- α is the probability of committing a Type I error and 1- α is the confidence level,
- β is the probability of committing a Type II error and 1 - β is the power of the test,

Since critical values from chi-square distribution are both dependent on the sample size n, an iterational algorithm should be employed.

The user is expected to enter:

- Sample Variance
- Population Variance
- Power of the test
- Confidence Level

and the program will output the estimated sample size.

Example

Example 7.12 on p. 124 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Variance and the Sample Size option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Variance

Sample Size

```
       Sample Variance =    2.6898
   Population Variance =    1.5000
     Power of the Test =    0.9000
      Confidence Level =    0.9500
 Estimated Sample Size =    50.7813
```

6.7.3.2. Power of the Test for Variance

Power of the test is defined as the following probability:

$$1-\beta = P\left(\chi^2_{1-\beta,\nu} \geq \chi^2_{\alpha,\nu}\frac{\sigma_0^2}{s^2}\right)$$

Since the first chi-square value on the right hand side of this equation depends on β, an iterational algorithm should be employed.

The user is expected to enter:

- Sample Size
- Sample Variance
- Population Variance
- Confidence Level

and the program will output the estimated chi-square statistic and its p-value, β.

Example

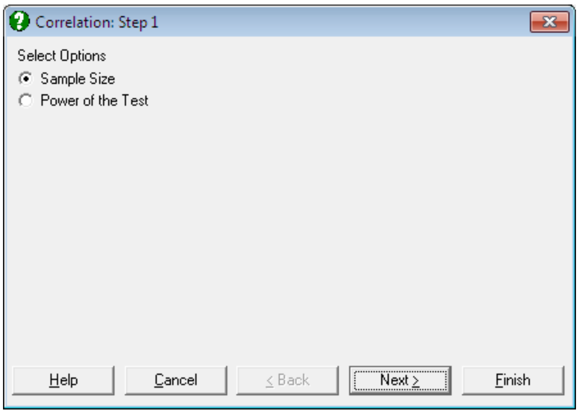
The example on p. 124 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Variance and the Power of the Test option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Variance

Power of the Test

```
         Sample Size =    8.0000
     Sample Variance =    2.6898
 Population Variance =    1.5000
    Confidence Level =    0.9500
  Power of the Test:
 Chi-Square Statistic =   7.8447
Right-Tail Probability =  0.3465
```

6.7.4. Correlation



Select Options
● Sample Size
○ Power of the Test

6.7.4.1. Sample Size for Correlation

The sample size is estimated using the following formula:

$$n = \left(\frac{Z_{\beta(1)} + Z_{\alpha(2)}}{z} \right)^2 + 3$$

where:

$$z = 0.5\mathrm{Log}\left(\frac{1+r}{1-r} \right)$$

- and z is the Fisher's z transformation of the correlation coefficient r,
- $Z_{\beta(1)}$ is the one-tailed critical value from the standard normal distribution,
- $Z_{\alpha(2)}$ is the (one or) two-tailed critical value from the standard normal distribution.

The user is expected to enter:

- Correlation Coefficient
- Power of the test
- Confidence Level
- 1 or 2 tailed test

and the program will output the estimated sample size.

Example

Example 19.5a on p. 388 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Correlation and the Sample Size option. Enter the following data at the next dialogue:

```
Sample Size and Power Estimation: Correlation

Sample Size
  Correlation Coefficient =    0.5000
          Power of the Test =    0.9900
          Confidence Level =    0.9500
          1 or 2 Tailed Test =    2.0000
   Estimated Sample Size =    63.9136
```

6.7.4.2. Power of the Test for Correlation

Power of the test is one minus the p-value of the following Z-statistic:

$$Z_{\beta(1)} = (z - z_{\alpha})\sqrt{n-3}$$

where:

$$z = 0.5\mathrm{Log}\left(\frac{1+r}{1-r}\right)$$

$$z_{\alpha} = 0.5\mathrm{Log}\left(\frac{1+r_{\alpha(2),n-2}}{1-r_{\alpha(2),n-2}}\right)$$

$$r_{\alpha(2),\nu} = \frac{1}{\sqrt{\left(\frac{n-2}{t^2_{\alpha(2),n-2}}\right)+1}}$$

and $t_{\alpha(2),n-2}$ is the (one or) two-tailed critical value from t-distribution with n-2 degrees of freedom.

The user is expected to enter:

- Sample Size
- Correlation Coefficient
- Confidence Level
- 1 or 2 tailed test

and the program will output the estimated Z-statistic and its p-value.

Example

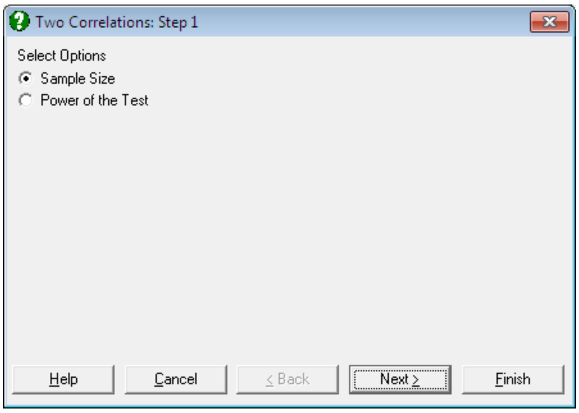
Example 19.4 on p. 388 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Correlation and the Power of the Test option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Correlation

Power of the Test

```
            Sample Size =    12.0000
Correlation Coefficient =    0.8700
       Confidence Level =    0.9500
      1 or 2 Tailed Test =    2.0000
     Power of the Test:
             Z-Statistic =    2.0300
      2-Tail Probability =    0.9788
```

6.7.5. Two Correlations



Two Correlations: Step 1

Select Options

● Sample Size

○ Power of the Test

6.7.5.1. Sample Size for Two Correlations

The sample size is estimated using the following formula:

$$n = 2\left(\frac{Z_{\beta(1)} + Z_{\alpha(2)}}{z_1 - z_2}\right)^2 + 3$$

where:

$$z_1 = 0.5\mathrm{Log}\left(\frac{1 + r_1}{1 - r_1}\right)$$

$$z_2 = 0.5\mathrm{Log}\left(\frac{1 + r_2}{1 - r_2}\right)$$

- $Z_{\alpha(2)}$ is the (one or) two-tailed critical value from the standard normal distribution for Type I error probability α.
- $Z_{\beta(1)}$ is the one-tailed critical value from the standard normal distribution for Type II error probability β.

The user is expected to enter:

- Correlation Coefficient 1
- Correlation Coefficient 2
- Power of the test
- Confidence Level
- 1 or 2 tailed test

and the program will output the estimated sample size.

This procedure assumes that the two sample sizes are equal. However, when there is a constraint on one of the sample sizes, the other can be found as follows:

$$n_2 = \frac{nn_1 + 3n_1 - 6n}{2n_1 - n - 3}$$

If such a constraint exists, then enter the given sample size in the N1 Given (optional) field. Otherwise this field should have a zero entry.

Example

Example 19.8 on p. 393 from Zar, J. H. (2010). The difference between two Fisher's z transforms is given as 0.5. We substitute this input with the correlation coefficients as 0.75 and 0.9, which give a difference of 0.5 between their respective Fisher's z transforms. Select Statistics 1 → Sample Size and Power Estimation → Two Correlations and the Sample Size option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Two Correlations

Sample Size

```
Correlation Coefficient 1 =    0.7500
Correlation Coefficient 2 =    0.9000
        Power of the Test =    0.9000
         Confidence Level =    0.9500
        1 or 2 Tailed Test =   2.0000
      N1 Given (optional) =    0.0000
    Estimated Sample Size =    87.3389
```

6.7.5.2. Power of the Test for Two Correlations

Power of the test is one minus the p-value of the following Z-statistic:

$$Z_{\beta(1)} = \frac{(z_1 - z_2)}{\sigma_{z_1 - z_2}} - Z_{\alpha(2)}$$

where:

$$z_1 = 0.5\mathrm{Log}\left(\frac{1 + r_1}{1 - r_1}\right)$$

$$z_2 = 0.5\mathrm{Log}\left(\frac{1 + r_2}{1 - r_2}\right)$$

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

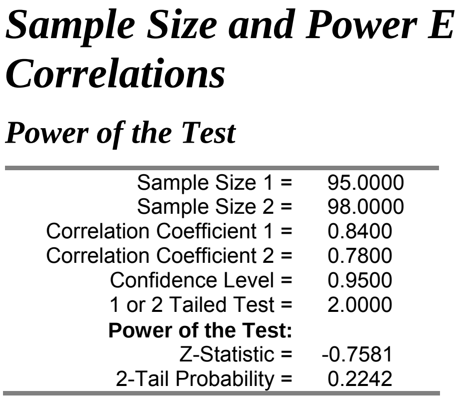
The user is expected to enter:

- Sample Size 1
- Sample Size 2
- Correlation Coefficient 1
- Correlation Coefficient 2
- Confidence Level
- 1 or 2 tailed test

and the program will output the estimated Z-statistic and its p-value.

Example

Example 19.7 on p. 392 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Two Correlations and the Power of the Test option. Enter the following data at the next dialogue:

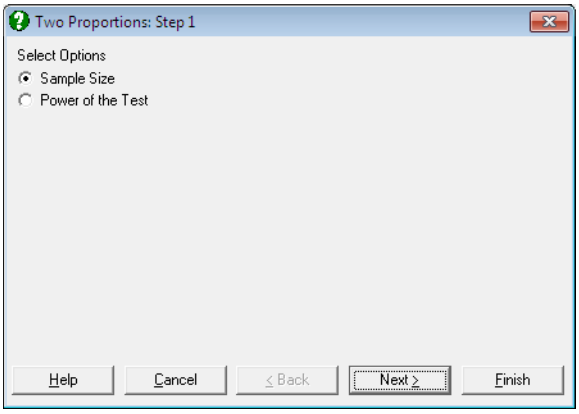


Sample Size and Power Estimation: Two Correlations

Power of the Test

```
           Sample Size 1 =    95.0000
           Sample Size 2 =    98.0000
 Correlation Coefficient 1 =  0.8400
 Correlation Coefficient 2 =  0.7800
        Confidence Level =    0.9500
        1 or 2 Tailed Test =  2.0000
       Power of the Test:
             Z-Statistic =   -0.7581
      2-Tail Probability =    0.2242
```

6.7.6. Two Proportions



Select Options

● Sample Size

○ Power of the Test

6.7.6.1. Sample Size for Two Proportions

One sample size (n_1) can be estimated assuming that the ratio of the two sample sizes ($r = n_2/n_1$) is known. The algorithm for power of the test for two proportions (see next section) is also used here, employing a half-interval search method.

The user is asked to enter:

- Proportion 1
- Proportion 2
- Power of the test
- Confidence Level
- 1 or 2 tailed test
- n_2/n_1

and the program will output the estimated sample size (n_1).

Example

Example 24.26 on p. 561 from Zar, J. H. (1999). Select Statistics 1 → Sample Size and Power Estimation → Two Proportions and the Sample Size option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Two Proportions

Sample Size

Proportion 1 =	0.4500
Proportion 2 =	0.2500
Power of the Test =	0.9000
Confidence Level =	0.9500
1 or 2 Tailed Test =	2.0000
N2 / N1 =	1.0000
Estimated Sample Size =	127.2343

Next, click [Last Procedure Dialogue], enter n_2 / n_1 = 3 and click [Finish] again:

Sample Size and Power Estimation: Two Proportions

Sample Size

Proportion 1 =	0.4500
Proportion 2 =	0.2500
Power of the Test =	0.9000
Confidence Level =	0.9500
1 or 2 Tailed Test =	2.0000
N2 / N1 =	3.0000
Estimated Sample Size =	83.1527

n_2 is found as 3 x 83.1527 = 249.4581.

6.7.6.2. Power of the Test for Two Proportions

Power of the test in comparing two proportions is given by the following formula:

$$1-\beta = P(Z \le A) + P(Z \ge B)$$

where:

$$A = \frac{-Z_{\alpha(2)}\sqrt{\dfrac{\bar{p}\bar{q}}{n_1}+\dfrac{\bar{p}\bar{q}}{n_2}}-(p_1-p_2)}{\sqrt{\dfrac{p_1q_1}{n_1}+\dfrac{p_2q_2}{n_2}}}$$

$$B = \frac{Z_{\alpha(2)}\sqrt{\frac{\overline{p}\overline{q}}{n_1} + \frac{\overline{p}\overline{q}}{n_2}} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$\overline{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$\overline{q} = 1 - \overline{p}$$

$$q_1 = 1 - p_1$$

$$q_2 = 1 - p_2$$

- $Z_{\alpha(2)}$ is the (one or) two-tailed critical value from the standard normal distribution for Type I error probability α.

The user is expected to enter:

- Sample Size 1
- Sample Size 2
- Proportion 1
- Proportion 2
- Confidence Level
- 1 or 2 tailed test

and the program will output the two Z-statistics A and B and the power of the test computed from the above equation. For a one-tailed test the power of the test is computed from:

$$1 - \beta = P(Z \le A)$$

Example

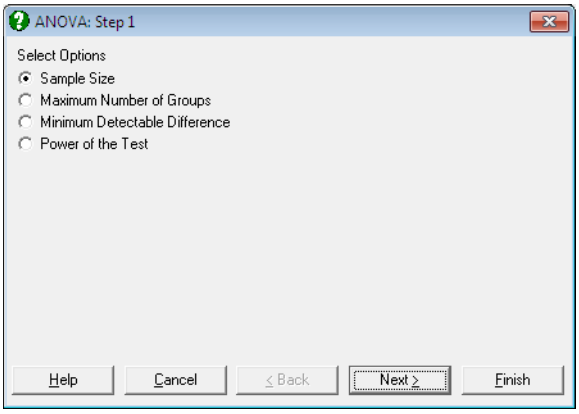
Example 24.16 on p. 554 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Two Proportions and the Power of the Test option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: Two Proportions

Power of the Test

```
       Sample Size 1 =   50.0000
       Sample Size 2 =   45.0000
        Proportion 1 =   0.7500
        Proportion 2 =   0.5000
    Confidence Level =   0.9500
   1 or 2 Tailed Test =  2.0000
  Power of the Test:
   2-Tail Probability =  0.6402
```

6.7.7. ANOVA



ANOVA: Step 1

Select Options

● Sample Size
○ Maximum Number of Groups
○ Minimum Detectable Difference
○ Power of the Test

All parameters input and estimated in this section are based on a single factor ANOVA. However, different types of ANOVA can also be accommodated simply by entering the relevant statistics in place of existing parameters. For instance, it is possible to apply the sample size and power of the test algorithms introduced below to a multi-way ANOVA problem, by treating each of the factors separately.

6.7.7.1. Sample Size for ANOVA

Estimation of the sample size is based on the following relationship:

$$\varphi = \sqrt{\frac{n\delta^2}{2ks^2}}$$

where:

- φ is the power of the test statistic the p-value of which can be determined by consulting the OC Curves published by Pearson and Hartley (1951), pp. 112-130.
- n is the sample size (i.e. number of cases in each group)
- k is the number of groups
- δ is the minimum detectable difference
- s^2 is the error mean square

However, since φ itself depends on n, an iterational algorithm should be employed and this is an extremely laborious process. Here we completely automate this process by reproducing the OC Curves with an iterational algorithm based on the noncentral F-distribution. First defining the noncentrality parameter as:

$$\lambda = k\varphi^2$$

and inserting this into the above equation we obtain:

$$\lambda = \frac{n\delta^2}{2s^2}$$

Then n is augmented until the following equality is obtained within the given tolerance limits:

$$F_{\alpha,\nu_1,\nu_2} = F'_{1-\beta,\nu_1,\nu_2,\lambda}$$

where:

- $\nu_1 = k - 1$ is the numerator degrees of freedom,
- $\nu_2 = k(n-1)$ is the denominator degrees of freedom,
- F_{α,ν_1,ν_2} is the critical value from F-distribution for Type I error level of α,
- $F'_{1-\beta,\nu_1,\nu_2,\lambda}$ is the critical value from noncentral F-distribution for power of the test corresponding to Type II error level of β.

The user is expected to enter:

- Number of Groups
- Minimum Detectable Difference
- Error Mean Square
- Power of the Test
- Confidence Level

and the program will output the estimated sample size.

Example

Example 10.6 on p. 211 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → ANOVA and the Sample Size option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: ANOVA

Sample Size

```
             Number of Groups =   4.0000
Minimum Detectable Difference =   3.5000
            Error Mean Square =   9.3833
            Power of the Test =   0.8000
             Confidence Level =   0.9500
                          Phi =   1.7000
        Estimated Sample Size =   17.7100
```

6.7.7.2. Maximum Number of Groups for ANOVA

Estimation of the maximum number of groups is similar to that of the sample size described in the previous section. An iterational algorithm is employed based on the noncentral F-distribution.

The user is expected to enter:

- Total number of cases (i.e. n x k)
- Minimum Detectable Difference
- Error Mean Square
- Power of the Test
- Confidence Level

and the program will output the estimated group size.

Example

Example 10.8 on p. 213 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → ANOVA and the Maximum Number of Groups option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: ANOVA

Maximum Number of Groups

Total Number of Cases (nk) =	50.0000
Minimum Detectable Difference =	4.5000
Error Mean Square =	9.3833
Power of the Test =	0.8000
Confidence Level =	0.9500
Phi =	1.6860
Number of Groups =	4.3567

6.7.7.3. Minimum Detectable Difference for ANOVA

Rearranging the equation given in section 6.7.7.1. Sample Size for ANOVA we obtain:

$$\delta = \sqrt{\frac{2s^2\lambda}{n}}$$

where:

$$\lambda = k\varphi^2$$

is the noncentrality parameter. Then the minimum detectable difference is estimated employing an iterational algorithm similar to the one in section 6.7.7.1. Sample Size for ANOVA.

The user is expected to enter:

- Sample Size per Group (n)
- Number of Groups (k)
- Error Mean Square
- Power of the Test
- Confidence Level

and the program will output the estimated φ and the minimum detectable difference computed according to the above equation.

Example

Example 10.7 on p. 212 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → ANOVA and the Minimum Detectable Difference option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: ANOVA

Minimum Detectable Difference

```
                   Sample Size =    10.0000
              Number of Groups =    4.0000
             Error Mean Square =    9.3833
             Power of the Test =    0.9000
              Confidence Level =    0.9500
                           Phi =    1.9883
 Minimum Detectable Difference =    5.4476
```

6.7.7.4. Power of the Test for ANOVA

Power of the test can be computed directly from the equation given in section 6.7.7.1. Sample Size for ANOVA where there is no need for an iterational solution.

The user is expected to enter:

- Sample Size per Group (n)
- Number of Groups (k)
- Error Mean Square
- Minimum Detectable Difference
- Confidence Level

and the program will output the estimated φ and its corresponding p-value.

Example

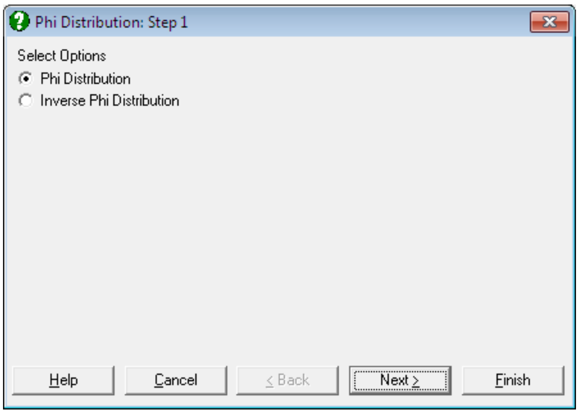
Example 10.5 on p. 209 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → ANOVA and the Power of the Test option. Enter the following data at the next dialogue:

Sample Size and Power Estimation: ANOVA

Power of the Test

```
                   Sample Size =    10.0000
              Number of Groups =    4.0000
             Error Mean Square =    7.5888
 Minimum Detectable Difference =    4.0000
              Confidence Level =    0.9500
                           Phi =    1.6234
             Power of the Test =    0.7349
```

6.7.8. Phi Distribution


Phi Distribution: Step 1

Select Options

● Phi Distribution

○ Inverse Phi Distribution

We should point out that phi is not a distribution as such, but a multi-parameter relationship based on the noncentral F-distribution.

The following two procedures are provided here as an alternative to OC Curves published by Pearson and Hartley (1951).

Here the significance level (α) can be any value between 0 and 1, whereas the OC Curves are limited to α = 0.05 and α = 0.01. Also, the accuracy of calculations here is far higher.

6.7.8.1. Phi Distribution

The user is expected to enter:

- Phi
- Numerator Degrees of Freedom
- Denominator Degrees of Freedom
- Confidence Level

and the program will output the estimated power of the test (1 - β).

Example

Figure B.1d on p. AppB 862 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Phi Distribution and the Phi Distribution option. Enter the following data at the next dialogue:

Sample Size and Power Estimation:

Phi Distribution

```
                           Phi =    3.0000
  Numerator Degrees of Freedom =    4.0000
Denominator Degrees of Freedom =    20.0000
              Confidence Level =    0.9900
             Power of the Test =    0.9899
```

6.7.8.2. Inverse Phi Distribution

The user is expected to enter:

- Power of the Test
- Numerator Degrees of Freedom
- Denominator Degrees of Freedom
- Confidence Level

and the program will output the estimated φ.

Example

Figure B.1g on p. AppB 865 from Zar, J. H. (2010). Select Statistics 1 → Sample Size and Power Estimation → Phi Distribution and the Inverse Phi Distribution option. Enter the following data at the next dialogue:

Sample Size and Power Estimation:

Inverse Phi Distribution

```
             Power of the Test =    0.9050
  Numerator Degrees of Freedom =    7.0000
Denominator Degrees of Freedom =    12.0000
              Confidence Level =    0.9500
                           Phi =    1.9980
```

6.8. Meta Analysis

Meta analysis is used to combine and analyse the results of several independent studies on a particular research topic. Often, different research results on the same topic are reported in terms of different statistical entities such as t-tests, correlations, risk ratios, etc. with different dispersion measures such as standard errors, p-values or confidence intervals. The aim of meta analysis is to translate these statistics back to a common measure (the effect size) and then to aggregate them taking into consideration the weight of each individual study.

Here we assume that a systematic review has been carried out and possible sources of bias are well understood. UNISTAT's intuitive user interface allows you to select many different types of data and mix them as input for any meta analysis. This is a powerful feature that needs to be used with caution. The user should determine beforehand which types of data can be combined in a meta analysis in a meaningful way.

It is possible run one study removed (OSR), cumulative (CUM) or subgroup analyses. You can select one or both of fixed effect (with inverse variance (IV) or Mantel-Haenszel (MH) weights) or random effect models.

The output includes a summary table with confidence intervals, significance tests and a forest diagram, Cochran's Q and I-square heterogeneity tests, Begg-Mazumdar Rank Correlation and Egger regression tests for publication bias and forest, funnel and precision plots.

6.8.1. Data Preparation

UNISTAT does not use a special data format requiring a major re-arrangement of your data. All you need to do is to enter your data in a standard spreadsheet format where each column contains a different statistic, i.e. odds ratios in one column, probabilities in another column, etc. In most cases, UNISTAT will not require you to make any prior calculations before entering your data.

For instance, if the analysis is based on a 2 x 2 table, you do not need to calculate each cell frequency to perform a meta analysis. UNISTAT will accept data in terms of sums and ratios of frequencies reported by individual studies.

Study Name	a	b	c	d	a+b	c+d	a/[a+b]	c/[c+d]
Study 1	19	18	164	157				
Study 2	11		13		73	41		
Study 3		17		24	82	57		
Study 4					53	54	0.4313	0.2517

For an unpaired two-sample analysis with continuous data there are many more types of data entry possibilities.

Study Name	No A	No B	Mean A	Mean B	Mean Diff	Std Dev A	Std Dev B	Std Err Pooled
S 1	43	42	43.2	47.6		2.43	7.89	
S 2	25	42	73.7	64.2				
S 3	37	52	23.4	25.5				
S 4	53	32	5.31	6.37				
S 5	41	32			5.43			
S 6	74	96	6.67	7.81				
S 7	87	76			-4.56			
S 8	23	21			7.65			
S 9	32	34			2.34			
S 10	13	16			23.6			4.57

(table continued)

	Var A	Var B	t-value	p-value	Tails	Lower Bound	Upper Bound	Conf Level
S 1								
S 2								
S 3	12.6	32.1						
S 4			3.456					
S 5			4.567					
S 6				0.0756	1			
S 7				0.1034	2			
S 8							9.02	0.95
S 9						1.23		0.90
S 10								

The above tables demonstrate only some of the data types supported by UNISTAT. For a full list of data options see the next section.

You can also mix different types of data in an analysis. For instance, in below hypothetical example, the data from unpaired two-sample analyses are entered in the first five columns and data from 2 x 2 tables entered in the last five columns. UNISTAT's variable selection interface allows you to define the first nine studies as unpaired samples and studies from S 10 to S 13 as 2 x 2 tables and run a meta analysis on all data.

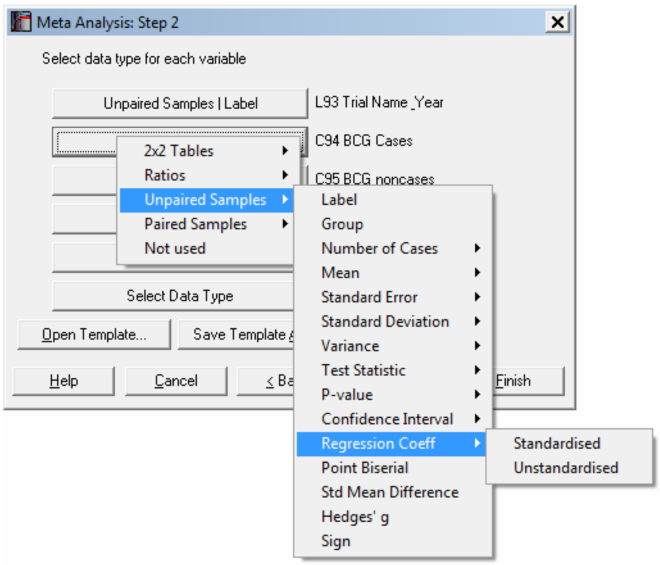
Study Name	No A	No B	Mean Diff	Std Err Pooled	Study Name	A	B	C	D
S 1	43	42	0.02	7.68	**S 10**	14	65	41	89
S 2	25	42	-2.32	6.98	**S 11**	23	79	37	67
S 3	37	52	-0.68	3.40	**S 12**	37	87	17	93
S 4	53	32	0.73	5.33	**S 13**	41	73	24	77
S 5	41	32	5.43	6.98					
S 6	74	96	2.66	3.45					
S 7	87	76	-4.56	3.21					
S 8	23	21	7.65	5.15					
S 9	32	34	2.34	5.87					

6.8.2. Input Data Types

UNISTAT classifies commonly used data types into four groups:

1) 2 x 2 Tables
2) Ratios
3) Unpaired Samples
4) Paired Samples

Once the columns of the data matrix to be used in the analysis are selected as [Variable]s in the Variable Selection Dialogue, a second dialogue facilitates assigning a specific input data type and a task to each variable.



This three-level menu system allows you to combine almost any type of study result entered in a simple spreadsheet format. Once the selections are made, you can save them to a file so that the selection process is not unnecessarily repeated. The default extension for meta analysis template file is .MTA.

The full extent of selection possibilities for these six data groups are given in the following tables.

2 x 2 Tables	Ratios	Unpaired Samples	Paired Samples
Label	Label	Label	Label
Group	Group	Group	Group
Cells	Odds Ratio	Number of Cases	Number of Pairs
A	Odds Ratio	Cases A	Mean
B	Ln(Odds Ratio)	Cases B	Mean A
C	Risk Ratio	Mean	Mean B
D	Risk Ratio	Mean A	Mean Diff
Sums	Ln(Risk Ratio)	Mean B	Standard Error
A+B	Rate Ratio	Mean Diff	Std Err A
C+D	Rate Ratio	Standard Error	Std Err B
Ratios	Ln(Rate Ratio)	Std Err A	Std Err Diff
A/(A+B)	Hazard Ratio	Std Err B	Std Err Corr
C/(C+D)	Hazard Ratio	Std Err Pooled	Standard Deviation
Chi-Square	Ln(Hazard Ratio)	Std Err Overall	Std Dev A
Chi-Square	Risk Difference	Standard Deviation	Std Dev B
Number of Cases	Rate Difference	Std Dev A	Std Dev Diff
Sign	Observed - Expected	Std Dev B	Std Dev Corr
Pairs	Standard Error	Std Dev Pooled	Variance
P pair	Standard Error	Std Dev Overall	Var A
Q pair	Ln(Std Err)	Variance	Var B
R pair	Variance	Var A	Var Diff
S pair	Variance	Var B	Var Corr
Rho	Ln(Variance)	Var Pooled	Test Statistic
	Confidence Interval	Var Overall	t-stat
	Lower Bound	Test Statistic	F-stat
	Upper Bound	t-stat	Z-stat
	Confidence Level	F-stat	P-value
		Z-stat	P (t-dist)
		P-value	P (F-dist)
		P (t-dist)	P (normal)
		P (F-dist)	Tails
		P (normal)	Confidence Interval
		Tails	Lower Bound
		Confidence Interval	Upper Bound
		Lower Bound	Confidence Level
		Upper Bound	Correlation
		Confidence Level	Fisher's Z
		Regression Coeff	Std Mean Diff
		Standardised	Sign
		Unstandardised	
		Point Biserial Corr	
		Std Mean Diff	
		Hedges' g	
		Sign	

6.8.2.1. 2 x 2 Tables

Normally, cell frequencies from a 2 x 2 table are entered for each study. However, UNISTAT also accepts data in the form of sums and ratios. If one or more cell frequencies are zero, 0.5 is added to all four frequencies. Optionally, if the data is paired, an external correlation ϱ (rho) can be entered.

The commonly used effect sizes for 2 x 2 tables are as follows:

Odds ratio:

$$\mathrm{OR} = \frac{\mathrm{ad}}{\mathrm{bc}}$$

$$\mathrm{SE}_{\mathrm{Ln(OR)}} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Peto odds ratio:

$$\mathrm{POR} = \mathrm{Exp}\left(\frac{\text{Observed - Expected}}{\text{Variance}}\right)$$

where:

$$\text{Observed} = a$$

$$\text{Expected} = \frac{(a+b)(a+c)}{n}$$

$$n = a + b + c + d$$

$$\mathrm{SE}_{\mathrm{Ln(POR)}} = \sqrt{\frac{n^2(n-1)}{(a+b)(c+d)(a+c)(b+d)}}$$

Risk ratio:

$$\mathrm{RR} = \frac{a/(a+b)}{c/(c+d)}$$

$$\mathrm{SE}_{\mathrm{Ln(RR)}} = \sqrt{\frac{b/a}{b+a} + \frac{d/c}{d+c}}$$

Risk difference:

$$\mathrm{RD} = \frac{a}{a+b} - \frac{c}{c+d}$$

$$SE_{RD} = \sqrt{\frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}}$$

Standardised mean difference:

$$d = \frac{\sqrt{3}}{\pi} Ln(OR)$$

$$SE_d = \frac{\sqrt{3}}{\pi} SE_{Ln(OR)}$$

Correlation from a chi-square statistic and sample size for a 2 x 2 table:

If a chi-square statistic and sample size from a 2 x 2 table is reported by a study, the correlation and its standard error are calculated as:

$$r = \sqrt{\frac{\chi^2}{n}}$$

$$SE_r = \frac{(1 - r^2)}{\sqrt{n-3}}$$

Event pairs:

Define each cell frequency from pairs data as:

A = P + Q
B = P + R
C = R + V
D = Q + V

For the calculation of other effect size measures based on d, see 6.8.2.3. Unpaired Samples.

6.8.2.2. Ratios

Given the odds ratio and its dispersion measure, the standardised mean difference d is calculated as above. Other ratios can only be compared within their own group.

6.8.2.3. Unpaired Samples

Select this option if the individual study is based on unpaired (or unmatched or independent or unequal size) samples consisting of continuous data.

The following data entry combinations are possible to calculate the standardised mean difference d, which is one of the most commonly used summary effect size for continuous data.

Sample sizes and means are given:

When the standard deviations are given, the standardised mean difference is calculated as:

$$d = \frac{\overline{X}_A - \overline{X}_B}{s_p}$$

where s_p is the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

The standard error of d is calculated as:

$$SE_d = \sqrt{\frac{1}{n_A} + \frac{1}{n_B} + \frac{d^2}{2(n_A + n_B)}}$$

When standard errors or variances are given for samples A and B, standard deviations are calculated first.

UNISTAT also accepts data where difference in means is given instead of individual means and pooled or overall standard deviation is given instead of individual standard deviations.

When the overall standard deviation is given, the pooled standard deviation is calculated as:

$$s_p = \sqrt{\frac{s_o^2(n_A + n_A - 1) - \left(\left(\overline{X}_A^2 + \overline{X}_B^2 - 2\overline{X}_A\overline{X}_B\right)n_A n_B\right)/(n_A + n_B)}{n_A + n_B}}$$

The overall standard error and variance are converted into overall standard deviation first.

Sample sizes and the test statistic are given:

t-statistic is given:

$$d = t\sqrt{\frac{n_A + n_B}{n_A n_B}}$$

The standard error of d is calculated as above.

F-statistic from a one-way ANOVA is given:

Find t-statistic as:

$$t = \sqrt{F}$$

and use the formula for t-statistic to calculate the standardised mean difference.

Z-statistic is given:

The Z-statistics can be used instead of t-statistic:

$$t = Z$$

and the formula for t-statistic is used to calculate the standardised mean difference.

Sample sizes and the p-value are given:

P-value from a t-test and its tail (1 or 2) are given: The t-value is found using the inverse t-distribution with $n_A + n_B - 2$ degrees of freedom. If the tail value is not specified, a two-tailed test is assumed.

P-value from an F-test is given: The F-value is found using the inverse F-distribution with 1 and $n_A + n_B - 2$ degrees of freedom.

P-value from a Z-test and its tail (1 or 2) are given: The Z-value is found using the inverse normal distribution. If the tail value is not specified, a two-tailed test is assumed.

Standardised regression coefficient:

Enter the standard deviation of dependent variable as Sample B standard deviation (or standard error or variance).

$$d = \frac{b}{s_p}$$

where b is the standardised regression coefficient and the pooled standard deviation is calculated as:

$$s_p = \sqrt{\frac{s_B^2\left(n_A + n_A - 1\right) - \left(b^2\ n_A n_B\right)/\left(n_A + n_B\right)}{n_A + n_B}}$$

The standard error of d is calculated as in the **Sample sizes and means are given** option above.

Unstandardised regression coefficient:

Enter the standard deviation of dependent variable as Sample B standard deviation (or standard error or variance). Standard deviation of sample A (which is assumed to be a binary variable containing 0s and 1s only) is calculated as:

$$s_A = \sqrt{\frac{n_A - n_A^2/\left(n_A + n_B\right)}{n_A + n_B}}$$

Given the unstandardised regression coefficient β, the standardised regression coefficient b and d are calculated as:

$$b = \beta\frac{s_B}{s_A}$$

$$d = \frac{b}{s_p}$$

where the pooled standard deviation is as calculated as above. The standard error of d is calculated as in the **Sample sizes and means are given** option above.

Point Biserial Correlation:

$$d = \frac{r}{\sqrt{\left(1 - r^2\right)\left(p\left(1 - p\right)\right)}}$$

where:

$$p = \frac{n_A}{n_A + n_B}$$

The standard error of d is calculated as in the **Sample sizes and means are given** option above. If a p-value is given for point biserial correlation, select the **P-value from a t-test** option above.

Other effect sizes based on standardised mean difference are calculated as follows.

Mean Difference:

$$MD = \overline{X}_A - \overline{X}_B$$

and its standard error is:

$$SE_{MD} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Hedges' g:

Hedges' correction factor for standardised mean difference d is:

$$J = 1 - \frac{3}{4df - 1}$$

where:

$$df = n_A + n_B - 2$$

Hedges' g is defined as:

$$g = Jd$$

The standard error of Hedges' g can be calculated in one of the two following ways, which can be selected from the intermediate inputs dialogue. The default and recommended SE is:

$$SE_g = J\,SE_d$$

and the alternative definition is:

$$SE_g = \sqrt{\frac{n_A + n_B}{n_A n_B} + \frac{g^2}{2(n_A + n_B - 3.94)}}$$

Odds Ratio:

The correction factor for log odds ratio is:

$$J = \frac{\pi}{\sqrt{3}}$$

The log odds ratio and its standard error can be calculated from standardised mean difference as:

$$\mathrm{Ln(OR)} = J\, d$$

$$SE_{\mathrm{Ln(OR)}} = J\, SE_d$$

The odds ratio is:

$$OR = e^{\mathrm{Ln(OR)}}$$

Correlation:

Given standardised mean difference and its standard error, the correlation and its SE are calculated as:

$$r = \frac{d}{\sqrt{d^2 + p'}}$$

$$SE_r = \frac{p'^2\, SE_d}{\sqrt{\left(d^2 + p'\right)^3}}$$

where:

$$p = \frac{n_A}{n_A + n_B}$$

and:

$$p' = \frac{1}{p(1-p)}$$

Fisher's Z:

Fisher's Z and its standard error are calculated from a given correlation as:

$$Z = \frac{1}{2} \mathrm{Ln}\left(\frac{1+r}{1-r}\right)$$

$$SE_Z = \frac{SE_r}{1-r^2}$$

6.8.2.4. Paired Samples

This option should be selected when the individual study is based on paired (or matched or equal size) samples consisting of continuous data. This group also includes correlation coefficients and pre/post correlations. If the correlation value is not available, the imputed r value is used instead. The default value for imputed r is 0.5.

When only a t-statistic and the sample size are given, there are two ways of calculating the effect size. If the t-statistic was originally reported for a correlation coefficient, then in the intermediate inputs dialogue choose Yes for Use paired t-test box.

The following data entry combinations are possible to calculate the standardised mean difference d. Other effect sizes based on standardised mean difference are calculated as in unpaired samples above.

Sample size, means and standard deviations are given:

The standardised mean difference is calculated as:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s_{Diff}} \sqrt{2(1-r)}$$

where:

$$s_{Diff} = \sqrt{s_1^2 + s_2^2 - 2rs_1^2 s_2^2}$$

When r is not available and it is substituted by the imputed r value of 0.5, the correction factor for the correlation (i.e. the term in square brackets) simply disappears and the standardised mean difference formula becomes identical to the unpaired samples case.

The standard error of d is calculated as:

$$SE_d = \sqrt{\frac{1}{n} + \frac{d^2}{2n}} \sqrt{2(1-r)}$$

When standard error or variance is given for the paired differences, standard deviations are calculated first.

UNISTAT also accepts mean difference and standard deviation instead of individual means and standard deviations.

Sample size and the test statistic are given:

t-statistic is given:

$$d = \frac{t}{\sqrt{n}} \sqrt{2(1-r)}$$

The standard error of d is calculated as above.

F-statistic from a one-way ANOVA is given:

Find t-statistic as:

$$t = \sqrt{F}$$

and use the above formula to find d.

Z statistic is given:

$$d = \frac{Z}{\sqrt{n - Z^2}} \sqrt{2(1-r)}$$

Sample size and the test statistic are given (paired t-test):

To use paired t-test for correlations you need to select this option from the intermediate inputs dialogue first. The correlation is calculated as:

$$r = \frac{t^2}{t^2 + n - 2}$$

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

The standard error of d is calculated as.

$$\mathrm{SE_d} = \frac{4\mathrm{SE_m^2}}{\left(1 - r^2\right)^3}$$

where:

$$\mathrm{SE_m} = \frac{1 - r^2}{\sqrt{n - 3}}$$

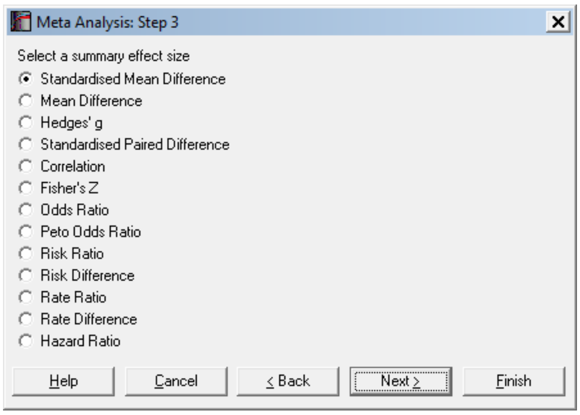
Sample sizes and p-values are given:

P-value from a t-test and its tail (1 or 2) are given: The t-value is found using the inverse t-distribution with $n_A + n_B - 2$ degrees of freedom. If the tail value is not specified, a two-tailed test is assumed.

P-value from an F-test is given: The F-value is found using the inverse F-distribution with 1 and $n_A + n_B - 2$ degrees of freedom.

P-value from a Z-statistic and its tail (1 or 2) are given: The Z-value is found using the inverse normal distribution. If the tail value is not specified, a two-tailed test is assumed.

6.8.3. Summary of Effect Sizes

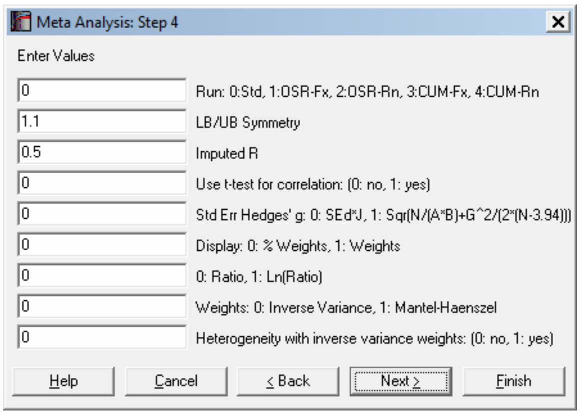


Meta Analysis: Step 3

Select a summary effect size

● Standardised Mean Difference
○ Mean Difference
○ Hedges' g
○ Standardised Paired Difference
○ Correlation
○ Fisher's Z
○ Odds Ratio
○ Peto Odds Ratio
○ Risk Ratio
○ Risk Difference
○ Rate Ratio
○ Rate Difference
○ Hazard Ratio

The summary effect size type is selected from the next dialogue. The program will estimate a fixed effect model and a random effects model for the mean effect size using a weighted average. By default, for each individual study the inverse of its variance is used as the weight for both models. It is also possible to select the Mantel-Haenszel weights for 2 x 2 tables.



Meta Analysis: Step 4

Enter Values

0	Run: 0:Std, 1:OSR-Fx, 2:OSR-Rn, 3:CUM-Fx, 4:CUM-Rn
1.1	LB/UB Symmetry
0.5	Imputed R
0	Use t-test for correlation: (0: no, 1: yes)
0	Std Err Hedges' g: 0: SEd*J, 1: Sqr(N/(A*B)+G^2/(2*(N-3.94)))
0	Display: 0: % Weights, 1: Weights
0	0: Ratio, 1: Ln(Ratio)
0	Weights: 0: Inverse Variance, 1: Mantel-Haenszel
0	Heterogeneity with inverse variance weights: (0: no, 1: yes)

Run analysis:

0: Standard: Individual effect sizes and their standard errors are displayed with both fixed and random effects models.

1: One study removed with fixed effect: For each study, that study is removed and the fixed effect summary for the rest of the studies is displayed in its place.
2: One study removed with random effects
3: Cumulative with fixed effect: For each study, the fixed effect summary for all studies up to and including that study is displayed in its place. If the studies are already sorted in chronological order, then the cumulative analysis provides the summary effects of all studies known to a particular study on its publication date.
4: Cumulative with random effects

LB/UB Symmetry: When both lower and upper bounds are given for a confidence interval, they must be consistent with each other. UNISTAT will calculate the central tendency using both LB and UB. If the two values are within 10% of each other they will be deemed consistent and the results will be displayed. Otherwise, the current study will be reported as having invalid data. You can adjust the consistency level by entering a different multiplier.

Imputed r: Calculation of effect size for paired samples requires knowing the correlation between them. However, this may not be readily available in most studies. In such cases, the program assumes a correlation of 0.5 by default. This value can be changed by the user. See section 6.8.2.4. Paired Samples for details.

Use t-test for correlation: If the t-statistic was originally reported for a study based on the correlation coefficient, then enter a non-zero value for this box. See section 6.8.2.4. Paired Samples for details.

Standard Error for Hedges' g: You can choose one of two alternative methods for calculating the standard error of Hedges' g. See section 6.8.2.3. Unpaired Samples for details.

Display relative or absolute weights: In Results and Forest Plot output, the weights can be displayed as relative weights in percentage terms or as absolute weights as used in calculations.

Ratio or Log(Ratio): This option will be available when one of odds, Peto odds, risk, rate and hazard ratio is selected as the summary effect size.

Weights: Inverse Variance or Mantel-Haenszel: The inverse variance method is the default for all types of effect size. Mantel-Haenszel weights can be calculated for the fixed effect model when one of odds or risk ratios or risk difference is selected as the summary effect size.

Heterogeneity with inverse variance weights: When Mantel-Haenszel weights are selected, the heterogeneity statistics are calculated with Mantel-Haenszel weights by default (see 6.8.4.2. Tests). You can, however, override this and force using inverse variance weights.

6.8.3.1. Fixed Effect Model

Inverse variance method

The weighted mean of individual effect sizes M is calculated as:

$$M = \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i}$$

where the weight Wi is the reciprocal of the effect size variance Vi for a study:

$$W_i = \frac{1}{V_i}$$

and the variance of the mean effect size M is:

$$V = \frac{1}{\sum_{i=1}^{n} W_i}$$

The Z-statistic is:

$$Z = \frac{M}{\sqrt{V}}$$

The relative weights are:

$$w_i = 100\frac{V}{W_i}$$

and the standardised residuals:

$$e_i = \frac{Y_i - M}{\sqrt{V}}$$

Mantel-Haenszel method for 2 x 2 tables

Odds ratio:

The summary odds ratio calculated as:

$$OR_{MH} = \frac{\sum_{i=1}^{n} W_{MHi} OR_i}{\sum_{i=1}^{n} W_{MHi}}$$

where the Mantel-Haenszel weight W_{MHi} is:

$$W_{MHi} = \frac{n_{Bi} + n_{Ci}}{n_i}$$

and the variance $Ln(OR_{MH})$ is:

$$SE_{Ln(OR_{MH})} = \sqrt{\frac{1}{2}\left(\frac{E}{R^2} + \frac{F+G}{RS} + \frac{H}{S^2}\right)}$$

where:

$$R = \sum \frac{n_{Ai} + n_{Di}}{n_i}$$

$$S = \sum \frac{n_{Bi} + n_{Ci}}{n_i}$$

$$E = \sum \frac{(n_{Ai} + n_{Di}) n_{Ai} n_{Di}}{n_i^2}$$

$$F = \sum \frac{(n_{Ai} + n_{Di}) n_{Bi} n_{Ci}}{n_i^2}$$

$$G = \sum \frac{(n_{Bi} + n_{Ci}) n_{Ai} n_{Di}}{n_i^2}$$

$$H = \sum \frac{(n_{Bi} + n_{Ci}) n_{Bi} n_{Ci}}{n_i^2}$$

Risk ratio:

The Mantel-Haenszel weight W_{MHi} is calculated as:

$$W_{MHi} = \frac{n_{Ci}\left(n_{Ai} + n_{Bi}\right)}{n_i}$$

and the variance $Ln(RR_{MH})$ is:

$$SE_{Ln(RR_{MH})} = \sqrt{\frac{P}{RS}}$$

where:

$$P = \sum \frac{\left(n_{Ai} + n_{Bi}\right)\left(n_{Ci} + n_{Di}\right)\left(n_{Ai} + n_{Ci}\right) - n_{Ai} n_{Ci} N_i}{n_i^2}$$

$$R = \sum \frac{n_{Ai}\left(n_{Ci} + n_{Di}\right)}{n_i}$$

$$S = \sum \frac{n_{Ci}\left(n_{Ai} + n_{Bi}\right)}{n_i}$$

Risk difference:

The Mantel-Haenszel weight W_{MHi} is:

$$W_{MHi} = \frac{\left(n_{Ai} + n_{Bi}\right)\left(n_{Ci} + n_{Di}\right)}{n_i}$$

and the variance $Ln(RD_{MH})$ is:

$$SE_{Ln(RD_{MH})} = \frac{\sqrt{J}}{K}$$

where:

$$J = \sum \frac{n_{Ai} n_{Bi}\left(n_{Ci} + n_{Di}\right)^3 + n_{Ci} n_{Di}\left(n_{Ai} + n_{Bi}\right)^3}{\left(n_{Ai} + n_{Bi}\right)\left(n_{Ci} + n_{Di}\right) n_i^2}$$

$$K = \sum \frac{\left(n_{Ai} + n_{Bi}\right)\left(n_{Ci} + n_{Di}\right)}{n_i}$$

6.8.3.2. Random Effects Model

Define the between-studies variance as introduced by DerSimonian R, Laird N. (1986):

$$\tau^2 = \frac{Q - df}{C}$$

where df = n – 1, Cochran's Q is:

$$Q = \sum W_i (Y_i - M)^2$$

and:

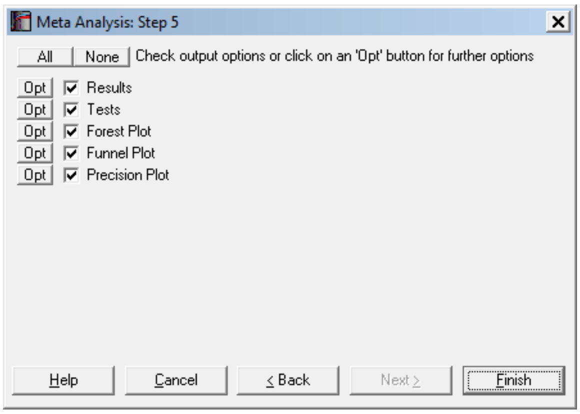
$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

Next, determine the total variance for each study as the sum of within-study variance V_i and the between-studies variance of τ^2:

$$V_i^* = V_i + \tau_i^2$$

The mean effect size, its variance, relative weights and standardised residuals are then computed as in the inverse variance method.

6.8.4. Output Options

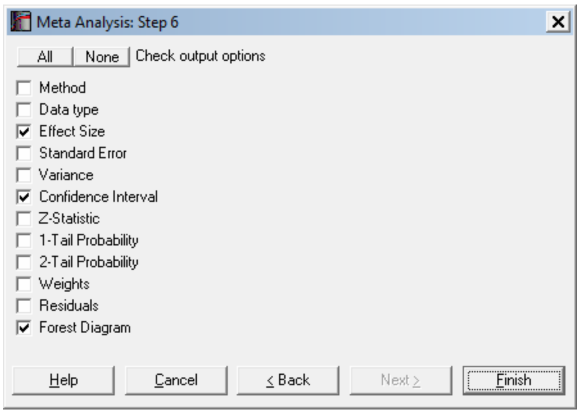

Meta Analysis: Step 5

Check output options or click on an 'Opt' button for further options

- ☑ Results
- ☑ Tests
- ☑ Forest Plot
- ☑ Funnel Plot
- ☑ Precision Plot

The first two output options have further selections.

6.8.4.1. Results

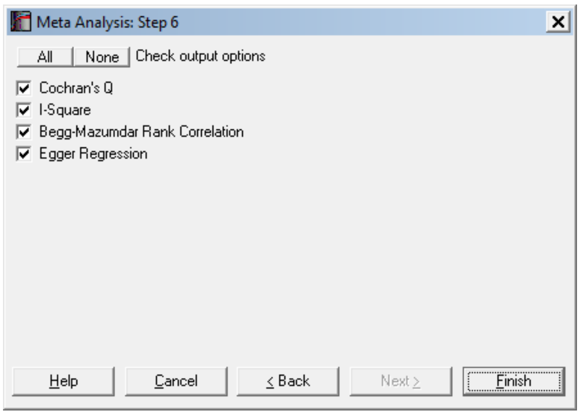

Meta Analysis: Step 6

Check output options

- □ Method
- □ Data type
- ☑ Effect Size
- □ Standard Error
- □ Variance
- ☑ Confidence Interval
- □ Z-Statistic
- □ 1-Tail Probability
- □ 2-Tail Probability
- □ Weights
- □ Residuals
- ☑ Forest Diagram

By default, only the effect size, its confidence limits and a character forest diagram are included in the output. Summary effects with fixed and random effects models are also displayed at the bottom of the table. It is possible to output the weights used for each model and residuals.

6.8.4.2. Tests


Meta Analysis: Step 6

All | None | Check output options

☑ Cochran's Q
☑ I-Square
☑ Begg-Mazumdar Rank Correlation
☑ Egger Regression

Tests for Heterogeneity

Cochran's Q: As already defined in section 6.8.3.2. Random Effects Model:

$$Q = \sum W_i (Y_i - M)^2$$

which is chi-square distributed with n – 1 degrees of freedom. The weights here are always the inverse variance weights but when the Mantel-Haenszel option is selected, the average is the Mantel-Haenszel average. You can override this and use the inverse variance average for M by entering 1 for **Weights: Inverse Variance or Mantel-Haenszel** in the intermediate inputs dialogue.

I-square: This is a percentage calculated as:

$$I^2 = 100 \frac{Q - df}{Q}$$

where df = n – 1.

Lower values of I^2 indicate less heterogeneity. The confidence intervals for I^2 are calculated as in Higgins & Thompson (2002).

Tests for Publication Bias

Forest, funnel and precision plots provide a visual guide whether the studies being analysed are biased in positive or negative direction. The following two tests provide a statistical measure for publication bias.

Begg-Mazumdar Rank Correlation: The Kendall's rank correlation coefficient is calculated (Begg & Mazumdar, 1994) between the standardised effect size differences:

$$(Y_i - M)\sqrt{W_i}, i = 1,...,n$$

and the standardised variances of effect sizes:

$$V_i - \frac{1}{\sum W_i}, i = 1,...,n$$

where the weights are the inverse variance weights.

Egger Regression: A bivariate linear regression (Egger, Smith, Schneider & Minder, 1997) is run between the standardised effect sizes:

$$Y_i \sqrt{W_i}, i = 1,...,n$$

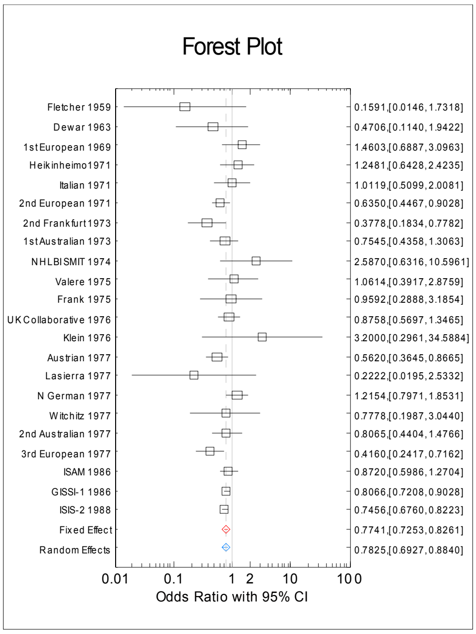
against their precisions:

$$\sqrt{W_i}, i = 1,...,n$$

where the weights are the inverse variance weights.

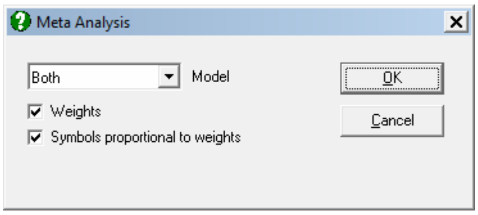
6.8.4.3. Forest Plot

A high-resolution forest plot is drawn with study names on the left Y-axis and on the right Y-axis, the summary effect, its confidence interval and the relative weights.



By default, UNISTAT produces a forest plot in landscape orientation. If, however, there are too many studies included in the analysis, you can select Export → Metafile Orientation → Portrait option from the graphics window menu.

Using the Edit → View dialogue, you can choose to display one, none or both fixed and random effects, weights and symbols proportional to the weights. When both fixed and random effects are selected, the fixed weights are displayed. Symbols are not drawn proportional to weights when the weights box is unchecked or the absolute weights are displayed (as opposed to percentage weights).


Meta Analysis

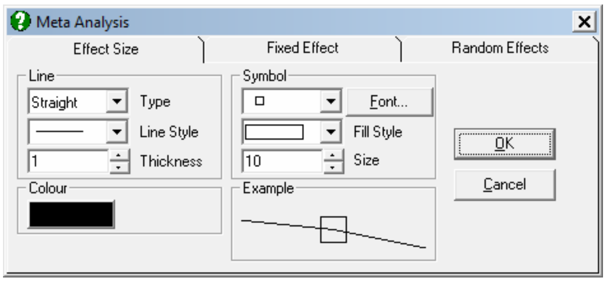
Both ▾ Model

☑ Weights

☑ Symbols proportional to weights

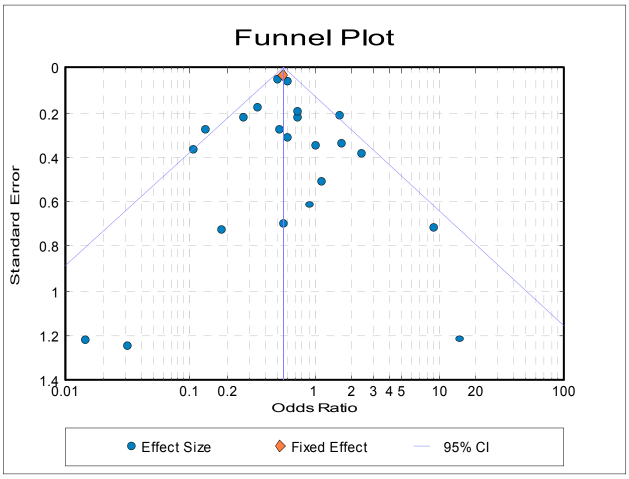
As in other graphics procedures, X-axis parameters can be changed from the Edit → Scale dialogue. Although Y-axis scale parameters cannot be changed, you can change the font size to accommodate for a large number of studies.

The three different types of symbols displayed in the graph can be changed using the Edit → Lines dialogue.

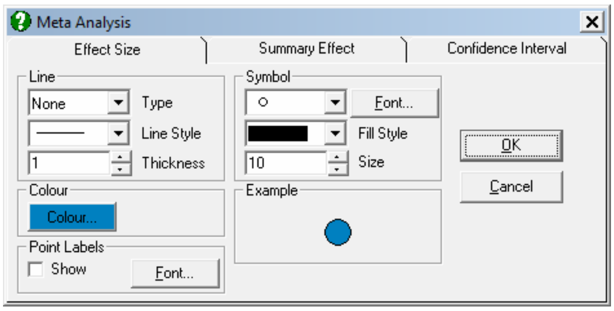


6.8.4.4. Funnel Plot

Standard error of individual studies are plotted against their effect size. A vertical line is drawn through the summary effect size, which can be fixed or random. Confidence interval is drawn at 95% level by default, but this can be changed by the user.

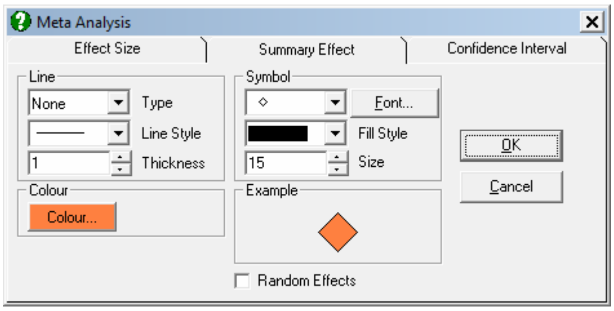


You can edit the lines and symbols and choose to display study names next to each symbol using the Edit → Lines dialogue.



Type: None
Line Style: ———
Thickness: 1
Symbol: ○
Fill Style: ■
Size: 10
□ Show

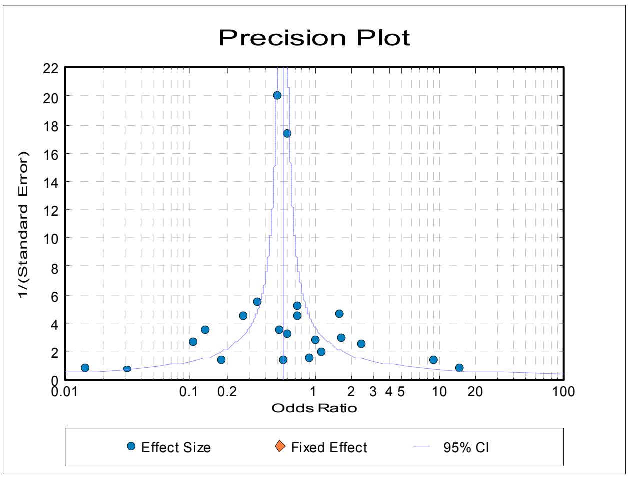
The Summary Effect tab on the same dialogue allows you to display the funnel plot using random effects, instead of the fixed effect.



Type: None
Line Style: ———
Thickness: 1
Symbol: ◇
Fill Style: ■
Size: 15
□ Random Effects

6.8.4.5. Precision Plot

This is similar to funnel plot, but the reciprocal of the individual study standard errors are plotted against their effect sizes.

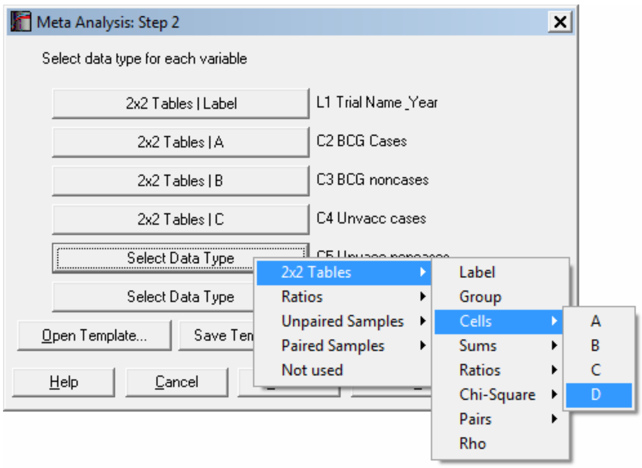

Precision Plot
1/(Standard Error)
22
20
18
16
14
12
10
8
6
4
2
0
0.01
0.1
0.2
1
2
3
4
5
10
20
100
Odds Ratio
● Effect Size
◆ Fixed Effect
95% CI

6.8.5. Examples

Example 1: Meta Analysis with Inverse Variance Weights

Data on the efficacy of BCG vaccine against tuberculosis is given for eleven studies in the form of 2 x 2 tables in Colditz et al. (1994), p. 699.

Open the data file META and select the first six columns as variables. In Excel Add-In Mode, if the program asks if there are case labels in column 1, select *No*. On the second dialogue, assign the following tasks to the selected variables.



Meta Analysis: Step 2

Select data type for each variable

2x2 Tables | Label — L1 Trial Name_Year
2x2 Tables | A — C2 BCG Cases
2x2 Tables | B — C3 BCG noncases
2x2 Tables | C — C4 Unvacc cases
Select Data Type — C5 Unvacc noncases
Select Data Type

On the next dialogue, select Risk Ratio as the summary effect size type. Click [Finish] to display results with the default options.

Meta Analysis

Results

<table>
<tr><th></th><th>Risk
Ratio</th><th>Lower
95%</th><th>Upper
95%</th><th>-2.5465   1.8762</th></tr>
<tr><td>Canada 1933</td><td>0.2049</td><td>0.0863</td><td>0.4864</td><td> -----+------ |</td></tr>
<tr><td>Northern USA 1935</td><td>0.4109</td><td>0.1343</td><td>1.2574</td><td>   -------+-----|-</td></tr>
<tr><td>Chicago 1941</td><td>0.2538</td><td>0.1494</td><td>0.4310</td><td>   ----+--- |</td></tr>
<tr><td>Georgia (Sch) 1947</td><td>1.5619</td><td>0.3737</td><td>6.5284</td><td>   -------|--+---------</td></tr>
<tr><td>Puerto Rico 1949</td><td>0.7122</td><td>0.5725</td><td>0.8860</td><td>   -+--|</td></tr>
<tr><td>Georgia (Com) 1950</td><td>0.9828</td><td>0.5821</td><td>1.6593</td><td>   ----|---</td></tr>
<tr><td>Madanapalle 1950</td><td>0.8045</td><td>0.5163</td><td>1.2536</td><td>   ---+-|-</td></tr>
<tr><td>UK 1950</td><td>0.2366</td><td>0.1793</td><td>0.3121</td><td>   --+-- |</td></tr>
<tr><td>South Africa 1965</td><td>0.6254</td><td>0.3926</td><td>0.9962</td><td>   ---+--|</td></tr>
<tr><td>Haiti 1965</td><td>0.1977</td><td>0.0784</td><td>0.4989</td><td>------+------ |</td></tr>
<tr><td>Madras 1968</td><td>1.0120</td><td>0.8946</td><td>1.1449</td><td>   -|-</td></tr>
<tr><td>Fixed Effect</td><td>0.7305</td><td>0.6668</td><td>0.8002</td><td>   -+ |</td></tr>
<tr><td>Random Effects</td><td>0.5080</td><td>0.3361</td><td>0.7679</td><td>   --+--- |</td></tr>
</table>

Cochran's Q

	Cochran's Q	Degrees of Freedom	Probability	Tau-square
Total	125.6261	10	0.0000	0.3818

I-Square

	I-square %	Lower 95%	Upper 95%
Total	92.0399	87.7435	94.8302
*		88.2994	94.1658

* CI using inverse noncentral chi-square function.

Begg-Mazumdar Rank Correlation

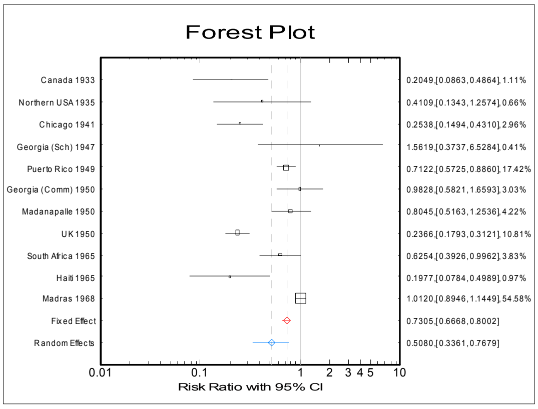
	Correlation Coefficient	Z-Statistic	1-Tail Probability	2-Tail Probability
Kendall Rank	-0.0909	-0.3892	0.3485	0.6971
Kendall Rank with CC	-0.0727	-0.3114	0.3777	0.7555

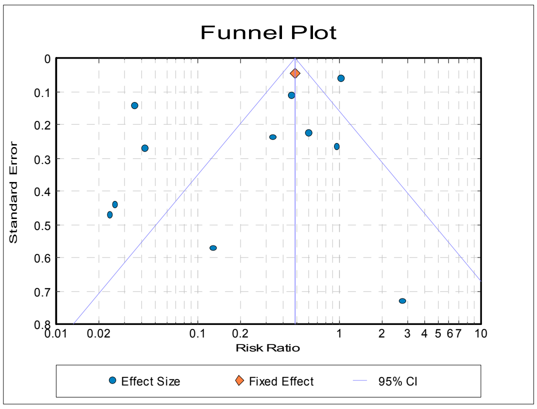
	Lower 95%	Upper 95%
Kendall Rank	-0.6551	0.5383
Kendall Rank with CC	-0.6445	0.5512

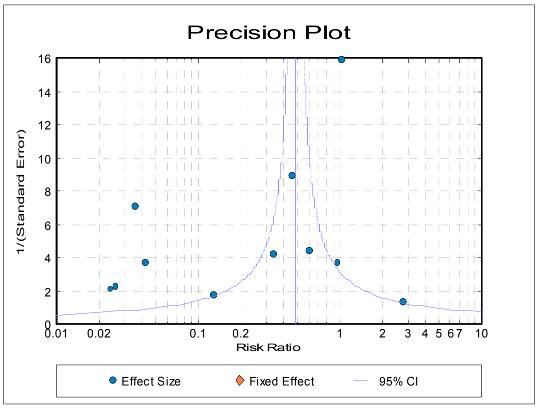
Egger Regression

	Coefficient	Standard Error	t-Statistic	Degrees of Freedom
Intercept	-2.6889	1.5541	-1.7303	9

	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
Intercept	0.0588	0.1176	-6.2045	0.8266







Example 2: Meta Analysis with Subgroups

Continuing from the last example, go back to the Variable Selection Dialogue, add *Group* (*S6*) to the variable list. On the second dialogue, assign the task 2 x 2 Tables / Group to this variable. In Output Options Dialogue (Step 5), uncheck the Funnel and Precision plot boxes and click the [Opt] button to the left of Results. Check Group and 2-Tail Probability and uncheck Forest Diagram.

Meta Analysis

Results

Selected by Group

	Group	Risk Ratio	Lower 95%	Upper 95%	2-Tail Probability
Georgia (Sch) 1947	Central	1.5619	0.3737	6.5284	0.5412
Georgia (Comm) 1950	Central	0.9828	0.5821	1.6593	0.9483
South Africa 1965	Central	0.6254	0.3926	0.9962	0.0482
Fixed Effect	Central	0.7947	0.5667	1.1144	0.1828
Random Effects	Central	0.8117	0.5430	1.2134	0.3090
Canada 1933	Northern	0.2049	0.0863	0.4864	0.0003
Northern USA 1935	Northern	0.4109	0.1343	1.2574	0.1191
Chicago 1941	Northern	0.2538	0.1494	0.4310	0.0000
UK 1950	Northern	0.2366	0.1793	0.3121	0.0000
Fixed Effect	Northern	0.2430	0.1928	0.3061	0.0000
Random Effects	Northern	0.2430	0.1928	0.3061	0.0000
Puerto Rico 1949	Tropical	0.7122	0.5725	0.8860	0.0023
Madanapalle 1950	Tropical	0.8045	0.5163	1.2536	0.3364
Haiti 1965	Tropical	0.1977	0.0784	0.4989	0.0006
Madras 1968	Tropical	1.0120	0.8946	1.1449	0.8494
Fixed Effect	Tropical	0.9045	0.8154	1.0033	0.0578
Random Effects	Tropical	0.7197	0.5000	1.0359	0.0767
Total Fixed Effect		0.7305	0.6668	0.8002	0.0000
Total Random Effects		0.5080	0.3361	0.7679	0.0013

Cochran's Q

Selected by Group

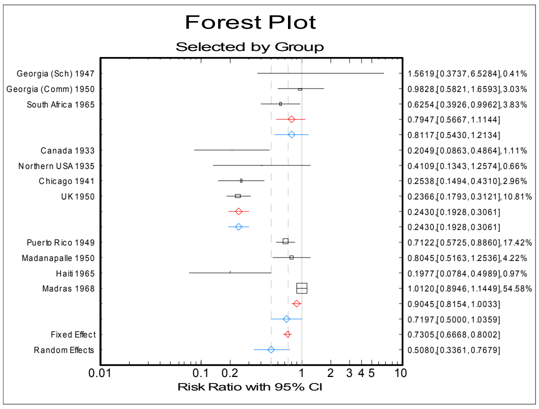
Group	Cochran's Q	Degrees of Freedom	Probability	Tau-square
Central	2.5072	2	0.2855	0.0277
Northern	1.0593	3	0.7869	0.0000
Tropical	18.4213	3	0.0004	0.0969
Within	21.9878	8	0.0049	
Between	103.6383	2	0.0000	
Total	125.6261	10	0.0000	0.3818

I-Square

Selected by Group

Group	I-square %	Lower 95%	Upper 95%
Central	20.2284	0.0000	91.7022
*		0.0000	77.9553
Northern	0.0000	0.0000	84.6878
*		0.0000	67.9090
Tropical	83.7145	58.8130	93.5607
*		43.4256	91.9141
Total	92.0399	87.7435	94.8302
*		88.2994	94.1658

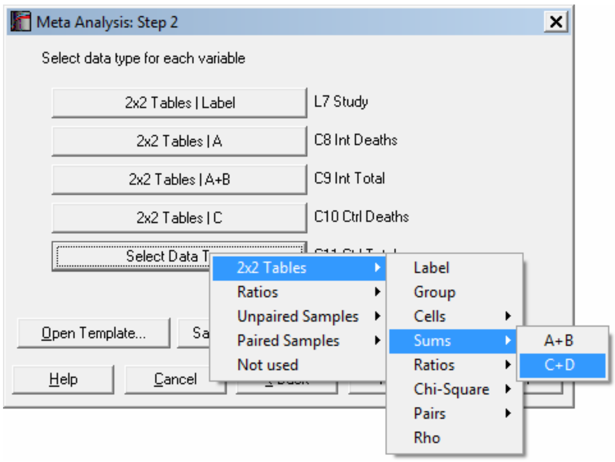
* CI using inverse noncentral chi-square function.



Example 3: Meta Analysis with Mantel-Haenszel Weights

Data on 22 randomised controlled trials of streptokinase in the prevention of death following myocardial infarction is given in the form of 2 x 2 tables in ISIS-2 (1988).

Open the data file META and select *Study* to *Ctrl Total* (*L7-C11*) as [Variable]s. In Excel Add-In Mode, if the program asks if there are case labels in column 1, select *No*. On the second dialogue, assign the following tasks to the selected variables.


Meta Analysis: Step 2

Select data type for each variable

2x2 Tables | Label — L7 Study
2x2 Tables | A — C8 Int Deaths
2x2 Tables | A+B — C9 Int Total
2x2 Tables | C — C10 Ctrl Deaths
Select Data T — C11 Ctrl Total

Next select Risk Ratio as the summary effect size type and on the next dialogue select the Mantel-Haenszel method. Click [Finish] to display results.

Meta Analysis

Results

<table>
<tr><th></th><th>Risk
Ratio</th><th>Lower
95%</th><th>Upper
95%</th><th>-3.5062 2.9695</th></tr>
<tr><td>Fletcher 1959</td><td>0.2292</td><td>0.0300</td><td>1.7499</td><td>---------+------|--</td></tr>
<tr><td>Dewar 1963</td><td>0.5714</td><td>0.1962</td><td>1.6647</td><td>-----+--|--</td></tr>
<tr><td>1st European 1969</td><td>1.3494</td><td>0.7429</td><td>2.4509</td><td>--|+---</td></tr>
<tr><td>Heikinheimo1971</td><td>1.2232</td><td>0.6688</td><td>2.2371</td><td>--|+--</td></tr>
<tr><td>Italian 1971</td><td>1.0105</td><td>0.5510</td><td>1.8531</td><td>---|--</td></tr>
<tr><td>2nd European 1971</td><td>0.7026</td><td>0.5338</td><td>0.9247</td><td>-+-|</td></tr>
<tr><td>2nd Frankfurt 1973</td><td>0.4571</td><td>0.2522</td><td>0.8282</td><td>--+---|</td></tr>
<tr><td>1st Australian 1973</td><td>0.7786</td><td>0.4780</td><td>1.2683</td><td>---+|-</td></tr>
<tr><td>NHLBI SMIT 1974</td><td>2.3774</td><td>0.6490</td><td>8.7086</td><td>--|---+-----</td></tr>
<tr><td>Valere 1975</td><td>1.0476</td><td>0.4809</td><td>2.2821</td><td>----|---</td></tr>
<tr><td>Frank 1975</td><td>0.9636</td><td>0.3316</td><td>2.8005</td><td>-----|----</td></tr>
<tr><td>UK Collaborative 1976</td><td>0.8956</td><td>0.6261</td><td>1.2809</td><td>-+|-</td></tr>
<tr><td>Klein 1976</td><td>2.5714</td><td>0.3394</td><td>19.4813</td><td>-----|---+---------</td></tr>
<tr><td>Austrian 1977</td><td>0.6080</td><td>0.4173</td><td>0.8861</td><td>-+--|</td></tr>
<tr><td>Lasierra 1977</td><td>0.2821</td><td>0.0340</td><td>2.3403</td><td>---------+-----|----</td></tr>
<tr><td>N German 1977</td><td>1.1609</td><td>0.8403</td><td>1.6038</td><td>-|--</td></tr>
<tr><td>Witchitz 1977</td><td>0.8125</td><td>0.2634</td><td>2.5062</td><td>-----+|----</td></tr>
<tr><td>2nd Australian 1977</td><td>0.8497</td><td>0.5369</td><td>1.3446</td><td>--+|-</td></tr>
<tr><td>3rd European 1977</td><td>0.5096</td><td>0.3327</td><td>0.7805</td><td>--+--|</td></tr>
<tr><td>ISAM 1986</td><td>0.8801</td><td>0.6195</td><td>1.2503</td><td>-+|-</td></tr>
<tr><td>GISSI-1 1986</td><td>0.8274</td><td>0.7491</td><td>0.9138</td><td>-+|</td></tr>
<tr><td>ISIS-2 1988</td><td>0.7690</td><td>0.7044</td><td>0.8395</td><td>-+|</td></tr>
<tr><td>Fixed Effect (MH)</td><td>0.7988</td><td>0.7546</td><td>0.8455</td><td>-+|</td></tr>
<tr><td>Random Effects (MH)</td><td>0.8112</td><td>0.7333</td><td>0.8973</td><td>-+|</td></tr>
</table>

Cochran's Q

	Cochran's Q (MH)	Degrees of Freedom	Probability	Tau-square
Total	30.4116	21	0.0840	0.0840

I-Square

	I-square (MH) %	Lower 95%	Upper 95%
Total	30.9474	0.0000	58.9457
*		0.0000	58.0966

* CI using inverse noncentral chi-square function.

Begg-Mazumdar Rank Correlation

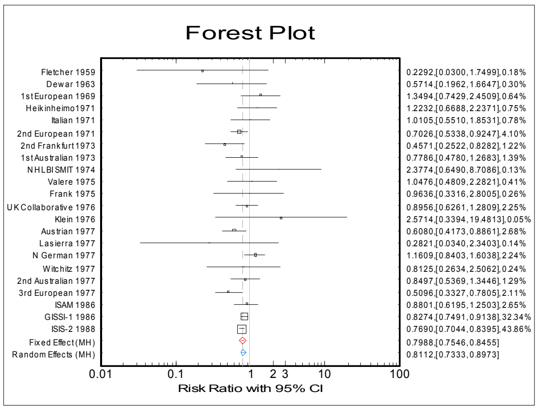
	Correlation Coefficient	Z-Statistic	1-Tail Probability	2-Tail Probability
Kendall Rank	0.6710	4.3707	0.0000	0.0000
Kendall Rank with CC	0.6753	4.3989	0.0000	0.0000

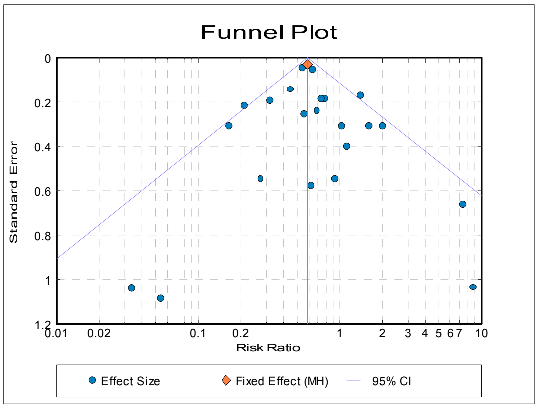
	Lower 95%	Upper 95%
Kendall Rank	0.3478	0.8517
Kendall Rank with CC	0.3547	0.8538

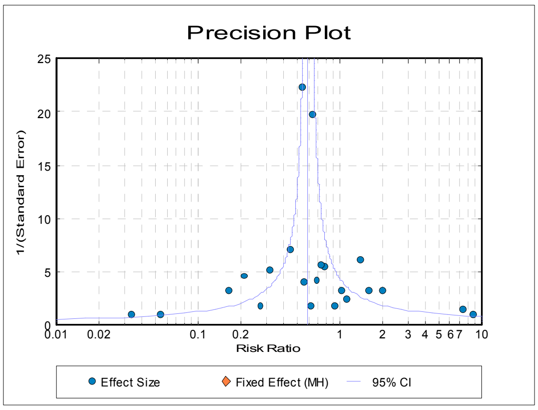
Egger Regression

	Coefficient	Standard Error	t-Statistic	Degrees of Freedom
Intercept	0.1670	0.3567	0.4681	20

	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
Intercept	0.3224	0.6448	-0.5771	0.9111







Example 4: Cumulative Meta Analysis

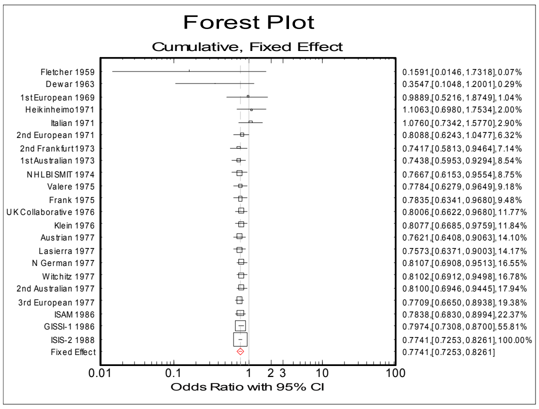
Using the data from the last example, select Odds Ratio as the summary effect size type and on the next dialogue select the inverse variance method for weights and select cumulative fixed effect (3: CUM-Fx) analysis to run. On the Output Options Dialogue select only the Results and Forest Plot output options and click [Opt] situated next to results. After selecting only the Effect Size, Standard Error, Z-Statistic and 2-Tail Probability options, come back to the Output Options Dialogue and click [Finish] to display results.

Meta Analysis

Results

Cumulative, Fixed Effect

	Odds Ratio	Lower 95%	Upper 95%	Z-Statistic	2-Tail Probability
Fletcher 1959	0.1591	0.0146	1.7318	-1.5091	0.1313
Dewar 1963	0.3547	0.1048	1.2001	-1.6666	0.0956
1st European 1969	0.9889	0.5216	1.8749	-0.0341	0.9728
Heikinheimo1971	1.1063	0.6980	1.7534	0.4298	0.6673
Italian 1971	1.0760	0.7342	1.5770	0.3756	0.7072
2nd European 1971	0.8088	0.6243	1.0477	-1.6068	0.1081
2nd Frankfurt 1973	0.7417	0.5813	0.9464	-2.4029	0.0163
1st Australian 1973	0.7438	0.5953	0.9294	-2.6044	0.0092
NHLBI SMIT 1974	0.7667	0.6153	0.9554	-2.3663	0.0180
Valere 1975	0.7784	0.6279	0.9649	-2.2855	0.0223
Frank 1975	0.7835	0.6341	0.9680	-2.2617	0.0237
UK Collaborative 1976	0.8006	0.6622	0.9680	-2.2962	0.0217
Klein 1976	0.8077	0.6685	0.9759	-2.2128	0.0269
Austrian 1977	0.7621	0.6408	0.9063	-3.0721	0.0021
Lasierra 1977	0.7573	0.6371	0.9003	-3.1504	0.0016
N German 1977	0.8107	0.6908	0.9513	-2.5713	0.0101
Witchitz 1977	0.8102	0.6912	0.9498	-2.5958	0.0094
2nd Australian 1977	0.8100	0.6946	0.9445	-2.6877	0.0072
3rd European 1977	0.7709	0.6650	0.8938	-3.4478	0.0006
ISAM 1986	0.7838	0.6830	0.8994	-3.4697	0.0005
GISSI-1 1986	0.7974	0.7308	0.8700	-5.0925	0.0000
ISIS-2 1988	0.7741	0.7253	0.8261	-7.7111	0.0000
Fixed Effect	0.7741	0.7253	0.8261	-7.7111	0.0000



UNISTAT Statistical Package

Chapter 7
Regression and Analysis of Variance

7.0. Overview

This module brings together various regression and Analysis of Variance procedures. It is possible to perform linear (ordinary least squares or OLS), polynomial, stepwise, nonlinear, Logit / Probit / Gompit, logistic, multinomial, Poisson and Box-Cox regressions and multi-way unbalanced analysis of variance and covariance with repeated measures and nested designs. Further options are provided for each type of analysis.

The procedures available in this module can be grouped in two broad categories according to the type of data they use. In order to avoid repetition under each section, these data types will be explained here.

7.0.1. Matrix Data

Raw data arranged in a matrix is the data type used by all regression procedures. An unlimited number of data columns can be selected as independent variables from the Variables Available list. Each row of the data matrix should correspond to the same case (such as time intervals, or names of patients in a hospital). Therefore, all columns are expected to have the same size.

Missing data handling: Any rows containing one or more missing observations are omitted. Here we refer to a row of the matrix defined by the selected set of variables, but not to a row of all columns in the data matrix. Therefore, depending on the combination of selected columns, different numbers of rows may be omitted.

An important special case for missing data handling is when more than one dependent variable is selected. In such cases, missing values will be omitted for each run (with a different dependent variable) separately. Consider, for instance, a hypothetical data set from which we select two dependent variables, where only the first one contains a missing value. Also assume that no other variables contain any missing values. In this case, the first run will report one case omitted due to missing values and in the second run (with the second dependent variable) no cases will be omitted.

It is also important to remember that in Stepwise Regression, rows containing missing values are omitted at the beginning. This means that rows will be omitted from the matrix defined by all variables selected for the stepwise analysis rather than the final configuration of variables that are included in the regression equation. This may be important when you run the final stepwise equation once again using the Linear Regression procedure (say, to obtain diagnostic statistics). In this case, it may be necessary to delete some rows manually.

Subsample selection: Most procedures requiring matrix data will also allow for selection of subsample of rows defined by one or more factor columns. A factor is a categorical variable that contains a limited number of distinct numeric or string values (levels). An unlimited number of factor columns may be selected from the Variables Available list. In case two or more factors are selected, you will be able to include in the analysis any rows defined by combinations of factor levels. More information on this topic can be found for each regression procedure.

The Nonlinear Regression procedure allows for selection of only one factor variable and the task assigned to this variable is not selection of subsamples of rows. For more information see 7.2.4. Nonlinear Regression.

Weights: Data in matrix form allows for selection of one column to be used as weights in the analysis of rest of the selected columns. When, for instance, rows of data correspond to different regions in a country, you may eliminate the effect of population differences by selecting a variable containing region populations as weights. In this case, the program will first normalise the weights column so that its sum is equal to the valid number of cases (the number of rows after omission of missing rows), and then run the regression after multiplying all selected columns by the square root of the normalised weights column.

A weighted regression is thus a regression on a different set of data without transforming the data columns in the original data set. Also considering the type of missing data handling (omission of rows containing missing observations), and creation of dummy and lag/lead variables, the user may not be sure about the final configuration of data on which the regression is run. This is the reason why the menu option Statistics 1 → Matrix Statistics is provided here. Under this menu option it is possible to compute descriptive statistics, zero order (Pearson) Correlation Coefficients, variance-covariance and moment matrices for the same data set which is used in the Regression Analysis. The weights option and missing data handling is exactly the same as in the Regression Analysis.

Stepwise Regression and Logit / Probit / Gompit do not support weight variables and weights in Logistic Regression, Multinomial Regression and Poisson Regression are frequency weights.

7.0.2. Categorical Data

All Analysis of Variance procedures require use of categorical variables to determine the group membership of a particular observation in a continuous data variable (see 7.3.0.1. ANOVA and GLM Data Format).

Missing data handling: Any rows containing one or more missing values are omitted. In case of Analysis of Variance procedures, this includes the factor columns as well as the continuous data variable.

As in matrix format data explained above, selection of more than one dependent variable also needs special consideration here. In such cases, missing values will be omitted for each run (with a different dependent variable) separately.

Factors: A column intended for classifying observations of another column is called a factor. A factor may be a numeric or string variable with a limited number of distinct values (levels). An unlimited number of data columns may be selected from the Variables Available list as factors by clicking on [Factor]. The order of selection is significant. All Analysis of Variance procedures also require the choice of one or more continuous data (dependent) variables which are marked by clicking on [Dependent]. When more than one [Dependent] variable is selected, the same model will be run with each data variable separately.

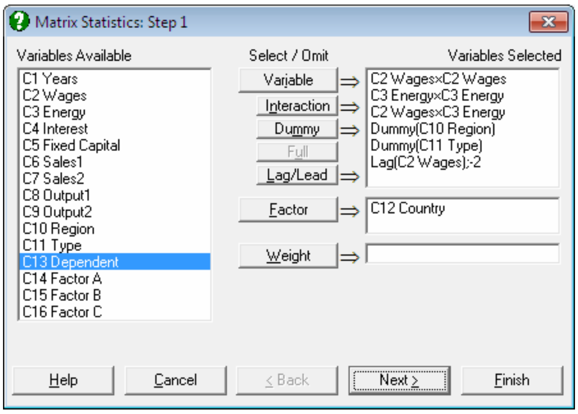
Levels: A numeric or string variable selected as a factor must have a limited number of distinct values, which are called *levels*. Levels can be any integer or floating point numbers, or short or long strings. The program will first scan a factor column and register the distinct values. If the number of such values is too large, then the program may run out of memory. In this case a message will be displayed and the procedure will be aborted.

7.1. Matrix Statistics

The Matrix Statistics procedure provides information on matrix data under the same assumptions as the regression procedures. This means that all column lengths must be equal and any rows containing at least one missing value are omitted.

It is possible to create interaction terms, dummy and lag/lead variables just as in Linear Regression, Stepwise Regression, Logistic Regression, Multinomial Regression, Poisson Regression and Cox Regression procedures. This feature provides summary statistics on the terms of the regression models selected or created. It also enables you to send the entire final raw data (**X**) matrix to the Output Medium, so that you can see the actual values of all terms in the model. In Stand-Alone Mode, this output can then be sent to the Data Processor and used as input data in other procedures if necessary.

7.1.1. Matrix Statistics Variable Selection



Matrix Statistics: Step 1

Variables Available

C1 Years
C2 Wages
C3 Energy
C4 Interest
C5 Fixed Capital
C6 Sales1
C7 Sales2
C8 Output1
C9 Output2
C10 Region
C11 Type
C13 Dependent
C14 Factor A
C15 Factor B
C16 Factor C

Variables Selected

C2 Wages×C2 Wages
C3 Energy×C3 Energy
C2 Wages×C3 Energy
Dummy(C10 Region)
Dummy(C11 Type)
Lag(C2 Wages);-2

Factor: C12 Country

Weight:

For further information on the tasks of the following buttons see 7.2.1.1. Linear Regression Variable Selection and 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables.

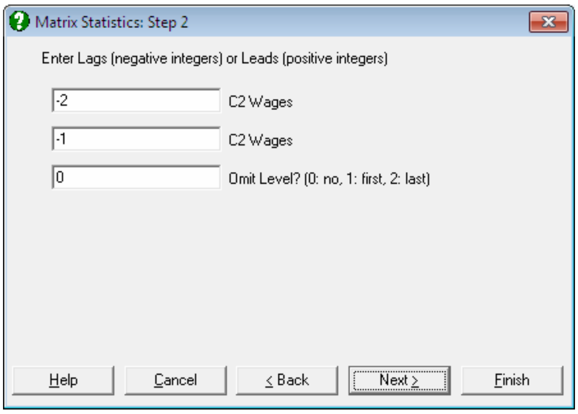
Variable: Click on [Variable] to select a column containing continuous numeric data.

Interaction: Use this button to create variables, which are the products of existing numeric variables. If only one variable is highlighted, then the new variable will be the product of the selected variable by itself.

Dummy: This button is used to create n or n – 1 new dummy (or indicator) variables for a factor column containing n levels. It is possible to include all n levels in the analysis or to omit the first or the last level in order to remove the inherent over-parameterisation of the model.

Full: This button becomes activated when two or more categorical variables are highlighted. Like the [Dummy] button, it is also used to create dummy variables. The only difference is that this button will create all necessary dummy variables and their interactions to specify a complete model.

Lag/Lead: This button is used to create new variables by shifting the rows of an existing variable up or down. When a lag variable is specified then a further dialogue will ask for the number of lags (or leads) for each item selected. Negative integers represent the lags and positive integers the leads.



Matrix Statistics: Step 2

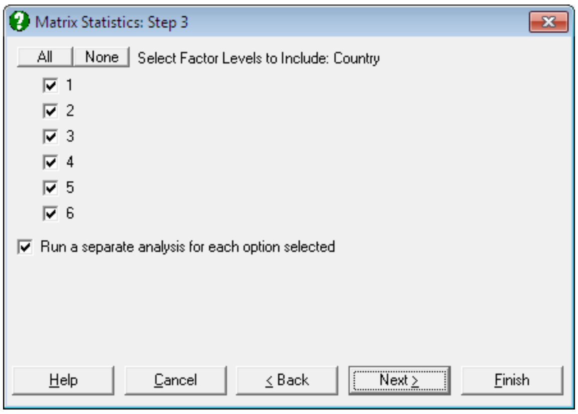
Enter Lags (negative integers) or Leads (positive integers)

-2 C2 Wages

-1 C2 Wages

0 Omit Level? (0: no, 1: first, 2: last)

Factor: Selection of a factor variable is optional. It is possible to select an unlimited number of factor variables to define the rows to be included in the analysis. This allows you to analyse subsamples of rows (cases) without having to extract them by data manipulation first. When one or more factor columns are selected a further dialogue will ask you which factor levels to include. It is also possible to run a single analysis on all selected rows combined, or to run a separate analysis for each selection.



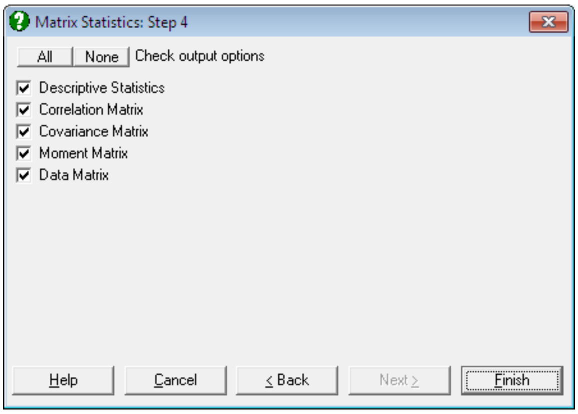
Select Factor Levels to Include: Country

☑ 1
☑ 2
☑ 3
☑ 4
☑ 5
☑ 6

☑ Run a separate analysis for each option selected

Weight: As in the Linear Regression procedure, a column may be selected as a weights variable. The program first normalises this column so that its sum is equal to the number of valid rows (after omitting missing rows), and then multiplies every row of the other selected columns by the square root of the corresponding row of the normalised weight column. The original data remains unchanged.

7.1.2. Matrix Statistics Output Options



Check output options

☑ Descriptive Statistics
☑ Correlation Matrix
☑ Covariance Matrix
☑ Moment Matrix
☑ Data Matrix

Descriptive Statistics: The following information on selected data columns is displayed in the form of a matrix: mean, standard deviation, standard error, variance, sum of squares, minimum and maximum.

This table is similar to the Data Processor's Information output, except that the present output is produced under the assumptions outlined above. If the original data contains columns with equal lengths, it has no missing data, and a weight column is not selected, then the numbers in two outputs will be the same.

Correlation Matrix: Zero order (Pearson) Correlation Coefficients between all possible pairs of selected columns are computed. Among other uses, this option may be helpful in choosing variables for the Regression Analysis.

Covariance Matrix: Covariances between all possible pairs of selected columns are computed. Diagonal elements are variances and off-diagonal elements are covariances.

Moment Matrix: Second moments (sum of squared differences from the mean) between all possible pairs of selected columns are computed.

Data Matrix: Check this option to send the raw data matrix to Output Medium. This is most useful when interaction terms or dummy or lag/lead variables are included in the regression model and you wonder what the program does exactly. In Stand-Alone Mode, the matrix can then be sent to Data Processor and the generated variables used as input data in other procedures.

7.1.3. Matrix Statistics Examples

Example 1

Example 20.1b on p. 422 from Zar, J. H. (2010).

Open REGRESS, select Statistics 1 → Matrix Statistics and select *temperature*, *cm*, *mm*, *min* and *ml* (*C1* to *C5*) as [Variable]s to obtain the following results:

Matrix Statistics

Valid Number of Cases: 33, 0 Omitted

Correlation Matrix

	temperature	cm	mm	min	ml
Temperature	1.0000	0.3287	0.1677	0.0519	-0.7308
Cm	0.3287	1.0000	-0.1455	0.1803	-0.2120
Mm	0.1677	-0.1455	1.0000	0.2413	-0.0554
Min	0.0519	0.1803	0.2413	1.0000	0.3127
Ml	-0.7308	-0.2120	-0.0554	0.3127	1.0000

Example 2

Open DEMODATA, select Statistics 1 → Matrix Statistics and select the following terms in the model:

- *Wages* × *Wages*
- *Energy* × *Energy*
- *Wages* × *Energy*
- Dummy(*Region*)
- Dummy(*Type*)
- Lag(C2 *Wages*);0
- Lag(C2 *Wages*);0

On the next dialogue, enter –2 and 2 for the number of lags and 2 for the Omit Level? field. In the following results, the Data Matrix output is abbreviated for space considerations.

Matrix Statistics

Valid Number of Cases: 50, 8 Omitted

Descriptive Statistics

	AVG	STD	SER	VAR
Wages × Wages	10463.2380	2475.8040	350.1316	6129605.2550
Energy × Energy	10263.4280	2774.2010	392.3313	7696191.0868
Wages × Energy	10354.7937	2601.7470	367.9426	6769087.2265
Region = 1	0.2200	0.4185	0.0592	0.1751
2	0.5000	0.5051	0.0714	0.2551
Type = 1	0.3000	0.4629	0.0655	0.2143
Lag(C2 Wages);-2	100.1280	12.6974	1.7957	161.2237
Lag(C2 Wages);2	103.0340	12.3198	1.7423	151.7778

	SUM	SSQ	MIN	MAX
Wages × Wages	523161.9000	5774318129.726	6593.4400	13924.0000
Energy × Energy	513171.3997	5644011072.654	6480.2500	14935.2841
Wages × Energy	517739.6860	5692772923.285	6536.6000	14176.3600
Region = 1	11.0000	11.0000	0.0000	1.0000
2	25.0000	25.0000	0.0000	1.0000
Type = 1	15.0000	15.0000	0.0000	1.0000
Lag(C2 Wages);-2	5006.4000	509180.7800	81.2000	116.1000
Lag(C2 Wages);2	5151.7000	538237.3700	81.3000	121.2000

Correlation Matrix

	Wages × Wages	Energy × Energy	Wages × Energy	Region = 1
Wages × Wages	1.0000	0.9607	0.9885	-0.1581
Energy × Energy	0.9607	1.0000	0.9917	-0.1349
Wages × Energy	0.9885	0.9917	1.0000	-0.1461
Region = 1	-0.1581	-0.1349	-0.1461	1.0000
2	0.0955	0.1285	0.1149	-0.5311
Type = 1	-0.0870	-0.0480	-0.0666	0.2845
Lag(C2 Wages);-2	0.9919	0.9694	0.9896	-0.1433
Lag(C2 Wages);2	0.9884	0.9140	0.9574	-0.1800

	2	Type = 1	Lag(C2 Wages);-2	Lag(C2 Wages);2
Wages × Wages	0.0955	-0.0870	0.9919	0.9884
Energy × Energy	0.1285	-0.0480	0.9694	0.9140
Wages × Energy	0.1149	-0.0666	0.9896	0.9574
Region = 1	-0.5311	0.2845	-0.1433	-0.1800
2	1.0000	-0.2182	0.1139	0.0769
Type = 1	-0.2182	1.0000	-0.0841	-0.0952
Lag(C2 Wages);-2	0.1139	-0.0841	1.0000	0.9688
Lag(C2 Wages);2	0.0769	-0.0952	0.9688	1.0000

Covariance Matrix

	Wages × Wages	Energy × Energy	Wages × Energy	Region = 1
Wages × Wages	6129605.2550	6598743.5394	6367140.5020	-163.7585
Energy × Energy	6598743.5394	7696191.0868	7157544.0234	-156.6458
Wages × Energy	6367140.5020	7157544.0234	6769087.2265	-159.0871
Region = 1	-163.7585	-156.6458	-159.0871	0.1751
2	119.4400	180.0707	151.0527	-0.1122
Type = 1	-99.6882	-61.6220	-80.1629	0.0551
Lag(C2 Wages);-2	31180.0560	34145.7844	32690.9879	-0.7614
Lag(C2 Wages);2	30147.2970	31238.5588	30688.3630	-0.9280

	2	Type = 1	Lag(C2 Wages);-2	Lag(C2 Wages);2
Wages × Wages	119.4400	-99.6882	31180.0560	30147.2970
Energy × Energy	180.0707	-61.6220	34145.7844	31238.5588
Wages × Energy	151.0527	-80.1629	32690.9879	30688.3630
Region = 1	-0.1122	0.0551	-0.7614	-0.9280
2	0.2551	-0.0510	0.7306	0.4786
Type = 1	-0.0510	0.2143	-0.4943	-0.5431
Lag(C2 Wages);-2	0.7306	-0.4943	161.2237	151.5460
Lag(C2 Wages);2	0.4786	-0.5431	151.5460	151.7778

Moment Matrix

	Wages × Wages	Energy × Energy	Wages × Energy	Region = 1	2	Type = 1	Lag(C2 Wages);-2	Lag(C2 Wages);2
Wages × Wages	115486362.5945	113855458.4657	114584468.8253	2141.4290	5348.6702	3041.2770	1078219.5494	1107613.6152
Energy × Energy	113855458.4657	112880221.4531	113290072.8809	2104.4413	5308.1833	3018.6389	1061119.3869	1088095.8276
Wages × Energy	114584468.8253	113290072.8809	113855458.4657	2122.1492	5325.4285	3027.8784	1068841.9538	1096970.4119
Region = 1	2141.4290	2104.4413	2122.1492	0.2200	0.0000	0.1200	21.2820	21.7580
2	5348.6702	5308.1833	5325.4285	0.0000	0.5000	0.1000	50.7800	51.9860
Type = 1	3041.2770	3018.6389	3027.8784	0.1200	0.1000	0.3000	29.5540	30.3780
Lag(C2 Wages);-2	1078219.5494	1061119.3869	1068841.9538	21.2820	50.7800	29.5540	10183.6156	10465.1034
Lag(C2 Wages);2	1107613.6152	1088095.8276	1096970.4119	21.7580	51.9860	30.3780	10465.1034	10764.7474

Data Matrix

	Wages × Wages	Energy × Energy	Wages × Energy	Region = 1	2	Type = 1	Lag(C2 Wages);-2	Lag(C2 Wages);2
1	14689.4400	14234.8761	14460.3720	0.0000	0.0000	0.0000	118.0000	*
2	14256.3600	14089.6900	14172.7800	1.0000	0.0000	0.0000	116.5000	*
3	13924.0000	13667.9481	13795.3800	0.0000	1.0000	0.0000	115.4000	121.2000
4	13572.2500	13218.1009	13394.0050	0.0000	0.0000	1.0000	114.9000	119.4000
5	13317.1600	12210.2500	12751.7000	1.0000	0.0000	0.0000	114.8000	118.0000
…	…	…	…	…	…	…	…	…

7.2. Regression Analysis

The Regression Analysis is used to estimate the coefficients B_0, ..., B_m of the equation:

$$Y = B_0 + B_1X_1 + ... + B_mX_m$$

given n observations on m independent variables X_1, ..., X_m and a dependent variable Y. The Stepwise Regression procedure also determines a subset of the selected variables which contribute significantly to the explanation of variation in the dependent variable.

It is possible to select any numeric column of data as the dependent variable and to select the columns to be included in the analysis as independent variables. A Regression Analysis can be performed by selecting one column as the dependent variable and at least one column as an independent variable. The program will not proceed unless this requirement is met. Regressions can also be run on a sub-set of cases as determined by a combination of factor columns. The Polynomial Regression procedure allows the choice of one independent variable, but will also require the degree of the polynomial to be entered.

The Variable Selection Dialogue contains a check box to include the constant term (or the intercept) in the analysis. The default is regression with constant as in the above equation. If this box is unchecked then the following equation without a constant term will be estimated:

$$Y = B_1X_1 + ... + B_mX_m.$$

An important feature of regression models without a constant term is that the method they employ for calculation of R-squared and adjusted R-squared values is fundamentally different from that of regression with a constant term. Therefore, R-squared values calculated for regressions with and without a constant term are not comparable.

The standard method of calculating the R-squared value for regressions including a constant term can be expressed as:

R-squared = 1 – Var(Residuals) / Var(Dependent)

where Var() stands for variance. However, this definition fails completely when the constant term is omitted from the model. A better definition, which applies to both types of regression, can be made by reference to the ANOVA of Regression table, where Ssq() stands for sum of squares:

R-squared = Ssq(Regression) / Ssq(Total)

There is also a slight difference between Linear Regression and Polynomial Regression on one hand and Stepwise Regression, Analysis of Variance and General Linear Model procedures on the other, in the way they handle the degrees of freedom in regressions without a constant. In line with the most common approach in the literature, we here also calculate the degrees of freedom as (n - m, m) in Stepwise Regression, Analysis of Variance and General Linear Model procedures and (n - m, m - 1) in the Linear Regression and Polynomial Regression procedures.

Also, although both groups of procedures operate in double precision, there may be a slight difference between their estimates on the same set of data. The reason for this is that two completely different algorithms are used in each case: the Linear Regression and Polynomial Regression procedures are based on the square root free version of the Cholesky decomposition originally suggested by Gentleman (1974, Applied Statistics, 23, pp. 448-454), whereas the Stepwise Regression, Analysis of Variance and General Linear Model procedures are based on the SWEEP algorithm by Jennrich (in Statistical Methods for Digital Computers, ed. Enslein, Ralston, Wilf, 1977, Wiley, pp. 58-75). The first algorithm is more accurate but the second is more suitable for Stepwise Regression and Analysis of Variance.

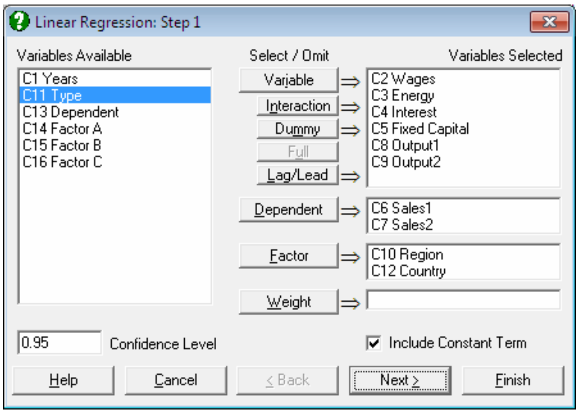
7.2.1. Linear Regression

The Linear Regression procedure is suitable for estimating weighted or nonweighted linear regression models with or without a constant term, including nonlinear models such as multiplicative, exponential or reciprocal regressions that can be linearised by logarithmic or exponential transformations. It is possible to run regressions without an independent variable, which is equivalent to running a noconstant regression against a unity independent variable.

Regressions may be run on a subset of cases as determined by the combination of levels of an unlimited number of factor columns. An unlimited number of dependent variables can be selected to run the same model on different dependent variables. It is also possible to include interaction terms, dummy and lag/lead variables in the model, without having to create them as data columns in the spreadsheet first.

The output includes a wide range of graphics and statistics options. In regression plots, it is possible to omit outliers interactively, by pressing down the right mouse button on a data point and then pressing <Delete>.

7.2.1.1. Linear Regression Variable Selection



Variables Available: C1 Years, C11 Type, C13 Dependent, C14 Factor A, C15 Factor B, C16 Factor C

Variables Selected: C2 Wages, C3 Energy, C4 Interest, C5 Fixed Capital, C8 Output1, C9 Output2

Dependent: C6 Sales1, C7 Sales2

Factor: C10 Region, C12 Country

Weight:

0.95 Confidence Level

☑ Include Constant Term

All columns selected for this procedure should have an equal size. The buttons on the Variable Selection Dialogue have the following tasks:

Variable: Click on [Variable] to select a column containing continuous numeric data as an independent variable.

Interaction: This button is used to create independent variables, which are the products of existing numeric variables. If only one variable is highlighted, then the new independent variable created will be the product of the selected variable by itself. If two or more variables are highlighted, then the new term will be the product of these variables. Maximum three-way interactions are allowed. Interactions of dummy variables or lags are not allowed. In order to create interaction terms for dummy variables, create interactions first, and then create dummy variables for them. For further information see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables.

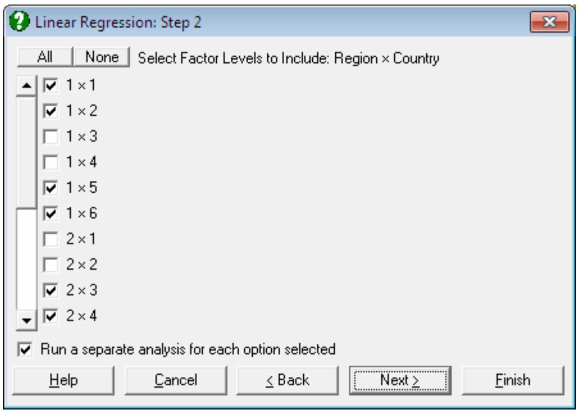
Dummy: This button is used to create n or n - 1 new independent (dummy or indicator) variables for a factor column containing n levels. Each dummy variable corresponds to a level of the factor column. A case in a dummy column will have the value of 1 if the factor contains the corresponding level in the same row, and 0 otherwise. If the selected variable is an interaction term, then dummy variables will be created for this interaction term. Up to three-way interactions are allowed and columns containing short or Long Strings can be selected as factors. It is possible to include all n levels or to omit the first or the last level in order to remove the inherent over-parameterisation of the model. For further information see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables.

Full: This button becomes activated when two or more categorical variables are highlighted. Like the [Dummy] button, it is also used to create dummy variables. The only difference is that this button will create all necessary dummy variables and their interactions to specify a complete model. For instance, if two categorical variables are highlighted, this button will create two sets of dummy variables representing the main effects and a third set representing the interaction term between the two factors. Maximum three-way interactions are supported.

Lag/Lead: This button is used to create new independent variables by shifting the rows of an existing variable up or down. When a lag variable is specified in the Variable Selection Dialogue, then a further dialogue will ask for the number of lags (or leads) for each item selected. Negative integers represent the lags and positive integers the leads. For further information see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables.

Dependent: It is compulsory to select at least one column containing numeric data as a dependent variable. When more than one dependent variable is selected, the analysis will be repeated as many times as the number of dependent variables, each time only changing the dependent variable and keeping the rest of selections unchanged.

Factor: This allows you to run regressions on subsamples of rows (cases). With some time series or panel data it is desirable to run regressions on some, rather than all rows of the data matrix. Although it is possible to extract subsets using a Data Processor function such as **If()** (see 3.4.2.7. Conditional Functions), Data → Recode Column, Subsample Save (see 2.4.1.6.3. Options) or Data → Select Row, it is much more convenient to use the selection facility provided here.



Select Factor Levels to Include: Region × Country

☑ 1 × 1
☑ 1 × 2
□ 1 × 3
□ 1 × 4
☑ 1 × 5
☑ 1 × 6
□ 2 × 1
□ 2 × 2
☑ 2 × 3
☑ 2 × 4

☑ Run a separate analysis for each option selected

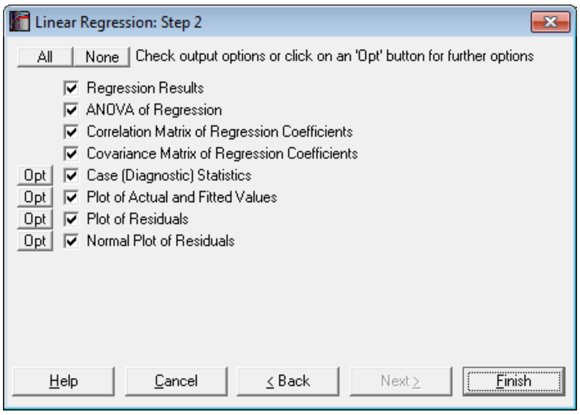
To make use of this facility the data matrix should contain at least one factor column. An unlimited number of factors can be selected. These can be numeric or String Data columns, but each column must contain a limited number of distinct values. Select the factor columns from the Variables Available list by clicking on [Factor]. Then the program will display a dialogue where all possible combinations of factor levels are displayed in a list of check boxes. For instance, if one factor containing three levels is selected, only three check boxes will be displayed representing each level. Only the rows of data matrix corresponding to the checked levels will be included in the regression. If there are two factors selected, say one having two levels and the other three, then the list will contain six check boxes, 1x1, 1x2, 1x3, 2x1, 2x2, 2x3. Suppose the check boxes 1x2 and 2x2 are checked. Then only those rows of the data matrix containing 1 in the first factor column and 2 in the second and 2 in the first factor and 2 in the second will be included in the regression. When more than one selection is made, it is possible to run a single Regression Analysis on all selected rows combined, or to run a separate analysis for each selection.

With subsample selection, the possibility of getting an Insufficient degrees of freedom message will increase considerably. Also, interpretation of the Durbin-Watson statistic may not be obvious.

Weight: It was mentioned at the beginning of this chapter that when a column is selected as a *weights* variable, the program will normalise this column so that its sum is equal to the number of valid cases, and then multiply each independent variable by its square root. In a weighted regression run with constant term included, the column of 1s in the X matrix should also be multiplied by the square root of weights, as the Regression Analysis considers the constant term just like any other coefficient. The algorithm used here produces exactly the same effect without having to take the square root of weights, thus achieving higher accuracy.

7.2.1.2. Linear Regression Output Options

An Output Options Dialogue will provide access to the following options. The Actual and Fitted Values, Residuals and Confidence Intervals for Mean and Actual Y Values options that existed in earlier version of UNISTAT have now been merged under the Case (Diagnostic) Statistics option. See 7.2.1.2.2. Linear Regression Case Output.


Linear Regression: Step 2

All | None | Check output options or click on an 'Opt' button for further options

☑ Regression Results
☑ ANOVA of Regression
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
Opt ☑ Case (Diagnostic) Statistics
Opt ☑ Plot of Actual and Fitted Values
Opt ☑ Plot of Residuals
Opt ☑ Normal Plot of Residuals

Multicollinearity: One of the basic assumptions of the method of least squares estimation is that no linear dependency exists between the regression variables. However, the present implementation will detect the variables causing multicollinearity and display them at the end of coefficients table. If you do not wish to display these variables enter the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
DispCollin=0
```

The rest of the coefficients will be determined as if the regression were run without the variables causing collinearity.

Perfect Fit: If the current set of independent variables fully explain the variation of the dependent variable, the program displays a restricted number of results options, main results themselves being confined to estimated regression coefficients only.

Perfect fit will be reported under the following two circumstances;

1) if R-squared > 0.99999 or
2) if the number of variables (including the dependent variable) is equal to the number of observations, in which case R-squared is not computed.

7.2.1.2.1. Linear Regression Coefficient Output

Regression Results: The main regression output displays a table for coefficients of the estimated regression equation, their standard errors, t-statistics, probability values (from the t-distribution) and confidence intervals for the significance level specified in the regression Variable Selection Dialogue. If any independent variables have been omitted due to multicollinearity, they are reported at the end of the table. If you do not wish to display these variables enter the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
DispCollin=0
```

When the model contains dummy variables with very long string values, output may look cramped. You can display the numbers of levels, instead of their string values, by entering the following line in the [Options] section of *Unistat65.ini*:

```
OLSFullLabel=0
```

In Stand-Alone Mode, the estimated regression coefficients can be saved to data matrix by clicking on the UNISTAT icon (which becomes visible on the right of the toolbar in Output Window and Data Processor after running a procedure). The same coefficients will also be saved automatically in the file POLYCOEF.TXT, in the order of *C0* (constant term, if any), *C1, C2, ..., Cr*.

The rest of the output consists of the following statistics: residual sum of squares, standard error of regression, mean and standard deviation of the dependent variable, R-squared, R-squared adjusted for the degrees of freedom, F-statistic and its tail probability, the Durbin-Watson statistic, press statistic and the log of likelihood function. The Durbin-Watson statistic will be adjusted for the number of gaps in data caused by missing values. The number of rows omitted due to missing values is reported if it is other than zero.

ANOVA of Regression: The total variation of the dependent variable is partitioned into the *Regression* (or explained) part which is due to the linear influence of independent variables and the *Error* (or unexplained) part which is expressed in residuals.

The F-value is the ratio of the mean squares for regression and mean squares for the error term. The null hypothesis of "no relationship between the dependent variable and independent variables as a whole" can be tested by means of the probability value reported.

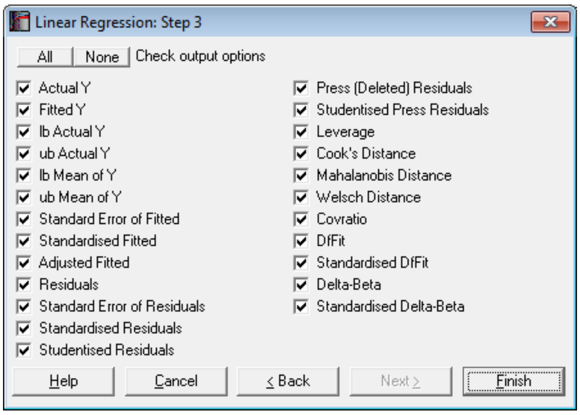
Individual contributions of independent variables to the regression (explained) sum of squares are also displayed. The sum of individual contributions will be equal to the regression sum of squares. However, it is important here to emphasise that these individual contributions are specific to the order in which independent variables enter into the regression equation. Normally, when two independent variables change place, their individual contributions to the regression sum of squares will also change.

Correlation Matrix of Regression Coefficients: This is a symmetric matrix with unity diagonal elements. It gives the correlation between the regression coefficients and is obtained by dividing the elements of $(X'X)^{-1}$ matrix by the square root of the diagonal elements corresponding to its row and column.

Covariance Matrix of Regression Coefficients: This option displays a symmetric matrix where diagonal elements are the variances and off-diagonal elements are the covariances of the estimated regression coefficients. This matrix is sometimes referred to as the dispersion matrix and it can also be obtained by multiplying $(X'X)^{-1}$ matrix by the estimated variance of the error terms.

7.2.1.2.2. Linear Regression Case Output

Case (Diagnostic) Statistics: These statistics are useful in determining the influence of individual observations on the overall fit of the model. Looking at *outliers* (cases with a large residual value) is an effective way of determining whether the model fitted explains well the variation in data. However, residuals alone cannot explain all types of influence of individual cases on the regression. Suppose, for instance, a data set where most observations are clustered together but only one point lies outside the cluster. Suppose also that the regression line passes near this point so that it does not have a large residual. Nevertheless, removing this single point from the regression may have substantial effects on the estimated coefficients (called *leverage*).



☑ Actual Y
☑ Fitted Y
☑ lb Actual Y
☑ ub Actual Y
☑ lb Mean of Y
☑ ub Mean of Y
☑ Standard Error of Fitted
☑ Standardised Fitted
☑ Adjusted Fitted
☑ Residuals
☑ Standard Error of Residuals
☑ Standardised Residuals
☑ Studentised Residuals
☑ Press (Deleted) Residuals
☑ Studentised Press Residuals
☑ Leverage
☑ Cook's Distance
☑ Mahalanobis Distance
☑ Welsch Distance
☑ Covratio
☑ DfFit
☑ Standardised DfFit
☑ Delta-Beta
☑ Standardised Delta-Beta

Most of the regression diagnostic statistics below measure such effects, which answer the question *what would happen if this case was removed from the regression.* Luckily, we do not need to estimate the entire model after deleting each case, but compute the same results by applying the following algebraic manipulations.

Let:

n = valid number of cases and
m = number of coefficients in the model.

Therefore:

m = 1 + number of independent variables (with constant) and
m = number of independent variables (with no constant).

Also, each row of the data matrix is defined as:

$X_i = 1+X_{1i},\ldots,X_{m-1i}$ for regressions with constant and
$X_i = X_{1i},\ldots,X_{mi}$ for regressions without constant.

Central to most diagnostic statistics are definitions of root mean square of residuals:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n-m}}$$

and the diagonal vector of the projection matrix:

$$h_i = X_i(X^{'}X)^{-1}X_i^{'}$$

Predictions (Interpolations): There are three conditions under which predictions will be computed for estimated Y values:

1) If, for a case, all independent variables are non-missing, but only the dependent variable is missing,
2) If a case does not contain missing values but it has been omitted from the analysis by Data Processor’s Data → Select Row function,
3) If a case does not contain missing values but it has been omitted from the analysis by selecting subsamples from the Variable Selection Dialogue (see 2.1.2. Categorical Data Analysis).

Such cases are not included in the estimation of the model. When, however, Case (Diagnostic) Statistics option is selected, the program will detect these cases and compute and display the fitted (estimated) Y values, as well as their confidence intervals and some other related statistics. Therefore, it will be a good idea to include the cases for which predictions are to be made in the data matrix during the data preparation phase, and then exclude them from the analysis by one of the above three methods. When a case is predicted, its label will be prefixed by an asterisk (*).

In Stand-Alone Mode, the spreadsheet function **Reg** can also be used to make predictions (see 3.4.2.6.3. UNISTAT Functions).

Statistics available under Case (Diagnostic) Statistics option are as follows.

Actual Y: Observed values of the dependent variable.

$$Y_i$$

Fitted Y: Estimated values of the dependent variable.

$$\hat{Y}_i = \sum_{k=1}^{m} b_k X_{ki}$$

Confidence Intervals for Actual Y-values:

$$X\beta \pm t_{\alpha/2} S\sqrt{1+X(X'X)^{-1}X'}$$

where X is the given vector of independent variable values, ß is the vector of estimated coefficients, $t_{\alpha/2}$ is the critical value from the t-distribution for an α / 2 level of significance and n - k degrees of freedom and S is the standard error of prediction. Any significance level can be entered from the Variable Selection Dialogue. The default is 95%.

The confidence intervals for bivariate regressions can also be plotted as an option on X-Y line plots (see 4.1.1.1.1. Line).

Confidence Intervals for Mean of Y:

$$X\beta \pm t_{\alpha/2} S\sqrt{X(X'X)^{-1}X'}$$

Standard Error of Fitted:

$$S_{\hat{Y}} = S\sqrt{h_i}$$

Standardised Fitted:

$$(\hat{Y}_i - \overline{\hat{Y}}) / \sigma_{\hat{Y}}$$

Adjusted Fitted:

Y_i – Press Residual_i

Residuals:

$$e_i = \hat{Y}_i - Y_i$$

Standard Error of Residuals:

$$S\sqrt{1-h_i}$$

Standardised Residuals:

$$e_i / S$$

Studentised (Jackknife) Residuals:

$$(e_i / S) / \sqrt{1 - h_i}$$

Press (Deleted) Residuals:

$$PR_i = e_i / (1 - h_i)$$

Press (deleted) residuals are defined as the change in a residual when this case is omitted from the analysis. An estimate can be computed from the above formula, without having to run n regressions.

Studentised Press Residuals:

$$PR_i / S_{(i)}$$

where:

$$S_{(i)} = \frac{1}{n - m - 1} \sqrt{\frac{(n - m)S^2}{1 - h_i} - PR_i^2}$$

Leverage:

h_i for regression without a constant term and

$h_i - \frac{1}{n}$ for regression with a constant term.

Cook's Distance:

$$(PR_i^2 h_i) / (mS^2)$$

Mahalanobis Distance:

nh_i for regression without a constant term and

$\left(h_i - \frac{1}{n}\right)(n - 1)$ for regression with a constant term.

Welsch Distance:

$$DfFit_i \sqrt{(n - 1) / (1 - h_i)}$$

Covratio:

$$(S_{(i)}/S)^{2m}/(1-h_i)$$

DfFit:

$$(h_i e_i)/(1-h_i)$$

Standardised DfFit:

$$\mathrm{DfFit}_i/\left(S_{(i)}\sqrt{h_i}\right)$$

Delta-Beta$_j$:

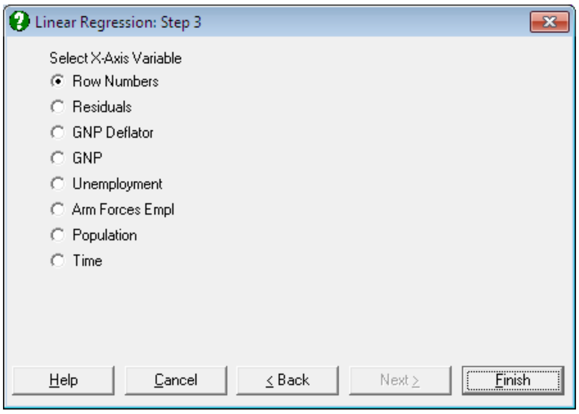
$$\mathrm{PR}_i(X'X)^{-1}X_i^j,\ \text{for } j=1,\ldots,m$$

Delta-beta is defined as the change in an estimated coefficient when a case is omitted from the analysis. Like in press residuals, an estimate can be computed from the above formula, without having to run n regressions.

Standardised Delta-Beta:

$$\mathrm{DfBeta}_j/\left(S_{(i)}\sqrt{(X'X)_j^{-1}}\right),\ \text{for } j=1,\ldots,m$$

Plot of Actual and Fitted Values: Select this option to plot actual and fitted Y values and their confidence intervals against row numbers (index), residuals or against any independent variable. A further dialogue will enable you to choose the X-axis variable from a list containing Row Numbers, Residuals and all independent variables.



Linear Regression: Step 3

Select X-Axis Variable

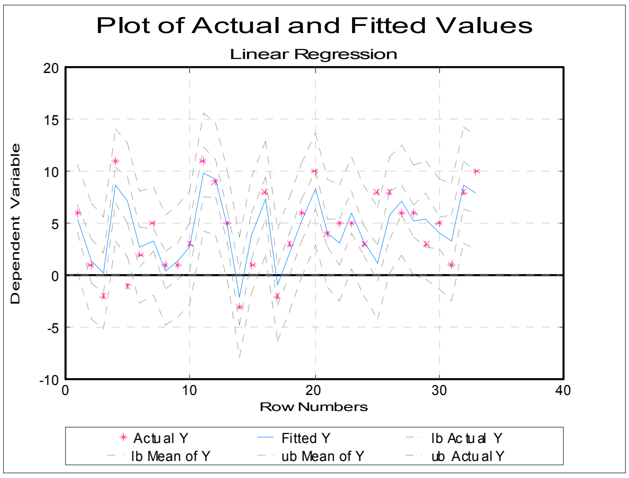
● Row Numbers
○ Residuals
○ GNP Deflator
○ GNP
○ Unemployment
○ Arm Forces Empl
○ Population
○ Time

By default, a line graph of the two series is plotted. However, since this procedure (like the plot of residuals) uses the X-Y Plots engine, it has almost all controls and options available for X-Y Plots, except for error bars and right Y-axes. This means that, as well as being able to edit all aspects of the graph, you can connect data points with lines, curves or display confidence intervals.

The data points on the graph will also respond to the right mouse button in the way X-Y Plots does; the point is highlighted, a panel displays information about the point and in Stand-Alone Mode, the row of the spreadsheet containing the data point is also highlighted (a procedure which is also known as *Brushing* or *Point identification*). While the point is highlighted you can press <Delete> to omit the particular row containing the point. The entire Regression Analysis will be run again without the deleted row. If you want to restore the original regression, you will need to take one of the following two actions depending on the way you run UNISTAT:

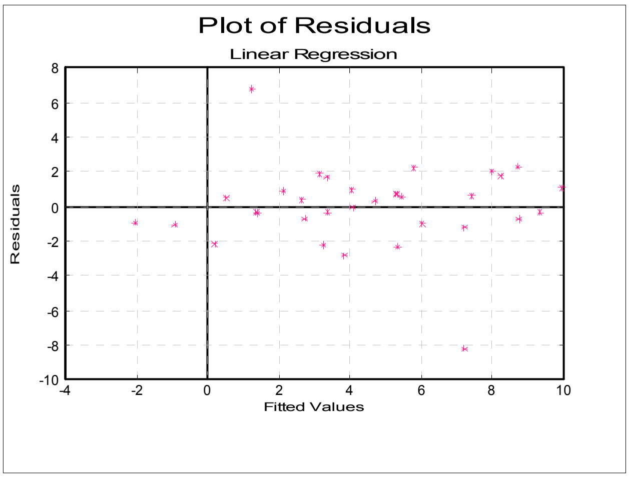
1. In Stand-Alone Mode, go back to Data Processor and delete or deactivate the Select Row column created by the program.

2. In Excel Add-In Mode, highlight a different block of data to remove the effect of the internal Select Row column.

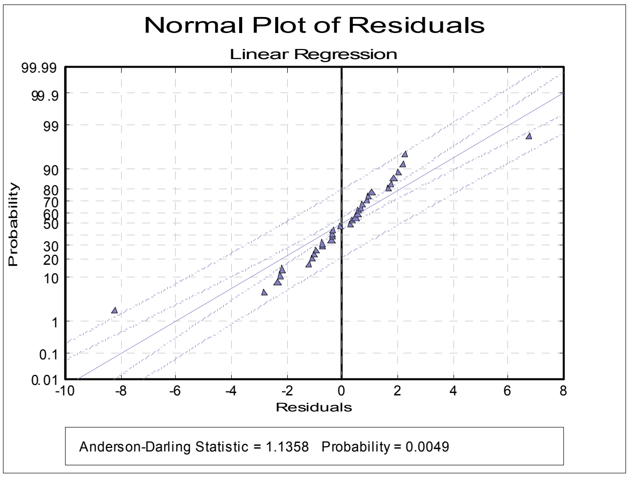


Plot of Residuals: Residuals can be plotted against row numbers (index), fitted values or against any independent variable. A further dialogue will enable you to choose the X-axis variable from a list containing Row Numbers, Fitted Values and all independent variables.

By default a scatter graph of residuals is plotted. For more information on available options see Plot of Actual and Fitted Values above.



Normal Plot of Residuals: Residuals are plotted against the normal probability (probit) axis. For more information on available options see Plot of Actual and Fitted Values above (also see 5.3.2. Normal Probability Plot).



7.2.1.3. Linear Regression Examples

Example 1

Table 5.1 on p. 134 from Tabachnick, B. G. & L. S. Fidell (1989).

Open REGRESS, select Statistics 1 → Regression Analysis → Linear Regression and select *Motiv*, *Qual* and *Grade* (*C6* to *C8*) as [Variable]s and *Compr* (*C9*) as [Dependent]. Select all output options to obtain the following results:

Linear Regression

Dependent Variable: Compr
Valid Number of Cases: 6, 0 Omitted

Regression Results

	Coefficient	Standard Error	t-Statistic	Probability	Lower 95%	Upper 95%
Constant	-4.7218	9.0656	-0.5208	0.6544	-43.7281	34.2845
Motiv	0.6583	0.8721	0.7548	0.5292	-3.0942	4.4107
Qual	0.2720	0.5891	0.4618	0.6896	-2.2627	2.8068
Grade	0.4160	0.6462	0.6438	0.5857	-2.3643	3.1964

Residual Sum of Squares =	30.3599
Standard Error =	3.8961
Mean of Y =	10.0000
Stand Dev of y =	4.5166
Correlation Coefficient =	0.8381
R-squared =	0.7024
Adjusted R-squared =	0.2559
F(3,2) =	1.5731
Probability of F =	0.4114
Durbin-Watson Statistic =	1.7838
log of likelihood =	-14.6736
Press Statistic =	661.8681

ANOVA of Regression

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Motiv	35.042	1	35.042	2.308	0.2680
Qual	30.306	1	30.306	1.996	0.2932
Grade	6.292	1	6.292	0.415	0.5857
Regression	71.640	3	23.880	1.573	0.4114
Error	30.360	2	15.180		
Total	102.000	5	20.400	1.344	0.4787

Correlation Matrix of Regression Coefficients

	Constant	Motiv	Qual	Grade
Constant	1.0000	-0.8485	0.1286	-0.1935
Motiv	-0.8485	1.0000	-0.1768	-0.1151
Qual	0.1286	-0.1768	1.0000	-0.7455
Grade	-0.1935	-0.1151	-0.7455	1.0000

Covariance Matrix of Regression Coefficients

	Constant	Motiv	Qual	Grade
Constant	82.1859	-6.7083	0.6870	-1.1338
Motiv	-6.7083	0.7606	-0.0908	-0.0649
Qual	0.6870	-0.0908	0.3471	-0.2838
Grade	-1.1338	-0.0649	-0.2838	0.4176

Case (Diagnostic) Statistics

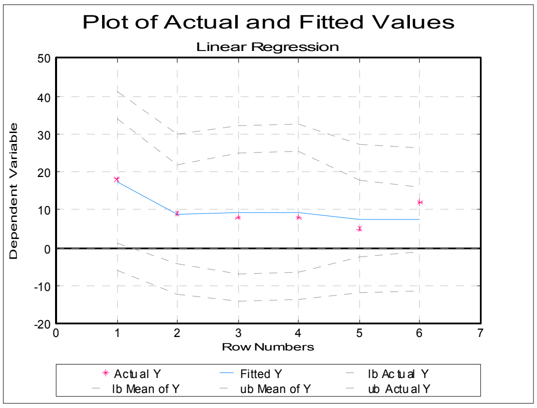
	Actual Y	Fitted Y	95% lb Actual Y	95% ub Actual Y	95% lb Mean of Y	95% ub Mean of Y	Standard Error of Fitted
1	18.0000	17.5675	-5.9769	41.1119	1.0353	34.0997	3.8423
2	9.0000	8.8399	-12.3120	29.9918	-4.0589	21.7387	2.9979
3	8.0000	9.0893	-13.9745	32.1531	-6.7510	24.9296	3.6815
4	8.0000	9.3562	-13.8431	32.5555	-6.6807	25.3932	3.7272
5	5.0000	7.6376	-11.9530	27.2281	-2.4997	17.7749	2.3561
6	12.0000	7.5095	-11.3204	26.3394	-1.0661	16.0851	1.9931

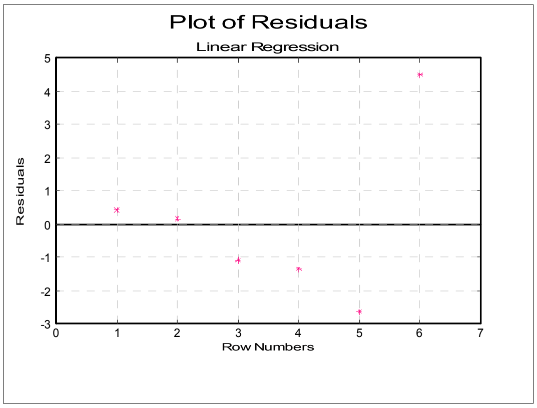
	Standard ised Fitted	Adjusted Fitted	Residuals	Standard Error of Residuals	Standardised Residuals	Studentised Residuals
1	1.9992	2.2363	0.4325	0.6454	0.1110	0.6702
2	-0.3065	8.6076	0.1601	2.4885	0.0411	0.0643
3	-0.2406	18.1667	-1.0893	1.2753	-0.2796	-0.8541
4	-0.1701	23.9868	-1.3562	1.1348	-0.3481	-1.1951
5	-0.6241	9.1581	-2.6376	3.1031	-0.6770	-0.8500
6	-0.6580	5.9179	4.4905	3.3478	1.1525	1.3413

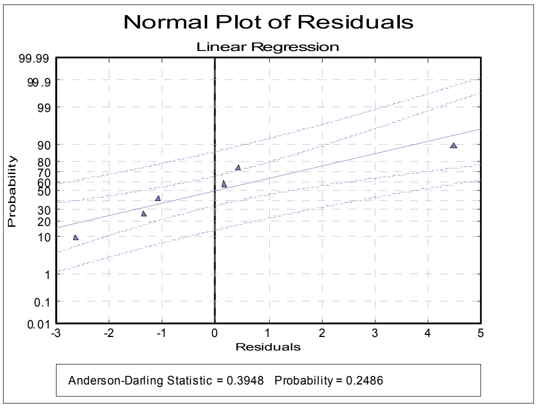
	Press (Deleted) Residuals	Studentised Press Residuals	Leverage	Cook's Distance	Mahalanobis Distance	Welsch Distance
1	15.7637	0.5382	0.8059	3.9802	4.0295	43.2522
2	0.3924	0.0455	0.4254	0.0015	2.1269	0.1920
3	-10.1667	-0.7578	0.7262	1.5199	3.6310	-14.9437
4	-15.9868	-1.5807	0.7485	3.8520	3.7425	-39.8563
5	-4.1581	-0.7520	0.1990	0.1041	0.9951	-1.6031
6	6.0821	2.9934	0.0950	0.1594	0.4751	4.6377

	Covratio	DfFit	Standardised DfFit	Delta-Beta Constant	Delta-Beta Motiv	Delta-Beta Qual
1	210.8306	15.3312	3.2040	-20.9957	1.0201	0.6412
2	38.8964	0.2323	0.0549	0.1779	0.0036	0.0319
3	24.3153	-9.0774	-2.1875	-13.0714	1.6339	0.3631
4	1.2590	-14.6305	-5.1916	13.6116	-2.1874	1.6812
5	4.1991	-1.5205	-0.5710	-3.3462	0.1201	-0.1731
6	0.0022	1.5916	1.7821	3.8897	-0.1289	-0.1386

	Delta-Beta Grade	Standardised Delta-Beta Constant	Standardised Delta-Beta Motiv	Standardised Delta-Beta Qual	Standardised Delta-Beta Grade
1	0.5191	-1.8597	0.9392	0.8739	0.6451
2	-0.0421	0.0139	0.0029	0.0384	-0.0461
3	-0.9077	-1.2792	1.6622	0.5468	-1.2463
4	-0.8213	1.9858	-3.3171	3.7744	-1.6809
5	0.2728	-0.3266	0.1218	-0.2600	0.3735
6	-0.0043	0.9575	-0.3298	-0.5251	-0.0149







Example 2

Table 4.3.1 on p. 296 from Elliot, M. A., J. S. Reisch, N. P. Campbell (1989). This data set is known as Longley's data and it is particularly sensitive to rounding-off errors and the regression algorithm used.

Open REGRESS, select Statistics 1 → Regression Analysis → Linear Regression and select *GNP Deflator*, *GNP*, *Unemployment*, *Arm Forces Empl*, *Population* and *Time* (*C10* to *C15*) as [Variable]s and *Total* (*C16*) as [Dependent]. Select only the Regression Results output option to obtain the following:

Linear Regression

Dependent Variable: Total
Valid Number of Cases: 16, 0 Omitted

Regression Results

	Coefficient	Standard Error	t-Statistic
Constant	-3482258.634596	890420.3836074	-3.910802918154
GNP Deflator	15.06187227137	84.91492577477	0.17737602823
GNP	-0.0358191792926	0.0334910077722	-1.069516317221
Unemployment	-2.020229803817	0.488399681652	-4.136427355941
Arm Forces Empl	-1.033226867174	0.214274163162	-4.821985310445
Population	-0.0511041056536	0.226073200069	-0.226051144664
Time	1829.151464614	455.4784991422	4.01588981271

	Probability	Lower 95%	Upper 95%
Constant	0.0036	-5496534.8253	-1467982.4439
GNP Deflator	0.8631	-177.0295	207.1533
GNP	0.3127	-0.1116	0.0399
Unemployment	0.0025	-3.1251	-0.9154
Arm Forces Empl	0.0009	-1.5179	-0.5485
Population	0.8262	-0.5625	0.4603
Time	0.0030	798.7848	2859.5181

Residual Sum of Squares =	836424.0555059
Standard Error =	304.854073562
Mean of Y =	65317
Stand Dev of y =	3511.96835597
Correlation Coefficient =	0.997736941572
R-squared =	0.995479004577
Adjusted R-squared =	0.992465007629
F(6,9) =	330.2853392354
Probability of F =	0.0000
Durbin-Watson Statistic =	2.559487689283
log of likelihood =	-110.7203479964
Press Statistic =	2886892.562947

WARNING! *Table 4.3.1 contains a misprint in row 13 of the (X5) variable. The above results have been obtained by using the correct value of 123366.*

Example 3

Example 20.1c on p. 426 from Zar, J. H. (2010).

Open REGRESS, select Statistics 1 → Regression Analysis → Linear Regression and select *temperature*, *cm*, *mm* and *min* (*C1* to *C4*) as [Variable]s and *ml* (*C5*) as [Dependent]. Select only the Regression Results and ANOVA of Regression output options to obtain the following results:

Linear Regression

Dependent Variable: ml
Valid Number of Cases: 33, 0 Omitted

Regression Results

	Coefficient	Standard Error	t-Statistic	Probability	Lower 95%	Upper 95%
Constant	2.9583	1.3636	2.1695	0.0387	0.1651	5.7515
temperature	-0.1293	0.0213	-6.0751	0.0000	-0.1729	-0.0857
cm	-0.0188	0.0563	-0.3338	0.7410	-0.1341	0.0965
mm	-0.0462	0.2073	-0.2230	0.8252	-0.4708	0.3784
min	0.2088	0.0670	3.1141	0.0042	0.0714	0.3461

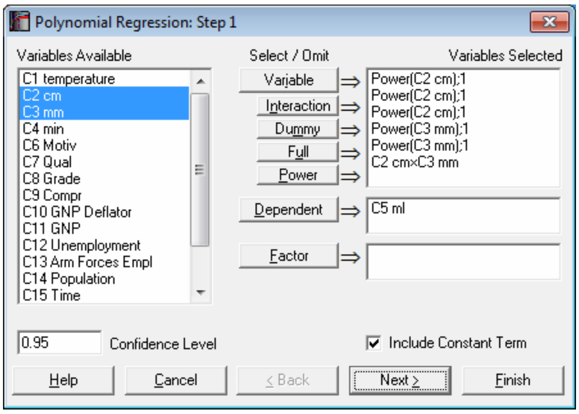
Residual Sum of Squares =	5.0299
Standard Error =	0.4238
Mean of Y =	2.4742
Stand Dev of y =	0.6789
Correlation Coefficient =	0.8117
R-squared =	0.6589
Adjusted R-squared =	0.6102
F(4,28) =	13.5235
Probability of F =	0.0000
Durbin-Watson Statistic =	1.9947
log of likelihood =	-15.9976
Press Statistic =	7.1248

ANOVA of Regression

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
temperature	7.876	1	7.876	43.845	0.0000
cm	0.013	1	0.013	0.073	0.7888
mm	0.086	1	0.086	0.478	0.4950
min	1.742	1	1.742	9.698	0.0042
Regression	9.717	4	2.429	13.524	0.0000
Error	5.030	28	0.180		
Total	14.747	32	0.461	2.565	0.0066

7.2.2. Polynomial Regression

Polynomials can be fitted on multivariate data. There are no restrictions on the degree of polynomials, but you need to remember that with high degree polynomials number overflow problems may occur.


Polynomial Regression: Step 1

Variables Available
C1 temperature
C2 cm
C3 mm
C4 min
C6 Motiv
C7 Qual
C8 Grade
C9 Compr
C10 GNP Deflator
C11 GNP
C12 Unemployment
C13 Arm Forces Empl
C14 Population
C15 Time

Variables Selected
Power(C2 cm);1
Power(C2 cm);1
Power(C2 cm);1
Power(C3 mm);1
Power(C3 mm);1
C2 cm×C3 mm

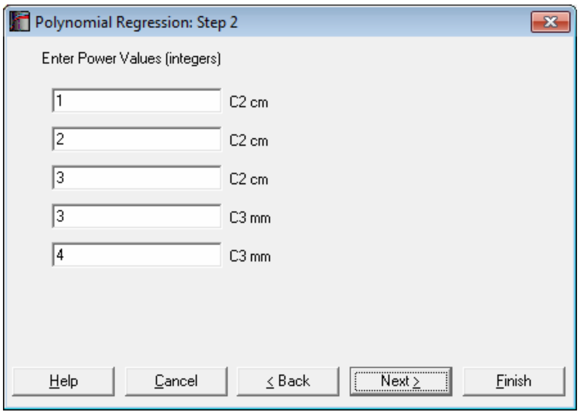
Dependent: C5 ml

Factor:

0.95 Confidence Level

☑ Include Constant Term

As in Linear Regression, it is possible to create interaction terms, dummy variables, select multiple dependent variables and run regressions on subsamples defined by several factor columns (see 7.2.1.1. Linear Regression Variable Selection). However, here, the Lag/Lead button is replaced with a Power button. Once you highlight a variable from the Variables Available list, you can create as many power terms for that variable as you wish, by clicking on the Power button. The power terms are created with a degree one. You can assign the desired powers on the next dialogue.



Polynomial Regression: Step 2

Enter Power Values (integers)

1 C2 cm

2 C2 cm

3 C2 cm

3 C3 mm

4 C3 mm

All output options available for the Linear Regression procedure will be available.

Fitted values for the Polynomial Regression are extremely sensitive to slight changes in coefficients. Therefore, use of the truncated coefficient values from the formatted output (as in text, Word or HTML) display is not recommended in reconstructing a fitted polynomial equation. To run predictions, you are advised to use the full precision Excel output or one of the two methods explained in section 7.2.1.2.2. Linear Regression Case Output. These functions use the full 16-digit precision of the estimated coefficients. The estimated coefficients will also be saved in full precision automatically in the file POLYCOEF.TXT, in the order they appear in the Regression Results output option.

Example 1

Open REGRESS, select Statistics 1 → Regression Analysis → Polynomial Regression. Highlight *cm* (*C2*) and click the Power button three times. Next, highlight *mm* (*C3*) and click the Power button twice. Then highlight both *cm* (*C2*) and *mm* (*C3*) and click the Interaction button once. Select *ml* (*C5*) as [Dependent]. On the Output Options Dialogue check only the Regression Results option.

Polynomial Regression

Dependent Variable: ml
Valid Number of Cases: 33, 0 Omitted

Regression Results

	Coefficient	Standard Error	t-Statistic	Probability	Lower 95%	Upper 95%
Constant	-46.8884	38.6993	-1.2116	0.2366	-126.4360	32.6592
cm^1	15.8446	12.3077	1.2874	0.2093	-9.4543	41.1436
cm^2	-1.5529	1.3422	-1.1570	0.2578	-4.3119	1.2060
cm^3	0.0539	0.0475	1.1341	0.2671	-0.0438	0.1515
mm^3	0.0464	0.1202	0.3856	0.7030	-0.2008	0.2935
mm^4	-0.0037	0.0132	-0.2795	0.7821	-0.0308	0.0234
cm x mm	-0.2194	0.2598	-0.8445	0.4061	-0.7535	0.3147

Residual Sum of Squares =	12.5474
Standard Error =	0.6947
Mean of Y =	2.4742
Standard Deviation of Y =	0.6789
Correlation Coefficient =	0.3862
R-squared =	0.1492
Adjusted R-squared =	-0.0472
F(6,26) =	0.7597
Probability of F =	0.6079
Durbin-Watson Statistic =	1.9130
Log of Likelihood =	-31.3034
Press Statistic =	17.5842

Example 2

Table 4.4.1 on p. 295 from Elliot, M. A., J. S. Reisch, N. P. Campbell (1989). The following results are given on p. 297.

Open REGRESS, select Statistics 1 → Regression Analysis → Polynomial Regression and select *X* (*C17*) as [Variable] and *Y* (*C18*) as [Dependent]. The following set of outputs has been obtained by using these variables with only changing the degree of polynomial. Here we will only print the estimated regression coefficients:

Polynomial Regression

Dependent variable: Y
Valid Number of Cases: 19, 0 Omitted

Regression results

	Coefficient	standard error	t-statistic	Probability
Constant	37.3890	0.43640	85.6754	0.0000
X^1	3.12686	0.15099	20.7087	0.0000

	Coefficient	standard error	t-statistic	Probability
Constant	40.3017	1.13349	35.5554	0.0000
X^1	0.66658	0.91352	0.72968	0.4761
X^2	0.45397	0.16688	2.72031	0.0151

	Coefficient	standard error	t-statistic	Probability
Constant	32.7673	3.11320	10.5253	0.0000
X^1	10.4109	3.90298	2.66743	0.0176
X^2	-3.38682	1.51356	-2.23765	0.0408
X^3	0.47011	0.18442	2.54915	0.0222

	Coefficient	standard error	t-statistic	Probability
Constant	6.92654	7.28551	0.95073	0.3579
X^1	55.8348	12.4946	4.46873	0.0005
X^2	-31.4866	7.60544	-4.14001	0.0010
X^3	7.76246	1.95731	3.96587	0.0014
X^4	-0.67507	0.18076	-3.73460	0.0022

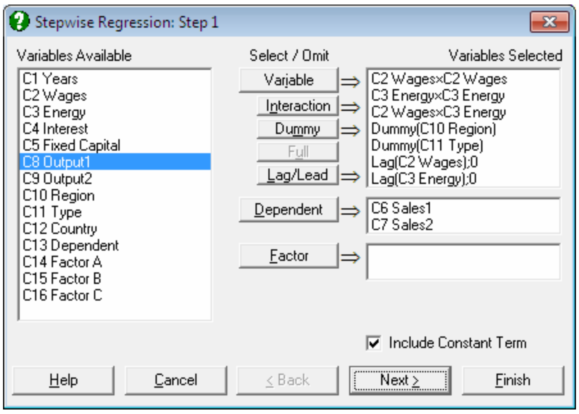
	Coefficient	standard error	t-statistic	Probability
Constant	36.2391	22.7989	1.58951	0.1360
X^1	-9.16153	49.5645	-0.18484	0.8562
X^2	23.3871	41.2381	0.56712	0.5803
X^3	-14.3460	16.4561	-0.87178	0.3991
X^4	3.59360	3.16090	1.13689	0.2761
X^5	-0.31740	0.23467	-1.35255	0.1993

	Coefficient	standard error	t-statistic	Probability
Constant	157.882	73.6834	2.14271	0.0533
X^1	-330.976	192.285	-1.72128	0.1109
X^2	364.043	201.286	1.80858	0.0956
X^3	-199.361	108.401	-1.83911	0.0908
X^4	58.1131	31.7588	1.82983	0.0922
X^5	-8.60699	4.81303	-1.78827	0.0990
X^6	0.50964	0.29560	1.72410	0.1103

7.2.3. Stepwise Regression

Stepwise Regression provides an answer to the question of which independent variables to include in the regression equation.

The simplest way to isolate the effects of various independent variables on the variation of dependent variable would be to start with one independent variable and run a series of regressions adding one independent variable at a time. An alternative would be to start with all independent variables and omit one at a time. Indeed, these are the two basic procedures most commonly used in Stepwise Regression, but with a difference. Rather than adding or omitting variables randomly it is possible to introduce a statistically meaningful criterion to rank the sequence. The enter/omit criteria used here are the F-to-enter, F-to-remove and Tolerance parameters.



Variables Selected
C2 Wages×C2 Wages
C3 Energy×C3 Energy
C2 Wages×C3 Energy
Dummy(C10 Region)
Dummy(C11 Type)
Lag(C2 Wages);0
Lag(C3 Energy);0

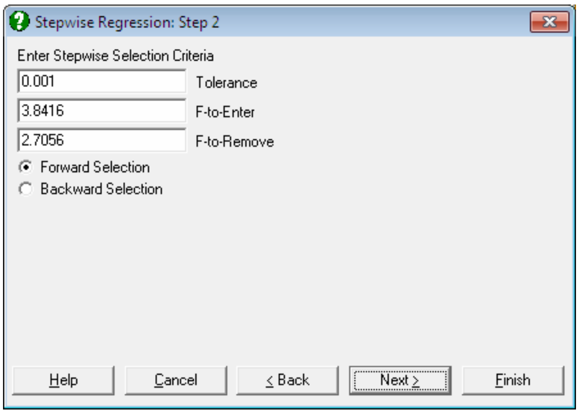
Dependent
C6 Sales1
C7 Sales2

☑ Include Constant Term

As in Linear Regression, it is possible to create interaction terms, dummy variables, lag/lead terms, select multiple dependent variables and run regressions on subsamples defined by several factor columns (see 7.2.1.1. Linear Regression Variable Selection). However, a weights option is not included. The set of independent variables selected or created are the candidates for inclusion in the regression equation. The stepwise procedure will not consider columns that are not in the Variables Selected list.

7.2.3.1. Stepwise Selection Criteria

The next dialogue is for selecting the Tolerance, F-to-enter and F-to-remove thresholds. One of Forward Selection or Backward Selection methods is also specified on this dialogue.



Stepwise Regression: Step 2

Enter Stepwise Selection Criteria

0.001 Tolerance

3.8416 F-to-Enter

2.7056 F-to-Remove

● Forward Selection

○ Backward Selection

The values suggested by the program are the most commonly used limits. Of course, it is possible to enter any value of choice by editing the number in the field. UNISTAT allows entry of F-values only as enter / remove thresholds. If you wish to enter tail probability values instead, the corresponding F-values can be calculated easily using the Statistics 1 → Distribution Functions → Critical Value procedure. The complement of the desired tail probability value (1 - α) should be entered in the Probability dialogue, and numerator and denominator degrees of freedom should be entered as 1 and 100,000 (representing infinity) respectively. The critical value obtained in this way can then be used in the Stepwise Regression procedure.

F-to-Enter: The F-to-enter statistic of an independent variable is the F-statistic for testing the significance of the regression coefficient it would have if it were in the regression equation. If this calculated value is above the one specified by the user, then the variable can enter the equation. The default value is 3.8416, corresponding to a tail probability value of 0.05 (with 1 and 100,000 degrees of freedom) and it must always be greater than the F-to-remove value. If you wish to change this default value permanently, enter and edit the following line in the [Options] section of *Documents**Unistat65**Unistat65.ini* file:

```
StepwiseFtoEnter=3.8416
```

F-to-Remove: The F-to-remove statistic of an independent variable which is already in the regression equation is the F-statistic for testing the significance of its regression coefficient. If this calculated value is below the one specified by the user then the variable is removed from the equation. The default value is 2.7056, corresponding to a tail probability value of 0.10 (with 1 and 100,000 degrees of freedom) and it must always be less than the F-to-enter value. If you wish to change this default value permanently, enter and edit the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
StepwiseFtoRemove=2.7056
```

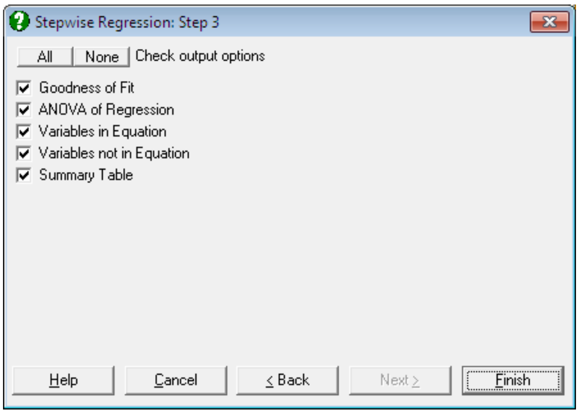
Tolerance: In order to avoid highly correlated variables and also to prevent accumulation of rounding errors, a Tolerance value is specified. The Tolerance of a variable which is not in the equation is defined as 1 - R-squared where R is the multiple correlation between the variable and all variables which are in the regression equation. If you wish to change this default value permanently, enter and edit the following line in the [Options] section of *Unistat65.ini*:

```
StepwiseTolerance=0.001
```

Forward/Backward Selection: If the Forward Selection method is employed, then the program will first run a regression with the most likely candidate, and then successively introduce other variables or omit existing ones. If the Backward Selection method is selected, then the program will first run a regression with all independent variables included and then proceed with the omission process. In this case, the output will also include a full regression output in the beginning.

It is important to emphasise that neither F-to-enter or F-to-remove, nor the Tolerance of a variable (either in the equation or not) remains the same when a variable is added to or removed from the regression equation. Therefore, whenever an addition or omission takes place, all variables, regardless of being in the equation or not, are made subject to the above checks. When the last of the independent variables is tried for entry or removal and no variables can be entered or removed, then the selection process is terminated.

7.2.3.2. Stepwise Regression Output Options



Check output options

☑ Goodness of Fit
☑ ANOVA of Regression
☑ Variables in Equation
☑ Variables not in Equation
☑ Summary Table

The full output can be substantial, as a large amount of statistics are reported for each step. These include the standard error, multiple correlation, R-squared, adjusted R-squared, change in R-squared, Analysis of Variance. The regression coefficient, its standard error, t-statistic, its tail probability and the calculated F-to-remove value are displayed for each independent variable. Partial correlation, Tolerance and F-to-enter values of variables which are not in the equation are also displayed.

At the end of the selection process, a summary table gives the multiple correlation, R-squared and F-statistic for each step.

7.2.3.3. Stepwise Regression Example

Example 20.1e on p. 436 from Zar, J. H. (2010).

Open REGRESS, select Statistics 1 → Regression Analysis → Stepwise Regression and select *temperature*, *cm*, *mm* and *min* (*C1* to *C4*) as [Variable]s and *ml* (*C5*) as [Dependent]. Select Backward Selection and accept the Tolerance levels given in the next dialogue to obtain the following output:

Stepwise Regression

Dependent Variable: ml
Valid Number of Cases: 33, 0 Omitted

Backward Selection

Tolerance: 0.001
F-to-Enter: 3.8416 (5.0%)
F-to-Remove: 2.7056 (10.0%)

All uncorrelated variables entered

Standard Error	Multiple Correlation	R-squared	Adjusted R-squared	Change in R-squared
0.4238	0.8117	0.6589	0.6102	0.6589

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Regression	9.717	4	2.429	13.524	0.0000
Error	5.030	28	0.180		

Variables in Equation	Coefficient	Std Error	t-Statistic	Prob	F-to-Remove
Constant	2.9583				
Temperature	-0.1293	0.0213	-6.0751	0.0000	36.9063
cm	-0.0188	0.0563	-0.3338	0.7410	0.1114
mm	-0.0462	0.2073	-0.2230	0.8252	0.0497
min	0.2088	0.0670	3.1141	0.0042	9.6979

Step 1: Variable Removed: mm

Standard Error	Multiple Correlation	R-squared	Adjusted R-squared	Change in R-squared
0.4168	0.8114	0.6583	0.6230	-0.0006

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Regression	9.708	3	3.236	18.625	0.0000
Error	5.039	29	0.174		

Variables in Equation	Coefficient	Std Error	t-Statistic	Prob	F-to-Remove
Constant	2.6725				
Temperature	-0.1305	0.0203	-6.4232	0.0000	41.2572
cm	-0.0154	0.0533	-0.2892	0.7745	0.0837
min	0.2045	0.0632	3.2356	0.0030	10.4694

Variables not in Equation	Partial Corr	Tolerance	F-to-Enter
mm	-0.0421	0.8518	0.0497

Step 2: Variable Removed: cm

Standard Error	Multiple Correlation	R-squared	Adjusted R-squared	Change in R-squared
0.4104	0.8108	0.6573	0.6345	-0.0010

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Regression	9.694	2	4.847	28.775	0.0000
Error	5.053	30	0.168		

Variables in Equation	Coefficient	Std Error	t-Statistic	Prob	F-to-Remove
Constant	2.5520				
Temperature	-0.1324	0.0189	-6.9993	0.0000	48.9907
min	0.2013	0.0613	3.2850	0.0026	10.7910

Variables not in Equation	Partial Corr	Tolerance	F-to-Enter
mm	-0.0261	0.9176	0.0198
cm	-0.0536	0.8652	0.0837

Summary Table

Dependent Variable: ml

Step	In/Out	Variable	Multiple Corr	R-squared	F-Stat	Prob
1	Out	mm	0.8114	0.6583	18.6251	0.0000
2	Out	cm	0.8108	0.6573	28.7748	0.0000

7.2.4. Nonlinear Regression

7.2.4.0. Overview

The Nonlinear Regression procedure provides a least-squares method of fitting a user-specified function to a suitable data set. This regression function would usually be based on a theoretical model of the system under analysis, and can be written in terms of any number of independent variables and any number of parameters (subject to program restrictions). The user provides initial estimates for the parameter values, and the program then adjusts these over a number of iterations in order to obtain the best fit to the data (that is, minimise the residual sum of squares). The program ends by displaying a large number of output options regarding the parameters and fitted curve.

In addition to providing numerical values for the parameters and some related statistics, Nonlinear Regression can also be used to compare regression models. For example, with data describing an exponential decay curve, it is possible to assess (in a statistical sense) if the curve is monophasic:

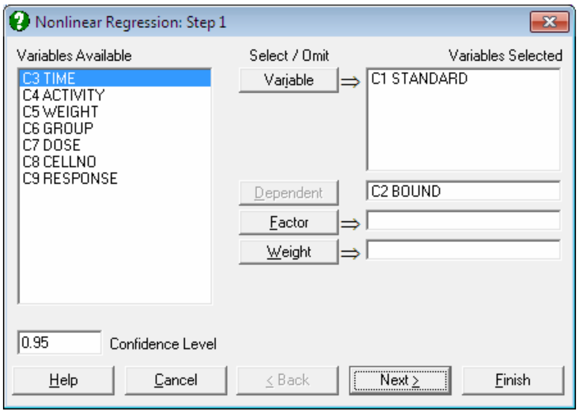
$$Y = P_1 \mathrm{Exp}(P_2 X)$$

or biphasic:

$$Y = P_1 \mathrm{Exp}(P_2 X) + P_3 \mathrm{Exp}(P_4 X)$$

The same method can be used to compare two or more sets of data and determine if differences exist between them. This is explained in detail later in this section. First we shall look at the dialogues and facilities of this procedure.

7.2.4.1. Nonlinear Regression Variable Selection



Variables Available: C3 TIME, C4 ACTIVITY, C5 WEIGHT, C6 GROUP, C7 DOSE, C8 CELLNO, C9 RESPONSE

Variables Selected: C1 STANDARD

Dependent: C2 BOUND

Factor:

Weight:

0.95 Confidence Level

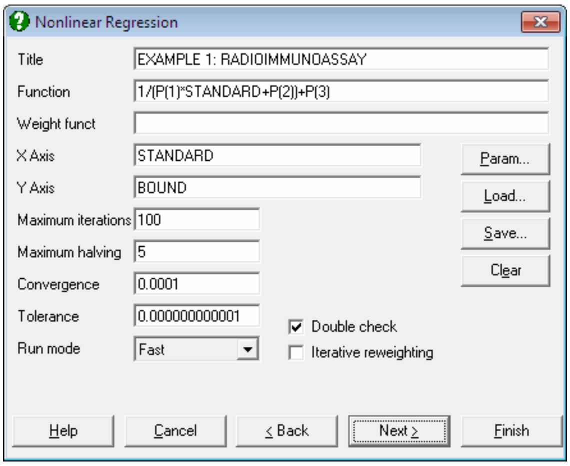
Dependent: Selection of a dependent variable containing numeric data is compulsory. Only one dependent variable can be selected.

Variable: At least one independent variable must be selected. It is not necessary for all independent variable columns selected in the Variables Selected list to be used in the analysis. The regression function is written as an expression involving variables *C1, C2,* ... or their labels, and parameters *P(1), P(2),* ... or their labels. Therefore, which columns are used depends on how the regression function is constructed.

Weight: A column of spreadsheet containing numeric data can be selected as weights. Cases with the highest weight value have the greatest effect on the fitted curve and those with the lowest weight have the least effect. If a weight variable is specified, then all entries in that data column must be greater than zero and not greater than 1.0E+10.

Factor: If a factor column is selected, the program can fit two or more curves simultaneously on the subgroups defined by the levels of the factor. This is useful when making comparisons between sets of data. The program can deal with up to eight curves, using the first eight levels of a factor.

7.2.4.2. Nonlinear Regression Setup Dialogue



Title: EXAMPLE 1: RADIOIMMUNOASSAY

Function: 1/(P(1)*STANDARD+P(2))+P(3)

Weight funct:

X Axis: STANDARD

Y Axis: BOUND

Maximum iterations: 100

Maximum halving: 5

Convergence: 0.0001

Tolerance: 0.000000000001

Run mode: Fast

☑ Double check

□ Iterative reweighting

This dialogue allows you to enter all the information needed to specify how the regression is to be performed. All this information can be saved in a Nonlinear Regression setup file and loaded back when needed, which makes repeating the same or similar analyses on several data files much easier.

7.2.4.2.1. Buttons

The four buttons provided have the following tasks:

Param: Shortened for parameters. It will activate the Parameter Setup Dialogue.

Load: This will invoke the standard windows dialogue for opening a file. You can load a previously saved Nonlinear Regression setup file. The default extension is .USN.

The names and column numbers of all variables selected from the spreadsheet (dependent variable, all independent variables and if any, weight and factor variables) are recorded in the setup file. When the setup file is subsequently loaded, these will be checked against the actual selections made from the Variable Selection Dialogue so that there are no inconsistencies between the setup file and the variables in the spreadsheet. If there are inconsistencies, these will be reported, but it will still be possible to run an analysis. In this case, the selections made from the Variable Selection Dialogue will override the definitions in the setup file and it will be up to you to ensure that the variables used in various functions do refer to the correct columns in the spreadsheet.

Save: All current selections made in the Setup and Parameter dialogues, as well as the variable assignments made in the Variable Selection Dialogue will be saved in a binary file. The default extension is .USN and you are strongly recommended to keep that unchanged.

The [Save] option always saves the current values of parameters. Therefore, if this option is selected after the convergence has been achieved then the file may be saved with convergence parameters. Or more importantly, for particularly difficult problems, after breaking the run after a number of iterations, the current values of the parameters can be saved in the setup file making it possible to continue to run the model subsequently from where you left.

Clear: This will clear all controls in both Setup and Parameter dialogues for a fresh restart.

7.2.4.2.2. Controls

Title: This is a line of text you may enter to annotate a particular Nonlinear Regression setup and will also be printed as a sub title (after the procedure name) in the output.

Function: The regression function is entered as an expression involving the independent variable(s) and parameters. The independent variables are referred to with their labels or column numbers (*C1, C2,* etc.), and the parameters as *P(1), P(2),* etc.

As an example, to fit an exponential decay curve with the independent variable in *C1*, the function can be written as:

*P(1)**Exp(*C1*P(2)*)

If *C1* has a label *TIME* and parameters 1 and 2 named *ACTIVITY0* and RATE respectively, then the expression can be written as:

*ACTIVITY0**Exp(*TIME*RATE*)

Weighting Function: Iterative reweighting can be used, which means that at each iteration the weights to be used in that iteration are calculated according to an expression (the weighting function) which is supplied by the user. This function is written in the same way as the regression function, except that a variable containing the fitted values *V(0)* can also be used.

The weight values generated in this way must be greater than zero and not greater than 1.0E+10, otherwise an error will be generated. The weighting function will only be used if the iterative reweighting option (see below) is enabled.

If a weights variable has been selected and iterative reweighting enabled, then an error will be reported. They cannot be used at the same time.

X-Axis: A transform expression for the horizontal axis of the graphical display must be entered. In the simplest case this will be a single variable (e.g. *TIME* or *C1* for the example described above), but any expression containing more than one independent variable can be used. This expression must not include parameters or the dependent variable.

Y-Axis: As above but applies to the vertical axis of the graphical display. The dependent variable must be included in the Y-axis expression.

Max Iterations: This is the maximum number of iterations to be performed (default value is 100). It is seldom necessary to alter this value, since the program will terminate when the best fit has been obtained and the residual sum of squares stops changing from one iteration to the next (see Converge below). Also, the program can be interrupted at any time. Setting the maximum number of iterations to zero has the same effect as selecting Run Mode = FINISH (see below). A negative value will cause an error to be reported.

Max Halving: At each iteration the program calculates a correction vector for each active parameter which is used to modify that parameter value. The new parameter values generated will usually result in a decrease in the residual sum of squares (RSS), but if the RSS increases then the correction vectors are halved and their signs reversed (subject to lower/upper limits on the parameters) until a decrease in the RSS is obtained or the maximum number of step halving has been performed. The default value is five and the minimum is zero. A negative value will cause an error to be reported. The setting for Max Halving is only relevant if Run Mode = STANDARD (see below).

Converge: The program will terminate when the absolute fractional change in the RSS over three successive iterations is less than the value specified for Converge. The default value is 0.0001, meaning that if the RSS changes by less than 0.01% over three iterations then it is assumed that the best fit (minimum RSS) has been obtained. The final calculations are made and the results are then displayed.

Tolerance: The pivoting tolerance value should not normally be changed from its default value of 1.0E-11. However, when fitting a function which is more complex (has more parameters) than necessary the program may become unstable, evidenced by the RSS increasing rather than decreasing. This does not immediately indicate a program failure, but if it persists for 10 iterations or more, then some corrective action will probably be needed. This can usually be accomplished by specifying different initial estimates and / or supplying appropriate lower / upper limits for the parameters. If the instability is still present then increasing the tolerance value to 1.0E-5 may be helpful.

Run Mode: This determines how the program operates. The four options available are:

1) **FAST:** The regression is performed using the Marquardt-Levenburg algorithm, which in most cases is much faster than the conventional Gauss-Newton method. However, the Marquardt-Levenburg method as implemented in this program is less suitable for *difficult* regression analyses and will fail if there is an excessively high correlation between two or more parameters. If this happens the program will automatically switch to STANDARD mode. Difficult regression analyses may include those in which a) for certain combinations of parameter values, a large change in RSS accompanies a very small change in the value(s) of one or more parameters; b) an inappropriate regression model has been specified, which might be done intentionally in order to compare it statistically with another model; and c) there is excessive correlation between the parameters.

2) **STANDARD:** The regression is performed using a slower but much more stable method based on the conventional Gauss-Newton algorithm, which can cope with difficult analyses as described above. Since the program will change from FAST to STANDARD mode itself when needed, it is not necessary to start in this mode. However, if an inappropriate regression model (particularly one that is more complex or has more parameters than necessary) is being used, then beginning the analysis in STANDARD mode may give somewhat different parameter values; the extra parameters tend not to change from their initial values in this mode.

3) **FINISH:** The final calculations are made on the current parameter values and the results are displayed. These statistics (the parameter standard errors, correlation matrix and predicted standard deviations of the fitted curve) are only valid if a good fit to the data has been obtained.

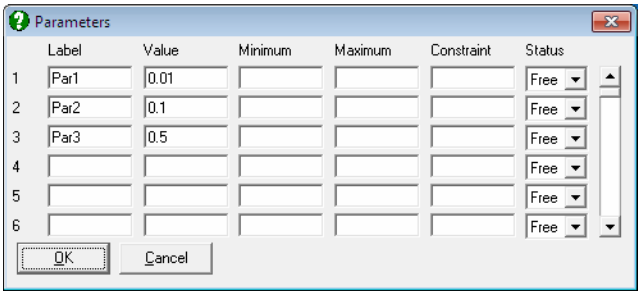
4) **MANUAL:** In this mode the user enters a set of parameter values and the program then shows the corresponding fitted curve and RSS. This process repeats until the user exits from this part of the program. The program does not alter the parameter values, except for evaluation of any parameter constraint expressions that have been defined. If weighting or iterative reweighting has been enabled then the weighted RSS is calculated. This mode can be used to obtain initial parameter estimates for the regression, or to examine the effect of changing the parameter values on the fitted curve.

Double Check: If the regression converges in FAST mode and the fitted curve satisfies a built-in goodness-of-fit test (see 7.2.4.6. Technical Details) then the final calculations are made and the results displayed. If the fitted curve fails the goodness-of-fit test then the STANDARD mode is entered automatically. However, if Double Check is enabled then the program will switch to STANDARD mode to see if any further improvement in fit can be obtained (which does sometimes happen). The default settings (Run Mode = FAST and Double Check = On) are recommended for most applications.

Iterative Reweighting: this enables iterative reweighting as explained above for weighting function.

7.2.4.2.3. Parameter Setup Dialogue

The program will determine the number of parameters used in the regression from the number of rows filled in by the user. The minimum is one and the maximum is fifty.



	Label	Value	Minimum	Maximum	Constraint	Status
1	Par1	0.01				Free
2	Par2	0.1				Free
3	Par3	0.5				Free
4						Free
5						Free
6						Free

Label: These are optional, and can be used instead of the parameter numbers *P(1), P(2),* etc.

Value: In Nonlinear Regression the user must specify an initial value for each parameter which the program will then adjust in order to minimise the residual sum of squares. These initial estimates should be close to the true (or best-fit) values, but exactly how close they need to be depends on various factors such as the amount or error in the data and the complexity of the regression function. The initial values might be determined from the theory underlying the Regression Analysis, prior knowledge of the system, or graphical examination of the data.

The initial parameter values may be set to zero, but this will slow the program down quite markedly. A blank field does not have a zero value.

On completion of the Regression Analysis, the final parameter values are written back to the Parameter Setup Dialogue where they can be edited if necessary and used as starting values for further analysis.

Minimum: This is the minimum value that each parameter can take during the regression. A blank field means that no lower limit is imposed.

Maximum: This sets the maximum value for each parameter.

Although optional, it is recommended that lower and upper limits should be assigned for each parameter as a precaution against arithmetic errors. For example, if the regression function includes the square root of a parameter, then that parameter should have a lower limit of zero. Also, if the theory underlying Regression Analysis indicates that some parameter could not have a negative value, then that parameter would also have a lower limit of zero. Otherwise, quite broad limits can be set; parameters should not be restricted to a very narrow range unless there is special reason to do so.

If lower and / or upper limits have been set, then the initial value must be within the defined range otherwise an error will be reported when the Regression Analysis is started.

Constraint: This is the constraint expression referred to above. The expression is written in terms of other parameters only and must not include any of the input variables. The constraint expression will only be used if the corresponding parameter status is set to Constrained.

Status: This control determines how the parameter will be handled during the regression. There are three possible values:

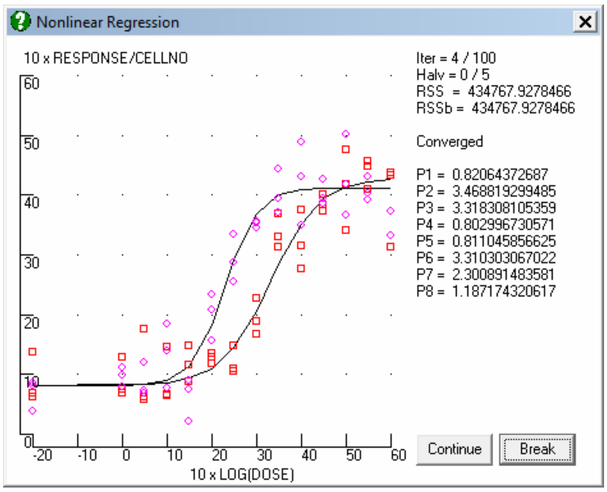
1) **Free:** The value of the parameter will be adjusted in an attempt to obtain the best fit.

2) **Fixed:** The parameter value is fixed (constant) and it will not change.

3) **Constrained:** The value of the parameter is determined by other parameter values according to an expression (a parameter constraint expression) which you must supply. For example, if you wanted to perform the regression such that parameter 5 always had the same value as parameter 3, then parameter 5 would have status Constrained and the constraint expression should be *P(3)*. If a parameter has status Constrained then the corresponding parameter constraint expression must be defined.

Parameters with status Free are termed active parameters. There must be at least one active parameter, otherwise an error will be reported.

7.2.4.3. Running the Program



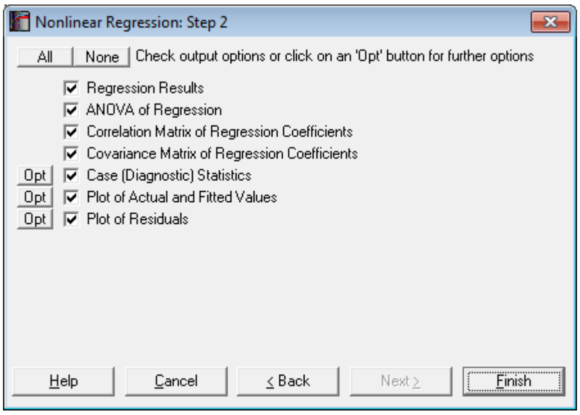
```
Iter = 4 / 100
Halv = 0 / 5
RSS  =  434767.9278466
RSSb =  434767.9278466

Converged

P1 =  0.82064372687
P2 =  3.468819299485
P3 =  3.318308105359
P4 =  0.802996730571
P5 =  0.811045856625
P6 =  3.310303067022
P7 =  2.300891483581
P8 =  1.187174320617
```

After filling in the Nonlinear Regression Setup Dialogue and the Parameter Setup Dialogue, click [Next] to run the model. The program will first check all the entries and report any inconsistencies found. Otherwise a new window is opened showing a bivariate plot of data according to the X-axis and Y-axis expressions defined in the Nonlinear Regression Setup Dialogue. At each iteration, the progress of fitting can be followed visually. The same window will also display the current values of number of iterations, residual sum of squares, as well as the current values of the parameters.

It is possible to interrupt or break the program while it is running by pressing <Escape>. However, the response may not be immediate, as the program will always finish the current iteration before stopping. When this happens, the Nonlinear Regression Setup Dialogue will be displayed. The current values can be saved to a regression setup file to continue later on, or change some of the parameters manually. When [Continue] is clicked, the program will resume iterations. Changing the regression function will result in a complete resetting of the regression run.

If the iterations continue with no sign of convergence, you may break the run, select the FINISH mode from the Nonlinear Regression Setup Dialogue and force the program to perform calculations necessary for output options.

7.2.4.4. Nonlinear Regression Output Options



Nonlinear Regression: Step 2

Check output options or click on an 'Opt' button for further options

☑ Regression Results
☑ ANOVA of Regression
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
Opt ☑ Case (Diagnostic) Statistics
Opt ☑ Plot of Actual and Fitted Values
Opt ☑ Plot of Residuals

After convergence has been achieved or the program is stopped before convergence and Run Mode = FINISH has been selected, the Output Options Dialogue will be displayed. For a sample printout of the full Nonlinear Regression output (see 7.2.4.5. Nonlinear Regression Examples).

The Actual and Fitted Values, Residuals, Confidence Intervals for Actual Y Values and Interpolation options that existed in earlier version of UNISTAT are now combined under the Case (Diagnostic) Statistics option.

Regression Results: The main regression results include a report on whether the convergence has been achieved and the result of a goodness-of-fit test, which will be either *OK* or *Failed test*.

All values related to parameters are displayed in a table. These include minimum, maximum, coefficient estimate, standard error and status (Free, Fixed or Constrained). Confidence intervals for regression coefficients are also displayed for the user-defined confidence level. The entire table can be sent to the spreadsheet for further analysis.

Statistics reported include residual sum of squares, root mean square error, number of parameters, number of active parameters, number of cases, degrees of freedom (number of cases minus number of active parameters), R-squared and Durbin-Watson statistic.

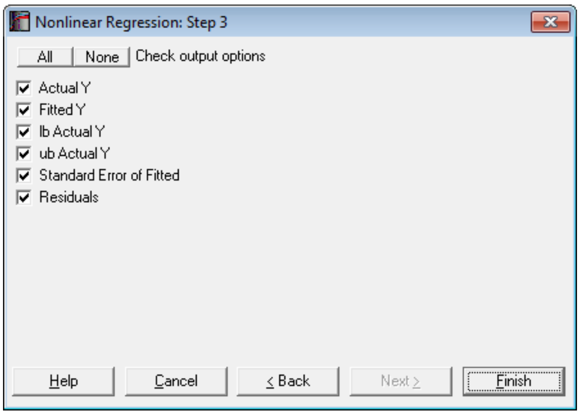
Standard errors are only available for parameters with status Free. If a factor column has been selected, then most of the statistics will be displayed for each sub group separately.

ANOVA of Regression: The ANOVA table compares the regression model with the null hypothesis that "Y = constant".

Correlation Matrix of Regression Coefficients: Only parameters with status Free are included in this table.

Covariance Matrix of Regression Coefficients: Only parameters with status Free are included in this table.

Case (Diagnostic) Statistics:



Nonlinear Regression: Step 3

All | None | Check output options

☑ Actual Y
☑ Fitted Y
☑ lb Actual Y
☑ ub Actual Y
☑ Standard Error of Fitted
☑ Residuals

Predictions (Interpolations): As of this version of UNISTAT, the Interpolation output option is removed and the predicted Y values can be generated as in Linear Regression (see 7.2.1.2.2. Linear Regression Case Output). If, for a case, all independent variables are non-missing, but only the dependent variable is missing, a value will be predicted for the fitted Y value from the estimated regression equation. When a case is predicted, its label will be prefixed by an asterisk (*).

Statistics available under Case (Diagnostic) Statistics option are as follows.

Actual Y: Observed values of the dependent variable.

$$Y_i$$

Fitted Y: Estimated values of the dependent variable.

$$\hat{Y}_i = \sum_{k=1}^{m} b_k X_{ki}$$

Confidence Intervals for Actual Y-values: Confidence intervals for actual Y values are displayed. The default confidence level is 95%, but any value can be entered from the Variable Selection Dialogue.

$$\hat{Y}_i \pm t_{\alpha/2} S_{\hat{Y}}$$

Where $t_{\alpha/2}$ is the critical value from the t-distribution for an $\alpha / 2$ level of significance and 1 degree of freedom.

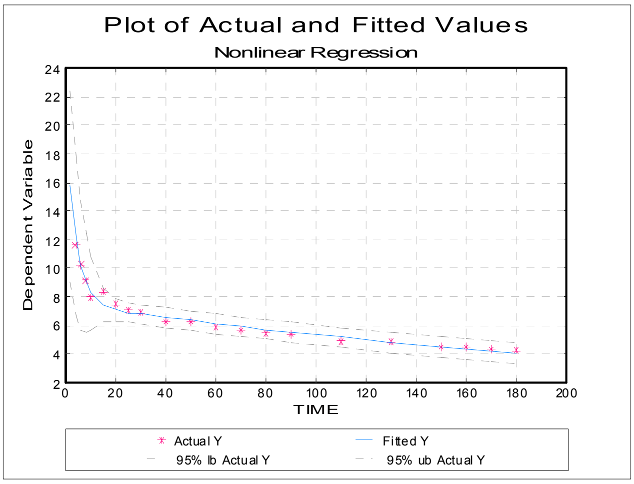
Standard Error of Fitted:

$$S_{\hat{Y}} = S\sqrt{h_i}$$

Residuals:

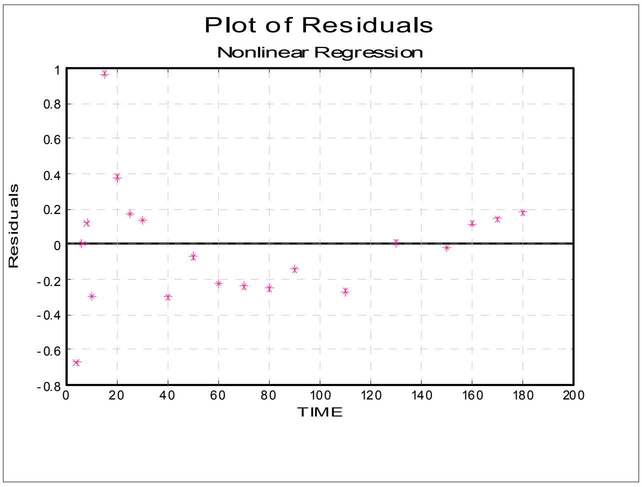
$$e_i = \hat{Y}_i - Y_i$$

Plot of Actual and Fitted Values: Actual and fitted Y values, as well as their confidence intervals, are plotted. The X-axis variable is as selected in the Nonlinear Regression Setup Dialogue.



As with plots in Linear Regression and Polynomial Regression, the data points here are also interactive, but pressing <Delete> when a point is highlighted will not omit this point and re-run the analysis.

Plot of Residuals: A scatter plot of residuals is drawn against the dependent variable.



7.2.4.5. Nonlinear Regression Examples

7.2.4.5.1. Example 1: Radioimmunoassay

This is a simple example based on a calibration curve for a type of assay. Open NONLINRG and select Statistics 1 → Regression Analysis → Nonlinear Regression. The data consists of 14 cases and two variables *STANDARD* and *BOUND*.

Select *STANDARD* (*C1*) as the independent variable by clicking [Variable], and *BOUND* (*C2*) as the [Dependent] variable. Then clicking [Next], the Nonlinear Regression Setup Dialogue will be opened. Click on the [Load] button and load the NONLIN1.USN setup file. The function involves three parameters:

$$\mathrm{BOUND} = \frac{1}{(\mathrm{P}(1) * \mathrm{STANDARD} + \mathrm{P}(2))} + \mathrm{P}(3)$$

which is entered as:

1/((P(1)*STANDARD+P(2))+P(3)

or could be entered as:

1/((P(1)*V(1)+P(2))+P(3)

The three parameters have initial values of 0.01, 0.1 and 0.1 respectively, all have status Free, and no lower or upper limits are set. With the default settings of Run Mode = FAST and Double Check = On, the program converges after 8 iterations, then switches to STANDARD mode and performs another 3 iterations, at which point it converges again. The following are the main results:

Nonlinear Regression

Regression Results

EXAMPLE 1: RADIOIMMUNOASSAY
BOUND = 1/(P(1)*STANDARD+P(2))+P(3)
Convergence achieved at iteration 11, No of halving = 0
Goodness of Fit: OK

Parameter	Minimum	Maximum	Estimate	Standard Error	Status	Lower 95%	Upper 95%
Par1			0.0279	0.0019	Free	0.0237	0.0321
Par2			1.0914	0.0171	Free	1.0538	1.1291
Par3			0.0160	0.0153	Free	-0.0177	0.0496

Residual Sum of Squares =	0.0018
Root Mean Square Error =	0.0126
Number of Parameters =	3
No of Active Parameters =	3
Number of Cases =	14
Degrees of Freedom =	11
R-squared =	0.9983
Durbin-Watson Statistic =	2.5336

ANOVA of Regression

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Regression	1.005	2	0.503	3141.719	0.0000
Error	0.002	11	0.000		
Total	1.007	13	0.077		

Correlation Matrix of Regression Coefficients

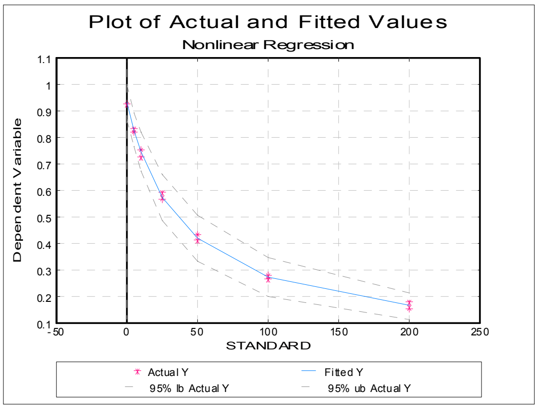
	1	2	3
1	1.0000	0.7328	0.9336
2	0.7328	1.0000	0.8761
3	0.9336	0.8761	1.0000

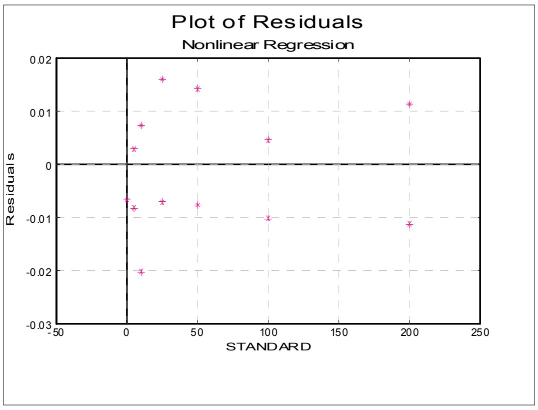
Covariance Matrix of Regression Coefficients

	1	2	3
1	0.0000	0.0000	0.0000
2	0.0000	0.0003	0.0002
3	0.0000	0.0002	0.0002

Case (Diagnostic) Statistics

	Actual Y	Fitted Y	95% lb Actual Y	95% ub Actual Y	Standard Error of Fitted	Residuals
1	0.9472	0.9322	0.8638	1.0006	0.0054	0.0150
2	0.9255	0.9322	0.8638	1.0006	0.0054	-0.0067
3	0.8201	0.8284	0.7637	0.8930	0.0051	-0.0083
4	0.8312	0.8284	0.7637	0.8930	0.0051	0.0028
5	0.7254	0.7457	0.6742	0.8172	0.0056	-0.0203
6	0.7530	0.7457	0.6742	0.8172	0.0056	0.0073
7	0.5679	0.5750	0.4885	0.6614	0.0068	-0.0071
8	0.5910	0.5750	0.4885	0.6614	0.0068	0.0160
9	0.4324	0.4181	0.3315	0.5048	0.0068	0.0143
10	0.4105	0.4181	0.3315	0.5048	0.0068	-0.0076
11	0.2782	0.2736	0.2024	0.3448	0.0056	0.0046
12	0.2634	0.2736	0.2024	0.3448	0.0056	-0.0102
13	0.1772	0.1659	0.1161	0.2156	0.0039	0.0113
14	0.1546	0.1659	0.1161	0.2156	0.0039	-0.0113





7.2.4.5.2. Example 2: Biphasic Exponential Decay

This example will illustrate the comparison of regression models. Open NONLINRG and select Statistics 1 → Regression Analysis → Nonlinear Regression. The data for this example consists of 21 cases and 3 variables.

Select *TIME* (*C3*) as the independent [Variable], *ACTIVITY* (*C4*) as the [Dependent] variable, and *WEIGHT* (*C5*) as the [Weight] variable. Click [OK] and after the Nonlinear Regression Setup Dialogue is displayed click on the [Load] button and load the NONLIN2.USN setup file.

The function involves 4 parameters (*INIT1, RATE1, INIT2, RATE2*) and describes the following biphasic exponential decay curve:

$$ACTIVITY = INIT1 * \text{Exp}(RATE1 * TIME) + INIT2 * \text{Exp}(RATE2 * TIME)$$

which could be entered, without variable or parameter names, as

P(1)*Exp(P(2)*C1)+P(3)*Exp(P(4)*C1)

The parameters have initial values of 10, -0.1, 5 and -0.01 respectively, all have status Free, and no lower or upper limits are set. The program will perform 8 iterations in FAST mode, then 3 iterations in STANDARD mode, and then the results are displayed.

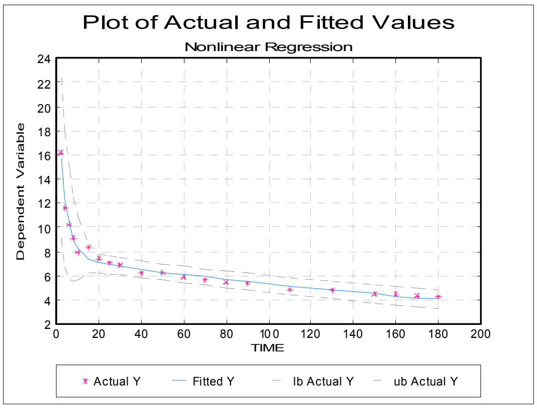
Nonlinear Regression

Regression Results

EXAMPLE 2: BIPHASIC EXPONENTIAL DECAY
ACTIVITY = INIT1*EXP(RATE1*TIME)+INIT2*EXP(RATE2*TIME)
Convergence achieved at iteration 11, No of halving = 0
Goodness of Fit: OK

Parameter	Minimum	Maximum	Estimate	Standard Error	Status	Lower 95%	Upper 95%
INIT1			14.0304	1.5996	Free	10.6554	17.4054
RATE1			-0.2627	0.0320	Free	-0.3303	-0.1952
INIT2			7.4943	0.1836	Free	7.1071	7.8816
RATE2			-0.0034	0.0002	Free	-0.0039	-0.0029

With normalised weights:	
Residual Sum of Squares =	1.4317
Root Mean Square Error =	0.2902
With absolute weights:	
Residual Sum of Squares =	40.0617
Root Mean Square Error =	1.5351
Number of Parameters =	4
No of Active Parameters =	4
Number of Cases =	21
Degrees of Freedom =	17
R-squared =	0.9863
Durbin-Watson Statistic =	0.8836



The rest of the output options are not reproduced here.

The output is essentially the same as for Example 1, except for weights. When weights are used, two RSS values are shown; the first is obtained with a normalised set of weights used for internal calculations and the second with the absolute weight values from the data file (or those calculated using the weighting function). The normalised weights have a mean value of 1.0 and are proportional to the absolute values. Since the RMS error is obtained from the RSS, there are also two entries for RMSerror. If the weights only indicate reliability of each data point relative to the others, it would be more appropriate to use the normalised RSS for reporting purposes or further calculations.

Although the fitted curve passes the goodness-of-fit test, the graphical display (plot of *ACTIVITY* against *TIME*) might suggest a slight systematic deviation of the data points about the curve. Interpretation of this would depend on the theoretical basis for the analysis, but suppose that a six parameter model was considered possible:

> *ACTIVITY* = *INIT1**Exp(*RATE1*TIME*)+*INIT2*
> *Exp(*RATE2*TIME*)+*INIT3**Exp(*RATE3*TIME*)

To establish if there is any statistical basis for accepting the six parameter model in preference to the four-parameter model, it is necessary to perform the regression again using the more complex function and then compare both sets of results on the basis of change in RSS with change in number of degrees of freedom for the regression using an extra-sum-of-squares test (sometimes referred to as an F-ratio test).

Altering the regression setup is quite simple: assign labels and initial values to parameters 5 & 6 of *INIT3* and *RATE3* and 1.0 and -0.001 respectively, and include these parameters in the regression function as shown above. Also, assign a maximum value of zero for each *RATE* parameter. The regression can now be performed (taking about 60 iterations) using the six parameter model. The results will be as follows.

Nonlinear Regression

Regression results

EXAMPLE 2: BIPHASIC EXPONENTIAL DECAY
Convergence achieved at iteration 52, no of halving = 0
Goodness of Fit Test: OK

Parameter	Minimum	Maximum	Estimate	Std Error
INIT1			16.8864	2.2845
RATE1		0	-0.4513	0.0806
INIT2			3.1163	0.5334
RATE2		0	-0.0438	0.0168
INIT3			6.4984	0.4217
RATE3		0	-0.0024	0.0004

With normalised weights:	
Residual Sum of Squares =	0.4571
Root Mean Square Error =	0.1746
With absolute weights:	
Residual Sum of Squares =	12.7893
Root Mean Square Error =	0.9234
Number of Parameters =	6
No of Active Parameters =	6
Number of Cases =	21
Degrees of Freedom =	15
R-squared =	0.9338
Durbin-Watson statistic =	2.0314

The null hypothesis is that “there is no difference in RSS between the two models”. The RSS has actually decreased by about 60%, but it is necessary to assign a significance level to this (the probability of rejecting the null hypothesis when it is true). For the following calculations it does not matter if the RSS values were obtained with normalised or absolute weights, except that you should use the same one throughout (normalised values are used below - this method is equally applicable to comparison of unweighted regression analyses). Defining;

Model 1 (4 parameters): RSS1 = 1.43174 with DF1 = 17
Model 2 (6 parameters): RSS2 = 0.45707 with DF2 = 15

an F-statistic is calculated as:

$$F = \frac{(RSS1 - RSS2)/(DF1 - DF2)}{(RSS2/DF2)}$$

with (DF1 - DF2) and DF2 degrees of freedom.

This gives F = 15.993 with 2 & 15 degrees of freedom. From statistical tables the associated p-value is less than 0.001, and consequently the null hypothesis can be rejected. That is, the six-parameter model might be said to give a significantly better fit.

In fact, a five-parameter model:

ACTIVITY = INIT1*Exp(RATE1*TIME)+INIT2 *Exp(*RATE2*TIME*)+*CONSTANT*

results in a significant decrease in RSS compared with the four-parameter model ($P < 0.001$), but there is only a slight difference in RSS between this and the six-parameter model ($P > 0.05$). This five-parameter model can be specified either by rewriting the regression function in terms of five parameters, or using the six-parameter model with *P(6)* = 0 and with status Fixed (parameter 6 is held constant with a value of zero).

7.2.4.5.3. Example 3: Dose-Response Curves

This example will show how the program can be used to fit two (or more) curves simultaneously and to make comparisons between them using the F-ratio method.

Open NONLINRG and select Statistics 1 → Regression Analysis → Nonlinear Regression. The data for this example consists of 84 cases and 4 variables. Select *GROUP* (*C6*) as the Factor, *DOSE* (*C7*) and *CELLNO* (*C8*) as the independent [Variable]s and *RESPONSE* (*C9*) as the [Dependent] variable. Click [OK] to proceed and when the Nonlinear Regression Setup Dialogue is displayed click on the [Load] button and load NONLIN3.USN regression setup file.

Suppose that in a pharmacology experiment, batches of cells are grown, subjected to some treatment or left untreated as the controls, then exposed to different concentrations of a drug and the response of the cells to that drug measured. The purpose of the experiment is to determine if the treatment has any effect on the drug response and if so, to describe that effect. The dose-response curve is sigmoid and can be described by the logistic function:

$$RESPONSE = BASAL + ACTIVE * \frac{DS1/(10\text{^}ES1)}{(1+DS1)/(10\text{^}ES1)}$$

where:

DS1=*DOSE^SLOPE*
ES1=*ED50*SLOPE*

DOSE represents the dose of the drug (an independent variable) and *RESPONSE* is the measured response (the dependent variable). The four parameters are *BASAL* (the minimum or resting response), *ACTIVE* (the activatible response or maximum rise above basal), *ED50* (the concentration of drug giving half-maximal activation) and *SLOPE* (which describes the steepness of the increase in response). Actually, *ED50* is the logarithm of the dose giving half-maximal activation; it is appropriate to estimate the logarithm of this parameter since it has a log-normal distribution (a plot of *ACTIVITY* against Log(*DOSE*) is symmetrical about the point *DOSE* = *ED50*).

Further, suppose that three experiments are performed, each involving a paired control and treated group. The cell preparation is homogeneous within each of the three experiments but can vary between experiments. The response is assumed to be proportional to the number of cells present, which is also measured (*CELLNO*, another independent variable). The function is now:

$$RESPONSE = \left[BASAL + ACTIVE * \frac{DS1/(10\text{^}ES1)}{(1+DS1)/(10\text{^}ES1)} \right] * CELLNO$$

and therefore the parameters *BASAL* and *ACTIVE* represent the response per unit cell number.

Since the data file represents two curves (control and treated), it is necessary to indicate which cases belong to each curve. A data grouping variable (*GROUP*) is included in the data file with values of 1 (control) or 2 (treated). The order in which the data points appear in the file has no effect on the regression.

Since a four-parameter function is used to describe one dose-response curve and the data describes two such curves, the regression will involve eight parameters. These are named *BASAL:1*, *ACTIVE:1*, *ED50:1*, *SLOPE:1* (for group 1), and *BASAL:2*, *ACTIVE:2*, *ED50:2* and *SLOPE:2* (for group 2). It is now necessary to write the regression function in such a way that the value of *GROUP* is used to select which parameters are used in evaluating the function. The function is written:

```
(P((GROUP-1)*4+1)+P((GROUP-1)*4+2)*
(DOSE^P((GROUP-1)*4+4)/10^(P((GROUP-1)*4+3)*
P((GROUP-1)*4+4)))/(1+DOSE^P((GROUP-1)*4+4)/
```

10^(P((GROUP-1)*4+3)*P((GROUP-1)*4+4))))*CELLNO

Factor variable *GROUP* is used to calculate the parameter subscripts. When *GROUP* = 1, the function is equivalent to:

(P(1)+P(2)*(DOSE^P(4)/10^(P(3)*P(4)))/
*(1+DOSE^P(4)/10^(P(3)*P(4))))*CELLNO*

and when *GROUP* = 2 it is equivalent to:

(P(5)+P(6)*(DOSE^P(8)/10^(P(7)*P(8)))/
*(1+DOSE^P(8)/10^(P(7)*P(8))))*CELLNO*

In the setup file, the following parameter constraints are defined but not enabled (initially all parameters have status Free);

BASAL:2 = BASAL:1
ACTIVE:2 = ACTIVE:1
ED50:2 = ED50:1
SLOPE:2 = SLOPE:1

These will be used later in making comparisons between the two curves.

Finally, the transforms for the graphical display are defined:

X-Axis = Log(*DOSE*)
Y-Axis = *RESPONSE/CELLNO*

The first regression analysis can now be performed. For the purposes of the following discussion, this model (8 active parameters) is model 1. The results are as follows:

Nonlinear Regression

Regression Results

```
EXAMPLE 3: DOSE-RESPONSE CURVES / MODEL 1
RESPONSE = (P((GROUP-1)*4+1)+P((GROUP-1)*4+2)*(DOSE^P((GROUP-
1)*4+4)/10^(P((GROUP-1)*4+3)*P((GROUP-1)*4+4)))/(1+DOSE^P((GROUP-
1)*4+4)/10^(P((GROUP-1)*4+3)*P((GROUP-1)*4+4))))*CELLNO
Convergence achieved at iteration 16, No of halving = 0
Goodness of Fit: OK
```

Parameter	Minimum	Maximum	Estimate	Standard Error	Status	Lower 95%	Upper 95%
BASAL:1	0.0000	10.0000	0.8215	0.1194	Free	0.5837	1.0592
ACTIVE:1	0.0000	10.0000	3.4658	0.2280	Free	3.0116	3.9200
ED50:1	1.0000	5.0000	3.3174	0.1232	Free	3.0720	3.5629
SLOPE:1	0.1000	5.0000	0.8061	0.1659	Free	0.4757	1.1365
BASAL:2	0.0000	10.0000	0.8111	0.1304	Free	0.5515	1.0707
ACTIVE:2	0.0000	10.0000	3.3104	0.1772	Free	2.9575	3.6632
ED50:2	1.0000	5.0000	2.3012	0.0923	Free	2.1173	2.4851
SLOPE:2	0.1000	5.0000	1.1878	0.2613	Free	0.6674	1.7083

Group	No	RSSQ(n/w)	R-squared	D-W stat	Fit
1	42	207905.2146	0.9795	1.9323	OK
2	42	226867.9214	0.9826	1.9267	OK

Residual Sum of Squares =	434773.1359
Root Mean Square Error =	75.6353
Number of Parameters =	8
No of Active Parameters =	8
Number of Cases =	84
Degrees of Freedom =	76
R-squared =	0.9381
Durbin-Watson Statistic =	1.9277

The rest of the output options are not reproduced here.

The output is largely as described previously except that a new list is included in the Regression Results for each data group; the number of cases, the RSS and the result of the goodness-of-fit test. If weighting has been used then the RSS values are calculated using normalised weights as explained above.

From the graphical display there appears to be little difference in basal or maximal (basal plus activatible) responses, but there does seem to be a difference in *ED50* and perhaps also in slope since, between minimum and maximum responses, the two curves are not completely parallel. To make a preliminary comparison between the two curves, we can determine if the difference between parameter values exceeds two standard errors of the difference, which would suggest a significant difference. That is, for parameters having values Pa and Pb with standard errors Ea and Eb, if:

$$|\mathrm{Pa} - \mathrm{Pb}| > 2\sqrt{\left(\mathrm{Ea}^2 + \mathrm{Eb}^2\right)}$$

then a genuine difference is likely. Using this method, the only apparent difference is between parameters 3 (*ED50:1*) and 7 (*ED50:2*). However, the F-ratio test is more reliable, and should be used in preference as demonstrated below.

To test the null hypothesis that "there is no difference of any kind between control and treated groups" (model 2), all four constraints are enabled. This gives a model with four active parameters, and results in RSS = 801471.4 with DF = 80. The goodness-of fit test returns Fail both for the entire data set and for both groups, which certainly indicates that this model is not appropriate. To compare models 1 & 2, the F-statistic is calculated as:

$$\mathrm{F} = \frac{(801471.4 - 434762.4)/(80-76)}{(434762.4/76)}$$

with (80-76) and 76 degrees of freedom. This gives F = 16.025 with 4 & 76 degrees of freedom and a p-value of less than 0.001. Therefore the null hypothesis is rejected and a genuine difference between the curves is established.

The preliminary analysis suggested no difference in *BASAL* or *ACTIVE* parameters between control and treated groups. To test this properly, the constraints for parameters 5 and 6 are enabled (thus *BASAL:2* = *BASAL:1* and *ACTIVE:2* = *ACTIVE:1*. This is model 3, and results in RSS = 438789.1 with DF = 79, and all goodness-of-fit tests are passed. Comparison of models 1 & 3 gives an F-statistic of 0.351 with 2 & 76 degrees of freedom, and a corresponding p-value of greater than 0.1. Thus the null hypothesis "no difference in *BASAL* or *ACTIVE* parameters" is not rejected.

With constraints *BASAL:2* = *BASAL:1*, *ACTIVE:2* = *ACTIVE:1* and *SLOPE:2* = *SLOPE:1* (model 4), the RSS is 446337.8 with DF = 79. Comparison of this with model 1 gives F = 0.927 with 3 & 76 degrees of freedom and a p-value of greater than 0.1. The null hypothesis "no difference in *BASAL*, *ACTIVE* or *SLOPE* parameters" is not rejected. It is sometimes necessary to carry out the comparisons in a stepwise manner. For example, having established no difference in *BASAL* or *ACTIVE* parameters, models 3 & 4 could be compared to further establish no difference in *SLOPE* parameters (which would give F = 1.341 with 1 & 78 degrees of freedom and $p > 0.1$, which is similar to the comparison of models 1 & 4).

Finally, with the single constraint *ED50:2 = ED50:1* (model 5, null hypothesis "no difference in *ED50* parameters"), we obtain RSS = 660274.5 with DF = 77. Comparison of this with model 1 gives F = 39.421 with 1 & 76 degrees of freedom and a p-value of less than 0.001, and therefore the null hypothesis is rejected.

Thus the preliminary analysis was correct in that the only difference between control and treated groups is a change in *ED50*. The null hypothesis for the experiment would be that there is no difference between control and treated groups (model 2). Comparison of this with model 4 (difference in *ED50* only) gives F = 62.857 with 1 & 79 degrees of freedom and a p-value of less than 0.001. Using the results obtained with model 4, it can be stated that the treatment had no effect on *BASAL*, *ACTIVE* or *SLOPE* parameters (0.835 ± 0.085, 3.319 ± 0.129, and 1.044 ± 0.156 respectively, $p > 0.1$) but reduced the *ED50* from 3.253 ± 0.091 to 2.322 ± 0.091 ($P < 0.001$ using the F-value of 62.857).

7.2.4.6. Technical Details

The program seeks the set of parameter values that minimise the residual sum of squares (RSS):

$$\sum(Y_{obs} - Y_{pred})^2$$

or:

$$\sum w(Y_{obs} - Y_{pred})^2$$

if weighting is used, where Y_{obs} and Y_{pred} are the observed and predicted values of the dependent variable, w is the weight, and the summation is over all cases. The predicted value Y_{pred} is obtained by evaluation of the regression function f(v, p), where v represents the set of independent variables and p the parameters.

The Gauss-Newton method is implemented in two forms. In both, the program obtains for each case the residual ($z, = Y_{obs} - Y_{pred}$), and for each case and each active parameter (p_i) the partial derivative

$$\partial Y_{pred} / \partial p_i$$

Using the approximation:

$$z = \sum q_i (\partial Y_{pred} / \partial p_i)$$

where the summation is over all active parameters and q_i represents a correction vector for each such parameter, solution of the set of linear equations so generated provides an estimate for each q_i. These correction vectors are used to generate a new set of parameter values for use in the next iteration, and this process is repeated until a termination condition is satisfied.

The partial derivatives are obtained arithmetically by evaluation of f(v, p) at parameter values p and p ± delta. The delta value is computed for each parameter as follows:

1) If p = 0, then delta starts at 1.0E-30 and increments in steps of 10 until a suitable change in RSS is obtained or delta = p/10 or a boundary constraint is met.

2) If $p \neq 0$, then delta starts at p/1000 and increments in steps of Sqr(10) until a suitable change in RSS is obtained or delta = p/Sqr(10) or a boundary constraint is met.

In FAST mode, the delta values are computed on entry and subsequently recomputed only for any parameter that changes sign. The partial derivatives are obtained at parameter values p and (p + delta) or (p - delta) depending on any boundary constraints. In STANDARD mode, the delta values are computed at each iteration and the partial derivatives are obtained at parameter values (p + delta) and (p - delta) subject to any boundary constraints. If a suitable change in RSS is not obtained then the parameter is omitted from the current iteration (in FAST mode the delta value will be recomputed at the next iteration).

In FAST mode, solution of the set of linear equations is achieved by a matrix inversion method incorporating the Marquardt-Levenburg procedure provided that excessive correlation between parameters is not detected. In the presence of such correlation or in STANDARD mode, a stepwise multivariate regression method is used with omission of parameters until any excessive correlation is removed.

The correction vectors obtained are used to update the parameter values; in STANDARD mode if a reduction in RSS is not achieved with the new set of parameter values then the correction vectors are halved and their signs reversed (subject to any boundary constraints) until a decrease in RSS is obtained or the maximum number of step halvings has been performed. At the end of each iteration, the RSS is compared with the recorded best-fit value and if necessary the best-fit value and associated parameter values are updated.

Parameters subject to an equality constraint are not included in the steps described above, but the constraints are evaluated at all stages where the parameter values are altered (e.g. in calculating the partial derivatives). The constraint expressions are evaluated in the order in which the parameters are numbered; therefore a constraint expression for parameter b should not use the result of constraint imposed on parameter a unless b > a.

Iteration continues until a) the maximum number of iterations has been performed, b) the convergence criterion is satisfied, c) the user interrupts the program, or an error occurs (termination conditions). Convergence is achieved when the absolute fractional change in RSS over 3 successive iterations is less than a user-specified value (Converge, default 0.0001). In a) or b), the parameters are reset to the recorded best-fit values. If convergence occurs in FAST mode and either the Double Check facility is enabled or the fitted curve fails the goodness-of-fit test, further iterations are performed in STANDARD mode until a termination condition occurs again. If termination conditions a) or b) are met in FAST mode, then the parameter standard errors, correlations and standard deviations of the y-predicted values are calculated using the partial derivatives obtained in the last iteration providing that the RSS in that iteration is within 1 ± Convergence of the best-fit value and the goodness-of-fit test is passed. Otherwise (or in STANDARD mode) the partial derivatives are recomputed after resetting the parameters to the best-fit values and before the parameter standard errors etc. are calculated.

Standard errors and correlations are not computed for parameters that have zero or negligible effect on the fitted curve, for which an excessive correlation is present, or which are held constant or are the subject of an equality constraint (status Fixed or Constrained).

The goodness-of-fit test consists of two parts; a) a Runs Test assuming a binomial distribution with $p = 0.5$ is applied to the signs of the residuals and fail registered at $\alpha \leq 0.005$; and b) the weighted mean and standard deviation of the residuals is calculated and Fail registered if zero is out with the interval mean ± 2 x standard deviations. The test always returns Fail if the number of cases is less than five.

For further reading see Ratkowski, D. A. (1983) and Wadsworth, H. M. (1990).

7.2.5. Logit / Probit / Gompit

Regressions with logit, probit, gompit (or complementary log log, cloglog) and loglog link functions can be estimated for models with binary dependent variables (dependent variables that consist of two values) as well as the aggregated models where data contains a variable on the number of positive (or negative) responses and another variable giving the total number of subjects. All regressions work with similar inputs, employ the same maximum likelihood method and share the same output format, but differ in the link (objective) function used.

The logit and Logistic Regression procedures are closely related. A logit analysis with a binary dependent variable will produce the same coefficients estimated by Logistic Regression. For such problems use the Logistic Regression procedure if you need;

1) Receiver Operating Characteristic (ROC) and Sensitivity and Specificity curves, the area enclosed under the ROC curves (AUC), their confidence intervals and comparisons,
2) statistics for diagnostic tests,
3) odds ratios and their confidence intervals,
4) a classification table for the predicted and observed group memberships,
5) and a wide range of Case (Diagnostic) Statistics.

On the other hand, use the Logit / Probit / Gompit procedure if you need;

1) marginal and average effects,
2) to use data in aggregated (i.e. not in binary) format,
3) to estimate the natural response rate or
4) to estimate a probit, gompit (cloglog) or loglog model.

Like other regression options, Logit / Probit / Gompit also allows for automatic creation of interaction terms and dummy variables.

7.2.5.1. Logit / Probit / Gompit Model Description

The link functions described here are also available as axis scaling options in UNISTAT graphics engine (see Scale Type).

Logit: The logit function is an odds ratio for a given probability value:

Logit(p) = Ln(p/(1-p))
Logit(0.025) = -3.66
Logit(0.95) = 2.94.

Probit: This is the inverse standard cumulative normal distribution:

Probit(p) = Φ^{-1}(p)
Probit(0.025) = -1.96
Probit(0.95) = 1.64

Gompit (Cloglog) : Unlike logit and probit, gompit is an asymmetric function:

Gompit(p) = Ln(-Ln(1-p))
Gompit(0.1) = -2.25
Gompit(0.9) = 0.834

Note that various sources interpret gompit, cloglog, loglog, nloglog in different ways. All these models are closely related to each other and one can obtain any of them by using the gompit link function.

For instance, if another source names complementary log log (cloglog) the function we call here gompit, one can switch between the two models by reversing the 0s and 1s in the dependent variable (in binary dependent variable models). This can be done without making any changes in data, by changing the encoding of the dependent variable on the Intermediate Inputs dialogue, by choosing the Max(Y) is encoded as 0 option.

Loglog:

Loglog(p) = -Ln(-Ln(p))
Loglog(0.1) = 2.25
Loglog(0.9) = -0.834

In models with binary dependent variable, one can switch between estimating a gompit model and a loglog model by reversing the 0s and 1s in the dependent variable and also reversing the signs of all independent variables (including the constant term).

A Newton-Raphson type maximum likelihood algorithm is employed to minimise the negative of the log likelihood function. The nature of this method implies that a solution (convergence) cannot always be achieved. In such cases, you are advised to edit the convergence parameters provided, in order to find the right levels for the particular problem at hand.

The logarithm of the likelihood function is:

$$L = \sum_{i=1}^{n} \left[r_i \mathrm{Ln}(F_i) + (s_i - r_i)\mathrm{Ln}(1 - F_i) \right]$$

and its first derivative:

$$\frac{\partial \mathrm{Ln(L)}}{\partial \beta_j} = \sum_{i=1}^{n} \left(r_i \frac{G_i}{F_i} + (s_i - r_i) \frac{-G_i}{1 - F_i} \right) x_i \text{, for } j = 1, \ldots, k.$$

where:

r_i is the number of responses,
s_i is the number of subjects,
F_i is the inverse link function and
G_i is the first derivative of F_i.

F_i and G_i are defined for each link function as follows:

Logit:

$$F_i = \frac{\mathrm{Exp}(\beta' x_i)}{1 + \mathrm{Exp}(\beta' x_i)}$$

$$G_i = F_i(1 - F_i)$$

Probit:

Normal cumulative probability function:

$$F_i = \Phi(\beta' x_i)$$

Normal density function:

$$G_i = \frac{1}{\sqrt{2\pi}} \mathrm{Exp}\left(-\frac{1}{2} (\beta' x_i)^2 \right)$$

Gompit (Cloglog):

$$F_i = 1 \text{ - } \mathrm{Exp}(\text{-}\mathrm{Exp}(\beta' x_i))$$

$$G_i = \mathrm{Exp}(\beta' x_i)(1 - F_i)$$

Loglog:

$$F_i = \mathrm{Exp}(\text{-}\mathrm{Exp}(\text{-}\beta' x_i))$$

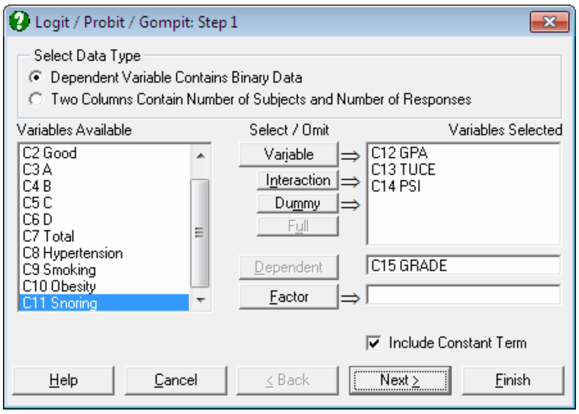
$$G_i = \mathrm{Exp}(\text{-}\beta' x_i)(1 - F_i)$$

With a binary dependent variable $r_i = y_i$ (0 or 1) and $s_i = 1$.

7.2.5.2. Logit / Probit / Gompit Variable Selection

Logit / Probit / Gompit can analyse data in two different formats:

1) Binary (casewise) data where the dependent variable is a binary variable (it consists of two values), and


Logit / Probit / Gompit: Step 1

Select Data Type
● Dependent Variable Contains Binary Data
○ Two Columns Contain Number of Subjects and Number of Responses

Variables Available: C2 Good, C3 A, C4 B, C5 C, C6 D, C7 Total, C8 Hypertension, C9 Smoking, C10 Obesity, C11 Snoring

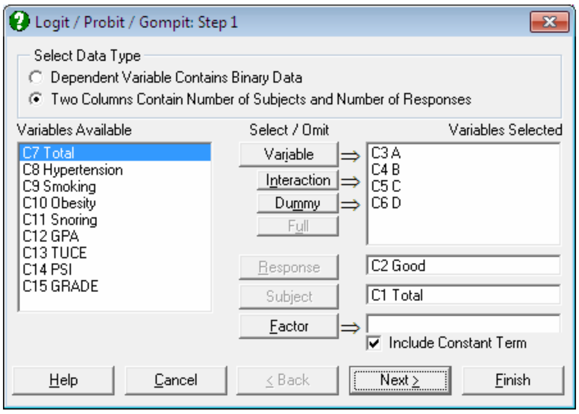
Variables Selected: C12 GPA, C13 TUCE, C14 PSI

Dependent: C15 GRADE

Factor:

☑ Include Constant Term

2) Aggregated (grouped) data, where similar cases are collapsed into groups to generate two columns, one containing the number of responses the other the total number of subjects in the group.


Logit / Probit / Gompit: Step 1

Select Data Type
○ Dependent Variable Contains Binary Data
● Two Columns Contain Number of Subjects and Number of Responses

Variables Available: C7 Total, C8 Hypertension, C9 Smoking, C10 Obesity, C11 Snoring, C12 GPA, C13 TUCE, C14 PSI, C15 GRADE

Variables Selected: C3 A, C4 B, C5 C, C6 D

Response: C2 Good

Subject: C1 Total

Factor:

☑ Include Constant Term

When the first data option is selected, the dependent variable should ideally contain only two distinct values (numeric or string). However, UNISTAT will accept any data column as the dependent variable and then, by default, encode internally those values which are equal to the minimum of this column as 0 and any other values as 1. Alternatively, you can make the program accept the maximum value as 1 and the rest as 0 by changing encoding of the dependent variable on the Intermediate Inputs dialogue by choosing the Max(Y) is encoded as 0 option.

This approach has the advantage and flexibility of running Logit / Probit / Gompit models on columns containing any type of categorical data. For instance, when a logit analysis is run on a column containing years 1995, 1996 and 1997, by default, UNISTAT will internally encode all 1995 entries as 0 and all 1996 and 1997 entries as 1. However, it is left to the user to ensure that the dependent variable selected contains sensible values.

The following is an example for the first data type, where there is one binary dependent variable and one independent variable:

Dependent	Independent
0	1.3
0	2.7
1	2.1
0	2.7
0	1.3
1	2.1
1	2.7
1	1.9
1	1.3
0	2.1
0	2.7
1	1.3
1	2.1
0	1.3
0	2.1
0	1.9

The same data set can be grouped (or collapsed) into the second (aggregated) format as follows:

Responses	Subjects	Independent
2	5	1.3
1	2	1.9
3	5	2.1
1	4	2.7

where the first variable is called the *response* variable (which represents the number of true values within the group), and the second the *subject* variable (which represents the total number of cases in that group).

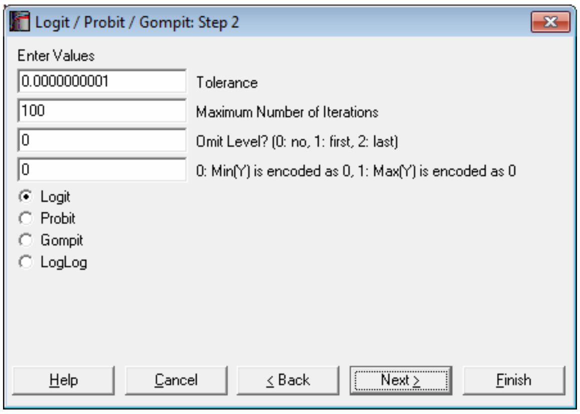
UNISTAT will first ask for the type of the dependent variable. If it is binary as described in (1) above, then select a dependent variable (by clicking on [Dependent]) which contains numeric or string categorical data, and any number of independent variables, which contain numeric data. If the data is in aggregated (or collapsed) from as described in (2) above, then select one column as *Response* (by clicking on [Response]) and one column as *Subjects* (by clicking on [Subject]). The following relation should hold for each case:

0 ≤ Response ≤ Subjects

Cases that do not conform to this will be considered as missing. As in Linear Regression, it is possible to create interaction terms and dummy variables, but not lag/lead terms (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

It is also possible to select a factor (categorical) variable (by clicking on [Factor]) in which case the program will perform the analysis on a sub group as defined by the user (see 7.2.1.1. Linear Regression Variable Selection).

Next, an Intermediate Inputs dialogue will pop up. When the dependent variable is binary, it will look as follows:


Logit / Probit / Gompit: Step 2

Enter Values

0.0000000001 Tolerance

100 Maximum Number of Iterations

0 Omit Level? (0: no, 1: first, 2: last)

0 0: Min(Y) is encoded as 0, 1: Max(Y) is encoded as 0

● Logit
○ Probit
○ Gompit
○ LogLog

Tolerance: This value is used to control the sensitivity of nonlinear minimisation procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

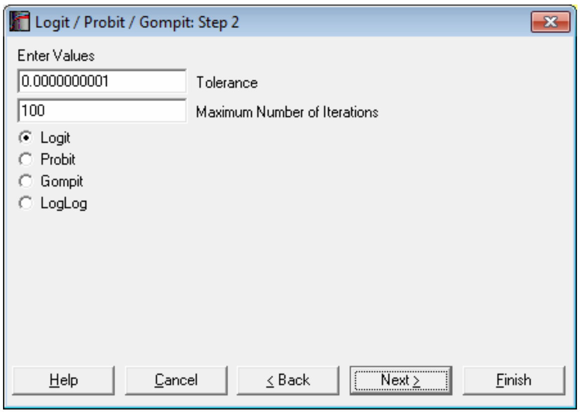
Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Omit Level: This box will appear only when one or more dummy variables have been included in the model from the Variable Selection Dialogue. Three options are available; (0) do not omit any levels, (1) omit the first level and (2) omit the last level. When no levels are omitted, the model will usually be over-parameterised (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

Dependent Variable Encoding: This box will appear only when the binary dependent variable option is selected. If this value is zero, the minimum of dependent variable is internally encoded as zero and any other value as 1. If the box value is nonzero, then the maximum of dependent variable is encoded as zero and any other value as 1.

The Link Function: Select the model to be estimated. It can be one of logit, probit, gompit (cloglog) or loglog (see 7.2.5.1. Logit / Probit / Gompit Model Description).

When the aggregated data option is selected, Step 2 will not have an encoding of the dependent variable option.



Logit / Probit / Gompit: Step 2

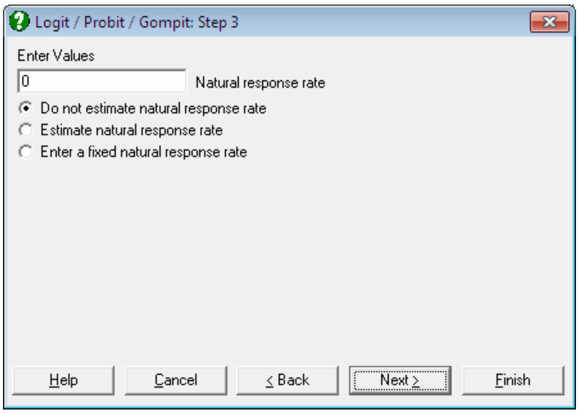
Enter Values

0.0000000001 Tolerance

100 Maximum Number of Iterations

● Logit
○ Probit
○ Gompit
○ LogLog

With the aggregated data option, a further dialogue is also displayed, asking whether a natural response rate is to be estimated or a fixed one will be given by the user. When a natural response rate is estimated, it will appear in the output just like any other estimated coefficient.



Logit / Probit / Gompit: Step 3

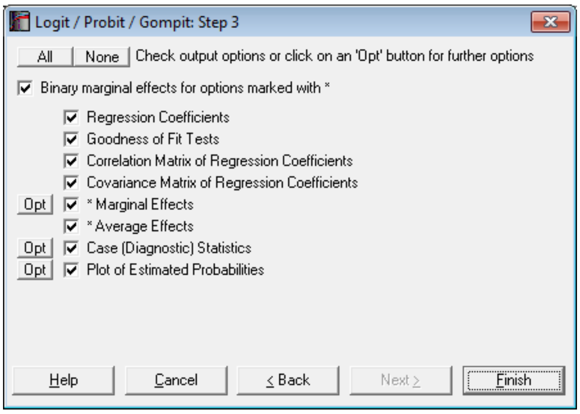
Enter Values

0 Natural response rate

● Do not estimate natural response rate
○ Estimate natural response rate
○ Enter a fixed natural response rate

7.2.5.3. Logit / Probit / Gompit Output Options

For a model with binary dependent variable, the following output options dialogue pops up.



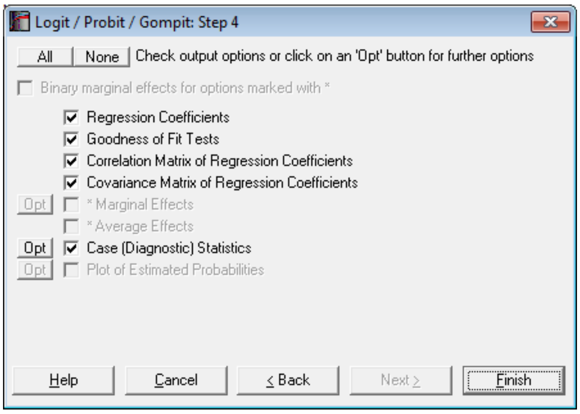
Logit / Probit / Gompit: Step 3

All None Check output options or click on an 'Opt' button for further options

☑ Binary marginal effects for options marked with *

☑ Regression Coefficients
☑ Goodness of Fit Tests
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
Opt ☑ * Marginal Effects
☑ * Average Effects
Opt ☑ Case (Diagnostic) Statistics
Opt ☑ Plot of Estimated Probabilities

If the number of subjects and responses are given, then some of the output options will not be available.



Check output options or click on an 'Opt' button for further options

□ Binary marginal effects for options marked with *

☑ Regression Coefficients
☑ Goodness of Fit Tests
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
□ * Marginal Effects
□ * Average Effects
☑ Case (Diagnostic) Statistics
□ Plot of Estimated Probabilities

Regression Coefficients:

The Z statistic is defined as:

$$Z_j = \frac{\beta_j}{\sigma_j}, j = 1, \ldots, k$$

The two-tailed normal probability value is:

$$p_j = 2\Phi(\text{Abs}(Z_j)), j = 1, \ldots, k$$

The confidence intervals for regression coefficients are computed from:

$$\beta_j \pm Z_{\alpha/2}\sigma_j, j = 1, \ldots, k$$

where each coefficient's standard error, σ_j, is the square root of the diagonal element of the covariance matrix.

Goodness of Fit Tests: See 7.2.6.4.1. Logistic Regression Results for details.

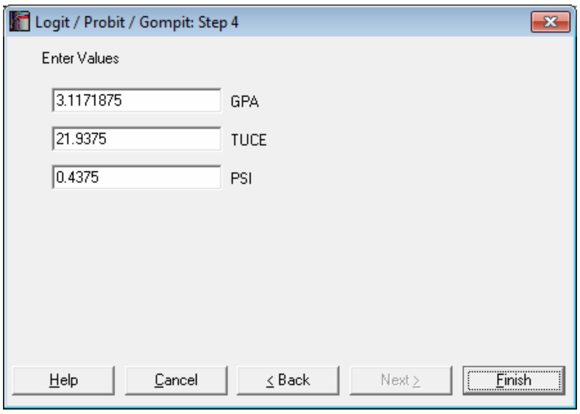
Correlation Matrix for Regression Coefficients: Correlations between the estimated coefficients are displayed.

Covariance Matrix for Regression Coefficients: Diagonal elements are the coefficient variances and off diagonal elements are the covariances between coefficients.

Marginal Effects: This option is available for only binary dependent variable models. Marginal effects measure the change in the estimated probability when a change is made in independent variables. In other words, it is the partial derivative of the prediction function with respect to x:

$$\mathrm{Marginal\,Effect}_j = \beta_j G_j$$

i.e., the estimated coefficient times its first derivative (slope) at x. You can enter values for the x-vector by clicking the [Opt] button situated to the left Marginal Effects output option. By default, the values displayed are the means of independent variables.


Logit / Probit / Gompit: Step 4

Enter Values

GPA: 3.1171875

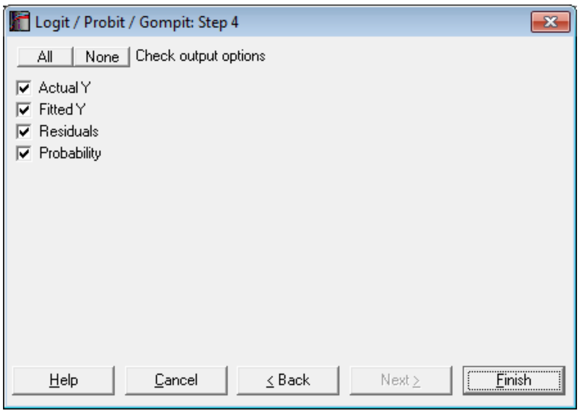
TUCE: 21.9375

PSI: 0.4375

For dummy variables, or in general when an independent variable consists only of 0s and 1s, a slightly different procedure can be used, by checking the box Binary marginal effects for options marked with *. For details see p. 733, Greene (2012).

Average Effects: This option is available for only binary dependent variable models. Average effects are the sample mean of marginal effects computed for each case of the data.

Case (Diagnostic) Statistics: Observed and expected responses, their differences and the expected probabilities are displayed. In models with a binary dependent variable, predictions are made for those cases where only the dependent variable is missing and no independent variables are missing. For further information see 7.2.6.4.2. Logistic Regression Case (Diagnostic) Statistics.



Check output options

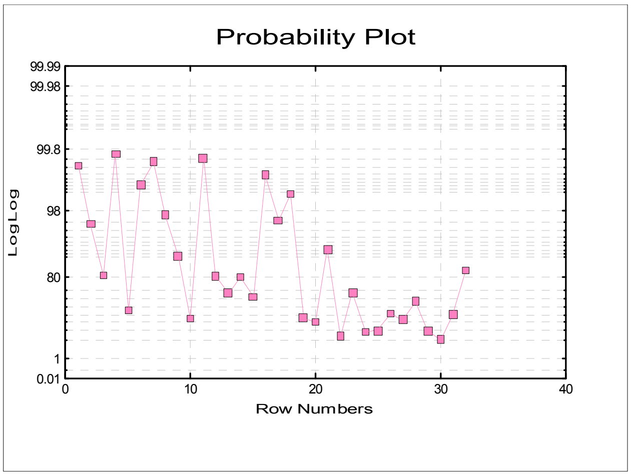
☑ Actual Y

☑ Fitted Y

☑ Residuals

☑ Probability

Plot of Estimated Probabilities: This option is available for only binary dependent variable models. The estimated probabilities are plotted against the row numbers, using the appropriate link function for the scaling of the Left-Y axis.



7.2.5.4. Logit / Probit / Gompit Examples

Example 1

Table 12.19 on p. 353 from Altman & Douglas (1991). Open LOGIT, select Statistics 1 → Regression Analysis → Logit / Probit / Gompit and select the data option Two Columns Contain Number of Subjects and Number of

Responses. Then select *Total* (*C7*) as [Subject], *Hypertension* (*C8*) as [Response] and *Smoking*, *Obesity* and *Snoring* (*C9* to *C11*) as [Variable]s. Select only the Regression Results output option to obtain the following results:

Logit / Probit / Gompit

Model selected: Logit
Response Variable: Hypertension
Subject Variable: Total
Valid Number of Cases: 8, 0 Omitted

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Constant	-2.3777	0.3802	-6.2540	0.0000	-3.1228	-1.6325
Smoking	-0.0678	0.2781	-0.2437	0.8075	-0.6129	0.4773
Obesity	0.6953	0.2851	2.4390	0.0147	0.1366	1.2541
Snoring	0.8719	0.3976	2.1932	0.0283	0.0927	1.6512

Goodness of Fit Tests

	-2 Log likelihood
Initial Model	411.4239
Final Model	398.9164

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	1.3643	4	0.8504
Likelihood Ratio	12.5075	3	0.0058

	Pseudo R-squared
McFadden	0.0304
Adjusted McFadden	0.0110
Cox & Snell	0.0285
Nagelkerke	0.0464

Go back to Variable Selection Dialogue, omit *Smoking* (*C9*) from the independent variable list and run the analysis again.

Logit / Probit / Gompit

Model selected: Logit
Response Variable: Hypertension
Subject Variable: Total
Valid Number of Cases: 8, 0 Omitted

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Constant	-2.3921	0.3757	-6.3662	0.0000	-3.1285	-1.6556
Obesity	0.6954	0.2851	2.4395	0.0147	0.1367	1.2541
Snoring	0.8655	0.3967	2.1819	0.0291	0.0880	1.6429

Goodness of Fit Tests

	-2 Log likelihood
Initial Model	411.4239
Final Model	398.9761

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	1.3854	5	0.9259
Likelihood Ratio	12.4478	2	0.0020

	Pseudo R-squared
McFadden	0.0303
Adjusted McFadden	0.0157
Cox & Snell	0.0283
Nagelkerke	0.0462

Example 2

Example 14.1 on p. 490 from Armitage & Berry (2002). Data given in Table 14.1 needs to be transformed into a suitable format where the main effects of the four factors *A*, *B*, *C* and *D* can be analysed. This is done by creating a new column for each factor such that it contains the value one if the factor occurs in the factor combination column and zero otherwise. The data matrix would then look like this:

Total	Good	A	B	C	D
477	84	0	0	0	0
231	75	1	0	0	0
63	13	0	1	0	0
94	35	1	1	0	0
150	67	0	0	1	0
378	201	1	0	1	0
32	16	0	1	1	0
169	102	1	1	1	0
12	2	0	0	0	1
13	7	1	0	0	1
7	4	0	1	0	1
12	8	1	1	0	1
11	3	0	0	1	1
45	27	1	0	1	1
4	1	0	1	1	1
31	23	1	1	1	1

Open LOGIT, select Statistics 1 → Regression Analysis → Logit / Probit / Gompit and select the data option Two Columns Contain Number of Subjects and Number of Responses. Then select *Total* (*C1*) as [Subject], *Good* (*C2*) as [Response] and *A*, *B*, *C*, *D* (*C3* to *C6*) as [Variable]s. Select all output options for the following results:

Logit / Probit / Gompit

Model selected: Logit
Response Variable: Good
Subject Variable: Total
Valid Number of Cases: 16, 0 Omitted

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Constant	-1.4604	0.0964	-15.1490	0.0000	-1.6494	-1.2715
A	0.6498	0.1154	5.6298	0.0000	0.4236	0.8760
B	0.3101	0.1222	2.5377	0.0112	0.0706	0.5496
C	0.9806	0.1107	8.8560	0.0000	0.7636	1.1976
D	0.4204	0.1910	2.2011	0.0277	0.0461	0.7946

Goodness of Fit Tests

	-2 Log likelihood
Initial Model	2306.7889
Final Model	2104.1204

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	13.6067	11	0.2555
Likelihood Ratio	202.6685	4	0.0000

	Pseudo R-squared
McFadden	0.0879
Adjusted McFadden	0.0835
Cox & Snell	0.1106
Nagelkerke	0.1106

Correlation Matrix of Regression Coefficients

	Constant	A	B	C	D
Constant	1.0000	-0.5095	-0.1961	-0.3929	-0.0716
A	-0.5095	1.0000	-0.1534	-0.2952	-0.0430
B	-0.1961	-0.1534	1.0000	-0.0014	-0.0810
C	-0.3929	-0.2952	-0.0014	1.0000	-0.0569
D	-0.0716	-0.0430	-0.0810	-0.0569	1.0000

Covariance Matrix of Regression Coefficients

	Constant	A	B	C	D
Constant	0.0093	-0.0057	-0.0023	-0.0042	-0.0013
A	-0.0057	0.0133	-0.0022	-0.0038	-0.0009
B	-0.0023	-0.0022	0.0149	-0.0000	-0.0019
C	-0.0042	-0.0038	-0.0000	0.0123	-0.0012
D	-0.0013	-0.0009	-0.0019	-0.0012	0.0365

Case (Diagnostic) Statistics

	Subjects	Responses	Expected Responses	Residuals	Probability
1	477	84	89.8673	-5.8673	0.1884
2	231	75	71.0915	3.9085	0.3078
3	63	13	15.1470	-2.1470	0.2404
4	94	35	35.4771	-0.4771	0.3774
5	150	67	57.3442	9.6558	0.3823
6	378	201	205.0245	-4.0245	0.5424
7	32	16	14.6455	1.3545	0.4577
8	169	102	104.4028	-2.4028	0.6178
9	12	2	3.1336	-1.1336	0.2611
10	13	7	5.2474	1.7526	0.4036
11	7	4	2.2764	1.7236	0.3252
12	12	8	5.7596	2.2404	0.4800
13	11	3	5.3365	-2.3365	0.4851
14	45	27	28.9549	-1.9549	0.6434
15	4	1	2.2493	-1.2493	0.5623
16	31	23	22.0422	0.9578	0.7110

Next go back to the Variable Selection Dialogue and check the Probit option. Select only the Regression Results output option.

Logit / Probit / Gompit

Model selected: Probit
Response Variable: Good
Subject Variable: Total
Valid Number of Cases: 16, 0 Omitted

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Constant	-0.8933	0.0561	-15.9286	0.0000	-1.0032	-0.7833
A	0.3963	0.0698	5.6740	0.0000	0.2594	0.5332
B	0.1890	0.0749	2.5238	0.0116	0.0422	0.3359
C	0.6027	0.0675	8.9292	0.0000	0.4704	0.7350
D	0.2584	0.1169	2.2106	0.0271	0.0293	0.4876

Example 3

Example 17.3 on p. 735 Greene (2012). Tables 17.1 and 17.2 display results for three link functions. As these tables contain some misprints, the user is advised to validate the results with the book's errata website.

Open LOGIT, select Statistics 1 → Regression Analysis → Logit / Probit / Gompit and select the data option Dependent Variable Contains Binary Data. Then select *GRADE* (*C15*) as [Dependent] and *GPA*, *TUCE* and *PSI* (*C12* to *C14*) as [Variable]s. Select Logit and all output options.

Logit / Probit / Gompit

Model selected: Logit
Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Constant	-13.0213	4.9313	-2.6405	0.0083	-22.6866	-3.3561
GPA	2.8261	1.2629	2.2377	0.0252	0.3508	5.3014
TUCE	0.0952	0.1416	0.6722	0.5014	-0.1823	0.3726
PSI	2.3787	1.0646	2.2344	0.0255	0.2922	4.4652

Goodness of Fit Tests

	-2 Log likelihood
Initial Model	41.1835
Final Model	25.7793

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	27.2571	27	0.4500
Likelihood Ratio	15.4042	3	0.0015

	Pseudo R-squared
McFadden	0.3740
Adjusted McFadden	0.1798
Cox & Snell	0.3821
Nagelkerke	0.5278

Correlation Matrix of Regression Coefficients

	Constant	GPA	TUCE	PSI
Constant	1.0000	-0.7343	-0.4960	-0.4494
GPA	-0.7343	1.0000	-0.2065	0.3181
TUCE	-0.4960	-0.2065	1.0000	0.0990
PSI	-0.4494	0.3181	0.0990	1.0000

Covariance Matrix of Regression Coefficients

	Constant	GPA	TUCE	PSI
Constant	24.3180	-4.5735	-0.3463	-2.3592
GPA	-4.5735	1.5950	-0.0369	0.4276
TUCE	-0.3463	-0.0369	0.0200	0.0149
PSI	-2.3592	0.4276	0.0149	1.1333

Marginal Effects

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%	X
GPA	0.5339	0.2370	2.2522	0.0243	0.0693	0.9984	3.1172
TUCE	0.0180	0.0262	0.6851	0.4933	-0.0334	0.0694	21.9375
*** PSI**	0.4565	0.1811	2.5213	0.0117	0.1016	0.8114	0.4375

x'B =	-1.0836
Predicted Probability =	0.2528
f*x'B =	-0.2047
Predicted Marginal Probability =	0.4490

* Binary Independent Variable

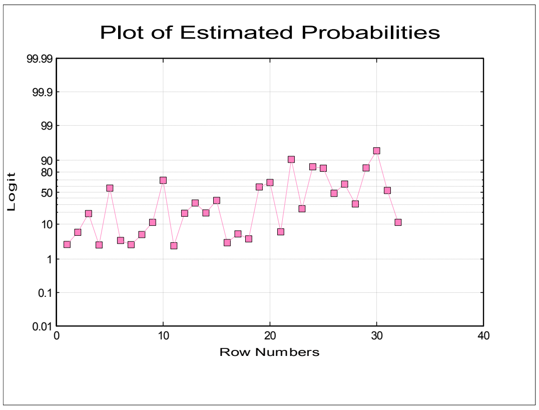
Average Effects

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
GPA	0.3626	0.1094	3.3130	0.0009	0.1481	0.5771
TUCE	0.0122	0.0178	0.6861	0.4927	-0.0227	0.0471
*** PSI**	0.3575	0.1420	2.5177	0.0118	0.0792	0.6358

* Binary Independent Variable

Case (Diagnostic) Statistics

	Actual Y	Fitted Y	Residuals	Probability
1	0	0.0266	-0.0266	0.0266
2	0	0.0595	-0.0595	0.0595
3	0	0.1873	-0.1873	0.1873
…	…	…	…	…
30	1	0.9453	0.0547	0.9453
31	0	0.5291	-0.5291	0.5291
32	1	0.1110	0.8890	0.1110



Now select Probit and Gompit link functions with only the Regression Results, Marginal Effects output options.

Logit / Probit / Gompit

Model selected: Probit
Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Constant	-7.4523	2.5425	-2.9311	0.0034	-12.4355	-2.4692
GPA	1.6258	0.6939	2.3431	0.0191	0.2658	2.9858
TUCE	0.0517	0.0839	0.6166	0.5375	-0.1127	0.2162
PSI	1.4263	0.5950	2.3970	0.0165	0.2601	2.5926

Marginal Effects

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%	X
GPA	0.5333	0.2325	2.2943	0.0218	0.0777	0.9890	3.1172
TUCE	0.0170	0.0271	0.6257	0.5315	-0.0362	0.0701	21.9375
*** PSI**	0.4644	0.1703	2.7274	0.0064	0.1307	0.7982	0.4375

x'B =	-0.6255
Predicted Probability =	0.2658
f*x'B =	-0.2052
Predicted Marginal Probability =	0.4187

* Binary Independent Variable

Logit / Probit / Gompit

Model selected: Gompit
Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Regression Results

	Coefficient	Standard	Z-	2-Tail	Lower	Upper
		Error	Statistic	Probability	95%	95%
Constant	-10.0314	3.4791	-2.8834	0.0039	-16.8502	-3.2126
GPA	2.2936	1.0350	2.2160	0.0267	0.2650	4.3221
TUCE	0.0412	0.1073	0.3835	0.7013	-0.1692	0.2515
PSI	1.5623	0.7305	2.1386	0.0325	0.1305	2.9940

Marginal Effects

	Coeffici	Standar	Z-	2-Tail	Lower	Upper	X
	ent	d Error	Statistic	Probability	95%	95%	
GPA	0.4775	0.2031	2.3509	0.0187	0.0794	0.8755	3.1172
TUCE	0.0086	0.0222	0.3865	0.6991	-0.0349	0.0520	21.9375
*** PSI**	0.3536	0.1610	2.1958	0.0281	0.0380	0.6693	0.4375

x'B =	-1.2956
Predicted Probability =	0.2395
f*x'B =	-0.2697
Predicted Marginal Probability =	0.5340

* Binary Independent Variable

Now select Probit and Gompit link functions, uncheck the Binary marginal effects box and only the Marginal Effects output option.

Logit / Probit / Gompit

Model selected: Logit
Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Marginal Effects

	Coeffici	Standard	Z-	2-Tail	Lower	Upper	X
	ent	Error	Statistic	Prob	95%	95%	
GPA	0.5339	0.2370	2.2522	0.0243	0.0693	0.9984	3.1172
TUCE	0.0180	0.0262	0.6851	0.4933	-0.0334	0.0694	21.9375
PSI	0.4493	0.1968	2.2837	0.0224	0.0637	0.8350	0.4375

x'B =	-1.0836
Predicted Probability =	0.2528
f*x'B =	-0.2047
Predicted Marginal Probability =	0.4490

Logit / Probit / Gompit

Model selected: Probit
Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Marginal Effects

	Coefficient	Standard Error	Z-Statistic	2-Tail Prob	Lower 95%	Upper 95%	X
GPA	0.5333	0.2325	2.2943	0.0218	0.0777	0.9890	3.1172
TUCE	0.0170	0.0271	0.6257	0.5315	-0.0362	0.0701	21.9375
PSI	0.4679	0.1876	2.4936	0.0126	0.1001	0.8357	0.4375

x'B =	-0.6255
Predicted Probability =	0.2658
f*x'B =	-0.2052
Predicted Marginal Probability =	0.4187

7.2.6. Logistic Regression

The Logistic Regression procedure is suitable for estimating Linear Regression models when the dependent variable is a binary (or dichotomous) variable, that is, it consists of two values such as *Yes* or *No*, or in general 0 and 1. In such cases, where the dependent variable has an underlying binomial distribution (and thus the predicted Y values should lie between 0 and 1) the Linear Regression procedure cannot be employed.

Like Linear Regression, Logistic Regression can be used to estimate models with or without a constant term and regressions may be run on a subset of cases as determined by the levels of an unlimited number of factor columns. An unlimited number of dependent variables (numeric or string) can be selected in order to run the same model on different dependent variables. It is also possible to include interaction terms, dummy and lag/lead variables in the model, without having to create them as spreadsheet columns first.

Logistic Regression is closely related to Logit / Probit / Gompit. For a brief discussion of similarities and differences of these two procedures see 7.2.5. Logit / Probit / Gompit.

As of this version of UNISTAT, a comprehensive implementation of ROC (Receiver Operating Characteristic) analysis is included in the Logistic Regression procedure. The two output options Classification by Group and ROC Analysis, as well as the two graphics options, will provide a complete ROC analysis output. It is possible to compute AUC (area under the curve) and plot ROC curves with covariates and plot multiple ROC curves with multiple comparisons between AUCs.

7.2.6.1. Logistic Regression Model Description

Logistic Regression employs the logit model as explained in Logit / Probit / Gompit (see 7.2.5.1. Logit / Probit / Gompit Model Description). However, the log of likelihood function for the logistic model can be expressed more explicitly as:

$$L = \sum_{i=1}^{n} \left[y_i \mathrm{Ln}(\mu_i) + (1 - y_i) \mathrm{Ln}(1 - \mu_i) \right]$$

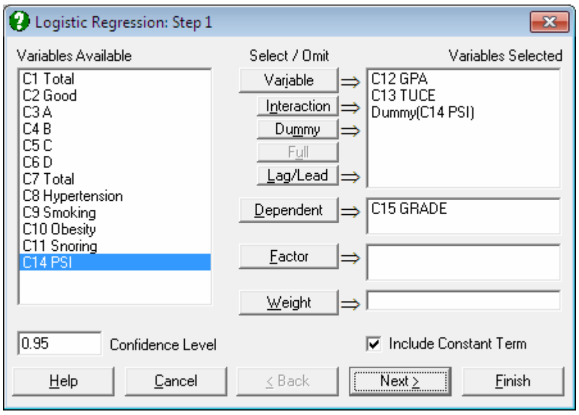
with first derivatives:

$$\frac{\partial \mathrm{Ln(L)}}{\partial \beta_j} = \sum_{i=1}^{n} (y_{\,i} - \mu_i)\; x_i,\ \text{for } j = 1, \ldots, m.$$

where:

$$\mu_i = \frac{\mathrm{Exp}(\beta^{'} x_i)}{1 + \mathrm{Exp}(\beta^{'} x_i)}$$

7.2.6.2. Logistic Regression Variable Selection



Logistic Regression: Step 1

Variables Available

C1 Total
C2 Good
C3 A
C4 B
C5 C
C6 D
C7 Total
C8 Hypertension
C9 Smoking
C10 Obesity
C11 Snoring
C14 PSI

Variables Selected

C12 GPA
C13 TUCE
Dummy(C14 PSI)

Dependent: C15 GRADE

Factor:

Weight:

0.95 Confidence Level

☑ Include Constant Term

As in Linear Regression, it is possible to create interaction terms, dummy variables, lag/lead terms, select multiple dependent variables (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables) and run regressions on subsamples defined by several factor columns with or without weights (see 7.2.1.1. Linear Regression Variable Selection).

It is compulsory to select at least one column containing numeric or String Data as a dependent variable. The program encodes the dependent variable internally such that, by default, the minimum value that occurs in the column is 0 and the rest are 1. It is possible to reverse this condition and encode the maximum of the dependent variable as 0 and the rest of the values as 1 using the Dependent Variable Encoding control in the Intermediate Inputs dialogue.

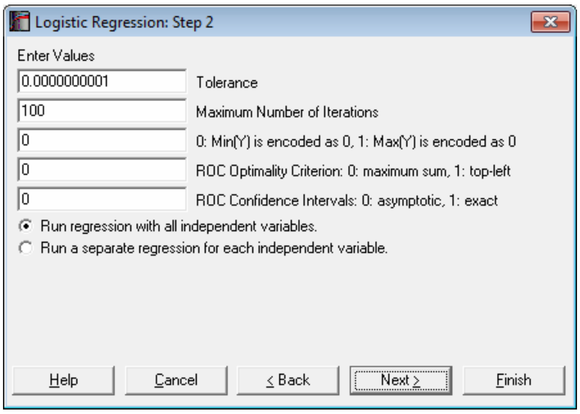
In case a categorical variable is not selected as the dependent variable, there may be too few 0s and too many 1s in the encoded dependent variable and a convergence may not be achieved.

When more than one dependent variable is selected, the analysis will be repeated as many times as the number of dependent variables, each time only changing the dependent variable and keeping the rest of selections unchanged.

When more than one independent variable is selected, you will be presented with the option to run a single analysis (see 7.2.6.3. Logistic Regression Intermediate Inputs) including all independent variables (which is the default case in earlier versions of UNISTAT) or to run a separate regression for each independent variable, while holding the dependent variable unchanged. The primary use of this option is to compare the areas enclosed under the ROC curves for each independent variable.

A column containing numeric data can be selected as a weights column. Unlike the Linear Regression procedure, however, weights here are frequency weights. All independent variables are multiplied by this column internally by the program.

7.2.6.3. Logistic Regression Intermediate Inputs



0.0000000001 Tolerance

100 Maximum Number of Iterations

0 0: Min(Y) is encoded as 0, 1: Max(Y) is encoded as 0

0 ROC Optimality Criterion: 0: maximum sum, 1: top-left

0 ROC Confidence Intervals: 0: asymptotic, 1: exact

● Run regression with all independent variables.

○ Run a separate regression for each independent variable.

The number and kind of controls that appear on this dialogue depend on the selections made in the previous dialogue. For instance, if a dummy or lag variable was created, the dialogue will display one or more other boxes (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables). The specific tasks of these controls are as follows:

Tolerance: This value is used to control the sensitivity of nonlinear minimisation procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Dependent Variable Encoding: By default, the program will internally encode the dependent variable values such that the minimum is 0 and the rest of the values are 1. If one is entered into this box, the program will encode the maximum value as 0 and the rest as 1.

Making a change in this control will normally reverse the signs of the estimated coefficients and will affect other output as well. If your aim in changing this control is to obtain the correct 2 x 2 table where the maximum value of the dependent variable corresponds to a positive state, e.g. the presence of a disease or *vice versa*, this can be done in the Output Options Dialogue more efficiently, without affecting the estimated coefficients and other output.

ROC Optimality Criterion: In Logistic Regression, the fitted Y is a continuous variable consisting of probability values. The estimated group membership (in terms of 0 and 1) is dependent on a critical cut-off probability which is also called the Classification threshold probability. The estimated group membership is 0 for any case with an estimated probability (fitted Y value) less than this critical probability and it is 1 otherwise.

In earlier versions of UNISTAT, the default value of Classification threshold probability was fixed at 0.5 and the user was allowed to change this value manually to play different what-if scenarios. As of this version of UNISTAT, the Classification threshold probability is estimated by the program using one of the following two methods:

Maximum sum of sensitivity and specificity: This is also known as the Youden's index and represents the point on the curve furthest away from the 45° line. It is defined as:

Max(Sensitivity + Specificity)

Point nearest to the top-left corner of ROC plot: This is given as:

Min((1 - Sensitivity)^2 + (1- Specificity)^2)

The estimated Classification threshold probability can be edited to observe the effect of different cut-off values on the 2 x 2 Table and Statistics For Diagnostic Tests output..

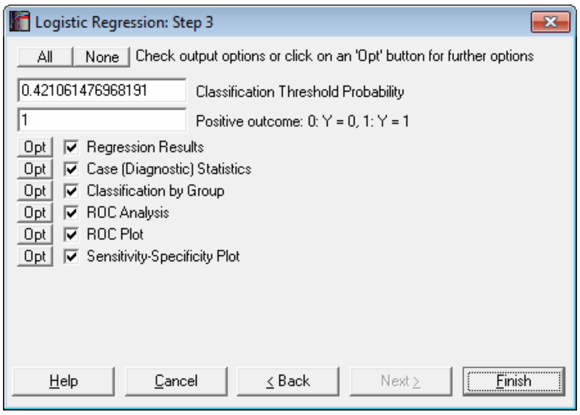
ROC Confidence Intervals: The ROC Table output option (see 7.2.6.4.4. ROC Analysis) can display all cases for a large number of test statistics together with their confidence intervals. Here you can choose the type of confidence intervals as:

0: Asymptotic normal (Wald), or
1: Exact binomial (Clopper-Pearson).

Run regression with all independent variables: This is the default option and produces one set of output with all independent variables included as in earlier versions of UNISTAT.

Run a separate regression for each independent variable: As of this version of UNISTAT, it is possible to run a separate regression for each independent variable, while holding the dependent variable unchanged. The primary use of this option is to compare the areas enclosed under the ROC curves for each independent variable.

7.2.6.4. Logistic Regression Output Options



Check output options or click on an 'Opt' button for further options

0.421061476968191 Classification Threshold Probability

1 Positive outcome: 0: Y = 0, 1: Y = 1

☑ Regression Results
☑ Case (Diagnostic) Statistics
☑ Classification by Group
☑ ROC Analysis
☑ ROC Plot
☑ Sensitivity-Specificity Plot

The Sensitivity-Specificity Plot option will not be available when Run a separate regression for each independent variable option is selected in the Intermediate Inputs dialogue.

Classification Threshold Probability: As described above for ROC Optimality Criterion, the estimated Classification threshold probability can be edited to try different scenarios. Changing this value will affect the output in

Classification by Group and the Predicted Group and misclassifications in the Case (Diagnostic) Statistics output.

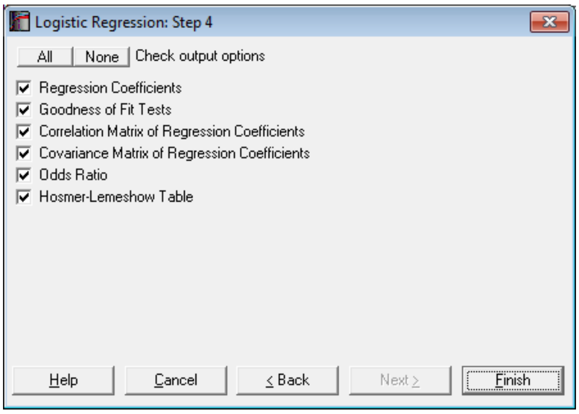
Positive Outcome: When calculating the Statistics For Diagnostic Tests output (e.g. sensitivity, specificity), UNISTAT assumes that the positive outcome is represented by 1 in the dependent variable and the true positive outcome of the test is represented in cell (1,1) of the 2 x 2 Table. Here you can control which value of the dependent variables represents the positive outcome, without affecting the rest of the Logistic Regression output.

Multicollinearity: Variables causing multicollinearity will be displayed with a zero coefficient at the end of the coefficients table. If you do not wish to display these variables enter the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
DispCollin=0
```

The rest of the coefficients will be determined as if the regression were run without the variables causing collinearity.

7.2.6.4.1. Logistic Regression Results



Check output options

☑ Regression Coefficients
☑ Goodness of Fit Tests
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
☑ Odds Ratio
☑ Hosmer-Lemeshow Table

The main regression output displays a table for coefficients of the estimated regression equation, their standard errors, Wald statistics, probability values and confidence intervals for the significance level specified in the Variable Selection Dialogue. If any independent variables have been omitted due to multicollinearity, they are reported at the end of the table with a zero coefficient.

Regression Coefficients:

The Wald statistic is defined as:

$$W_j = \frac{\beta_j^2}{\sigma_j^2}, j = 1, \ldots, m.$$

and has a chi-square distribution with one degree of freedom.

The confidence intervals for regression coefficients are computed from:

$$\beta_j \pm Z_{\alpha/2}\sigma_j, j = 1, \ldots, m,$$

where each coefficient's standard error, σ_j, is the square root of the diagonal element of the covariance matrix.

Goodness of Fit Tests:

-2 Log-Likelihood for Initial Model: This is -2 times the value when all independent variables are excluded from the model:

$$\beta_j = 0, j = 1, \ldots, m.$$

-2 Log-Likelihood for Final Model: This is -2 times the value of the log likelihood function when convergence is achieved.

Likelihood Ratio: This is a test statistic for the null hypothesis that "all regression coefficients for covariates are zero". It is equal to -2 times the difference between the initial and final model likelihood values and has a chi-square distribution with k degrees of freedom (the number of independent variables in the model).

Goodness of Fit: This is also known as Pearson's chi-square statistic and is for the observed versus expected number of responses. It has a chi-square distribution with n - k degrees of freedom (the number of valid cases minus the number of independent variables, including the constant term, if any).

$$C = 2\sum_{i=1}^{n} y_i \mathrm{Ln}\left(\frac{y_i}{\lambda_i}\right)$$

Hosmer-Lemeshow Test: This is a test for lack of fit. The observations are sorted according to their fitted Y values (estimated probabilities) in ascending order. The identical cases of independent variables are formed into blocks.

Then the cases are grouped into approximately ten classes without splitting the blocks.

The test statistic is defined as:

$$X^2_{HL} = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j(1 - E_j / n_j)}$$

with g - 2 degrees of freedom, where:

g is the number of classes,
n$_j$ is the number of observations in the jth class,
O$_j$ is the observed number of cases in the jth class,
E$_j$ is the expected number of cases in the jth class.

Pseudo R-squared: In Logistic Regression (as well as in other maximum likelihood procedures), an R-squared statistic as in Linear Regression is not available. This is because Logistic Regression employs an iterative maximum likelihood estimation method. Equivalent statistics to test the goodness of fit have been proposed using the initial (L_0) and maximum (L_1) likelihood values.

McFadden:

$$R^2_{McF} = 1 - \left(\frac{L_1}{L_0} \right)$$

Adjusted McFadden:

$$R^2_{AdjMcF} = 1 - \left(\frac{L_1 - m}{L_0} \right)$$

Cox & Snell:

$$R^2_{CS} = 1 - \left(\frac{L_1}{L_0} \right)^{\left(\frac{2}{n}\right)}$$

Nagelkerke:

$$R^2_N = \frac{R^2_{CS}}{1 - L_0^{\left(\frac{2}{n}\right)}}$$

Correlation Matrix of Regression Coefficients: This is a symmetric matrix with unity diagonal elements. The off-diagonal elements give correlations between regression coefficients.

Covariance Matrix of Regression Coefficients: This is a symmetric matrix where the square roots of the diagonal elements are the parameter standard errors. The off-diagonal elements are covariances between the regression coefficients.

Odds Ratio: Values of the odds ratio indicate the influence of one unit change in a covariate on the regression. It is defined as:

$$\mathrm{Exp}(\beta_j),\ j = 1, \ldots, m.$$

The standard error of the odds ratio is found as:

$$\sigma_j \mathrm{Exp}(\beta_j)$$

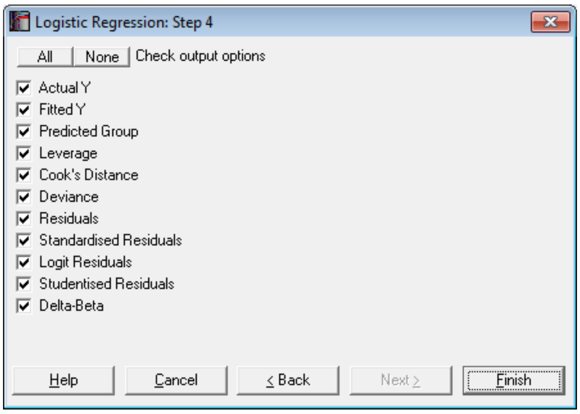
where σ_j is the ith coefficient's standard error, and its confidence intervals as:

$$\mathrm{Exp}(\beta_j \pm Z_{\alpha/2}\sigma_j)$$

which are simply the exponential of the coefficient confidence intervals.

Hosmer-Lemeshow Table: The contingency table described above in Hosmer-Lemeshow Test is displayed. The observed and expected values for both values of the independent variable are listed for all classes.

7.2.6.4.2. Logistic Regression Case (Diagnostic) Statistics



Check output options

☑ Actual Y
☑ Fitted Y
☑ Predicted Group
☑ Leverage
☑ Cook's Distance
☑ Deviance
☑ Residuals
☑ Standardised Residuals
☑ Logit Residuals
☑ Studentised Residuals
☑ Delta-Beta

Case statistics are useful to determine the influence of individual observations on the overall fit of the model. For further information see 7.2.1.2.2. Linear Regression Case Output.

Predictions (Interpolations): There are three conditions under which predictions will be computed for estimated Y values:

1) If, for a case, all independent variables are non-missing, but only the dependent variable is missing,
2) if a case does not contain missing values but it has been omitted from the analysis by Data Processor's Data → Select Row function and
3) if a case does not contain missing values but it has been omitted from the analysis by selecting subsamples from the Variable Selection Dialogue (see 2.1.2. Categorical Data Analysis).

Such cases are not included in the estimation of the model. When, however, Case (Diagnostic) Statistics option is selected, the program will detect these cases and compute and display the fitted (estimated) Y values, as well as their confidence intervals and some other related statistics. Therefore, it will be a good idea to include the cases for which predictions are to be made in the data matrix during the data preparation phase, and then exclude them from the analysis by one of the above three methods. When a case is predicted, its label will be prefixed by an asterisk (*).

In Stand-Alone Mode, the spreadsheet function **Reg** can also be used to make predictions (see 3.4.2.6.3. UNISTAT Functions).

This will give the logit of predicted values, which can be transformed back as:

$$\mu_i = \frac{\mathrm{Exp}(\hat{y}_i)}{1 + \mathrm{Exp}(\hat{y}_i)}$$

Statistics available under Case (Diagnostic) Statistics option are as follows.

Case Labels: If row labels exist in data, they are displayed as case labels. Otherwise the row numbers are displayed. If a fitted Y value is predicted (see Fitted Y Values below) its label is marked by an asterisk (*). The misclassified cases (i.e. where an actual Y value differs from a predicted group) are divided into two groups as False Positive and False Negative and are marked by F+ and F- respectively.

Actual Y: Encoded values of the dependent variable (y_i) are displayed.

Fitted Y: These are the estimated values for the dependent variable.

$$\mu_i = \frac{\text{Exp}(\beta' x_i)}{1 + \text{Exp}(\beta' x_i)}, i = 1, \ldots, n$$

where:

$$\beta' x_i = \sum_{j=1}^{m} \beta_j x_{ji}$$

If, for a case, all dependent variables are nonmissing, but the dependent variable is missing, the fitted (i.e. predicted) Y value will be displayed (see 7.2.6. Logistic Regression). Such cases are marked by a single asterisk (*) in their label.

Predicted Group: If $\mu_i \geq$ cut-off (classification threshold) probability, then the group is 1, and 0 otherwise.

Leverage:

$$h_i = \hat{V}^{1/2} X_i (X' \hat{V} X)^{-1} X_i' \hat{V}^{1/2}$$

where the vector:

$$\hat{V} = \text{Diag}\{\mu_i (1 - \mu_i)\}$$

Cook's Distance:

$$D_i = \frac{z_i^2 h_i}{1 - h_i}$$

where z_i^2 is the standardised residual as defined below.

Deviance:

$G_i = d_i$ if $y_i > \mu_i$ and $G_i = -d_i$ otherwise, where:

$$d_i = \sqrt{-2\left(y_i \text{Ln}(y_i) + (1 - y_i)\text{Ln}(1 - y_i)\right)}$$

Residuals:

$$e_i = y_i - \mu_i$$

Standardised Residuals:

$$z_i = \frac{e_i}{\sqrt{\mu_i(1-\mu_i)}}$$

Logit Residuals:

$$\tilde{e}_i = \frac{e_i}{\mu_i(1-\mu_i)}$$

Studentised Residuals:

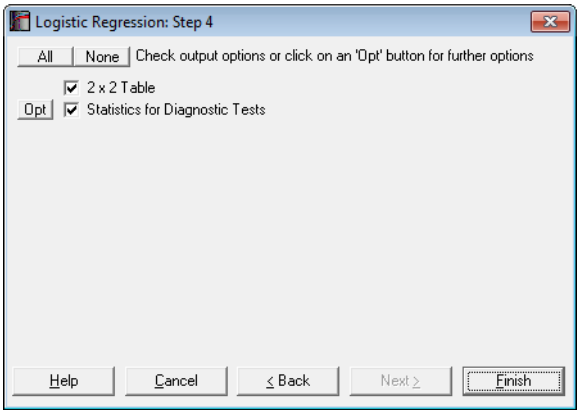
$$G_i^* = \frac{G_i}{\sqrt{1-h_i}}$$

Delta-Beta:

$$\Delta\beta_j = \frac{(X'\hat{V}X)^{-1}X_i^j e_i}{1-h_i} \text{ for } j=1,\ldots,m$$

Delta-beta is defined as the change in an estimated coefficient when a case is omitted from the analysis. An estimate can be computed from the above formula without having to run n regressions.

7.2.6.4.3. Classification by Group



Logistic Regression: Step 4

Check output options or click on an 'Opt' button for further options

☑ 2 x 2 Table

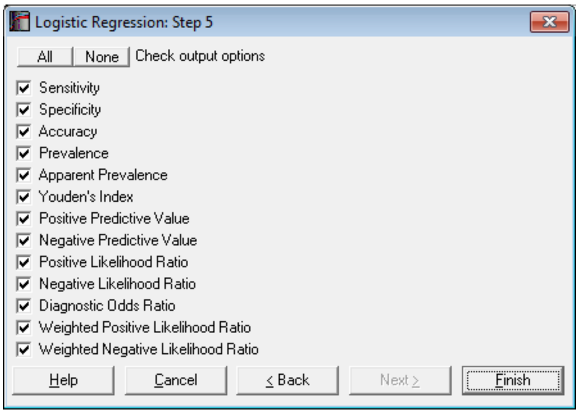
☑ Statistics for Diagnostic Tests

2 x 2 Table: The observed group membership (the actual Y in terms of 0 and 1) is tabulated against the estimated group membership. By default, it is assumed that the dependent variable value 1 represents the positive outcome The estimated group membership is 1 for any case with an estimated probability (fitted Y value) greater than or equal to the Classification Threshold Probability and it is 0 otherwise. If 1 does not represent the positive outcome of the test, then you can change this by entering 0 in the Positive Outcome box on the Output Options Dialogue. This will not force a re-estimation of the model.

	Positive Actual	**Negative Actual**	**Total**
Positive Estimate	TP	FP	TP + FP
Negative Estimate	FN	TN	FN + TN
Total	TP + FN	FP + TN	TOTAL

The table entries are defined as:

TP: True Positive: Correct acceptance,
TN: True Negative: Correct rejection,
FP: False Positive: False alarm (Type I error),
FN: False Negative: Missed detection (Type II error).



☑ Sensitivity
☑ Specificity
☑ Accuracy
☑ Prevalence
☑ Apparent Prevalence
☑ Youden's Index
☑ Positive Predictive Value
☑ Negative Predictive Value
☑ Positive Likelihood Ratio
☑ Negative Likelihood Ratio
☑ Diagnostic Odds Ratio
☑ Weighted Positive Likelihood Ratio
☑ Weighted Negative Likelihood Ratio

Statistics for Diagnostic Tests: These are the tests to determine how good a diagnostic method is, for instance, in detecting a positive outcome (i.e. sensitivity) or a negative outcome (i.e. specificity). Many of the statistics displayed here are proportions and their confidence intervals are computed employing the Wald (asymptotic) and Clopper-Pearson (exact) methods for binomial proportions (see 6.4.3.2. Binomial Test). Confidence intervals for likelihood ratios are computed as in Simel D., Samsa G., Matchar D. (1991).

The tests covered under this topic are also available in other procedures. When the data consists of two binary variables *Actual* and *Estimate*, you can use the Paired Proportions (see 6.4.5.6. Statistics for Diagnostic Tests) or Cross-Tabulation procedures. Alternatively, when you have an already formed 2 x 2 table, you can use the Contingency Table procedure.

Sensitivity: True positive rate or the probability of diagnosing a case as positive when it is actually positive.

TP / (TP + FN)

Specificity: True negative rate or the probability of diagnosing a case as negative when it is actually negative.

TN / (TN + FP)

Accuracy: The rate of correctly classified or the probability of true positive results, including true positive and true negative.

Sensitivity * Prevalence + Specificity * (1 - Prevalence)

(TP + TN) / TOTAL

Prevalence: The actual positive rate.

(TP + FN) / TOTAL

Apparent Prevalence: The estimated positive rate.

(TP + FP) / TOTAL

Youden's Index: Confidence intervals are calculated as in Bangdiwala S.I., Haedo A.S., Natal M.L. (2008).

Sensitivity + Specificity
TP / (TP + FN) + TN / (FP + TN)

Positive Predictive Value: PPV

TP / (TP + FP)

Negative Predictive Value: NPV

TN / (FN + TN)

Positive Likelihood Ratio: LR+

Sensitivity / (1 - Specificity)
(TP / (TP + FN)) / (1 - (TN / (FP + TN)))

Negative Likelihood Ratio: LR-

(1 – Sensitivity) / Specificity
(1 - (TP / (TP + FN))) / (TN / (FP + TN))

Diagnostic Odds Ratio: Confidence intervals are calculated as in Scott I.A., Greenburg P.B., Poole P.J. (2008).

Positive Likelihood Ratio / Negative Likelihood Ratio
(TP * TN) / (FP * FN)

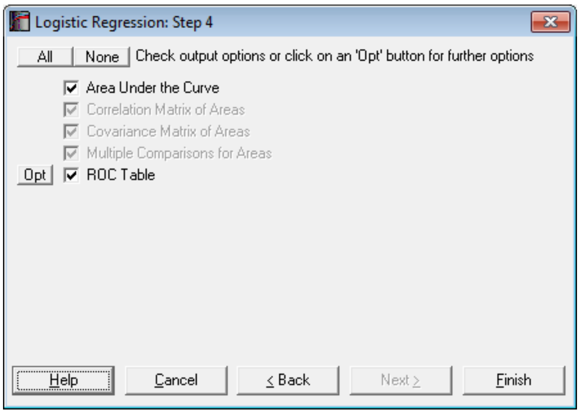
Weighted Positive Likelihood Ratio: WLR+. LR+ is weighted by prevalence.

(Prevalence * Sensitivity) / ((1-Prevalence)(1-Specificity))
TP / FP

Weighted Negative Likelihood Ratio: WLR-. LR- is weighted by prevalence.

(Prevalence (1-Sensitivity)) / ((1-Prevalence) Specificity)
FN / TN

7.2.6.4.4. ROC Analysis

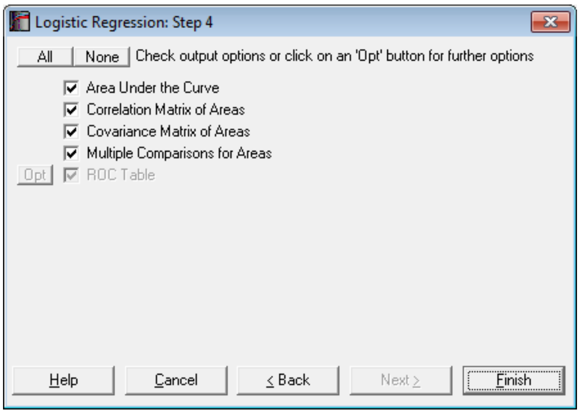

Logistic Regression: Step 4

Check output options or click on an 'Opt' button for further options

☑ Area Under the Curve
☑ Correlation Matrix of Areas
☑ Covariance Matrix of Areas
☑ Multiple Comparisons for Areas
☑ ROC Table

ROC analysis is widely used in assessing the statistical significance of diagnostic laboratory tests. Sensitivity and specificity values are computed for all fitted Y values. The best cut-off point is determined according to the ROC Optimality Criterion selected in the Intermediate Inputs dialogue. This can be Maximum sum of sensitivity and specificity (i.e. the Youden's index) or the value nearest to the top-left corner of ROC curve. The ROC curve is obtained by plotting the sensitivity values against 1 - specificity.

The output options for multiple ROC curves will not be available unless the Run a separate regression for each independent variable option is selected in the Intermediate Inputs dialogue.


Logistic Regression: Step 4

Check output options or click on an 'Opt' button for further options

☑ Area Under the Curve
☑ Correlation Matrix of Areas
☑ Covariance Matrix of Areas
☑ Multiple Comparisons for Areas
☑ ROC Table

Area Under the Curve (AUC): The area enclosed under the ROC curve is calculated by employing the algorithm developed by Delong E.R., Delong D.M., Clarke-Pearson D.L. (1998), which is based on the Mann-Whitney U test statistic. This not only produces an identical result to the area calculated by the trapezoidal rule, but also provides the standard errors (covariances) necessary to statistically compare AUCs, based on the nonparametric U-distribution (see 6.4.1.1. Mann-Whitney U Test).

AUC is a measure of the predictive power of the model. A value of 0.5 (which means that the curve is a 45° line) shows that the model has no power. A value of 1 (the theoretical maximum) means the full, 100% explanatory power.

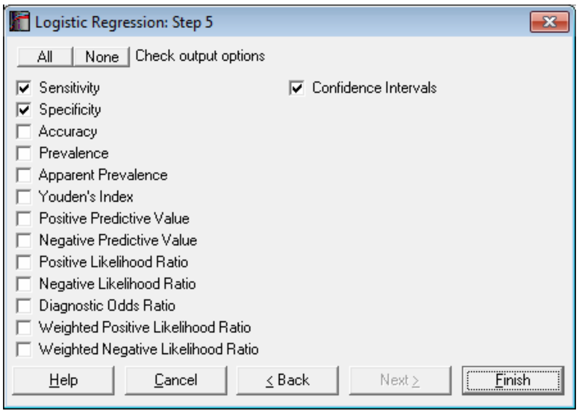
The output includes AUCs, their standard errors, tail probabilities and the confidence intervals (asymptotic normal).

Correlation and Covariance Matrices for Areas: These options are available only when the Run a separate regression for each independent variable option is selected in the Intermediate Inputs dialogue.

Multiple Comparisons for Areas: This option is available only when the Run a separate regression for each independent variable option is selected in the Intermediate Inputs dialogue. The difference between all possible pairs of AUCs, their standard errors and confidence intervals are displayed.

The output includes the difference between AUCs, their standard errors, tail probabilities and confidence intervals (asymptotic normal) and a further chi-square test with 1 degree of freedom.

ROC Table: All Statistics for Diagnostic Tests and their confidence intervals are computed for all fitted Y values (classification threshold probabilities). By default, only the sensitivity and specificity values and their confidence intervals are displayed. However, the user can choose to display any statistic with or without confidence intervals. The case (row) corresponding to the Classification Threshold Probability (the best cut-off point) is marked by an asterisk.



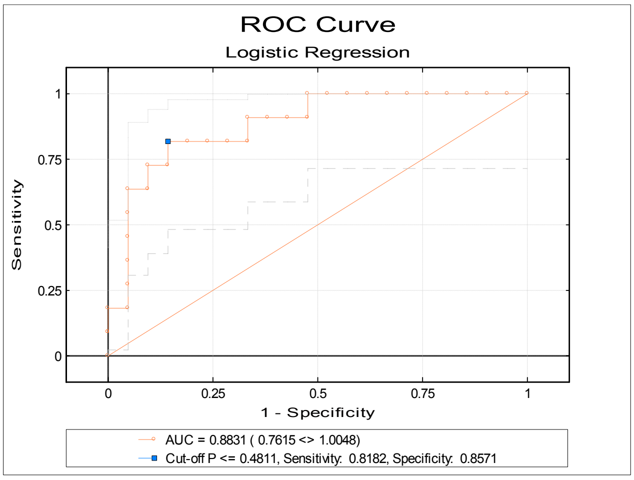
☑ Sensitivity
☑ Specificity
□ Accuracy
□ Prevalence
□ Apparent Prevalence
□ Youden's Index
□ Positive Predictive Value
□ Negative Predictive Value
□ Positive Likelihood Ratio
□ Negative Likelihood Ratio
□ Diagnostic Odds Ratio
□ Weighted Positive Likelihood Ratio
□ Weighted Negative Likelihood Ratio

☑ Confidence Intervals

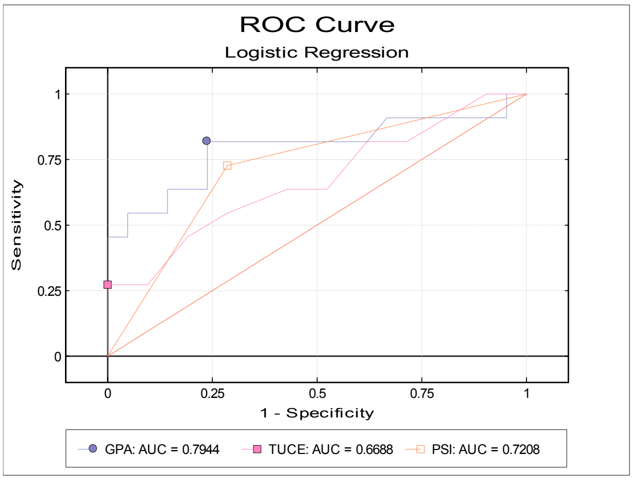
7.2.6.4.5. Receiver Operating Characteristic (ROC) Plot

Sensitivity and specificity values are computed for all classification threshold (cut-off) probabilities. For each probability, the sensitivity value is plotted against 1 - specificity. The best cut-off point is marked by a symbol on the curve.

When only one ROC curve is plotted, its confidence intervals are also displayed on the graph. The AUC, its confidence interval, sensitivity and specificity values are displayed in the legend. If there is only one independent variable, then its value corresponding to the best cut-off point is also displayed. When there are multiple independent variables, the best cut-off probability is displayed instead. The display of confidence intervals can be switch on or off from Edit → Data Series dialogue.

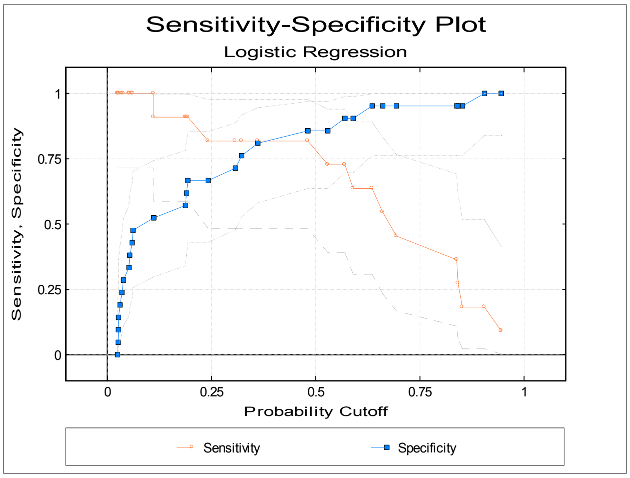


If the Run a separate regression for each independent variable option is selected in the Intermediate Inputs dialogue (i.e. if multiple ROCs are compared), the ROC plot is not output for each independent variable separately. In this case only one plot is drawn at the end of the output, with multiple ROC curves and without confidence intervals. For each curve, the area enclosed under the curve (AUC) is displayed in the legend. The best cut-off point is also marked by a symbol on each curve.



7.2.6.4.6. Sensitivitiy-Specificity Plot

This plot is similar to the ROC plot, except that sensitivity and specificity values are plotted against all values of the classification probability. This plot is not available when the Run a separate regression for each independent variable option is selected in the Intermediate Inputs dialogue.



7.2.6.5. Logistic Regression Examples

Example 1

Open LOGIT and select Statistics 1 → Regression Analysis → Logistic Regression. From the Variable Selection Dialogue select *GPA*, *TUCE* and *PSI* (*C12* to *C14*) as [Variable]s and *GRADE* (*C15*) as [Dependent] and. On the Intermediate Inputs dialogue select the Run regression with all independent variables option.

Some of the following results have been shortened to save space.

Logistic Regression

Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Regression Coefficients

	Coefficient	Standard Error	Wald Statistic	Probability	Lower 95%	Upper 95%
Constant	-13.0213	4.9313	6.9724	0.0083	-22.6866	-3.3561
GPA	2.8261	1.2629	5.0074	0.0252	0.3508	5.3014
TUCE	0.0952	0.1416	0.4519	0.5014	-0.1823	0.3726
PSI	2.3787	1.0646	4.9926	0.0255	0.2922	4.4652

Goodness of Fit Tests

	-2 Log likelihood
Initial Model	41.1835
Final Model	25.7793

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Pearson	27.2571	27	0.4500
Likelihood Ratio	15.4042	3	0.0015
Hosmer-Lemeshow	7.4526	8	0.4887

	Pseudo R-squared
McFadden	0.3740
Adjusted McFadden	0.1798
Cox & Snell	0.3821
Nagelkerke	0.5278

Correlation Matrix of Regression Coefficients

	Constant	GPA	TUCE	PSI
Constant	1.0000	-0.7343	-0.4960	-0.4494
GPA	-0.7343	1.0000	-0.2065	0.3181
TUCE	-0.4960	-0.2065	1.0000	0.0990
PSI	-0.4494	0.3181	0.0990	1.0000

Covariance Matrix of Regression Coefficients

	Constant	GPA	TUCE	PSI
Constant	24.3180	-4.5735	-0.3463	-2.3592
GPA	-4.5735	1.5950	-0.0369	0.4276
TUCE	-0.3463	-0.0369	0.0200	0.0149
PSI	-2.3592	0.4276	0.0149	1.1333

Odds Ratio

	Odds Ratio	Standard Error	Lower 95%	Upper 95%
GPA	16.8797	21.3181	1.4202	200.6239
TUCE	1.0998	0.1557	0.8334	1.4515
PSI	10.7907	11.4874	1.3393	86.9380

Hosmer-Lemeshow Table

	Actual Y = 0	Expected Y = 0	Actual Y = 1	Expected Y = 1	Total
1	4	3.8965	0	0.1035	4
2	3	2.8964	0	0.1036	3
3	3	2.8353	0	0.1647	3
4	2	2.7165	1	0.2835	3
5	2	2.4295	1	0.5705	3
6	4	2.7678	0	1.2322	4
7	1	1.4199	2	1.5801	3
8	1	1.1139	2	1.8861	3
9	0	0.6265	3	2.3735	3
10	1	0.2977	2	2.7023	3

Case (Diagnostic) Statistics

	Actual Y	Fitted Y	Predicted Group	Leverage	Cook's Distance	Deviance
1	0.0000	0.0266	0.0000	0.0390	0.0011	-0.2321
2	0.0000	0.0595	0.0000	0.0545	0.0036	-0.3503
3	0.0000	0.1873	0.0000	0.0889	0.0225	-0.6440
…	…	…	…	…	…	…
30	1.0000	0.9453	1.0000	0.0853	0.0054	0.3353
F+ 31	0.0000	0.5291	1.0000	0.1171	0.1490	-1.2273
F- 32	1.0000	0.1110	0.0000	0.1299	1.1953	2.0966

	Residuals	Standardised Residuals	Logit Residuals	Studentised Residuals
1	-0.0266	-0.1652	-1.0273	-0.2368
2	-0.0595	-0.2515	-1.0633	-0.3602
3	-0.1873	-0.4800	-1.2304	-0.6747
…	…	…	…	…
30	0.0547	0.2405	1.0578	0.3506
F+ 31	-0.5291	-1.0600	-2.1237	-1.3061
F- 32	0.8890	2.8296	9.0065	2.2477

F+: False Positive
F-: False Negative

2 x 2 Table

Estimated \ Actual	Positive	Negative	Total
Positive	9	3	12
	81.82%	14.29%	
Negative	2	18	20
	18.18%	85.71%	
Total	11	21	32
	100.00%	100.00%	

Classification Threshold Probability =	0.4211

Statistics for Diagnostic Tests

Confidence Intervals: Row 1: Asymptotic Normal, Row 2: Exact Binomial

	Value	Standard Error	Lower 95%	Upper 95%
Sensitivity	0.8182	0.1163	0.5903	1.0000
			0.4822	0.9772
Specificity	0.8571	0.0764	0.7075	1.0000
			0.6366	0.9695
Accuracy	0.8438	0.0642	0.7179	0.9696
			0.6721	0.9472
Prevalence	0.3438	0.0840	0.1792	0.5083
			0.1857	0.5319
Apparent Prevalence	0.3750	0.0856	0.2073	0.5427
			0.2110	0.5631
Youden's Index	0.6753			
			0.1188	0.9467
Positive Predictive Value	0.7500	0.1250	0.5050	0.9950
			0.4281	0.9451
Negative Predictive Value	0.9000	0.0671	0.7685	1.0000
			0.6830	0.9877
Positive Likelihood Ratio	5.7273		1.9371	16.9334
Negative Likelihood Ratio	0.2121		0.0596	0.7547
Diagnostic Odds Ratio	27.0000		3.8033	191.6750
Weighted Positive Likelihood Ratio	3.0000		1.0678	8.4284
Weighted Negative Likelihood Ratio	0.1111		0.0295	0.4183

Logistic Regression

Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Area Under the Curve

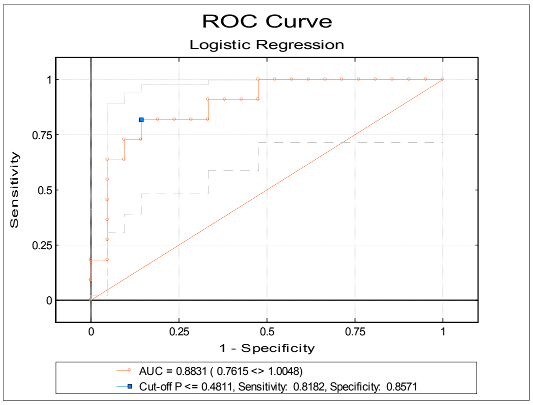
	AUC	Standard Error	Z-Statistic	1-Tail Probability	2-Tail Probability	Lower 95%	Upper 95%
	0.8831	0.0621	6.1721	0.0000	0.0000	0.7615	1.0048

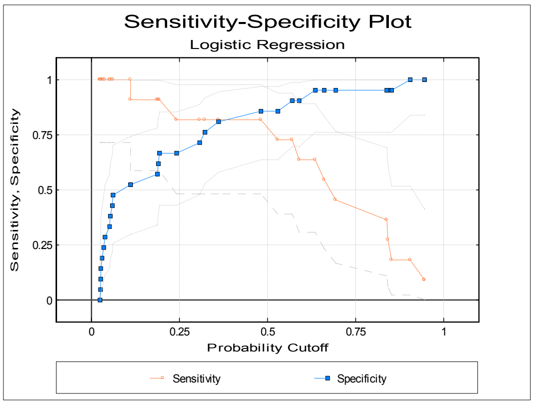
ROC Table

* marks the best cut-off case.
Optimality criterion: Max(Sensitivity + Specificity)
Confidence Intervals: Asymptotic Normal

	Cut-off P <=	Sensitivity	Lower 95%	Upper 95%	Specificity	Lower 95%	Upper 95%
1	0.9453	0.0909	0.0000	0.2608	1.0000	1.0000	1.0000
2	0.9048	0.1818	0.0000	0.4097	1.0000	1.0000	1.0000
3	0.8521	0.1818	0.0000	0.4097	0.9524	0.8613	1.0000
...	…	…	…	…	…	…	…
10	0.5699	0.7273	0.4641	0.9905	0.9048	0.7792	1.0000
11	0.5291	0.7273	0.4641	0.9905	0.8571	0.7075	1.0000
* 12	0.4811	0.8182	0.5903	1.0000	0.8571	0.7075	1.0000
...	…	…	…	…	…	…	…
30	0.0265	1.0000	1.0000	1.0000	0.0952	0.0000	0.2208
31	0.0259	1.0000	1.0000	1.0000	0.0476	0.0000	0.1387
32	0.0245	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000

Range of best Classification Threshold Probability = 0.3610 <> 0.4811





Example 2

Continuing from the above example, this time select the Run a separate regression for each independent variable option on the Intermediate Inputs dialogue. Also uncheck the first three output options Regression Results, Case (Diagnostic) Statistics and Classification by Group.

Logistic Regression

Dependent Variable: GRADE
Minimum of dependent variable is encoded as 0 and the rest as 1.
Valid Number of Cases: 32, 0 Omitted

Area Under the Curve

	AUC	Standard Error	Z-Statistic	1-Tail Prob	2-Tail Prob	Lower 95%	Upper 95%
GPA	0.7944	0.1001	2.9421	0.0016	0.0033	0.5983	0.9905
TUCE	0.6688	0.1075	1.5702	0.0582	0.1164	0.4581	0.8796
PSI	0.7208	0.0867	2.5477	0.0054	0.0108	0.5509	0.8906

Correlation Matrix of Areas

	GPA	TUCE	PSI
GPA	1.0000	0.2215	-0.2182
TUCE	0.2215	1.0000	-0.0997
PSI	-0.2182	-0.0997	1.0000

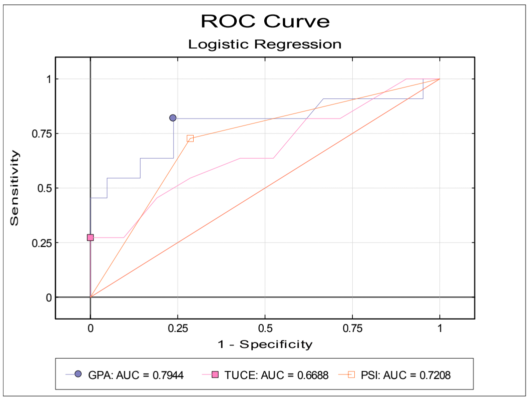
Covariance Matrix of Areas

	GPA	TUCE	PSI
GPA	0.0100	0.0024	-0.0019
TUCE	0.0024	0.0116	-0.0009
PSI	-0.0019	-0.0009	0.0075

Multiple Comparisons for Areas

	Difference	Standard	Z-Statistic	1-Tail	2-Tail
		Error		Probability	Probability
GPA - TUCE	0.1255	0.1296	0.9684	0.1664	0.3328
GPA - PSI	0.0736	0.1460	0.5042	0.3071	0.6141
TUCE - PSI	-0.0519	0.1447	0.3591	0.3598	0.7195

	Lower	Upper	Chi-Square	Right-Tail
	95%	95%	Statistic	Probability
GPA - TUCE	-0.1285	0.3796	0.9378	0.3328
GPA - PSI	-0.2125	0.3597	0.2542	0.6141
TUCE - PSI	-0.3355	0.2316	0.1289	0.7195



7.2.7. Multinomial Regression

The Multinomial Regression procedure (which is also known as Multinomial Logistic or Polytomous regression) is suitable for estimating models where the dependent variable is a categorical variable. If the dependent variable contains only two categories, its results are identical to that of Logistic Regression. Therefore, Multinomial Regression can be considered as an extension of Logistic Regression.

In Multinomial Regression, a set of coefficients are estimated for each category of the dependent variable. This makes its output format fundamentally different from other types of regression (see 7.2.7.3. Multinomial Regression Output Options). Since the estimated model is over-determined, the coefficients are scaled according to one of the categories. By default, the most frequent category is selected as the base category, though this can be changed by the user. The estimated coefficients are displayed relative to the base category and their exponentials are called Relative Risk Ratios.

Multinomial Regression is also closely related to Discriminant Analysis in the sense that both procedures are used to estimate the membership of cases to the groups defined by a categorical variable (see 8.2. Discriminant Analysis). Multinomial Regression should be preferred when the list of independent variables contains dummy variables. Discriminant Analysis is more powerful than Multinomial Regression when all independent variables are continuous and meet the assumptions of multivariate normality.

Predictions (interpolations) and multicollinearity are handled as in other regression options (see, for instance, Logistic Regression). Predicted cases are identified by an asterisk in Classification by Case, Probabilities and Index Values output options (see 7.2.7.3. Multinomial Regression Output Options).

7.2.7.1. Multinomial Regression Model Description

The multinomial logistic function is an extension of the logit function discussed in Logit / Probit / Gompit (see 7.2.5.1. Logit / Probit / Gompit Model Description).

Let J + 1 be the number of distinct categories in the dependent variable and assume that the category 0 is selected as the base category. Then the probabilities given by the multinomial logistic function are:

$$P(Y=j)=\frac{\mathrm{Exp}(\beta_j^{'}x_i)}{1+\sum_{k=1}^{J}\mathrm{Exp}(\beta_k^{'}x_i)}\text{ for } j=1,\ldots,J \text{ and}$$

$$P(Y=0)=\frac{1}{1+\sum_{k=1}^{J}\mathrm{Exp}(\beta_k^{'}x_i)}\text{ for the base category.}$$

where $\beta_j^{'}$ is the vector of estimated coefficients for the jth category and x_i is the ith case (row) of the data matrix.

The relative risk ratio for case i relative to the base category is:

$$\frac{P_{ij}}{P_{i0}}=\mathrm{Exp}(\beta_j^{'}x_i)\text{for } j=1,\ldots,J \text{ and } i=1,\ldots,n.$$

The logarithm of the likelihood function is given as:

$$L=\sum_{i=1}^{n}\sum_{j=0}^{J}d_{ij}\mathrm{Ln}(P(Y=j))$$

and the first derivatives as:

$$\frac{\partial \mathrm{Ln}(L)}{\partial\beta_i}=\sum_{i=1}^{n}(d_{ij}-P_{ij})\,x_i\text{ for } j=1,\ldots,J,$$

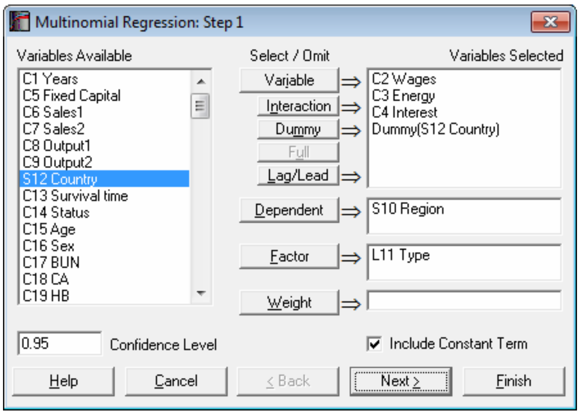
where:

$d_{ij} = 1$ if $Y_i = j$ and

$d_{ij} = 0$ otherwise.

A Newton-Raphson type maximum likelihood algorithm is employed to minimise the negative of the log likelihood function. The nature of this method implies that a solution (convergence) cannot always be achieved. In such cases, you are advised to edit the convergence parameters provided, in order to find the right levels for the particular problem at hand.

7.2.7.2. Multinomial Regression Variable Selection



Variables Available: C1 Years, C5 Fixed Capital, C6 Sales1, C7 Sales2, C8 Output1, C9 Output2, S12 Country, C13 Survival time, C14 Status, C15 Age, C16 Sex, C17 BUN, C18 CA, C19 HB

Variables Selected: C2 Wages, C3 Energy, C4 Interest, Dummy(S12 Country)

Dependent: S10 Region

Factor: L11 Type

Weight:

0.95 Confidence Level

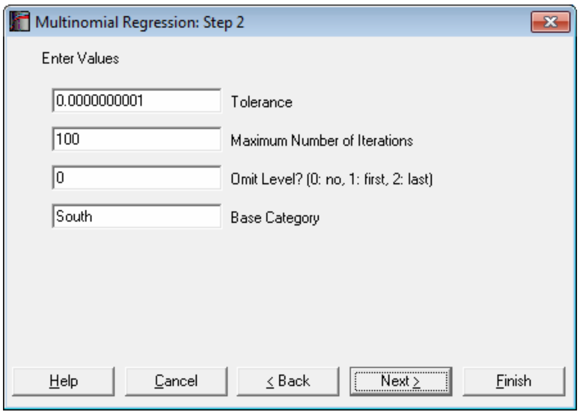
☑ Include Constant Term

Like Logistic Regression, Multinomial Regression can be used to estimate models with or without a constant term, with or without weights and regressions can be run on a subset of cases as determined by the levels of an unlimited number of factor columns. An unlimited number of dependent variables (numeric or string) can be selected in order to run the same model on different dependent variables. It is also possible to include interaction terms, dummy and lag/lead variables in the model, without having to create them as spreadsheet columns first (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

It is compulsory to select at least one categorical data column containing numeric or String Data as a dependent variable. When more than one dependent variable is selected, the analysis will be repeated as many times as the number of dependent variables, each time only changing the dependent variable and keeping the rest of selections unchanged.

A column containing numeric data can be selected as a weights column. Weights are frequency weights and all independent variables are multiplied by this column internally by the program.

An intermediate inputs dialogue is displayed next.


Multinomial Regression: Step 2

Enter Values

0.0000000001 Tolerance

100 Maximum Number of Iterations

0 Omit Level? (0: no, 1: first, 2: last)

South Base Category

Tolerance: This value is used to control the sensitivity of nonlinear minimisation procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

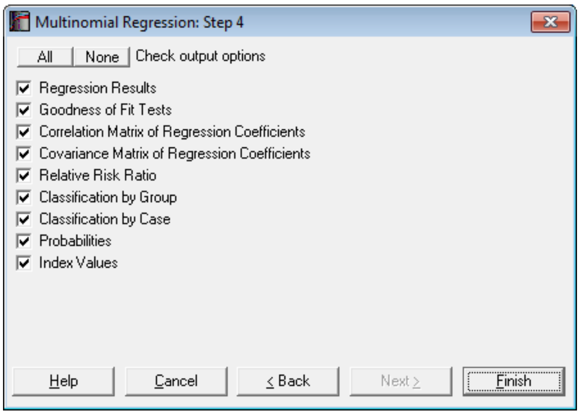
Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Omit Level: This field will appear only when one or more dummy variables have been included in the model from the Variable Selection Dialogue. Three options are available; (0) do not omit any levels, (1) omit the first level and (2) omit the last level. When no levels are omitted, the model will usually be over-parameterised (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

Base Category: By default, the most frequent category of the dependent variable is suggested as the base category. To scale the estimated coefficients according to a particular category, enter this category here. The Relative Risk Ratios will be relative to this category.

7.2.7.3. Multinomial Regression Output Options

When the calculations are finished an Output Options Dialogue will provide access to the following options.



Check output options

☑ Regression Results
☑ Goodness of Fit Tests
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
☑ Relative Risk Ratio
☑ Classification by Group
☑ Classification by Case
☑ Probabilities
☑ Index Values

Regression Results: The main regression output displays a table of estimated coefficients for each category of the dependent variable, except for the base category. Standard errors, Wald statistics, probability values and confidence intervals are also displayed for the estimated regression coefficients.

Wald Statistic: This is defined as:

$$W_{ij} = \frac{\beta_{ij}^2}{\sigma_{ij}^2}, i = 1, \ldots, k, j = 1, \ldots, J,$$

and has a chi-square distribution with one degree of freedom.

Confidence Intervals: The confidence intervals for regression coefficients are computed from:

$$\beta_{ij} \pm Z_{\alpha/2}\sigma_{ij}, i = 1, \ldots, k, j = 1, \ldots, J,$$

where k is the number of independent variables in the model and each coefficient's standard error, σ_{ij}, is the square root of the diagonal element of covariance matrix.

Goodness of Fit Tests: See 7.2.6.4.1. Logistic Regression Results for details.

Relative Risk Ratio:

Values of the relative risk ratio indicate the influence of one unit change in a covariate on the regression.

$$\text{Exp}(\beta_{ij}), i = 1, \ldots, k, j = 1, \ldots, J.$$

The standard error of the relative risk ratio is:

$$\sigma_{ij}\mathrm{Exp}(\beta_{ij})$$

where σ_{ij} is the standard error of the i^th^ independent variable for the j^th^ category of the dependent variable. Coefficient confidence intervals are:

$$\mathrm{Exp}(\beta_{ij} \pm Z_{\alpha/2}\sigma_{ij})$$

which are simply the exponential of the coefficient confidence intervals.

Correlation Matrix of Regression Coefficients: This is a symmetric matrix with unity diagonal elements. The off-diagonal elements give correlations between the regression coefficients.

Covariance Matrix of Regression Coefficients: This is a symmetric matrix where diagonal elements are the square of parameter standard errors. The off-diagonal elements are covariances between the regression coefficients.

Classification by Group: A table is formed with rows and columns corresponding to observed and estimated group memberships respectively. Each cell displays the number of elements and its percentage with respect to the row total. The diagonal elements are the cases classified correctly and the off-diagonal elements are the misclassified cases.

Classification by Case: The observed and estimated group memberships are listed and pairs with disagreement are marked. The estimated group is the one with the highest probability level.

Probabilities: These are calculated as in the first two equations of Multinomial Regression Model Description. The predicted values are marked with two asterisks.

Index Values: These are the fitted values given as:

$$\beta'_{ij}x_i \text{ for } j = 0, \ldots, J \text{ and } i = 1, \ldots, n.$$

The predicted values are marked with two asterisks.

7.2.7.4. Multinomial Regression Examples

Open ANOVA and select Statistics 1 → Regression Analysis → Multinomial Regression. From the Variable Selection Dialogue select *Period* (*C29*) as [Dependent], *Score* (*C26*) as [Variable] and *Noise* (*S30*) as [Dummy]. On Step 2 dialogue enter 1 for Omit Level and leave other entries unchanged. Check all output options to obtain the following output. Some tables have been shortened.

Multinomial Regression

Dependent Variable: Period
Base Category: 1
Valid Number of Cases: 54, 0 Omitted

Regression Results

	Coefficient	Standard Error	Wald Statistic	Probability	Lower 95%	Upper 95%
Period = 2 Constant	4.2452	1.9258	4.8596	0.0275	0.4708	8.0196
Score	-0.0818	0.0358	5.2104	0.0225	-0.1519	-0.0116
Noise = Low	-0.4319	0.7490	0.3325	0.5642	-1.8999	1.0362
Period = 3 Constant	8.9299	2.3890	13.9723	0.0002	4.2476	13.6122
Score	-0.1924	0.0503	14.6346	0.0001	-0.2910	-0.0938
Noise = Low	-1.1811	0.9236	1.6352	0.2010	-2.9914	0.6292

Goodness of Fit Tests

	-2 Log likelihood
Initial Model	118.6501
Final Model	91.6682

	Chi-Square Statistic	Degrees of Freedom	Right-Tail Probability
Likelihood Ratio	26.9820	4	0.0000

	Pseudo R-squared
McFadden	0.2274
Adjusted McFadden	0.1768
Cox & Snell	0.3933
Nagelkerke	0.4424

Correlation Matrix of Regression Coefficients

	Period = 2 Constant	Score	Noise = Low	Period = 3 Constant	Score	Noise = Low
Period = 2 Constant	1.0000	-0.9642	-0.4271	0.6576	-0.5719	-0.2615
Score	-0.9642	1.0000	0.2573	-0.6220	0.5703	0.1609
Noise = Low	-0.4271	0.2573	1.0000	-0.2553	0.1441	0.5618
Period = 3 Constant	0.6576	-0.6220	-0.2553	1.0000	-0.9635	-0.4907
Score	-0.5719	0.5703	0.1441	-0.9635	1.0000	0.3417
Noise = Low	-0.2615	0.1609	0.5618	-0.4907	0.3417	1.0000

Covariance Matrix of Regression Coefficients

	Period = 2 Constant	Score	Noise = Low	Period = 3 Constant	Score	Noise = Low
Period = 2 Constant	3.7085	-0.0665	-0.6160	3.0251	-0.0554	-0.4651
Score	-0.0665	0.0013	0.0069	-0.0532	0.0010	0.0053
Noise = Low	-0.6160	0.0069	0.5610	-0.4568	0.0054	0.3886
Period = 3 Constant	3.0251	-0.0532	-0.4568	5.7072	-0.1158	-1.0828
Score	-0.0554	0.0010	0.0054	-0.1158	0.0025	0.0159
Noise = Low	-0.4651	0.0053	0.3886	-1.0828	0.0159	0.8531

Relative Risk Ratio

	Relative Risk Ratio	Standard Error	Lower 95%	Upper 95%
Period = 2 Score	0.9215	0.0330	0.8590	0.9885
Noise = Low	0.6493	0.4863	0.1496	2.8184
Period = 3 Score	0.8249	0.0415	0.7475	0.9104
Noise = Low	0.3069	0.2835	0.0502	1.8761

Classification by Group

Observed\Predicted	1	2	3
1	12 66.67%	4 22.22%	2 11.11%
2	7 38.89%	5 27.78%	6 33.33%
3	0 0.00%	5 27.78%	13 72.22%

Correctly Classified = 55.56%

Classification by Case

	Actual Group	Misclassified	Estimated Group	Probability
1	1	*	2	0.4327
2	1		1	0.4552
3	1		1	0.6290
4	2	*	3	0.4842
5	2	*	1	0.4283
6	2	*	1	0.5585
...	…	…	…	…

Probabilities

	1	2	3
1	0.2456	0.4327	0.3217
2	0.4552	0.4170	0.1279
3	0.6290	0.3251	0.0459
4	0.1413	0.3745	0.4842
5	0.4283	0.4258	0.1459
6	0.5585	0.3689	0.0726
7	0.0235	0.1661	0.8105
8	0.0953	0.3229	0.5819
9	0.2700	0.4383	0.2917
10	0.0716	0.2858	0.6426
11	0.1595	0.3896	0.4509
12	0.3744	0.4383	0.1873
13	0.0328	0.1969	0.7703
14	0.0953	0.3229	0.5819
15	0.2952	0.4417	0.2631
...	…	…	…

Index Values

	1	2	3
1	0.0000	0.5663	0.2698
2	0.0000	-0.0877	-1.2698
3	0.0000	-0.6600	-2.6169
4	0.0000	0.9751	1.2320
5	0.0000	-0.0059	-1.0773
6	0.0000	-0.4147	-2.0396
7	0.0000	1.9561	3.5414
8	0.0000	1.2204	1.8094
9	0.0000	0.4846	0.0774
10	0.0000	1.3839	2.1943
11	0.0000	0.8934	1.0396
12	0.0000	0.1576	-0.6924
13	0.0000	1.7926	3.1565
14	0.0000	1.2204	1.8094
15	0.0000	0.4028	-0.1151
...	…	…	…

7.2.8. Poisson Regression

The Poisson Regression procedure is suitable for models where the dependent variable is a frequency (count) variable consisting of nonnegative integers. The exponential of estimated regression coefficients are called Incidence Rate Ratios, which give the estimated rate at which events occur. This rate can be multiplied by an Exposure variable to obtain the expected frequencies, which enters the model with a coefficient constrained as 1.

Predictions (interpolations) and multicollinearity are handled as in other regression options (see, for instance, Logistic Regression). Predicted cases are identified by an asterisk in Case (Diagnostic) Statistics output options (see 7.2.8.3. Poisson Regression Output Options). The spreadsheet **Reg** function (see 3.4.2.6.3. UNISTAT Functions) will give the natural log of predicted values, which should be exponentiated to obtain the expected frequencies.

7.2.8.1. Poisson Regression Model Description

Poisson Regression assumes that actual frequencies y_i are drawn from a Poisson distribution with parameters λ_i, i = 1, …, n. The associated probabilities are given as:

$$P(Y_i = y_i) = \frac{\mathrm{Exp}(\lambda_i)\lambda_i^{y_i}}{y_i!} \text{ for } = 0, 1, 2, \ldots$$

where:

$$\lambda_i = \mathrm{Exp}(\beta' x_i), i = 1, \ldots, n$$

is known as the loglinear model.

The logarithm of the likelihood function is given as:

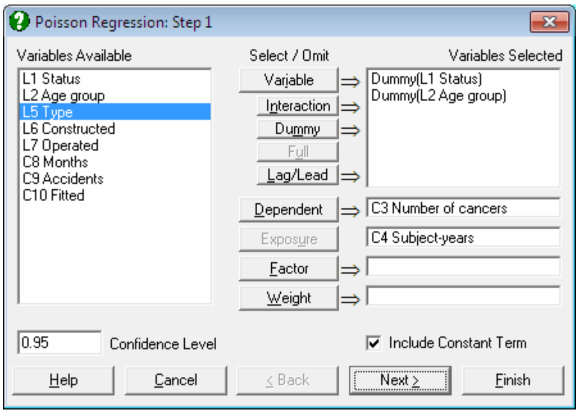
$$L = \sum_{i=1}^{n} \left[-\lambda_i + y_i \beta' x_i - \mathrm{Ln}(y_i!) \right]$$

and the first derivatives are:

$$\frac{\partial \mathrm{Ln}(L)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \lambda_i)\, x_i, j = 1, \ldots, k.$$

A Newton-Raphson type maximum likelihood algorithm is employed to minimise the negative of the log likelihood function. The nature of this method implies that a solution (convergence) cannot always be achieved. In such cases, you are advised to edit the convergence parameters provided and try again.

7.2.8.2. Poisson Regression Variable Selection



Poisson Regression: Step 1

Variables Available

L1 Status
L2 Age group
L5 Type
L6 Constructed
L7 Operated
C8 Months
C9 Accidents
C10 Fitted

Variables Selected

Dummy(L1 Status)
Dummy(L2 Age group)

Dependent: C3 Number of cancers

Exposure: C4 Subject-years

Factor:

Weight:

0.95 Confidence Level

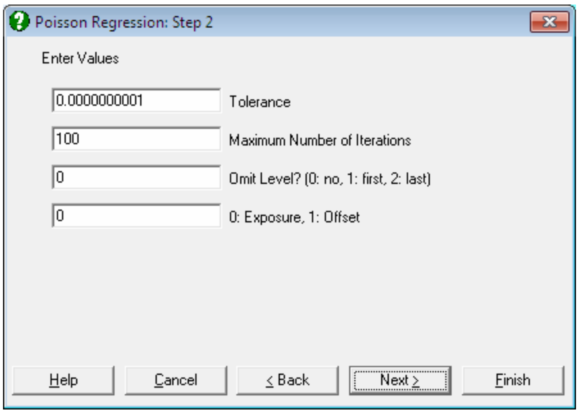
☑ Include Constant Term

As in other regression procedures, Poisson Regression can be used to estimate models with or without a constant term, with or without weights and regressions can be run on a subset of cases as determined by the levels of an unlimited number of factor columns. An unlimited number of dependent variables can be selected in order to run the same model on different dependent variables. It is also possible to include interaction terms, dummy and lag/lead variables in the model, without having to create them as spreadsheet columns first (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

It is compulsory to select at least one numeric data column containing frequency counts as a dependent variable. When more than one dependent variable is selected, the analysis will be repeated as many times as the number of dependent variables, each time only changing the dependent variable and keeping the rest of selections unchanged.

A column containing numeric data can be selected as a weights column. Weights are frequency weights and all independent variables are multiplied by this column internally by the program.

An intermediate inputs dialogue is displayed next.



Poisson Regression: Step 2

Enter Values

0.0000000001 Tolerance

100 Maximum Number of Iterations

0 Omit Level? (0: no, 1: first, 2: last)

0 0: Exposure, 1: Offset

Tolerance: This value is used to control the sensitivity of nonlinear minimisation procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Omit Level: This field will appear only when one or more dummy variables have been included in the model from the Variable Selection Dialogue. Three options are available; (0) do not omit any levels, (1) omit the first level and (2) omit the last level. When no levels are omitted, the model will usually be over-parameterised (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

Exposure / Offset: This field will appear only if a column has been assigned the task of [Exposure] in Variable Selection Dialogue. When the value of this field 0 the variable selected (E) will enter the model as Exposure:

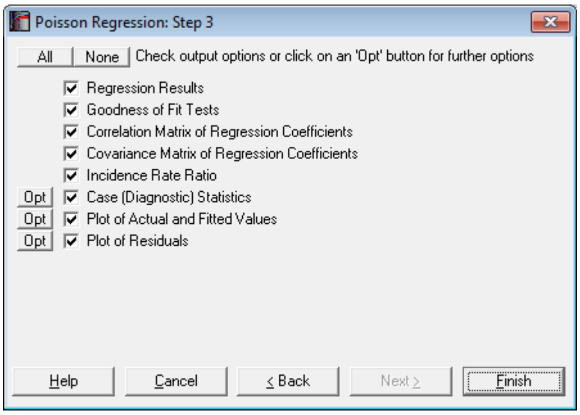
$$\lambda_i = E_i \mathrm{Exp}(\beta' x_i) = \mathrm{Exp}(\mathrm{Ln}(E_i) + \beta' x_i),\ i = 1, \ldots, n$$

and for any other value it will enter the model as Offset:

$$\lambda_i = \mathrm{Ln}(E_i)\mathrm{Exp}(\beta' x_i) = \mathrm{Exp}(E_i + \beta' x_i),\ i = 1, \ldots, n.$$

7.2.8.3. Poisson Regression Output Options

The Actual and Fitted Values, Residuals and Confidence Intervals for Mean and Actual Y Values options that existed in earlier version of UNISTAT have now been merged with the Case (Diagnostic) Statistics option. See 7.2.1.2.2. Linear Regression Case Output.



Poisson Regression: Step 3

Check output options or click on an 'Opt' button for further options

☑ Regression Results
☑ Goodness of Fit Tests
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
☑ Incidence Rate Ratio
☑ Case (Diagnostic) Statistics
☑ Plot of Actual and Fitted Values
☑ Plot of Residuals

Regression Results: The main regression output displays a table of estimated coefficients for each category of the dependent variable, except for the base category. Standard errors, Wald statistics, probability values and confidence intervals are also displayed for the estimated regression coefficients.

Wald Statistic: This is defined as:

$$W_i = \frac{\beta_i^2}{\sigma_i^2}, i = 1, \ldots, k.$$

and has a chi-square distribution with one degree of freedom.

Confidence Intervals: The confidence intervals for regression coefficients are computed from:

$$\beta_i \pm Z_{\alpha/2}\sigma_i, i = 1, \ldots, k.$$

where k is the number of independent variables in the model and each coefficient's standard error, σ_i, is the square root of the diagonal element of covariance matrix.

Goodness of Fit Tests: See 7.2.6.4.1. Logistic Regression Results for details.

Correlation Matrix of Regression Coefficients: This is a symmetric matrix with unity diagonal elements. The off-diagonal elements give correlations between the regression coefficients.

Covariance Matrix of Regression Coefficients: This is a symmetric matrix where diagonal elements are the square of parameter standard errors. The off-diagonal elements are covariances between the regression coefficients.

Incidence Rate Ratio: Values of the incidence rate ratio indicate the influence of one unit change in a covariate on the regression.

$$\mathrm{Exp}(\beta_i),\ i = 1, \ldots, k.$$

The standard error of the incidence rate ratio is:

$$\sigma_i \mathrm{Exp}(\beta_i)$$

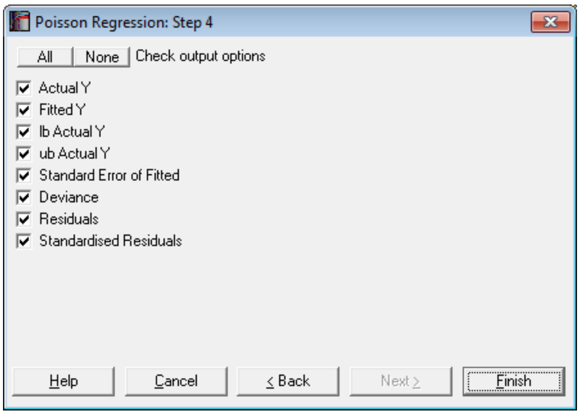
where σ_i is the standard error of the i^{th} independent variable for the j^{th} category of the dependent variable. Coefficient confidence intervals are:

$$\mathrm{Exp}(\beta_i \pm Z_{\alpha/2}\sigma_i)$$

which are simply the exponential of the coefficient confidence intervals.

Case (Diagnostic) Statistics: Case statistics are useful to determine the influence of individual observations on the overall fit of the model. For further information see 7.2.1.2.2. Linear Regression Case Output.

Statistics available under this option are defined as follows.



Poisson Regression: Step 4

Check output options

☑ Actual Y
☑ Fitted Y
☑ lb Actual Y
☑ ub Actual Y
☑ Standard Error of Fitted
☑ Deviance
☑ Residuals
☑ Standardised Residuals

Actual Y: Observed values of the dependent variable.

Fitted Y: Also known as expected values:

$$\lambda_i = E_i \text{Exp}(\beta' x_i), i = 1, \ldots, n$$

Standard Error of Fitted:

$$s_i = \sqrt{X_i (X'X)^{-1} X_i'}$$

Confidence Intervals of Fitted:

$$\lambda_i \pm Z_{\alpha/2} s_i$$

Deviance:

$$d_i = 2\left[y_i \text{Ln}\left(\frac{y_i}{\lambda_i}\right) - \left(y_i - \lambda_i\right)\right]$$

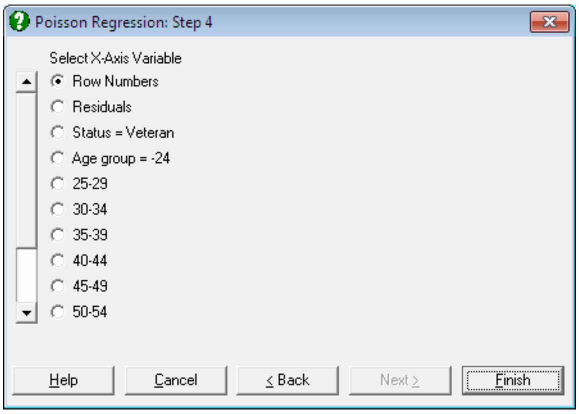
Residuals:

$$e_i = y_i - \lambda_i$$

Standardised Residuals:

$$z_i = \frac{e_i}{\sqrt{\lambda_i}}$$

Plot of Actual and Fitted Values: Select this option to plot actual and fitted Y values against row numbers (index), residuals or against any independent variable. A further dialogue will enable you to choose the X-axis variable from a list.



Poisson Regression: Step 4

Select X-Axis Variable

● Row Numbers
○ Residuals
○ Status = Veteran
○ Age group = -24
○ 25-29
○ 30-34
○ 35-39
○ 40-44
○ 45-49
○ 50-54

By default, a line graph of the two series is plotted. However, since this procedure (like the plot of residuals) uses the X-Y Plots engine, it has almost all controls and options available for X-Y Plots, except for error bars and right Y-axes.

The data points on the graph will also respond to the right mouse button in the way X-Y Plots does; the point is highlighted, a panel displays information about the point and in Stand-Alone Mode, the row of the spreadsheet containing the data point is also highlighted (a procedure which is also known as *Brushing* or *Point identification*). While the point is highlighted you can press <Delete> to omit the particular row containing the point. The entire Regression Analysis will be run again without the deleted row. If you want to restore the original regression, you will need to take one of the following two actions depending on the way you run UNISTAT:

1. In Stand-Alone Mode, go back to the Data Processor and delete or deactivate the Select Row column created by the program.

2. In Excel Add-In Mode, highlight a different block of data to remove the effect of the internal Select Row column.

Plot of Residuals: Residuals can be plotted against row numbers (index), fitted values or against any independent variable. A further dialogue will enable you to choose the X-axis variable from a list containing Row Numbers, Fitted Values and all independent variables.

By default a scatter graph of residuals is plotted. For more information on available options see Plot of Actual and Fitted Values above.

7.2.8.4. Poisson Regression Examples

Example 1

Example 14.4 on p. 501 Armitage & Berry (2002). The aim is to assess whether there is a significant difference in cancer risk between veterans and non-veterans. The servicemen are divided into 11 age groups and their experience is given in terms of subject-years.

Open POISSON and select Statistics 1 → Regression Analysis → Poisson Regression. Select *Status* and *Age group* (*L1* and *L2*) as [Dummy], *Number of cancers* (*C3*) as [Dependent] and *Subject-years* (*C4*) as [Exposure]. On Step 2 dialogue enter 1 for Omit Level and leave other entries unchanged. Uncheck all output options except for Regression Results and click [Finish].

Armitage and Berry include the logarithm of the *Subject-years* variable in the model as an Offset variable, which is equivalent to an Exposure variable.

Regression results show that the p-value of Status = Veteran variable is 0.9493. As this is much greater than 5%, we can conclude that there is no significant difference in cancer risk between veterans and non-veterans.

Poisson Regression

Dependent Variable: Number of cancers
Exposure: Subject-years
Valid Number of Cases: 22, 0 Omitted

Regression Results

	Coefficient	Standard Error	Wald Statistic	Significance	Lower 95%	Upper 95%
Constant	-9.3248	0.2045	2079.0648	0.0000	-9.7257	-8.9240
Status = Veteran	-0.0035	0.0555	0.0040	0.9493	-0.1123	0.1052
Age group = 25-29	0.6793	0.2325	8.5373	0.0035	0.2236	1.1350
30-34	1.3711	0.2177	39.6626	0.0000	0.9444	1.7978
35-39	1.9396	0.2121	83.6581	0.0000	1.5240	2.3553
40-44	2.0343	0.2161	88.6203	0.0000	1.6108	2.4579
45-49	2.7266	0.2222	150.5414	0.0000	2.2910	3.1621
50-54	3.2029	0.2206	210.7121	0.0000	2.7704	3.6353
55-59	3.7162	0.2178	291.1924	0.0000	3.2894	4.1430
60-64	4.0927	0.2177	353.4845	0.0000	3.6660	4.5193
65-69	4.2362	0.2242	356.9375	0.0000	3.7967	4.6757
70-	4.3637	0.2274	368.3262	0.0000	3.9181	4.8094

Example 2

Example 19.21 on p. 932 Greene (1997). The number of accidents per service month is given for a sample of ship types. There are five types of ships constructed in four different time periods and observed in two time periods.

Open POISSON and select Statistics 1 → Regression Analysis → Poisson Regression. From the Variable Selection Dialogue select *Type*, *Constructed* and *Operated* (*C5-C7*) as [Dummy], *Accidents* (*C9*) as [Dependent] and *Months* (*C8*) as [Exposure]. On Step 2 dialogue enter 1 for Omit Level and leave other entries unchanged. Check all output options to obtain the following output. Some tables have been shortened to save space.

Poisson Regression

Dependent Variable: Accidents
Exposure: Months
Valid Number of Cases: 34, 6 Omitted

Regression Results

	Coeffi	Standar	Wald	Signi	Lower	Upper
	cient	d Error	Statistic	ficance	95%	95%
Constant	-6.4029	0.2175	866.444	0.0000	-6.8292	-5.9765
Type = Type B	-0.5447	0.1776	9.4055	0.0022	-0.8928	-0.1966
Type C	-0.6888	0.3290	4.3818	0.0363	-1.3337	-0.0439
Type D	-0.0743	0.2906	0.0654	0.7981	-0.6438	0.4952
Type E	0.3205	0.2358	1.8485	0.1740	-0.1415	0.7826
Constructed = C 65-69	0.6958	0.1497	21.6190	0.0000	0.4025	0.9892
C 70-74	0.8175	0.1698	23.1665	0.0000	0.4846	1.1503
C 75-79	0.4450	0.2332	3.6397	0.0564	-0.0122	0.9021
Operated = O 75-79	0.3839	0.1183	10.5357	0.0012	0.1521	0.6156

Goodness of Fit Tests

	-2 Log likelihood
Initial Model	244.1948
Final Model	136.8291

	Chi-Square	Degrees of	Right-Tail
	Statistic	Freedom	Probability
Pearson	38.9626	24	0.0276
Likelihood Ratio	107.3657	8	0.0000

	Pseudo R-squared
McFadden	0.4397
Adjusted McFadden	0.3660
Cox & Snell	0.9575
Nagelkerke	0.9582

Correlation Matrix of Regression Coefficients

	Constant	Type = Type B	Type C	Type D	Type E
Constant	1.0000	-0.8115	-0.3783	-0.3706	-0.4680
Type = Type B	-0.8115	1.0000	0.4331	0.4469	0.5698
Type C	-0.3783	0.4331	1.0000	0.2377	0.3132
Type D	-0.3706	0.4469	0.2377	1.0000	0.3349
Type E	-0.4680	0.5698	0.3132	0.3349	1.0000
Constructed = C 65-69	-0.4847	0.0862	0.0360	0.0277	-0.0052
C 70-74	-0.5509	0.2722	0.0458	0.0284	-0.0377
C 75-79	-0.4015	0.2285	0.0968	-0.0962	0.0476
Operated = O 75-79	-0.2164	0.0256	-0.0030	-0.0047	0.0263

	Constructed = C 65-69	C 70-74	C 75-79	Operated = O 75-79
Constant	-0.4847	-0.5509	-0.4015	-0.2164
Type = Type B	0.0862	0.2722	0.2285	0.0256
Type C	0.0360	0.0458	0.0968	-0.0030
Type D	0.0277	0.0284	-0.0962	-0.0047
Type E	-0.0052	-0.0377	0.0476	0.0263
Constructed = C 65-69	1.0000	0.6337	0.4758	-0.1196
C 70-74	0.6337	1.0000	0.5494	-0.2629
C 75-79	0.4758	0.5494	1.0000	-0.3150
Operated = O 75-79	-0.1196	-0.2629	-0.3150	1.0000

Covariance Matrix of Regression Coefficients

	Constant	Type = Type B	Type C	Type D	Type E
Constant	0.0473	-0.0314	-0.0271	-0.0234	-0.0240
Type = Type B	-0.0314	0.0315	0.0253	0.0231	0.0239
Type C	-0.0271	0.0253	0.1083	0.0227	0.0243
Type D	-0.0234	0.0231	0.0227	0.0844	0.0229
Type E	-0.0240	0.0239	0.0243	0.0229	0.0556
Constructed = C 65-69	-0.0158	0.0023	0.0018	0.0012	-0.0002
C 70-74	-0.0204	0.0082	0.0026	0.0014	-0.0015
C 75-79	-0.0204	0.0095	0.0074	-0.0065	0.0026
Operated = O 75-79	-0.0056	0.0005	-0.0001	-0.0002	0.0007

	Constructed =	C 70-74	C 75-79	Operated
	C 65-69			= O 75-79
Constant	-0.0158	-0.0204	-0.0204	-0.0056
Type = Type B	0.0023	0.0082	0.0095	0.0005
Type C	0.0018	0.0026	0.0074	-0.0001
Type D	0.0012	0.0014	-0.0065	-0.0002
Type E	-0.0002	-0.0015	0.0026	0.0007
Constructed = C 65-69	0.0224	0.0161	0.0166	-0.0021
C 70-74	0.0161	0.0288	0.0218	-0.0053
C 75-79	0.0166	0.0218	0.0544	-0.0087
Operated = O 75-79	-0.0021	-0.0053	-0.0087	0.0140

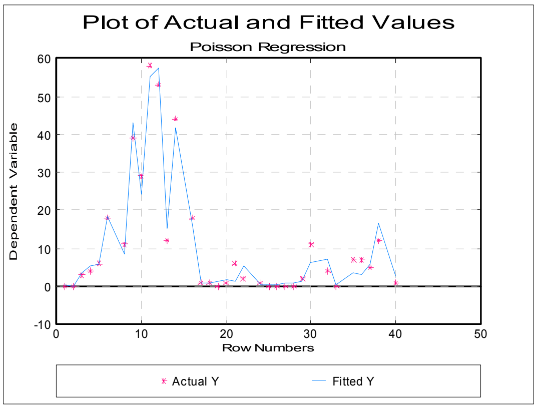
Incidence Rate Ratio

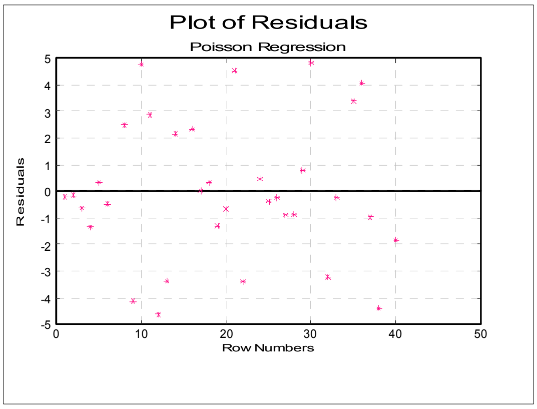
	Incidence	Standard	Lower	Upper
	Rate Ratio	Error	95%	95%
Type = Type B	0.5800	0.1030	0.4095	0.8215
Type C	0.5022	0.1652	0.2635	0.9571
Type D	0.9284	0.2697	0.5253	1.6408
Type E	1.3779	0.3248	0.8680	2.1871
Constructed = C 65-69	2.0054	0.3001	1.4956	2.6890
C 70-74	2.2647	0.3846	1.6235	3.1592
C 75-79	1.5604	0.3640	0.9879	2.4648
Operated = O 75-79	1.4679	0.1736	1.1642	1.8509

Case (Diagnostic) Statistics

	Actual Y	Fitted Y	95% lb	95% ub	Standard Error
			Actual Y	Actual Y	of Fitted
1	0.0000	0.2104	-0.2159	0.6367	0.2175
2	0.0000	0.1532	-0.2858	0.5922	0.2240
3	3.0000	3.6382	3.2553	4.0210	0.1953
…	…	…	…	…	…

	Deviance	Residuals	Standardised
			Residuals
1	0.4208	-0.2104	-0.4587
2	0.3064	-0.1532	-0.3914
3	0.1191	-0.6382	-0.3346
…	…	…	…


Plot of Actual and Fitted Values
Poisson Regression
Dependent Variable
60
50
40
30
20
10
0
-10
0
10
20
30
40
50
Row Numbers
Actual Y
Fitted Y


Plot of Residuals
Poisson Regression
Residuals
5
4
3
2
1
0
-1
-2
-3
-4
-5
0
10
20
30
40
50
Row Numbers

7.2.9. Box-Cox Regression

The ordinary least squares regression assumes normal distribution of residuals. When this is not the case, the Box-Cox Regression procedure may be useful (see Box, G. E. P. and Cox, D. R. 1964). It will transform the dependent variable using the Box-Cox Transformation function and employ maximum likelihood estimation to determine the optimal level of the power parameter lambda. In order to run a Box-Cox Regression, the dependent variable should not contain any non positive values.

Variable selection and Ordinary Least Squares Output dialogues for this procedure are identical to that of Linear Regression. There is a separate output dialogue for maximum likelihood estimation, which contains diagnostic and graphic options for the estimation parameters and normal probability plots of data before and after the transformation. It is possible to run a Box-Cox Regression without any independent variables. In this case the results will be similar to that of Data Transformation procedure with Box-Cox option (available under Statistics 2 → Quality Control menu).

7.2.9.1. Box-Cox Regression Model Description

Box-Cox Regression will transform the dependent variable as follows:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0$$

$$y^{(\lambda)} = \mathrm{Ln}(y) \text{ if } \lambda = 0$$

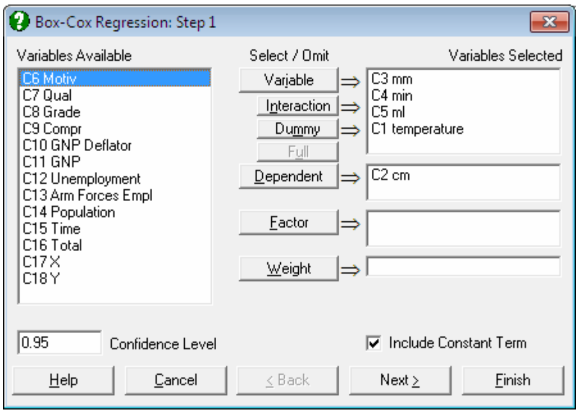
and determine the optimal value of lambda by maximising the following log-likelihood function:

$$L^{(\lambda)} = -\frac{n}{2}\mathrm{Ln}(\hat{\sigma}^2_{(\lambda)}) + (\lambda - 1)\sum_{i=1}^{n}\mathrm{Ln}(y_i)$$

where $\hat{\sigma}^2_{(\lambda)}$ is the estimate of the least squares variance using the transformed y variable.

A golden section minimisation algorithm is employed to minimise the negative of the log likelihood function within the range of -3 ≤ λ ≤ 3. These limits can be changed by the user if necessary and the changes will be stored by the program.

7.2.9.2. Box-Cox Regression Variable Selection



Box-Cox Regression: Step 1

Variables Available
C6 Motiv
C7 Qual
C8 Grade
C9 Compr
C10 GNP Deflator
C11 GNP
C12 Unemployment
C13 Arm Forces Empl
C14 Population
C15 Time
C16 Total
C17 X
C18 Y

Variables Selected
C3 mm
C4 min
C5 ml
C1 temperature

Dependent: C2 cm

0.95 Confidence Level

☑ Include Constant Term

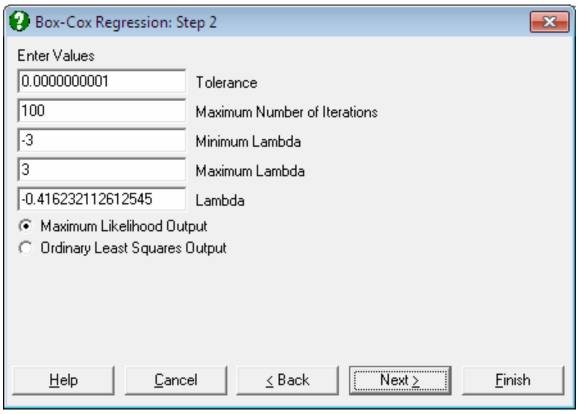
As in Linear Regression, Box-Cox Regression can be used to estimate models with or without a constant term, with or without weights and regressions can be run on a subset of cases as determined by the levels of an unlimited number of factor columns. An unlimited number of dependent variables can be selected in order to run the same model on different dependent variables. It is also possible to include interaction terms, dummy and lag/lead variables in the model, without having to create them as spreadsheet columns first (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

It is compulsory to select at least one numeric data column as a dependent variable. When more than one dependent variable is selected, the analysis will be repeated as many times as the number of dependent variables, each time only changing the dependent variable and keeping the rest of selections unchanged.

You can transform a single variable without using any predictor (independent) variables. In this case the results will be similar to that of the Data Transformation procedure with Box-Cox option (available under Statistics 2 → Quality Control menu). A column containing numeric data can be selected as a weights column.

An intermediate inputs dialogue is displayed next.

7.2.9.3. Box-Cox Regression Intermediary Inputs



Enter Values

0.0000000001 Tolerance

100 Maximum Number of Iterations

-3 Minimum Lambda

3 Maximum Lambda

-0.416232112612545 Lambda

● Maximum Likelihood Output

○ Ordinary Least Squares Output

Tolerance: This value is used to control the sensitivity of minimisation procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

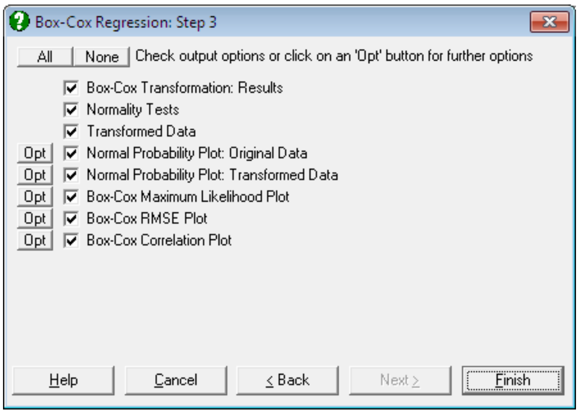
Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Minimum Lambda: Limits for the range where the optimum lambda will be searched can be set. Change this value if the optimal lambda cannot be found within the specified range. If the lambda displayed is the same or very near to this minimum, change it to a smaller value. When the limit is changed, a re-calculation is forced and lambda is estimated again.

Maximum Lambda: Change this value if the optimal lambda cannot be found within the specified range. If the lambda displayed is the same or very near to this maximum, change it to a higher value. When the limit is changed, a re-calculation is forced and lambda is estimated again.

Lambda: You can override the estimated lambda and enter your own value here. You may wish to do this to use a round power value (like -1, -0.5, 0.5, 2). If the estimated lambda is changed, confidence intervals and chi-squared tests for lambda will not be available.

7.2.9.4. Box-Cox Regression Maximum Likelihood Output Options



Check output options or click on an 'Opt' button for further options

☑ Box-Cox Transformation: Results
☑ Normality Tests
☑ Transformed Data
Opt ☑ Normal Probability Plot: Original Data
Opt ☑ Normal Probability Plot: Transformed Data
Opt ☑ Box-Cox Maximum Likelihood Plot
Opt ☑ Box-Cox RMSE Plot
Opt ☑ Box-Cox Correlation Plot

Box-Cox Transformation:

Results: The first output option displays results for the maximum likelihood estimation (see 9.3.7.2. Box-Cox Transformation).

Lambda with Confidence Limits: The confidence interval for optimum lambda is based on the likelihood ratio statistic and it is defined as:

$$f(y,\hat{\lambda}) \geq f(y,\hat{\lambda}) - \frac{\chi^2_{\alpha,1}}{2}$$

Values corresponding to lower and upper bound of lambda are computed separately using an iterational procedure.

Transformation Formula: The equation applied in transforming the dependent variable is displayed. The same equation is also printed on a separate line with estimated parameter values, in a format suitable for cell calculations in Excel. You can simply copy this equation, replace the variable x with a cell reference and run interpolations.

Likelihood Ratio Test: In Box-Cox Regression, this test performed by evaluating the regression equation for lambda fixed at λ_1 = -1, 0 and 1.

$$L = 2\left[f(y,\hat{\lambda}) - f(y,\lambda_1)\right]$$

which is chi-square distributed with one degree of freedom.

Normality Tests: Anderson-Darling Test of normality is performed on the original and transformed dependent variable thus allowing you to judge whether the transformation was useful. No or a small increase in the tail probability indicates that Box-Cox Transformation was not useful.

Transformed Data: The original and transformed dependent variable values and their group membership (if any) are sorted and displayed in a table. If you wish to display the unsorted values, you can use the Case (Diagnostic) Statistics output in Ordinary Least Squares Output option.

If you are using UNISTAT in Stand-Alone Mode, click on the UNISTAT icon on the Output Medium Toolbar to send all output to UNISTAT spreadsheet. In Excel Add-In Mode select the output matrix as data for further calculations.

Normal Probability Plot: Original Data: A Normal Probability Plot of the original data is displayed together with Anderson-Darling Test results in the legend. You can compare this graph with the next one to visualise the improvement provided by the transformation.

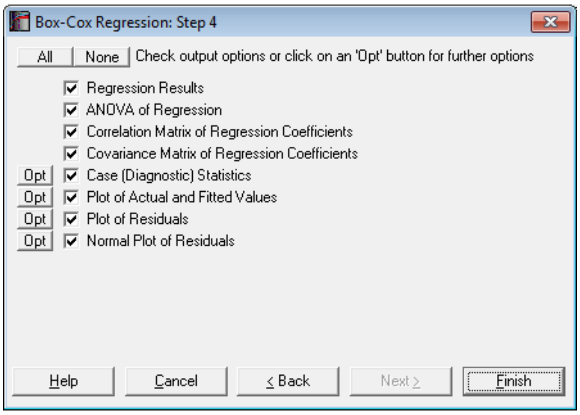
Normal Probability Plot: Transformed Data: A Normal Probability Plot of the transformed data is displayed together with Anderson-Darling Test results in the legend. You can compare this graph with the previous one to visualise the improvement provided by the transformation.

Box-Cox Maximum Likelihood Plot: Log likelihood values are plotted against the specified range of lambda. Lambda and its confidence limits are indicated by vertical lines. A horizontal line is drawn for the log likelihood value corresponding to confidence limits.

Box-Cox Root Mean Square Error Plot: Root mean square error (RMSE) of regression is plotted against lambda.

Box-Cox Correlation Plot: Values of the regression correlation coefficient are plotted against lambda.

7.2.9.5. Box-Cox Regression Ordinary Least Squares Output Options



Check output options or click on an 'Opt' button for further options

☑ Regression Results
☑ ANOVA of Regression
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
☑ Case (Diagnostic) Statistics
☑ Plot of Actual and Fitted Values
☑ Plot of Residuals
☑ Normal Plot of Residuals

All output options are as in Linear Regression. The transformed dependent variable is used. The unsorted values for the transformed dependent variable can be accessed from the Case (Diagnostic) Statistics output option.

7.2.9.6. Box-Cox Regression Example

Open REGRESS and select Statistics 1 → Regression Analysis → Box-Cox Regression. From the Variable Selection Dialogue select *temperature*, *mm, min* and *ml* (*C1, C3-C5*) as [Variable]s and *cm* (*C2*) as [Dependent]. On Step 2 leave convergence parameters unchanged.

The Maximum Likelihood Output option generates the following output.

Box-Cox Regression

Box-Cox Transformation: Results

Variables Selected: cm

	Value	Lower 95%	Upper 95%
Lambda	-0.4162	-2.7747	1.7807

Box-Cox Transformation:
y = (y ^ Lambda - 1) / Lambda
y = (POWER(y, -0.416232112612545) - 1) / -0.416232112612545

Lambda	Chi-Square	DoF	Probability
-1	0.2511	1	0.6163
0	0.1318	1	0.7165
1	1.5689	1	0.2104

Log of Likelihood =	-8.0864

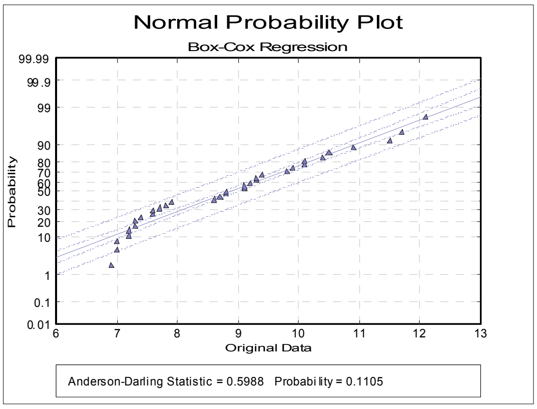
Normality Tests

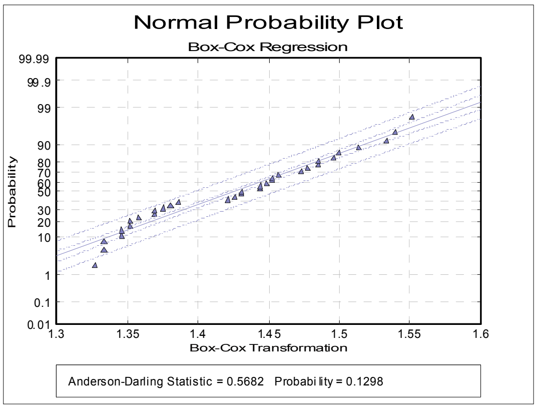
Smaller probabilities indicate non-normality.

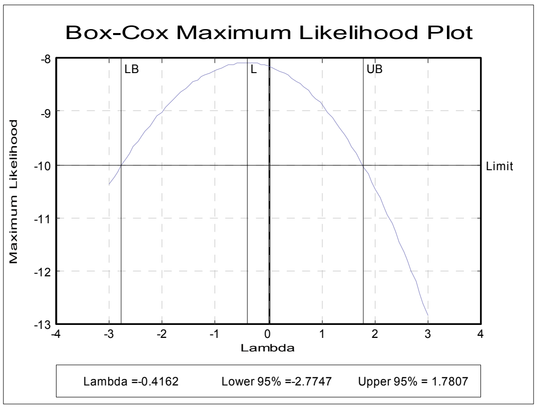
	A-D Stat	Probability
Original Data	0.5988	0.1202
Transformed Data	0.5682	0.1434

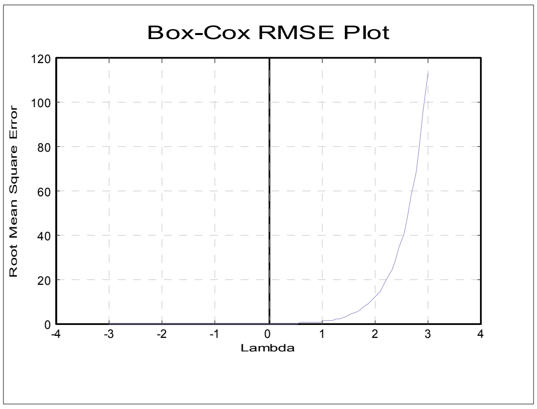
Transformed Data

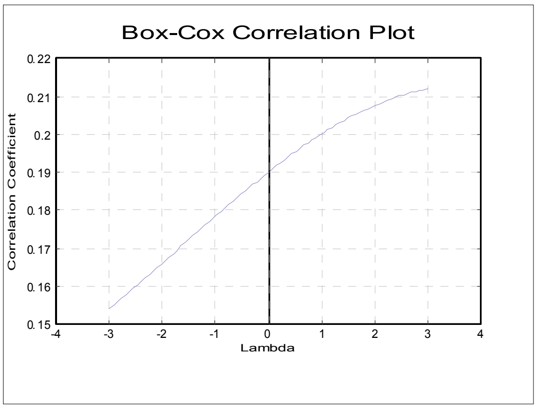
	Original Data	Transformed Data
1	6.9000	1.3273
2	7.0000	1.3337
3	7.0000	1.3337
…	…	…
31	11.5000	1.5332
32	11.7000	1.5394
33	12.1000	1.5514




Normal Probability Plot
Box-Cox Regression
Probability
99.99
99.9
99
90
80
70
60
50
30
20
10
1
0.1
0.01
1.3
1.35
1.4
1.45
1.5
1.55
1.6
Box-Cox Transformation
Anderson-Darling Statistic = 0.5682   Probability = 0.1298


Box-Cox Maximum Likelihood Plot
Maximum Likelihood
-8
-9
-10
-11
-12
-13
LB
L
UB
Limit
-4
-3
-2
-1
0
1
2
3
4
Lambda
Lambda =-0.4162   Lower 95% =-2.7747   Upper 95% = 1.7807


Box-Cox RMSE Plot
Root Mean Square Error
120
100
80
60
40
20
0
-4
-3
-2
-1
0
1
2
3
4
Lambda



Select the Ordinary Least Squares Output option and check only the Regression Results option to obtain the following ordinary least squares regression output.

Box-Cox Regression

Dependent Variable: cm
Valid Number of Cases: 33, 0 Omitted

Regression Results

	Coefficient	Standard Error	t-Statistic	Significance	Lower 95%	Upper 95%
Constant	1.6405	0.1886	8.6991	0.0000	1.2542	2.0268
temperature	0.0054	0.0048	1.1318	0.2673	-0.0044	0.0153
mm	-0.0436	0.0302	-1.4439	0.1599	-0.1055	0.0183
min	0.0125	0.0114	1.0896	0.2852	-0.0110	0.0359
ml	-0.0052	0.0285	-0.1813	0.8574	-0.0636	0.0532

Residual Sum of Squares =	0.1150
Standard Error =	0.0641
Mean of Y =	1.4271
Standard Deviation of Y =	0.0664
Correlation Coefficient =	0.4305
R-squared =	0.1854
Adjusted R-squared =	0.0690
F(4,28) =	1.5927
Significance of F =	0.2038
Durbin-Watson Statistic =	1.4016
Press Statistic =	0.1800

7.3. Analysis of Variance and General Linear Model

7.3.0. Overview

As we have seen in Chapter 6, t-test is the appropriate procedure for testing whether two samples belong to the same population (where the null hypothesis tested is expressed as “H_0: $\mu_1 = \mu_2$”). When we need to test whether three (or more) samples belong to the same population (the null hypothesis “H_0: $\mu_1 = \mu_2 = \mu_3$”), it may look as though performing a series of t-tests between all possible pairs of samples (the three null hypotheses “H_0: $\mu_1 = \mu_2$”, “H_0: $\mu_1 = \mu_3$” and “H_0: $\mu_2 = \mu_3$”), would solve the problem. Unfortunately this is not the case, since each t-test is associated with a confidence level (say 0.95) and when three are performed in a row, the final confidence level would drop to 0.95 x 0.95 x 0.95 = 0.86. Or in other words, the chance of rejecting the null hypothesis when it is in fact true (Type I error) would increase to 14%. As the number of samples to test increases, the chance of introducing an error would also increase. Analysis of Variance (ANOVA) was designed to overcome this problem by R A Fisher in the 1920s.

The data for a simple ANOVA problem consists of a number of measurements taken from a number of different groups. An example would be the weights of a sample of five people from four different regions of the country. The criterion used in grouping (country) is called a *factor* and each group (North, South, East West) a *level* of the factor. If an ANOVA problem has only one factor, then it is called a one-way ANOVA. There may, however, be another factor defined on the same set of measurements, such as sex (a factor with two levels), which makes the problem a two-way ANOVA. In this case, we can compare the means of groups defined by each factor separately (the main effects) and also compare the means of groups defined by the combinations of the two factor levels (interactions), males in the North, females in the East, etc. In theory, there are no limits to the number of factors and the number of interactions that can be defined in an ANOVA design. ANOVA and GLM procedures assume that each factor has a maximum of 2000 levels, though this number may be increased by entering and editing the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] group:

```
MaxFactorLevel=2000
```

Also see 3.2.12. Long String Table.

Let k be the number of groups and n_i the number of observations in group i for i = 1,…, k. in a one-way ANOVA problem. Let us also define the total number of observations as:

$$N = \sum_{i=1}^{k} n_i$$

the mean of group i as:

$$\overline{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

and the grand mean as:

$$\overline{y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{N}$$

If we express the deviation of an observation from the grand mean as the sum of its deviation from its own group mean (within-group) plus the deviation of its group mean from the grand mean (between-group) we have:

$$y_{ij} - \overline{y} = (y_{ij} - \overline{y}_i) + (\overline{y}_i - \overline{y})$$

If we then take the squares of both sides, sum over i and j and rearrange, we obtain:

$$\sum_{i,j} (y_{ij} - \overline{y})^2 = \sum_{i,j} (y_{ij} - \overline{y}_i)^2 + \sum_{i,j} (\overline{y}_i - \overline{y})^2$$

In other words:

Total Ssq = Within-Groups Ssq + Between-Groups Ssq

where Ssq stands for sum of squares.

Our aim is to test the null hypothesis that “all means are equal”. Therefore, the entity we are interested in is the Between-Groups Ssq. Since the Within-Groups Ssq term represents the rest of variation in data, we can also call it the *Error Term.* We construct the ANOVA table as follows:

	Sum of Squares (Ssq)	Degrees of Freedom	Mean Squares (MSQ)	F-Statistic	Probability
Factor	Between-Groups Ssq	k - 1	Between-Groups Ssq / (k - 1)	Between-Groups MSQ / Within-Groups MSQ	P-value for $F_{(k-1)(N-k)}$
Error	Within-Groups Ssq	N - k	Within-Groups Ssq / (N - k)		
Total	Total Ssq	N - 1	Total Ssq / (N - 1)		

The F-statistic for the *Factor* is the test statistic. The associated one-tail probability from the F-distribution is calculated with (k - 1) and (N - k) degrees of freedom and it is reported in the last column of the ANOVA table. If this p-value is less than or equal to a given confidence level (usually 5%), then we reject the null hypothesis "H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$", or in other words, we conclude that the k samples tested do not belong to the same population.

Note that all we can conclude using ANOVA is whether the k population means are all equal or not. If they are not equal, then ANOVA does not tell us which ones are different. In order to find out which pairs or groups of population means are different, one of the Multiple Comparisons tests should be used. If the ANOVA design is more complicated than one-way, then the General Linear Model procedure may also be used to find out the significantly different groups (see 7.3.2.3. GLM Output Options).

7.3.0.1. ANOVA and GLM Data Format

In order to analyse ANOVA and GLM designs in a general purpose statistical program, the data should be organised in an accurate and logical way. The approach adopted in almost all serious statistical packages involves stacking all measurement data (the explanatory variable) in a single column, and expressing the various group memberships of these observations in separate corresponding categorical data columns (factors). The user will often face the problem of having to convert a published data table into this format.

For instance, consider the Randomised Block Designs ANOVA example given in Table 5-4 p. 140 by Montgomery, D. C. (1991). Measurements on *Hardness* are given in the following format:

	Coupon (Block)			
Type of Tip	**1**	**2**	**3**	**4**
1	-2	-1	1	5
2	-1	-2	3	4
3	-3	-1	0	2
4	2	1	5	7

This should be entered into UNISTAT as follows:

Hardness	Tip	Coupon
-2	1	1
-1	1	2
1	1	3
5	1	4
-1	2	1
-2	2	2
3	2	3
4	2	4
-3	3	1
-1	3	2
0	3	3
2	3	4
2	4	1
1	4	2
5	4	3
7	4	4

If the data has already been entered into a spreadsheet in the form of a table, you do not have to retype it in the above format manually. UNISTAT's own spreadsheet Data Processor provides a number of functions that will help you to do the transformation automatically. The Data → Stack Columns procedure can be used to stack the hardness measurements in a column and create the *Tip* column automatically (see 3.3.9. Stack Columns).

You may also use the function **Level()** which is designed to generate factor columns containing regular (balanced) levels automatically (see 3.4.2.5. Statistical Functions). To do this, first create the data column and then enter the function **Level(4);B** in a blank column (to create the *Tip* column) and **Level(4)** into the next blank column (to create the *Coupon* column).

For the analysis, select *Tip* and *Coupon* as [Factor]s and *Hardness* as [Dependent].

Let us also consider a more complex example known as Graeco Latin Squares Designs given in Table 5-20 p. 168 by Montgomery, D. C. (1991).

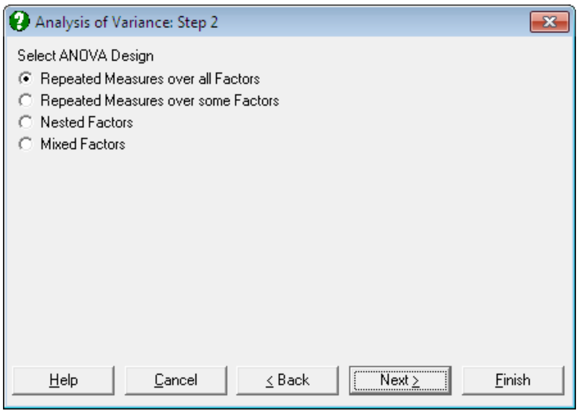
Batches of Raw	Operators				
Material	1	2	3	4	5
1	Aα = -1	Bχ = -5	Cε = -6	Dβ = -1	Eδ = -1
2	Bβ = -8	Cδ = -1	Dα = 5	Eχ = 2	Aε = 11
3	Cχ = -7	Dε = 13	Eβ = 1	Aδ = 2	Bα = -4
4	Dδ = 1	Eα = 6	Aχ = 1	Bε = -2	Cβ = -3
5	Eε = -3	Aβ = 5	Bδ = -5	Cα = 4	Dχ = 6

This table is entered into UNISTAT as follows:

Operator	Batch	Formulation	Test assemblies	Coded Data
1	1	A	a	-1
2	1	B	c	-5
3	1	C	e	-6
4	1	D	b	-1
5	1	E	d	-1
1	2	B	b	-8
2	2	C	d	-1
3	2	D	a	5
4	2	E	c	2
5	2	A	e	11
1	3	C	c	-7
2	3	D	e	13
3	3	E	b	1
4	3	A	d	2
5	3	B	a	-4
1	4	D	d	1
2	4	E	a	6
3	4	A	c	1
4	4	B	e	-2
5	4	C	b	-3
1	5	E	e	-3
2	5	A	b	5
3	5	B	d	-5
4	5	C	a	4
5	5	D	c	6

where *Formulation*, *Batch*, *Operator* and *Test assemblies* are the factors and *Coded Data* is the dependent variable.

7.3.0.2. ANOVA Designs


Analysis of Variance: Step 2

Select ANOVA Design

● Repeated Measures over all Factors
○ Repeated Measures over some Factors
○ Nested Factors
○ Mixed Factors

It is possible to test a large number of experimental designs using UNISTAT's Analysis of Variance and GLM procedures. In this section we shall describe seven major designs and demonstrate how we can solve these problems using UNISTAT with the help of published examples.

7.3.0.2.1. Randomised Block Designs

In many ANOVA problems, it is desirable to control the variability from known nuisance factors (blocks). We want to remove this variability from the error sum of squares to increase the power of the test. For example if we wish to determine the effectiveness of different fertilisers on a particular crop, we might try each fertiliser on the crop in a number of different fields. But the soil in each field might not be of the same quality and this would add variability to the results. As a result, the experimental error will reflect both the random error and the variability between fields. A better design would be the randomised block design, where each fertiliser is tested in each field. A randomised block design is said to be *complete* if all the treatments are used in all the blocks.

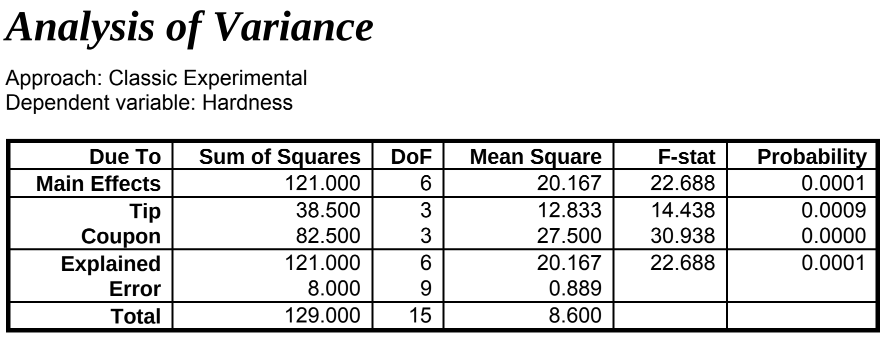
	Block 1	Block 2		Block b
Treatment 1	y_{11}	y_{12}	…	y_{1b}
Treatment 2	y_{21}	y_{22}	…	y_{2b}
	.	.		.
	.	.		.
Treatment a	y_{a1}	y_{a2}	…	y_{ab}

These designs can be constructed in UNISTAT using one of ANOVA or GLM procedures.

Example

Table 5-4 on p. 140 from Montgomery, D. C. (1991). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Hardness* and two factor columns *Tip* and *Coupon* created to keep track of the group memberships. Therefore, the resulting data matrix should have 16 rows and 3 columns.

Open ANOVA, select Statistics 1 → ANOVA and GLM → Analysis of Variance, and select *Tip* (*C2*) and *Coupon* (*C3*) as [Factor]s and *Hardness* (*C1*) as [Dependent]. Then select Classic Experimental Approach and no interaction terms to obtain the following ANOVA table:



Analysis of Variance

Approach: Classic Experimental
Dependent variable: Hardness

Due To	Sum of Squares	DoF	Mean Square	F-stat	Probability
Main Effects	121.000	6	20.167	22.688	0.0001
Tip	38.500	3	12.833	14.438	0.0009
Coupon	82.500	3	27.500	30.938	0.0000
Explained	121.000	6	20.167	22.688	0.0001
Error	8.000	9	0.889		
Total	129.000	15	8.600		

The result is the F-statistic and its probability value for *Tip*. Since the probability 0.0009 is much smaller than 5%, we reject the null hypothesis and conclude that the means differ significantly. In other words, the conclusion is that the type of tip affects the hardness values. The F-statistic and its probability for *Coupon* is not strictly meaningful, but informally we can see that the *Coupon* were an important source of variation in the resulting hardness, and power of the analysis has been increased by using a Randomised Block Designs.

7.3.0.2.2. Repeated Measures Designs

It is often the case that repeated measurements are taken on the same subject. This typically happens when the subjects are people and measurements are taken from the same people at different times, following a particular treatment.

If the subject receives different treatments and the treatments are administered in a random order, then it is possible to regard the experiment as having a Randomised Block Designs with the subject as a blocking factor. If the subject receives different treatments in a specified order, then it is possible to regard the experiment as a Crossover Designs. If each subject receives the same treatment a number of times, this should be considered as a repeated measures design.

In a repeated measures design, the total sum of squares can be partitioned into the Between Subjects sum of squares and the Within Subjects sum of squares. It is assumed that these terms are statistically independent. This means that the residual sum of squares can also be partitioned into Error Between Subjects and Error Within Subjects. These terms can be used to find the various factor F-ratios to increase the power of the analysis.

7.3.0.2.2.1. Repeated Measures over all Factors

This is the case when each subject only receives one level of each factor. The two-factor experiment of this kind is shown schematically below.

		c_1	c_2	…	c_r
a_1	b_1	X_{111}	X_{112}	…	X_{11r}
	⋮	⋮	⋮		⋮
	b_q	X_{1q1}	X_{1q2}	…	X_{1qr}
⋮	⋮				
a_p	b_1	X_{p11}	X_{p12}	…	X_{p1r}
	⋮	⋮	⋮		⋮
	b_q	X_{pq1}	X_{pq2}	…	X_{pqr}

Each row represents a group of subjects and each subject is measured on r occasions. This means that the factors A and B sum of squares are contained in the Between Subjects sum of squares. Only the *Trial* (factor C) sum of squares is contained in the Within Subjects sum of squares.

This can be done using the Analysis of Variance (ANOVA) procedure and selecting the Repeated Measures over all Factors option or using the General Linear Model (GLM). In order to use the GLM procedure, an additional factor column must be created in the spreadsheet to give information about the repeated measure (the *Trial* factor). In ANOVA with Repeated Measures over all Factors, this additional factor is created internally by the program, assuming a repeated measure across all the factors.

Example 1: Using ANOVA with Repeated Measures over all Factors

Table 7.4-3 on p. 341 from Winer, B. J. (1970). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Score* and three factor columns created *Subject*, *Anxiety* and *Tension*.

Open ANOVA and select Statistics 1 → ANOVA and GLM → Analysis of Variance. Select *Score* (*C24*) as [Dependent], *Subject* (*C23*) as [Repeated], *Anxiety*

(*C21*) and *Tension* (*C22*) as [Factor]s. From the next two dialogues select the Repeated Measures over all Factors and Classic Experimental Approach options and include all interactions at the last dialogue.

Analysis of Variance

Design: Repeated Measures over all Factors
Approach: Classic Experimental
Dependent Variable: Score

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Between Subjects	181.000	11	16.455		
Anxiety	10.083	1	10.083	0.978	0.3517
Tension	8.333	1	8.333	0.808	0.3949
Anxiety × Tension	80.083	1	80.083	7.766	0.0237
Error Between	82.500	8	10.313		
Within Subjects	1077.000	36	29.917		
Trial	991.500	3	330.500	152.051	0.0000
Anxiety × Trial	8.417	3	2.806	1.291	0.3003
Tension × Trial	12.167	3	4.056	1.866	0.1624
Anxiety × Tension × Trial	12.750	3	4.250	1.955	0.1477
Error Within	52.167	24	2.174		
Total	1258.000	47	26.76596		

Example 2: Using General Linear Model

To use the GLM procedure with this example, an extra factor column *Trial* must be created. This is defined as the number of times a measurement is made on a particular subject. So the second measurement taken with subject i would result in a 2 in the *Trial* column, the third measurement with subject j would results in a 3 in the *Trial* column, etc.

Anxiety	Tension	Subject	Trial
1	1	1	1
1	2	1	2
2	1	1	3
2	2	1	4
1	1	2	1
1	2	2	2
2	1	2	3
2	2	2	4

So the two factors, *Anxiety* and *Tension* play no part in the calculation of the *Trial* column.

Open ANOVA and select Statistics 1 → ANOVA and GLM → General Linear Model. Select *Score* (*C24*) as [Dependent], *Subject* (*C23*) as [Repeated]. Select the following terms as factors, and at the following dialogue select the F-Statistic denominators as shown.

C21 Anxiety	Error Between C23 Subject
C22 Tension	Error Between C23 Subject
C21 Anxiety × C22 Tension	Error Between C23 Subject
C25 Trial	Error Within C23 Subject
C21 Anxiety × C25 Trial	Error Within C23 Subject
C22 Tension × C25 Trial	Error Within C23 Subject
C21 Anxiety × C22 Tension × C25 Trial	Error Within C23 Subject

From the Output Options Dialogue select only the ANOVA option to obtain the following results.

General Linear Model

ANOVA

Dependent Variable: Score

Due To	Sum of Squares	DoF	Mean Square	F-Stat		Prob
Constant	4800.000	1	4800.000	35.644		0.0000
Between Subjects	181.000	11	16.455			
Anxiety	10.083	1	10.083	0.978	a	0.3517
Tension	8.333	1	8.333	0.808	a	0.3949
Anxiety × Tension	80.083	1	80.083	7.766	a	0.0237
Error Between	82.500	8	10.313			
Within Subjects	1077.000	36	29.917			
Trial	991.500	3	330.500	152.051	b	0.0000
Anxiety × Trial	8.417	3	2.806	1.291	b	0.3003
Tension × Trial	12.167	3	4.056	1.866	b	0.1624
Anxiety × Tension × Trial	12.750	3	4.250	1.955	b	0.1477
Error Within	52.167	24	2.174			
Explained	1123.333	15	74.889	17.795		0.0000
Error	134.667	32	4.208			
Total	1258.000	47	26.766			

R-squared =	0.8930
Adjusted R-squared =	0.8428

a F-Statistic: Error Between
b F-Statistic: Error Within

In this particular example, ANOVA with the Repeated Measures over all Factors option and GLM produce exactly the same results. This would not always be the case since GLM always adopts the Regression Approach and with ANOVA you can select different approaches. However with a balanced design (as above) all three approaches will give the same result.

7.3.0.2.2.2. Repeated Measures over some Factors

This is the case when each subject receives only one level of some factors but all levels of other factors. The three-factor experiment of this kind is shown schematically below.

	b_1			…	b_q		
	c_1	…	c_r	…	c_1	…	c_r
a_1	X_{111}	…	X_{11r}	…	X_{1q1}		X_{1qr}
a_2	X_{211}	…	X_{21r}	…	X_{2q1}		X_{2qr}
⋮	⋮		⋮		⋮		⋮
a_p	X_{p11}	…	X_{p1r}	…	X_{pq1}		X_{pqr}

Each row represents a group of subjects and each subject only receives one level of factor A. However each subject receives all levels of factors B and C. This means that only the factor A sum of squares is contained in the Between Subjects sum of squares. The factor B and C sum of squares is contained in the Within Subjects sum of squares.

Example

Table 7.3-3 on p. 324 from Winer, B. J. (1970). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Score* and five factor columns created *Subject*, *Noise*, *Period*, *Dial* and *SubjWGrps*.

The artificial factor *SubjWGrps* has been set up to partition the Error Within Subjects, and these partitions are used for the F-statistic ratios in this example. The factor *SubjWGrps* is equivalent to *Subject(Noise)* nest term. However since terms of the form *FactorA* × *FactorB(FactorC)* cannot be specified, *SubjWGrps* needs to be built explicitly. The table below shows how this is done. Each combination of *Subject* and *Noise* results in a different level. However, when each level of *Noise* is met for the first time this is pooled into level 0.

Subject	Noise	Count	SubjWGrps
1	High	1	0
2	High	2	2
3	High	3	3
4	Low	4	0
5	Low	5	5
6	Low	6	6

The *SubjWGrps* column calculated here has nothing to do with the *Trial* column calculated in the previous example (see 7.3.0.2.2.1. Repeated Measures over all Factors). In fact it can informally be considered as the opposite of the *Trial* column from the previous example.

Open ANOVA and select Statistics 1 → ANOVA and GLM → General Linear Model. Select *Score* (*C26*) as [Dependent] and *Subject* (*C27*) as [Repeated]. Select the following terms as factors, and on the following dialogue select the F-statistic denominators as shown.

C30 Noise	Error Between C27 Subject
C29 Period	C29 Period × C31 SubjWGrps
C29 Period × C30 Noise	C29 Period × C31 SubjWGrps
C29 Period × C31 SubjWGrps	Error Term
C28 Dial	C28 Dial × C31 SubjWGrps
C28 Dial × C30 Noise	C28 Dial × C31 SubjWGrps
C28 Dial × C31 SubjWGrps	Error Term
C28 Dial × C29 Period	C28 Dial × C29 Period × C31 SubjWGrps
C28 Dial × C29 Period × C30 Noise	C28 Dial × C29 Period × C31 SubjWGrps
C28 Dial × C29 Period × C31 SubjWGrps	Error Term

The following results are obtained:

General Linear Model

ANOVA

Dependent Variable: Score

Due To	Sum of Squares	DoF	Mean Square	F-Stat		Prob
Constant	105868.167	1	105868.167	35.782		0.0000
Between Subjects	2959.278	5	591.856			
Noise	468.167	1	468.167	0.751	a	0.4348
Error Between	2491.111	4	622.778			
Within Subjects	6965.556	48	145.116			
Period	3722.333	2	1861.167	63.389	b	0.0000
Period × Noise	333.000	2	166.500	5.671	b	0.0293
Dial	2370.333	2	1185.167	89.823	c	0.0000
Dial × Noise	50.333	2	25.167	1.907	c	0.2102
Dial × Period	10.667	4	2.667	0.336	d	0.8499
Dial × Period × Noise	11.333	4	2.833	0.357	d	0.8357
Error Within	467.556	32	14.611			
Period × SubjWGrps	234.889	8	29.361			
Dial × SubjWGrps	105.556	8	13.194			
Dial × Period × SubjWGrps	127.111	16	7.944			
Explained	6966.167	17	409.775	4.986		0.0000
Error	2958.667	36	82.185			
Total	9924.833	53	187.261			

a F-Statistic: Error Between
b F-Statistic: Period × SubjWGrps
c F-Statistic: Dial × SubjWGrps
d F-Statistic: Dial × Period × SubjWGrps

The Between Subjects sum of squares and the Within Subjects sum of squares partition the Total sum of squares. The Error Between and the Error Within partition the Total Error sum of squares. And the *Period* × *SubjWGrps*, *Dial* × *SubjWGrps* and *Dial* × *Period* × *SubjWGrps* errors partition the Error Within.

So, in the example output above, *Period* × *SubjWGrps*, *Dial* × *SubjWGrps* and *Dial* × *Period* × *SubjWGrps* are the terms that are not included in the model. The factor *Noise* is Between Subjects. The terms *Period*, *Period* × *Noise*, *Dial*, *Dial* × *Noise*, *Dial* × *Period*, and *Dial* × *Period* × *Noise* are Within Subjects. All the Between Subjects and Within Subjects terms are included in the model.

7.3.0.2.3. Latin Squares Designs

Consider an experiment to compare k treatments in which there are two nuisance factors (blocks) each at k levels (this is more common than it sounds). A complete factorial design with one observation at each level would need k^3 observations, but a *Latin Square* needs only k-Squared observations. Consider the following design with k = 5. The treatments are A, B, C, D and E, the two other sources of variation are represented by the rows and columns of the table.

	Column				
Row	**1**	**2**	**3**	**4**	**5**
1	A	B	C	D	E
2	E	A	B	C	D
3	D	E	A	B	C
4	C	D	E	A	B
5	B	C	D	E	A

Only k-Squared (= 25) observations are made, since at each combination of a row and a column only one of the five treatments is used. Each treatment occurs in each row and column precisely once. It is assumed that there are no interactions between the three factors.

These designs are constructed in UNISTAT using the ANOVA procedure. Treatment, columns and rows are selected as factors and all interaction terms are omitted. The main result is the F-statistic on the treatment.

Example

Table 5-11 on p. 159 from Montgomery, D. C. (1991). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Coded Data* and three factor columns *Operator*, *Batch* and *Formulation* created to keep track of the group memberships.

Open ANOVA and select Statistics 1 → ANOVA and GLM → Analysis of Variance, *Operator* (*C4*), *Batch* (*C5*) and *Formulation* (*C6*) as [Factor]s and *Coded Data* (*C7*) as [Dependent]. Then select Classic Experimental Approach and omit all interaction terms to obtain the following ANOVA table:

Analysis of Variance

Approach: Classic Experimental
Dependent Variable: Coded Data

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Main Effects	548.000	12	45.667	4.281	0.0089
Operator	150.000	4	37.500	3.516	0.0404
Batch	68.000	4	17.000	1.594	0.2391
Formulation	330.000	4	82.500	7.734	0.0025
Explained	548.000	12	45.667	4.281	0.0089
Error	128.000	12	10.667		
Total	676.000	24	28.167		

The result is the *Formulation* F-statistic and its tail probability. The *Batch* and *Operator* variables are nuisance factors (blocks) which are removed from the error sum of squares to increase the power of the test. The F-statistic and probability for *Batch* and *Operator* are not strictly meaningful, but informally we can see that the power of the test has been increased by including them in the design.

7.3.0.2.4. Graeco Latin Squares Designs

Consider a Latin square design with another factor added. Denote the levels of this extra factor by Greek letters. If each Latin letter appears once and only once with each Greek letter then the design is called a Graeco-Latin square.

	Column			
Row	1	2	3	4
1	Aα	Bβ	Cχ	Dδ
2	Bδ	Aχ	Dβ	Cα
3	Cβ	Dα	Aδ	Bχ
4	Dχ	Cδ	Bα	Aβ

A Graeco-Latin square design exists for all k ≥ 3 except for k = 6. The Graeco-Latin square design allows investigation of four factors (rows, columns, Latin letters and Greek letters), each at k levels with only k-Squared observations.

These designs are constructed in UNISTAT using ANOVA. Selecting the Latin letters, Greek letters, columns and rows as factors. All interaction terms are omitted. The main result is the F-statistic on the Latin letters.

Example

Table 5-20 on p. 168 from Montgomery, D. C. (1991). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Coded Data* and four factor columns *Operator*, *Batch*, *Formulation* and *Test* created to keep track of the group memberships.

Open ANOVA, select Statistics 1 → ANOVA and GLM → Analysis of Variance, select *Operator* (*C4*), *Batch* (*C5*), *Formulation* (*C6*) and *Test assemblies* (*C8*) as [Factor]s and *Coded Data* (*C7*) as [Dependent]. Then select Classic Experimental Approach and no interaction terms (the default) to obtain the following ANOVA table:

Analysis of Variance

Approach: Classic Experimental
Dependent Variable: Coded Data

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Main Effects	610.000	16	38.125	4.621	0.0171
Operator	150.000	4	37.500	4.545	0.0329
Batch	68.000	4	17.000	2.061	0.1783
Formulation	330.000	4	82.500	10.000	0.0033
Test assemblies	62.000	4	15.500	1.879	0.2076
Explained	610.000	16	38.125	4.621	0.0171
Error	66.000	8	8.250		
Total	676.000	24	28.167		

The result is the *Formulation* F-statistic and its probability. The *Batch*, *Operator* and *Test assemblies* variables are nuisance factors (blocks) which are removed from the error sum of squares to increase the power of the test. The F-statistic and its probability for *Batch*, *Operator* and *Test assemblies* are not strictly meaningful, but informally we can see that the power of the test has been increased by including them in the design.

7.3.0.2.5. Split-Plot Designs

In some experimental designs, one of the factors may be a sub-unit of another factor. For example a field may be divided into main plots and these main plots split into sub plots. If one factor is allocated to the main plots and another factor to their sub plots, then we have a Split-Plot design. The sub plots factor is compared against the variation between sub plots, the main plots factor is compared against the variation between the main plots. The variation within the main plots (between the sub plots) is likely to be less than the variation between the main plots. So the sub plots factor is tested with more power. The main plots factor is said to be confounded with blocks.

These designs are also called split-unit designs, in which case the terms main units and sub units are used instead of main plots and sub plots. Some examples of main plots and sub plots are as follows:

Main Plot	Sub Plot
Days	Hours within day
Subject	Occasion with the subject
Field	Area within the field

These designs are constructed in UNISTAT using ANOVA with the Repeated Measures over some Factors option. Select the main plot as the first factor and the sub plot as the first repeated measure.

Example

Example 9.5 on p. 266 from Armitage & Berry (2002). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Swabs* and three factor columns *Crowding*, *Status* and *Family* created to keep track of the group memberships.

Open ANOVA, select Statistics 1 → ANOVA and GLM → Analysis of Variance and select *Crowding* (*C9*) and *Status* (*C10*) as [Factor]s, *Family* (*C11*) as [Repeated] and *Swabs* (*C12*) as [Dependent]. From the next two dialogues select the Repeated Measures over some Factors and Classic Experimental Approach options. At the interaction terms dialogue check the only interaction term *Crowding* × *Status* to obtain the following ANOVA table:

Analysis of Variance

Design: Repeated Measures over some Factors
Approach: Classic Experimental
Dependent Variable: Swabs

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Main Effects	2004.156	6	334.026	13.214	0.0000
Crowding	470.489	2	235.244	5.223	0.0190
Error (Family)	675.600	15	45.040		
Status	1533.667	4	383.417	15.167	0.0000
2 Way Interactions	72.400	8	9.050	0.358	0.9384
Crowding × Status	72.400	8	9.050	0.358	0.9384
Explained	2076.556	14	148.325	5.868	0.0000
Error	1516.733	60	25.279		
Total	4268.889	89	47.965		

Crowding is compared against the between family variation instead of the overall error term. This increases the power of the test on the *Crowding* effect. The interaction between *Crowding* and *Status* is not significant, so we might consider removing it from the model.

7.3.0.2.6. Nested Designs

Consider an experiment where the levels of one factor (child) are different depending on the level of another factor (parent). For example the parent factor may be *Country* and the child factor *Region*. The north of England is not related to the north of Germany and thus *Region* is a nested factor of *Country*.

Nested designs resemble factorial designs with certain cells missing. This is because one factor is nested under another so that not all combinations of the two factors are observed.

These designs are constructed in UNISTAT using ANOVA selecting the parent as a factor and the child as the corresponding repeated measure. Then the Nested Factors option is selected from the next dialogue.

Example

Table 13-3 on p. 443 from Montgomery, D. C. (1991). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Purity* and two factor columns *Supplier* and *Batches* created to keep track of the group memberships.

Open ANOVA and select Statistics 1 → ANOVA and GLM → Analysis of Variance. Select *Supplier* (*C13*) as [Factor], *Batches* (*C14*) as [Repeated] and *Purity* (*C15*) as [Dependent]. Then select Nested Factors and Classic Experimental Approach from the next two dialogues to obtain the following results:

Analysis of Variance

Design: Nested Factors
Approach: Classic Experimental
Dependent Variable: Purity

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Main Effects	84.972	11	7.725	2.927	0.0135
Supplier	15.056	2	7.528	2.853	0.0774
Batches(Supplier)	69.917	9	7.769	2.944	0.0167
Explained	84.972	11	7.725	2.927	0.0135
Error	63.333	24	2.639		
Total	148.306	35	4.237		

We conclude that the difference between *Batches* is a source of variation.

7.3.0.2.7. Crossover Designs

Crossover Designs occur when subjects are reused, typically in time. For instance, in an experiment designed to compare the effects of drugs A and B, half the sample (chosen at random) take drug A and the remaining half take drug B at the start of the experiment. Sometime later the first sample now take drug B and the second sample take drug A. It is important that effect of the first drug taken does not carryover and affect the performance of the second drug taken. If this does happen it is called the carryover effect.

In UNISTAT the crossover design is analysed in two steps. The first step tests whether the carryover effect is significant. If the carryover effect is not significant then a standard ANOVA can be used on the remaining factors. If the carryover effect is significant then analysis should be restricted to the first trial, and in future experiments a larger time period left between the trials.

The significance of the carryover effect is tested using a Split-Plot design (see 7.3.0.2.5. Split-Plot Designs) of the treatment order against the subjects. To do this a factor column needs to be created which represents the order in which the treatments were given. The easiest way to do this is to have a string column with characters representing each treatment in the order they were given, say, for 3 treatments A, B and C. The column would contain ABC, ACB, BAC, BCA, CAB and CBA as required. When this sequence factor column is defined, it will be selected as the first factor and the subjects as a repeated measure. The sequence should not be significant. If it is not, continue to analyse the full data. If it is significant, then it may only be possible to use the results from the first trial.

Example

Table 11.5 on p. 380 from Bolton, S. (1990). The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *AUC* and four factor columns *Period*, *Subject*, *Sequence* and *Treatment* created to keep track of the group memberships.

The first step is to test for any carryover (*Sequence*) effects. This is a Split-Plot design (see 7.3.0.2.5. Split-Plot Designs), with *Sequence* against *Subject*. Open ANOVA and select Statistics 1 → ANOVA and GLM → Analysis of Variance. Select *Sequence* (*C18*) as [Factor], *Subject* (*C17*) as [Repeated] and *AUC* (*C20*) as [Dependent]. Then select the Repeated Measures over some Factors and Classic Experimental Approach options to obtain the following ANOVA table:

Analysis of Variance

Design: Repeated Measures over some Factors
Approach: Classic Experimental
Dependent Variable: AUC

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Main Effects	4620.375	1	4620.375	1.590	0.2313
Sequence	4620.375	1	4620.375	1.187	0.3016
Error(Subject)	38940.083	10	3894.008		
Explained	4620.375	1	4620.375	1.590	0.2313
Error	34870.500	12	2905.875		
Total	78430.958	23	3410.042		

The critical value is the F-statistic and its probability for *Sequence*. This shows that the sequence is not significant, so there are no significant crossover effects in the data. We can then proceed to analyse the full data set. This is done by selecting ANOVA and *Subject* (*C17*), *Period* (*C16*) and *Treatment* (*C19*) as [Factor]s and *AUC* (*C20*) as [Dependent]. Select Classic Experimental Approach and no interaction terms:

Analysis of Variance

Approach: Classic Experimental
Dependent Variable: AUC

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Main Effects	67760.875	13	5212.375	4.885	0.0083
Subject	43560.458	11	3960.042	3.711	0.0240
Period	13490.042	1	13490.042	12.643	0.0052
Treatment	10710.375	1	10710.375	10.038	0.0100
Explained	67760.875	13	5212.375	4.885	0.0083
Error	10670.083	10	1067.008		
Total	78430.958	23	3410.042		

This shows that *Subject*, *Period* and *Treatment* are all significant.

7.3.1. Analysis of Variance

This section concentrates on operational aspects of the procedure ANOVA. The following types of experimental design that can be estimated using this procedure have been explained in detail above with solved examples (see 7.3.0.2. ANOVA Designs):

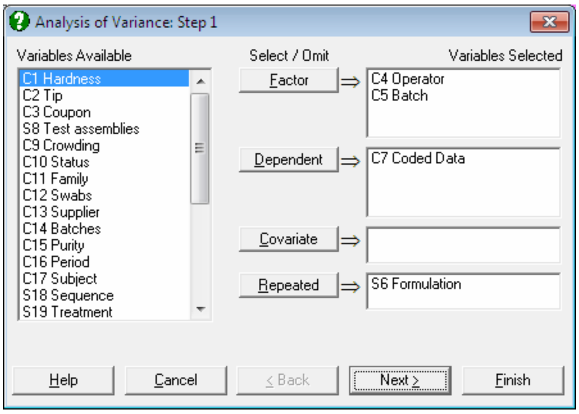
1) Randomised Block Designs
2) Repeated Measures Designs
3) Latin Squares Designs
4) Graeco Latin Squares Designs
5) Split-Plot Designs
6) Nested Designs
7) Crossover Designs

Of the four ANOVA dialogues described in this section, two are conditional upon selection of variables in the first dialogue. The ANOVA Design dialogue will not appear unless a [Repeated] column is selected and the interaction selection dialogue will not appear unless two or more [Factor] columns are selected.

UNISTAT's ANOVA procedure can cope with unbalanced designs where the number of observations in cells are not necessarily equal. Data may contain missing values and / or missing cells. Factor columns may be numeric, string, date or time variables and need not be sorted. It is possible to select an unlimited number of dependent variables, factors, repeated measures and covariates. In case more than one dependent variable is selected, the analysis will be repeated for each dependent variable with rest of the settings unchanged.

It is possible to analyse simple factorial, repeated measures, nested and mixed designs using the ANOVA procedure, whose output consists of an Analysis of Variance table. The output options of the more powerful General Linear Model procedure include table of means, coefficients, fitted values and residuals and their plots. It is also possible to perform multiple comparison tests on the GLM model fitted (see 7.3.2. General Linear Model).

7.3.1.1. ANOVA Variable Selection



Factor ⇒ C4 Operator, C5 Batch
Dependent ⇒ C7 Coded Data
Covariate ⇒
Repeated ⇒ S6 Formulation

To run an Analysis of Variance, at least one [Factor] column and one [Dependent] column should be selected. [Covariate] and [Repeated] selections are optional.

The variable selection for General Linear Model is slightly different from that of Analysis of Variance. The descriptions of the [Dependent] and [Covariate] buttons are similar, however the [Factor] and [Repeated] buttons are handled differently in GLM (see 7.3.2.1. GLM Variable Selection).

Factor: A factor is a categorical variable (numeric or string) which classifies the data into a number of groups. Analysis of Variance compares the mean value of data in these groups. At least one factor is required and an unlimited number of factors can be selected.

Dependent: It is compulsory to select at least one column containing numeric data. When more than one data variable is selected, the analysis will be repeated as many times as the number of data variables, each time only changing the data variable and keeping the rest of selections unchanged.

Covariate: A covariate is a continuous (noncategorical) numeric variable which may be included in the analysis in order to remove its effect on the dependent variable. Any number of covariates can be selected from the Variables Available list by clicking on [Covariate]. When covariates are included in the analysis, an adjustment is made in the sum of squares for them before any other factors.

Repeated: This button is used to select a repeated measure column or a nested sub-factor on another factor column. The number of repeated measures selected must be less than or equal to the number of factors selected. The repeated measures and Nested Factors are calculated in the same way, but repeated measures are used as an error term to test the significance of the factor and Nested Factors are used as part of the model.

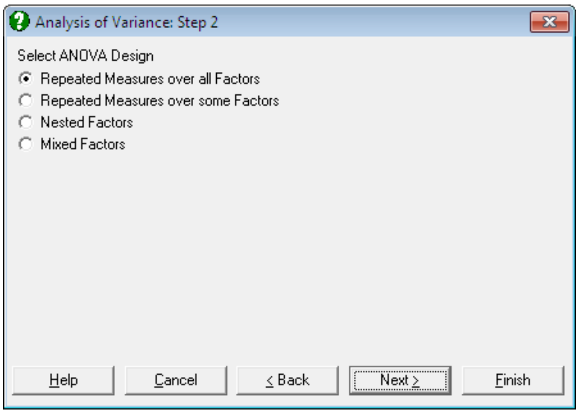
In ANOVA with Repeated Measures over all Factors (see 7.3.1.2.1. Repeated Measures over all Factors) only one factor column is selected with this button. This column defines the repeated measures on all factors.

In ANOVA with Repeated Measures over some Factors (see 7.3.1.2.2. Repeated Measures over some Factors) this button is used to specify repeated measures on the corresponding factors. The first column selected is considered as a repeated measure on the first factor, the second column selected as a repeated measure on the second factor, etc.

In ANOVA with Nested Factors (see 7.3.1.2.3. Nested Factors) this button is used to specify Nested Factors of the main factors. The first column selected is considered as a nested factor of the first factor, the second column selected as nested factor of the second factor, etc.

In ANOVA with Mixed Factors (see 7.3.1.2.4. Mixed Factors) this button is used to specify both repeated measures and Nested Factors. A subsequent dialogue is used to select the columns as repeated measures or Nested Factors.

7.3.1.2. ANOVA Designs



Select ANOVA Design

● Repeated Measures over all Factors
○ Repeated Measures over some Factors
○ Nested Factors
○ Mixed Factors

This dialogue does not appear if no columns were selected as [Repeated] in the Variable Selection Dialogue.

7.3.1.2.1. Repeated Measures over all Factors

When this option is selected, the factor selected as [Repeated] will be used to partition the total sum of squares into a between sum of squares and a within sum of squares term. An extra factor called *Trial* will be added to the model. The *Trial* sum of squares will be considered to be contained in the Within Subjects sum of squares. All other factors will be considered to be contained in the Between Subjects sum of squares. These are the assumptions for a repeated measures design across all specified factors. See 7.3.0.2.2. Repeated Measures Designs and 7.3.0.2.2.1. Repeated Measures over all Factors for a detailed description of these designs.

7.3.1.2.2. Repeated Measures over some Factors

With this option, the factors selected as [Repeated] will be used to specify repeated measures on the main factors. This behaves differently from the Repeated Measures over all Factors option, where only one repeated measure column should be selected.

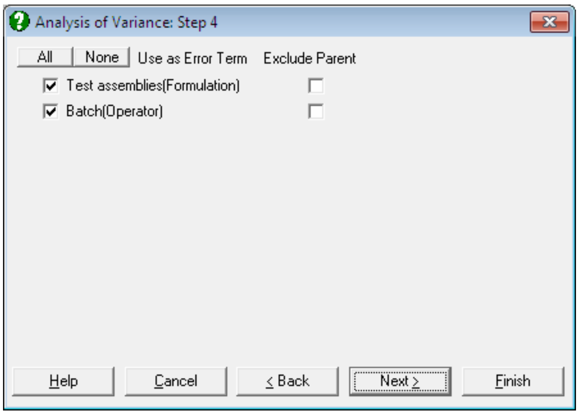
The first repeated measure selected is considered as a repeated measure on the first factor, the second repeated measure as a repeated measure on the second factor, etc. Repeated measures appear in the ANOVA table as Error(*column name*) directly under the factor it refers to. The value of the mean square is used as the error term to test the main factor for which it was selected. This procedure is used for Split-Plot designs (see 7.3.0.2.5. Split-Plot Designs).

7.3.1.2.3. Nested Factors

The factor columns selected as [Repeated] are used here to specify Nested Factors under the main factors. The first nested factor selected is considered as a nested factor of the first factor, the second as a nested factor of the second factor, etc.

Nested Factors appear as *nested factor label*(*main factor label*) in the ANOVA table. This procedure is used for Nested Factors designs (see 7.3.0.2.6. Nested Designs).

7.3.1.2.4. Mixed Factors

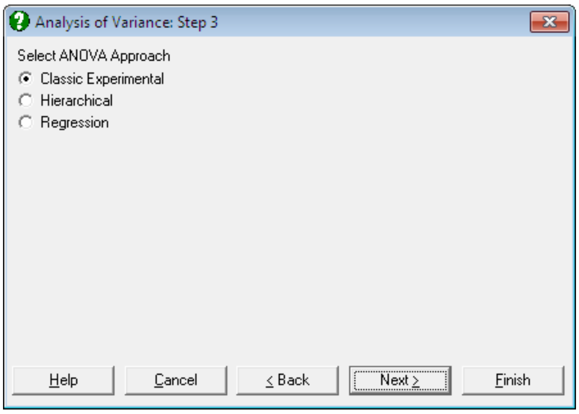


Use as Error Term	Exclude Parent
☑ Test assemblies(Formulation)	□
☑ Batch(Operator)	□

The factor columns selected as [Repeated] are used to specify either repeated measures (see 7.3.1.2.1. Repeated Measures over all Factors) or Nested Factors (see 7.3.1.2.3. Nested Factors) under the main factors. A further dialogue is used to select each term as a repeated measure or a nested factor. If the term is used as an error term, then it is a repeated measure. Otherwise it is a nested factor. The Exclude Parent check box is only applied to Nested Factors and should be left clear. Repeated measures appear in the ANOVA table as Error(*column name*) directly under the factor it refers to. Nested Factors appear as *nested factor label*(*main factor label*).

This procedure is used for Split-Plot designs (see 7.3.0.2.5. Split-Plot Designs), Nested Factors designs (see 7.3.0.2.6. Nested Designs) and mixtures of the two.

7.3.1.3. ANOVA Approaches


Analysis of Variance: Step 3

Select ANOVA Approach

● Classic Experimental
○ Hierarchical
○ Regression

Different approaches may be employed to compute the sum of squares figures displayed in ANOVA tables. The following are the three approaches supported by UNISTAT's ANOVA procedure. The General Linear Model procedure only supports the Regression Approach.

Classic Experimental Approach: This is probably the most common approach used in unbalanced design Analysis of Variance. It is also called the weighted means solution. For instance, in a three way ANOVA, the sum of squares computed for factors A, B, C and their interactions are calculated after the following adjustments: the main effects after the effects of all the other factors, two-way interactions after all the main effects and all the other two-way effects, and the three-way interactions after all the main effects, all the two-way effects and all the other three-way interactions.

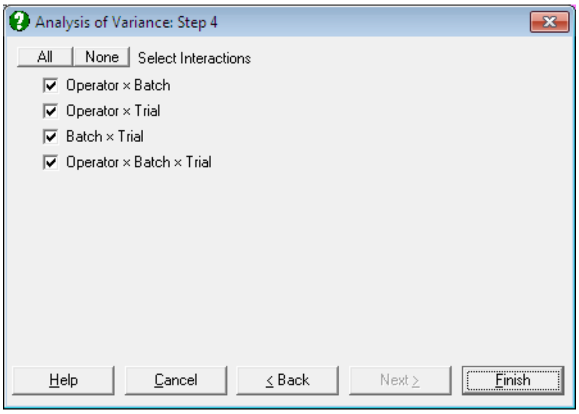
Hierarchical Approach: This is also called the forward sequential solution. Adjustments for the main effects are sequential. For instance, in a three way ANOVA, no adjustment is made for factor A, factor B is adjusted for A, and factor C is adjusted for A and B. The two-way interactions are not sequential. Each one is adjusted for the other two, but not for the main effects. The three-way interaction is adjusted for all main effects and two-way effects. So this approach differs from Classic Experimental Approach only in the way the main effects are calculated.

Regression Approach: This is also called the mean of cell means solution. All effects are computed after an adjustment is made for all other effects. For instance, in a three way ANOVA, a main effect is calculated after an adjustment is made for all other main effects, all two-way effects, the three-way effect, and if any, all covariates and the repeated measure.

The following table shows how the three approaches partition the main effects in a three way analysis of variance.

Effect	Classic Experimental	Hierarchical	Regression
A	A adjusted for B, C	A	A adjusted for full model
B	B adjusted for A, C	B adjusted for A	B adjusted for full model
C	C adjusted for A, B.	C adjusted for A, B	C adjusted for full model

7.3.1.4. ANOVA Interaction Selection



Select Interactions

☑ Operator × Batch
☑ Operator × Trial
☑ Batch × Trial
☑ Operator × Batch × Trial

This dialogue does not appear if at least two [Factor] columns have been selected in the Variable Selection Dialogue.

The interaction terms to be included in the model are selected here. If higher order interactions are included, then all the related lower order interactions are automatically included. That is, if ABC is selected then AB, AC and BC are all included in the model automatically. Maximum three way interactions are possible.

7.3.2. General Linear Model

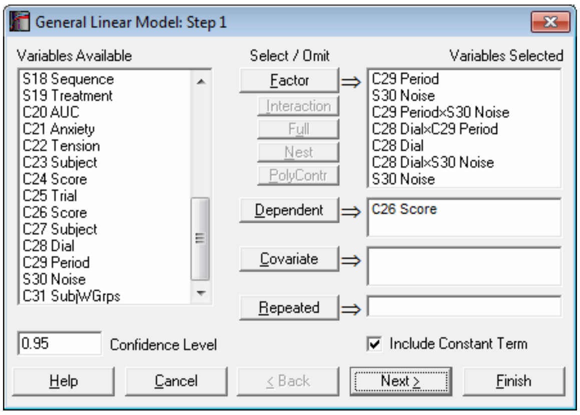
UNISTAT's GLM procedure can handle unbalanced designs where the number of observations in cells are not necessarily equal. Data may contain missing values and / or missing cells. Factor columns may be numeric, string, date or time variables and need not be sorted. It is possible to select an unlimited number of dependent variables, factors, repeated measures and covariates. When more than one dependent variable is selected, the analysis will be repeated for each dependent variable with the rest of the settings unchanged.

The GLM procedure assumes that each factor has a maximum of 1000 levels, though this number may be increased by the user, if necessary. It is possible to analyse simple factorial, repeated measures, nested and mixed designs. The output options include table of means, coefficients, fitted values and residuals and their plots. It is also possible to perform multiple comparison tests on the GLM model fitted.

This procedure allows more flexibility in defining the ANOVA model to be used. All the previous models can be built using the General Linear Model procedure. Also, this procedure can handle 4 way and above interactions / nested terms and allows the specification of the F-ratio terms.

The GLM procedure always adopts the Regression Approach to sum of squares. This means the factors can be selected in any order and the same sum of squares will be obtained.

7.3.2.1. GLM Variable Selection



Variables Available
S18 Sequence
S19 Treatment
C20 AUC
C21 Anxiety
C22 Tension
C23 Subject
C24 Score
C25 Trial
C26 Score
C27 Subject
C28 Dial
C29 Period
S30 Noise
C31 SubjWGrps

Variables Selected
C29 Period
S30 Noise
C29 Period×S30 Noise
C28 Dial×C29 Period
C28 Dial
C28 Dial×S30 Noise
S30 Noise

Dependent: C26 Score

0.95 Confidence Level

☑ Include Constant Term

The variable selection for General Linear Model is slightly different from the ANOVA procedures. When a selection is made from the Variables Available list on the left, the variable remains there, allowing it to be selected again. The [Dependent] and [Covariate] buttons work as before (see 7.3.1.1. ANOVA Variable Selection). The [Factor], [Interaction], [Full] and [Nest] all add factors to the factor list box. They do this in the following way:

Factor: is used to select a single factor variable. If multiple variables are highlighted in the Variables Available list, then all these variables will be added to the factors list.

Interaction: is used to select an interaction term. This will only be active when multiple variables are selected in the Variables Available list. A single factor of the form *FactA* × *FactB* × ... is added. You can always highlight multiple nonadjacent variables by pressing on the [Ctrl] key and then clicking on the desired variable.

A special cross sign is used between the variables in an interaction term. If there is a problem with this character on a non-English operating system, you can enter and edit the following line in *Documents\Unistat65\ Unistat65.ini* file under [Options] to display any other character (say x):

```
InteractionCross=×
```

This character also appears in all ANOVA and GLM output with interaction terms.

Full: is used to add all the factors and interactions for a fully factorial model. This will only be active when multiple variables are selected in the Variables Available list. You can always highlight multiple nonadjacent variables by pressing on the [Ctrl] key and then clicking on the desired variable.

Nest: is used to create a nested term. This will only be active when a variable is selected in the Variables Available list and a factor is highlighted in the factors list on the right. When selected, this will create a nested factor of the form *FactA(FactB)*. This can be repeated to create a multilevel nested factor e.g. *FactA(FactB(FactC))*. Nests on interactions, such as *FactA(FactB × FactC)*, are not allowed.

PolyContr: This button is used to request orthogonal polynomial contrasts for factors or interactions of factors. When a factor or interaction term is selected on the Variables Selected list, the [PolyContr] button will be enabled. The selected factors should contain minimum two and maximum six levels. Otherwise the polynomial contrasts cannot be computed. Contrasts for interaction terms are obtained from the dummy variables created for the main factors.

Polynomial contrasts should be used when the factor levels are equally spaced, although UNISTAT will not check for this condition and compute contrasts anyway.

The following table shows the coefficients used in dummy variables for factors with two to six levels. Each coefficient is divided by the Divisor displayed on the right of the table.

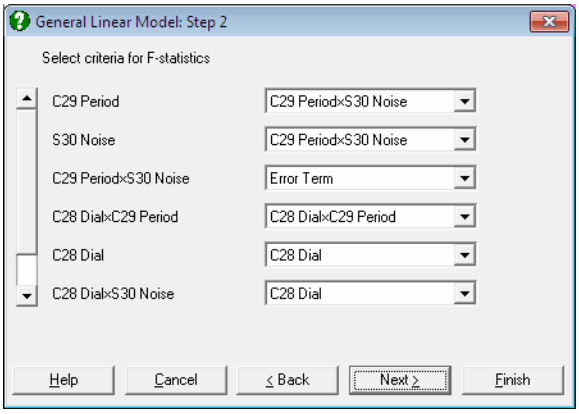
Factor	Polynomial							
Levels	**Degree**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Divisor**
2	1	-1	1					2
3	1	-1	0	1				2
	2	1	-2	1				6
4	1	-3	-1	1	3			20
	2	1	-1	-1	1			4
	3	-1	3	-3	1			20
5	1	-2	-1	0	1	2		10
	2	2	-1	-2	-1	2		14
	3	-1	2	0	-2	1		10
	4	1	-4	6	-4	1		70
6	1	-5	-3	-1	1	3	5	70
	2	5	-1	-4	-4	-1	5	84
	3	-5	7	4	-4	-7	5	180
	4	1	-3	2	2	-3	1	28
	5	-1	5	-10	10	-5	1	252

Dependent: It is compulsory to select at least one column containing numeric data. When more than one data variable is selected, the analysis will be repeated as many times as the number of data variables, each time only changing the data variable and keeping the rest of selections unchanged.

Covariate: A covariate is a continuous (noncategorical) numeric variable that may be included in the analysis in order to remove its effect on the dependent variable. Any number of covariates can be selected from the Variables Available list by clicking on [Covariate]. When covariates are included in the analysis, an adjustment is made in the sum of squares for them before any other factors.

Repeated: This button is used to select a factor to define Within and Between Subjects terms. These can be used in the F-ratio selection to create a number of different repeated measure designs. If a variable is selected as [Repeated], then at least one factor must use the Error Within term in its F-ratio (see 7.3.2.2. GLM F-Ratio Selection).

7.3.2.2. GLM F-Ratio Selection


General Linear Model: Step 2

Select criteria for F-statistics

C29 Period	C29 Period×S30 Noise
S30 Noise	C29 Period×S30 Noise
C29 Period×S30 Noise	Error Term
C28 Dial×C29 Period	C28 Dial×C29 Period
C28 Dial	C28 Dial
C28 Dial×S30 Noise	C28 Dial

This dialogue allows the denominator in the F-ratio of each factor to be specified independently. The possible selections include:

Residual: The denominator is given by the total error of the model. This is the default in GLM and it is also the denominator used in ANOVA.

None: The F-statistic and its tail probability are not calculated.

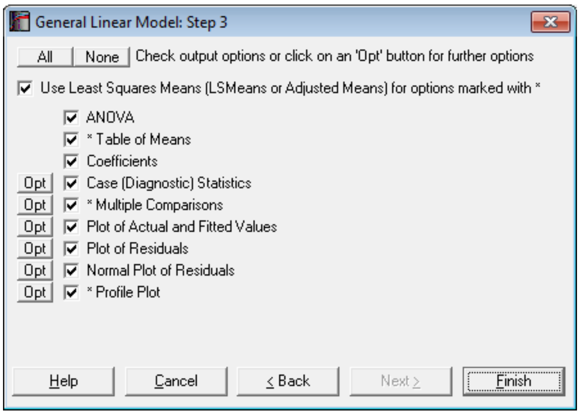
Error Term: The factor is not included in the model - also known as *Fixed Factor*. That is, the sum of squares calculated is not included in the explained sum of squares and hence the residual sum of squares is not adjusted for this factor. The sum of squares is reported separately from the factors included in the model. R-squared and adjusted R-squared are not reported for a model that includes error terms.

Any Factor's Mean Square: The denominator is given by the mean square of the selected factor.

Error Between Subjects: This option will only be available when a column is selected as [Repeated] in the variable selection. The denominator is given by the Error Between Subjects mean square. This value is calculated by the sum of squares of the selected column minus the sum of square of all the factors with denominator selected as Error Between Subjects. So the mean square value depends on which factors are selected as Error Between Subjects.

Error Within Subjects: This option will only be available when a column is selected as [Repeated] in the variable selection. This term partitions the residual sum of squares with the Error Between Subjects. So the mean square value depends on which factors are selected as Error Between Subjects.

7.3.2.3. GLM Output Options



Check output options or click on an 'Opt' button for further options

☑ Use Least Squares Means (LSMeans or Adjusted Means) for options marked with *

☑ ANOVA
☑ * Table of Means
☑ Coefficients
☑ Case (Diagnostic) Statistics
☑ * Multiple Comparisons
☑ Plot of Actual and Fitted Values
☑ Plot of Residuals
☑ Normal Plot of Residuals
☑ * Profile Plot

The main output of the GLM procedure is the ANOVA table. However, GLM also reports a wide range of diagnostic statistics, *post hoc* tests and plots.

As of this version of UNISTAT, three more options have been added to the Output Options Dialogue. It is now possible to include any or all three output tables for each test.

Use Least Squares Means: The following three output options are preceded by an asterisk in the Output Options Dialogue;

- Table of Means,
- Multiple Comparisons and
- Profile Plot.

The common point for these options is that they display information based on cell and / or marginal means (where a cell is defined as a unique combination of factor levels used in the model and a marginal mean is the mean of a group of cell means). It is possible to display these output options based on the least squares means (LSMeans or adjusted means) instead of arithmetic means. To do this check the Use Least Squares Means box at the top of the dialogue. Then these three output options will use the least squares means and this will be indicated on the output. Note that the least squares means output will not be available when a nested term exists in the model.

Normally, in balanced designs arithmetic and least squares means will not be different. However, in unbalanced designs with more than one effect, the arithmetic mean of a group may not represent an appropriate response for that group, since it does not take other effects in the model into account. Least squares means can also be described as within-group means adjusted for the other effects in the model. In other words, least squares means are predicted population margins which estimate the marginal means over a balanced population.

The least squares means are computed as:

$$\text{LSMeans} = \text{L}\beta$$

where L is the hypothesis matrix and β is the vector of estimated regression coefficients as displayed in the third output option:

$$\beta = (\text{X}'\text{X})^{-1}\text{XY}$$

The standard error of LSMeans is defined as:

$$\text{SE} = \sqrt{\text{L}(\text{X}'\text{X})^{-1}\text{L}'\text{*MSE}}$$

where MSE is the mean square error for the model.

To save the hypothesis matrix L in a file enter the following line under the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
    GLMSaveLMatrix=1
```

The matrix will be saved in:

..\Documents\Unistat65\GLMLMatrix.txt

To save the $(\text{X}'\text{X})^{-1}$ matrix enter the following line under the [Options] section of *Unistat65.ini*:

```
GLMSaveInvXpXMatrix=1
```

The matrix will be saved in the file:

..\Documents\Unistat65\GLMInvXpXMatrix.txt

Some of the GLM Output Options (such as coefficients, fitted values, residuals and their plots) are based on the underlying regression model for the GLM model. These are normally identical to the results obtained from the Linear Regression procedure if the same model is constructed using dummy variables (see 7.3.2.4. GLM Example below).

ANOVA: The factors in the model appear at the top of the table. Any factors that have been calculated but are not in the model appear at the bottom of the table in a section separated with a dividing line. These values may be used as denominators in F-statistic calculations. If a repeated measure has been specified, then the factor will be split into Between Subjects and Within Subjects terms.

*** Table of Means:** Number of cases, mean, standard deviation, standard error and the lower and upper limits of the confidence interval are displayed for each cell and marginal mean of the model. Missing cells are omitted from the analysis at the outset. This output option is similar to the Table of Means procedure available under Tests for ANOVA, but has the advantage of matching exactly the model specified in the GLM procedure. When the Use Least Squares Means box is checked, LSMeans will be displayed instead of arithmetic means.

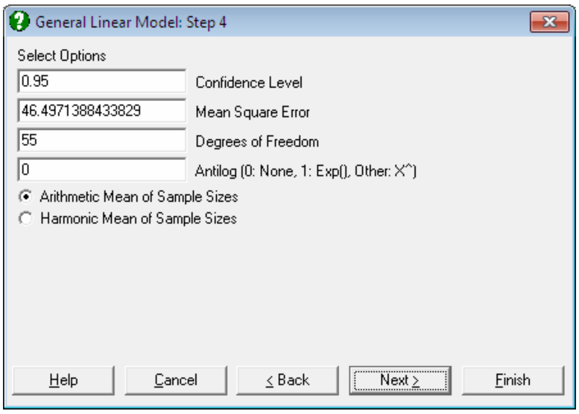
Coefficients: Estimated coefficients for the underlying regression model are displayed. The Row Labels of this table are identical to that of Table of Means output above. Variables causing multicollinearity will be displayed with a zero coefficient at the end of the coefficients table. If you do not wish to display these variables enter the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
DispCollin=0
```

Case (Diagnostic) Statistics: The actual, fitted and residual Y values for the underlying regression model are displayed.

*** Multiple Comparisons:** This output option is similar to the Multiple Comparisons procedure available under Tests for ANOVA, but has a major advantage. While the latter is always based on a one-way ANOVA model, this option takes the Mean Square Error and its Degrees of Freedom from the estimated GLM model. Note that both procedures give you the opportunity to override the suggested Mean Square Error and its Degrees of Freedom.

When the Use Least Squares Means box is checked, comparisons will be made between LSMeans instead of arithmetic means.



General Linear Model: Step 4

Select Options

0.95 Confidence Level

46.4971388433829 Mean Square Error

55 Degrees of Freedom

0 Antilog (0: None, 1: Exp(), Other: X^)

● Arithmetic Mean of Sample Sizes

○ Harmonic Mean of Sample Sizes

When the data used in multiple comparisons is already transformed with a function like e or 10 based logarithm, the results need to be transformed back to the original scale and the user is faced with the task of applying back transformations manually. The Antilog box allows the user to specify which back-transformation is to be applied in the output. The output values that are affected by this control are:

- means,
- difference between means, and
- lower and upper confidence limits for difference between means.

Let X be the value entered into the Antilog box.

- If X = 0 then no back-transformation is performed.
- If X = 1 then the natural antilog of the output value Y, Exp(Y) is displayed.
- If 1 < X ≤ 16, then X-based power of Y, X^Y is displayed.

When a back-transformation base value is specified, the columns affected will be marked by an asterisk in the output.

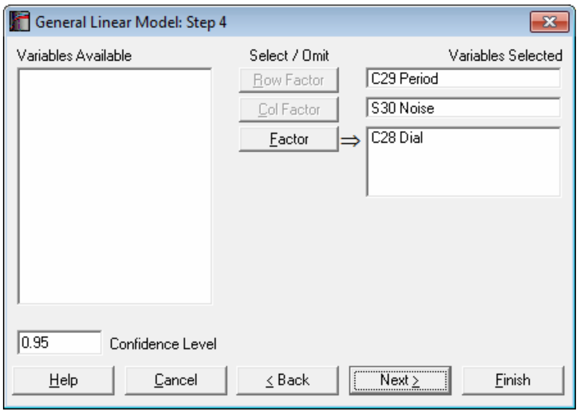
Plot of Actual and Fitted Values: Select this option to plot actual and fitted Y values against row numbers (index).

Plot of Residuals: Residuals are plotted against row numbers (index).

Normal Plot of Residuals: Residuals are plotted against the normal probability (probit) axis.

* **Profile Plot:** The cell means for an unlimited number of factors can be plotted. If only one factor is selected, the plot is similar to the Means Plot with Error Bars option available in X-Y Plots (see 4.1.1.3. Means Plot). When the Use Least Squares Means box is checked, LSMeans will be plotted instead of arithmetic means.

If two or more factors are selected, the first factor's levels are represented on the X-Axis and the second factor's levels as separate lines. This plot is known as Interaction Plot and it is different from a Means Plot with two factors (where interaction terms are plotted either on the X-Axis or as separate lines). The Interaction Plot is useful for comparing means of interaction terms. If the lines are parallel, then it can be concluded that there is no interaction between the two factors.

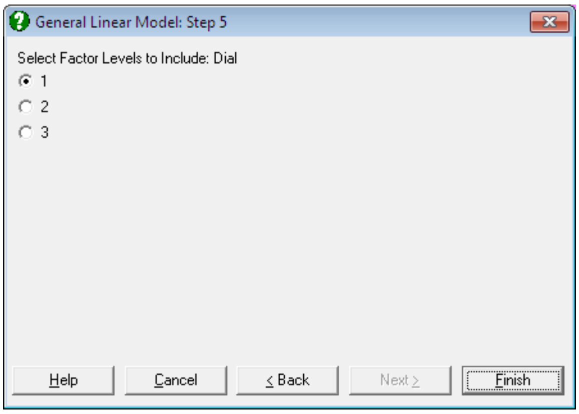


General Linear Model: Step 4

Variables Selected:
Row Factor: C29 Period
Col Factor: S30 Noise
Factor: C28 Dial

Confidence Level: 0.95

To change the default factor selections, you can click on the [Opt] button situated to the left of this output option. A further Variable Selection Dialogue is displayed. It is compulsory to select a [Row Factor] variable, in which case a Means Plot is displayed for the selected factor. If an optional [Column Factor] is also selected, then an Interaction Plot is displayed.



General Linear Model: Step 5

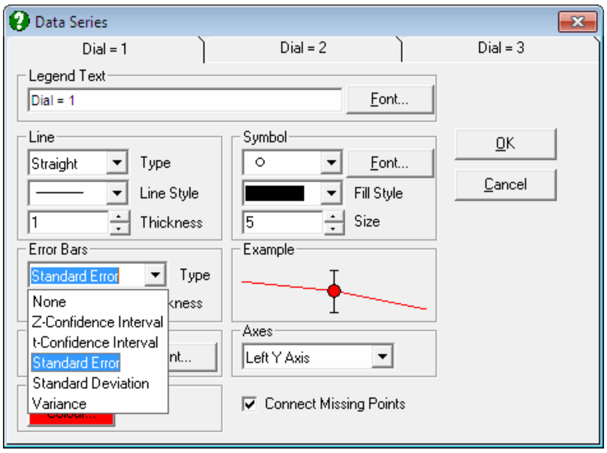
Select Factor Levels to Include: Dial

● 1

○ 2

○ 3

A third [Factor] button allows selection of an unlimited number of further factor columns. If at least one such factor is selected, then the program displays a further dialogue showing the levels of this factor. If two or more factors are selected then the dialogue shows the combinations of all selected factors. Only one level (or level combination) needs to be selected and the final plot is drawn for this selected subsample. When [Factor] columns are selected, selection of a [Column Factor] still remains optional.



Data Series

Dial = 1 | Dial = 2 | Dial = 3

Legend Text: Dial = 1

Line — Type: Straight; Line Style; Thickness: 1

Symbol — ○; Fill Style; Size: 5

Error Bars — Type: Standard Error

None
Z-Confidence Interval
t-Confidence Interval
Standard Error
Standard Deviation
Variance

Axes: Left Y Axis

☑ Connect Missing Points

In earlier versions of UNISTAT, error bars for means plot represented only the standard error of mean. Now, (as in Means Plot), the Error Bars control on the Edit → Data Series dialogue allows selecting one of the following dispersion measures: None, t-interval, Z-interval, Standard Error, Standard Deviation, Variance

7.3.2.4. GLM Example

Published examples using GLM have already been introduced in section 7.3.0.2. ANOVA Designs. Here we shall demonstrate the full GLM output with a simple example. Next, we shall run the same model using the Linear Regression procedure and emphasise the similarities and differences.

Open DEMODATA and select Statistics 1 → ANOVA and GLM → General Linear Model. Highlight *Region* (*C10*) and *Type* (*C11*) on the Variables Available list and then click on the button [Full]. This will add two main factors and an interaction term to the Variables Selected list. Also select *Output2* (*C9*) as [Dependent]. Accept default values suggested in the next two dialogues. Some of the following results have been shortened due to space considerations.

General Linear Model

ANOVA

Dependent Variable: Output2

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	669862.660	1	669862.660	281.165	0.0000
Region	70.934	2	35.467	0.774	0.4664
Type	1.122	1	1.122	0.024	0.8762
Region × Type	174.832	2	87.416	1.908	0.1586
Explained	206.738	5	41.348	0.902	0.4866
Error	2382.454	52	45.816		
Total	2589.193	57	45.424		

R-squared =	0.0798
Adjusted R-squared =	-0.0086

Table of Means

	Cases	Mean	Standard Deviation	Standard Error	Lower 95%	Upper 95%
whole sample	58	107.4679	6.7398	0.8850	105.6958	109.2401
Region = 1	14	106.1586	6.9131	1.8476	102.1671	110.1501
2	26	107.8269	6.7597	1.3257	105.0966	110.5572
3	18	107.9678	6.8330	1.6106	104.5698	111.3658
Type = 1	16	107.0837	6.9366	1.7342	103.3875	110.7800
2	42	107.6143	6.7430	1.0405	105.5130	109.7156
Region × Type = 1 × 1	7	103.6271	7.7791	2.9402	96.4327	110.8216
1 × 2	7	108.6900	5.2991	2.0029	103.7892	113.5908
2 × 1	5	111.5200	1.7219	0.7701	109.3820	113.6580
2 × 2	21	106.9476	7.2320	1.5782	103.6557	110.2396
3 × 1	4	107.5875	7.3881	3.6940	95.8315	119.3435
3 × 2	14	108.0764	6.9572	1.8594	104.0594	112.0934

Coefficients

	Coefficient	Standard Error	t-Statistic	Probability	Lower 95%	Upper 95%
Constant	108.0764	1.8090	59.7426	0.0000	104.4463	111.7065
Region = 1	0.6136	3.1333	0.1958	0.8455	-5.6740	6.9011
2	-1.1288	2.3355	-0.4833	0.6309	-5.8153	3.5576
3	*					
Type = 1	-0.4889	3.8375	-0.1274	0.8991	-8.1895	7.2117
2	*					
Region × Type = 1 × 1	-4.5739	5.2742	-0.8672	0.3898	-15.1574	6.0096
1 × 2	*					
2 × 1	5.0613	5.1060	0.9912	0.3262	-5.1848	15.3074
2 × 2	*					
3 × 1	*					
3 × 2	*					

* omitted due to multicollinearity

Case (Diagnostic) Statistics

	Actual Y	Fitted Y	Residuals
1	103.3600	108.0764	-4.7164
2	105.1300	108.6900	-3.5600
3	105.9700	106.9476	-0.9776
…	…	…	…
56	93.1100	103.6271	-10.5171
57	93.2000	106.9476	-13.7476
58	93.4700	108.0764	-14.6064

Dunnett

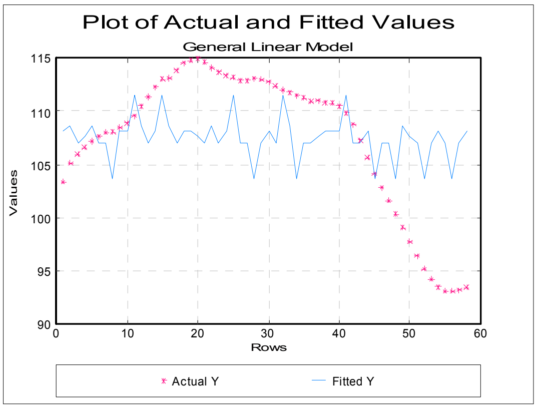
For Output2, classified by Region
Control Group: 1, Two-Tailed Test
Mean Square Error: 45.8164292353523, Degrees of Freedom: 52
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

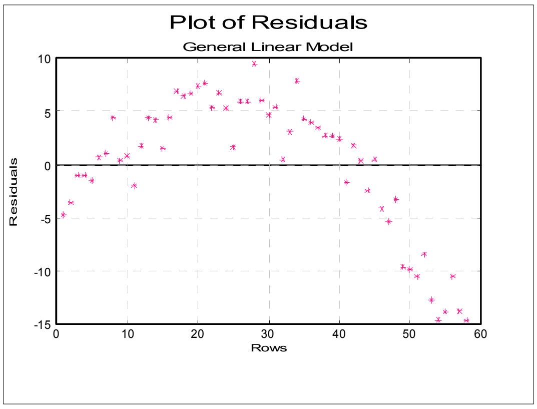
Group	Cases	Mean	1	
1	14	106.1586		I
2	26	107.8269		I
3	18	107.9678		I

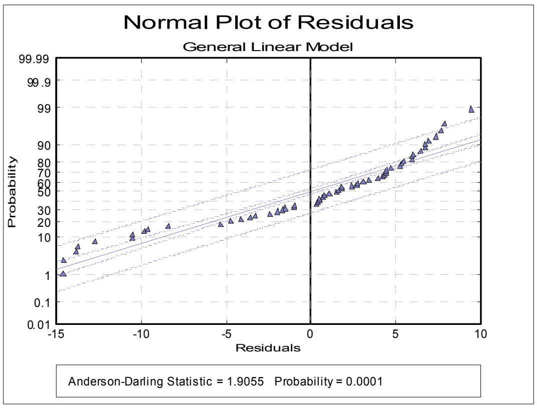
Comparison	Difference	Standard Error	q Stat	Table q	Probability
3 - 1	1.8092	2.4120	0.7501	2.2581	0.6578
2 - 1	1.6684	2.2438	0.7435	2.2581	0.6623

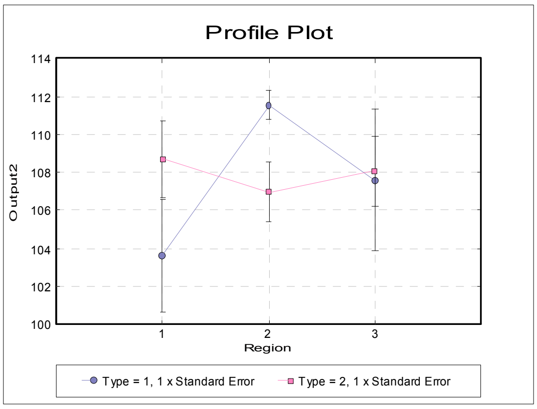
Comparison	Lower 95%	Upper 95%	Result
3 - 1	-3.6375	7.2559	
2 - 1	-3.3985	6.7352	

Homogeneous Subsets:
Group 1: 1 2 3
Pooled mean = 107.4679
95% Confidence Interval = 105.6845 <> 109.2514









Now we run the same model using Linear Regression.

Open DEMODATA and select Statistics 1 → Regression Analysis → Linear Regression. Highlight *Region* (*C10*) and *Type* (*C11*) on the Variables Available list and then click on the button [Full]. This will add two main factors and an interaction term to the Variables Selected list. Also select *Output2* (*C9*) as [Dependent]. From the Output Options Dialogue select only the Regression Results and ANOVA of Regression. The following results are obtained.

The Regression Results option produces exactly the same results as the GLM model. In the ANOVA of Regression table, Regression, Error and Total terms are also identical to Explained, Error and Total of GLM's ANOVA table. Moreover, if we add the two interaction Ssq's (1x1 and 2x1) we obtain the *Region x Type* Ssq of GLM. Main interactions are, however, different. This is because the GLM Procedure adopts the Regression Approach when it computes the sum of squares, whereas Linear Regression's sum of squares decomposition is sequential. See 7.3.1.3. ANOVA Approaches.

Linear Regression

Dependent Variable: Output2
Valid Number of Cases: 58, 0 Omitted

Regression Results

* omitted due to multicollinearity

	Coefficient	Standard Error	t-Statistic	Probability	Lower 95%	Upper 95%
Constant	108.0764	1.8090	59.7426	0.0000	104.4463	111.7065
Region = 1	0.6136	3.1333	0.1958	0.8455	-5.6740	6.9011
2	-1.1288	2.3355	-0.4833	0.6309	-5.8153	3.5576
Type = 1	-0.4889	3.8375	-0.1274	0.8991	-8.1895	7.2117
Region × Type = 1 × 1	-4.5739	5.2742	-0.8672	0.3898	-15.1574	6.0096
2 × 1	5.0613	5.1060	0.9912	0.3262	-5.1848	15.3074
Region = 3	*					
Type = 2	*					
Region × Type = 1 × 2	*					
2 × 2	*					
3 × 1	*					
3 × 2	*					

Residual Sum of Squares =	2382.4543
Standard Error =	6.7688
Mean of Y =	107.4679
Stand Dev of y =	6.7398
Correlation Coefficient =	0.2826
R-squared =	0.0798
Adjusted R-squared =	-0.0086
F(5,52) =	0.9025
Probability of F =	0.4866
Durbin-Watson Statistic =	0.1641
log of likelihood =	-190.2131
Press Statistic =	2916.1898

ANOVA of Regression

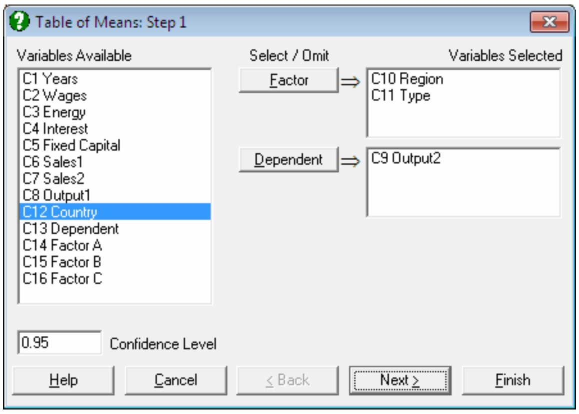
Due To	**Sum of Squares**	**DoF**	**Mean Square**	**F-Stat**	**Prob**
Region = 1	31.639	1	31.639	0.691	0.4098
2	0.211	1	0.211	0.005	0.9462
Type = 1	0.057	1	0.057	0.001	0.9721
Region × Type = 1 × 1	129.815	1	129.815	2.833	0.0983
2 × 1	45.017	1	45.017	0.983	0.3262
Regression	206.738	5	41.348	0.902	0.4866
Error	2382.454	52	45.816		
Total	2589.193	57	45.424	0.991	0.5142

7.4. Tests for ANOVA

Multifactor Table of Means, Homogeneity of Variance Tests, Multiple Comparisons, Regression with Replicates (linearity test) and Heterogeneity of Regression (intercept and slope) tests can be accessed under this topic.

7.4.1. Table of Means

The Table of Means procedure can be considered as a Break-Down analysis with interaction terms. Observations of the dependent variable which belong to every possible combination of factor levels are grouped together and the number of cases, mean, standard deviation, standard error and the lower and upper limits of the confidence interval for the mean are displayed. Unlike the Break-Down table, here factor combinations follow the sequence of ANOVA tables, that is, first one-way effects, then two-way interactions, then three-way interactions, etc.

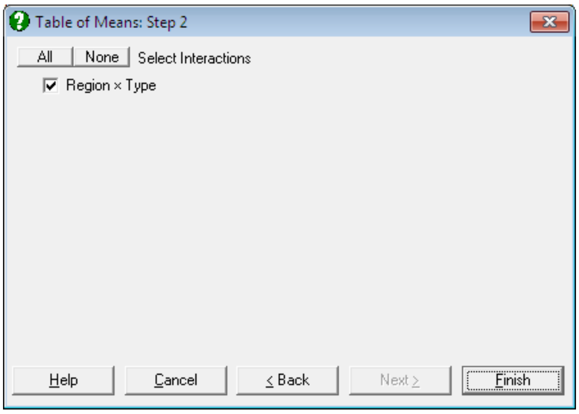


Factor ⇒ C10 Region, C11 Type

Dependent ⇒ C9 Output2

0.95 Confidence Level

At least one factor and one data column should be selected from the Variables Available list. An unlimited number of factors can be selected in exactly the same way as for an ANOVA procedure.



☑ Region × Type

It is optional to include two-way or higher interactions within the analysis. A further dialogue allows you to include the desired interaction terms in the analysis by clicking on the corresponding check boxes.

Example

Open DEMODATA, select Statistics 1 → Tests for ANOVA → Table of Means and select *Region* (*C10*) and *Type* (*C11*) as [Factor]s and *Output2* (*C9*) as [Dependent].

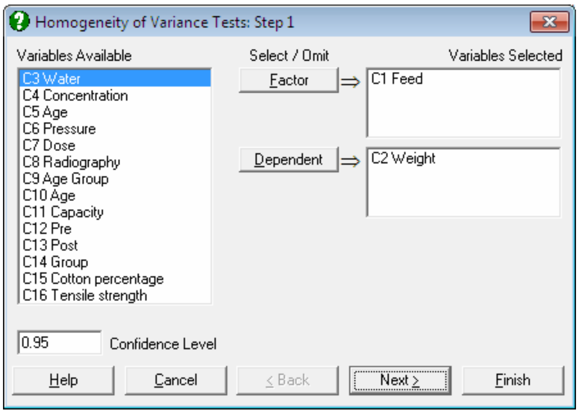
Table of Means

For Output2

	Cases	Mean	Standard Deviation	Standard Error	Lower 95%	Upper 95%
whole sample	58	107.4679	6.7398	0.8850	105.6958	109.2401
Region = 1	14	106.1586	6.9131	1.8476	102.1671	110.1501
2	26	107.8269	6.7597	1.3257	105.0966	110.5572
3	18	107.9678	6.8330	1.6106	104.5698	111.3658
Type = 1	16	107.0837	6.9366	1.7342	103.3875	110.7800
2	42	107.6143	6.7430	1.0405	105.5130	109.7156
Region × Type = 1 × 1	7	103.6271	7.7791	2.9402	96.4327	110.8216
1 × 2	7	108.6900	5.2991	2.0029	103.7892	113.5908
2 × 1	5	111.5200	1.7219	0.7701	109.3820	113.6580
2 × 2	21	106.9476	7.2320	1.5782	103.6557	110.2396
3 × 1	4	107.5875	7.3881	3.6940	95.8315	119.3435
3 × 2	14	108.0764	6.9572	1.8594	104.0594	112.0934

7.4.2. Homogeneity of Variance Tests

One of the assumptions of the Analysis of Variance is that variances of the subgroups of data (defined by factor levels) are equal. Four tests are provided here to test whether this is the case. The null hypothesis tested is “all population variances are equal”, against the alternative hypothesis “all population variances are not equal”. If the null hypothesis is rejected, then it is possible to perform a number of multiple comparisons to determine which pairs of subgroups have significantly different variances.



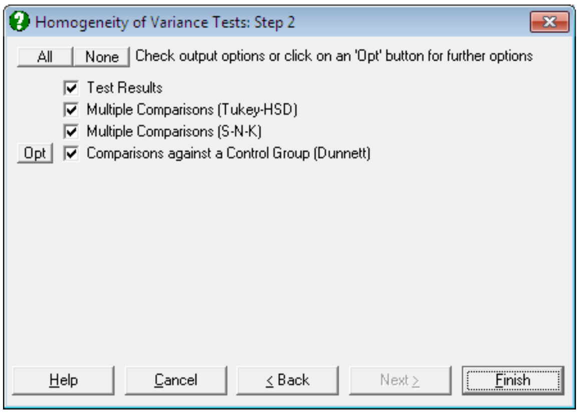
Homogeneity of Variance Tests: Step 1

Factor ⇒ C1 Feed

Dependent ⇒ C2 Weight

0.95 Confidence Level

The type of output depends on the number of factors selected. When a single factor variable is selected, a further Output Options Dialogue will allow you to perform multiple comparison tests for variances. If two or more factor variables are selected, then the Output Options Dialogue will not appear and the program will display the test results immediately.


Homogeneity of Variance Tests: Step 2

Check output options or click on an 'Opt' button for further options

☑ Test Results
☑ Multiple Comparisons (Tukey-HSD)
☑ Multiple Comparisons (S-N-K)
☑ Comparisons against a Control Group (Dunnett)

7.4.2.1. Homogeneity of Variance Test Results

Five alternative test statistics are computed, and by default, all are reported on the same table. If more than one factor column is selected, an overall test is also performed on the subgroups defined by combinations of all levels in all factors. For instance, if two factors are selected, the test is performed on subgroups defined by two-way interactions. If you wish to see exactly which subgroups are tested, you can use the Table of Means procedure first. It is possible to control the statistics displayed in the output by including and editing the following line in the [Options] section of *Documents*\Unistat65\Unistat65.ini file:

```
HomoVariance=0
```

This parameter takes the following values:

-1: Overall test only
0: All five test statistics and the overall test (default)
1: Bartlett's Chi-square Test
2: Bartlett-Box F test
3: Cochran's C
4: Hartley's F Test
5: Levene's F Test

Bartlett’s Chi-Square Test

This is the original form of the homogeneity of variance test as introduced by Bartlett (see Zar, J. H. (2010), p. 220).

All subgroups of the dependent variable defined by the selected factor are formed. Groups with a zero variance are omitted and the counts (n_j) and variances s_j^2 of the remaining groups are determined. The test statistic is calculated as follows:

$$B = (n - m)\mathrm{Log}(P) - Q$$

where m is the number of subgroups with non zero variances, n is the total number of cases within these subgroups and:

$$P = \sum \frac{s_j^2}{n - m}$$

$$Q = \sum (n_j - 1)\mathrm{Log}(s_j^2)$$

For a more accurate chi-square distribution the following term is computed:

$$C = 1 + \frac{1}{3(m-1)}\left(\sum \frac{1}{n_i} - \frac{1}{n-m}\right)$$

and the modified test statistic is obtained as:

$$B_C = \frac{B}{C}$$

which is approximately chi-square distributed with m - 1 degrees of freedom.

As Bartlett's chi-square test does not perform well when the population distributions are not normal, the following modification is widely regarded as a more powerful replacement.

Bartlett-Box F-test

The test statistic B is calculated as in Bartlett's chi-square test, with the exception in the definition of:

$$P = \sum \frac{(n_j - 1)S_j^2}{n - m}$$

and:

$$C = 1 + \frac{1}{3(m-1)}\left(\sum \frac{1}{n_i} - \frac{m}{n-m}\right)$$

which is F-distributed with m - 1 and R degrees of freedom, where:

$$R = \frac{m+1}{C^2}$$

Cochran's C

The test statistic is obtained by dividing the maximum subgroup variance by the sum of all subgroup variances.

$$C = \frac{s_{max}}{\sum s_i}$$

An approximate tail probability is computed from the F-distribution:

$$F = \frac{(n-1)C}{1-C}$$

with n/m - 1, (n/m - 1)(m - 1) degrees of freedom and it is multiplied by m.

Hartley's F Test

The test statistic is obtained by dividing the maximum subgroup variance by the minimum subgroup variance.

$$F = \frac{s_{max}}{s_{min}}$$

A lookup table for critical values of Hartley's F-values (Pearson & Hartley, 1954) is included for a limited range of degrees of freedom, number of subgroups and significance levels. The valid range is as follows:

- $4 \leq$ degrees of freedom ≤ 10
- number of subgroups = 4, 6, 8, 9, 10, 12
- all subgroups have equal number of observations
- significance level = 0.05 or 0.01

The computed F-value is compared with the table value at 0.05 and 0.01 levels and the result is reported as:

- .0500 > if $F > F_{.05}$
- .0100 > if $F_{.05} \geq F > F_{.01}$
- .0100 < if $F \leq F_{.01}$

Levene's F Test

This test has the advantage of being less sensitive to deviations from normality and is widely accepted as the most powerful homogeneity of variance test. The test statistic, which has an F distribution with (n - k) and (k - 1) degrees of freedom, is computed as follows:

$$F = \frac{(n-k)\sum_{i=1}^{k} n_i(\overline{Z}_i - \overline{Z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \overline{Z}_i)^2}$$

where:

$$Z_{ij} = \left| X_{ij} - \sum_{j=1}^{n_i} X_{ij} \right| \qquad \overline{Z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij} \qquad \overline{Z} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{Z}_i$$

For a two sample version of this test see 6.1.1.6. Levene's F-Test.

7.4.2.2. Multiple Comparisons Among Variances

This option is available only when a single factor variable is selected from the Variables Available list. If this is the case, a further dialogue will allow you to perform Tukey-HSD, Student-Newman-Keuls and Dunnett multiple comparison tests for variances.

For each test the q statistic is calculated as:

$$q = \frac{\mathrm{Log}(s_B^2) - \mathrm{Log}(s_A^2)}{SE}$$

Tukey-HSD test for variances

All possible pairs are compared. Therefore, m(m - 1)/2 comparisons are made. The standard error is calculated as:

$$SE = \sqrt{\frac{1}{n_B - 1} + \frac{1}{n_A - 1}}$$

For details see 7.4.3.2. Tukey-HSD.

Student-Newman-Keuls test for variances

This test is identical to Tukey-HSD test except for the way the tabulated q values are computed.

For details see Student-Newman-Keuls.

Dunnett test for variances

This option will display two further dialogues, (1) selection of the control subgroup and (2) one or two-tailed test. All subgroups are compared with the control subgroup and therefore only M - 1 comparisons are made. The standard error is calculated as:

$$SE = \sqrt{\frac{2}{n_{ctrl} - 1} + \frac{2}{n_A - 1}}$$

For details see 7.4.3.8. Dunnett test.

7.4.2.3. Homogeneity of Variance Examples

Example 1

Example 10.13 on p. 222 from Zar, J. H. (2010). The null hypothesis “all four feeds used have the same variance” is tested at a 95% confidence level.

The table format given in the book can be transformed into the factor format by using UNISTAT’s Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Weight* and a factor column *Feed* created to keep track of the group memberships.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Homogeneity of Variance Tests and select *Feed* (*C1*) as [Factor] and *Weight* (*C2*) as [Dependent] to obtain the following results:

Homogeneity of Variance Tests

For Weight

	Test Statistic	Probability
classified by Feed		
Bartlett's Chi-square Test	0.4752	0.9243
Bartlett-Box F Test	0.1610	0.9226
Cochran's C (max var / sum var)	0.3059	1.0000
Hartley's F (max var / min var)	1.9967	
Levene's F Test	0.5816	0.6361

According to Bartlett's chi-square test the tail probability is far greater than 5% and therefore the null hypothesis is not rejected. The probability value for Hartley's F is not reported here, as this example does not fulfil the strict criteria outlined above.

Example 2

Running the above example after entering the HomoVariance=5 line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file and selecting the Test Results and Comparisons against a Control Group (Dunnett) options only, the following output is obtained.

Homogeneity of Variance Tests

For Weight

	Test statistic	Probability
classified by Feed **Levene's F Test**	0.0335	0.9914

Comparisons against a Control Group (Dunnett)

Method: 95% Dunnett interval.
Control Group: 1, Two-Tailed Test
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	N-1	Ln(variance)	1	
3	3	3.1342		I
4	4	3.5131		I
2	4	3.5340		I
1	4	3.6262		I

Comparison	Difference	Standard Error	q Stat	Table q	Probability
2 - 1	-0.0922	1.0000	0.0922	2.3533	0.9995
4 - 1	-0.1131	1.0000	0.1131	2.3533	0.9990
3 - 1	-0.4920	1.0801	0.4555	2.3533	0.9433

Comparison	Lower 95%	Upper 95%	Result
2 - 1	-2.4455	2.2611	
4 - 1	-2.4663	2.2402	
3 - 1	-3.0338	2.0499	

Homogeneous Subsets:	
Group 1:	3 4 2 1

Example 3

Open DEMODATA and select Statistics 1 → Descriptive Statistics → Homogeneity of Variance Tests and from the Variable Selection Dialogue select *Output2* (*C9*) as [Variable] and *Region* (*C10*) and *Type* (*C11*) as [Factor]s. Selecting the Test Results output option and setting HomoVariance=-1 in *Documents*\Unistat65\Unistat65.ini file, the following output is obtained:

Homogeneity of Variance Tests

For Output2

	Test statistic	Probability
classified by Region		
classified by Type		
Overall		
Bartlett's Chi-square Test	7.7995	0.1676
Bartlett-Box F Test	1.5719	0.1648
Cochran's C (max var / sum var)	1.6238	0.8317
Hartley's F (max var / min var)	20.4095	
Levene's F Test	1.3014	0.2777

7.4.3. Multiple Comparisons

Multiple comparisons (also known as multiple range, *post hoc* or *a posteriori* tests) are designed to compare all possible pairs of means of a group of subsamples. These tests are usually performed after an ANOVA, where the null hypothesis "all population means are equal" is rejected. The null hypotheses "mean A is equal to mean B", "mean A is equal to mean C", etc. can be tested for all k(k - 1)/2 pairs of k subgroups. It is also possible to test all subgroups against a control subgroup (Dunnett test), in which case only k - 1 comparisons are made.

Comparisons are performed on subgroups of a single factor column, corresponding to a one-way analysis. However, the user can perform comparisons based on any ANOVA model by entering the Mean Square Error (MSE) and its Degrees of Freedom manually.

All tests are of the following general form:

$$[\bar{x}_i - \bar{x}_j] < R(a,g,f) * S_{\bar{x}}$$

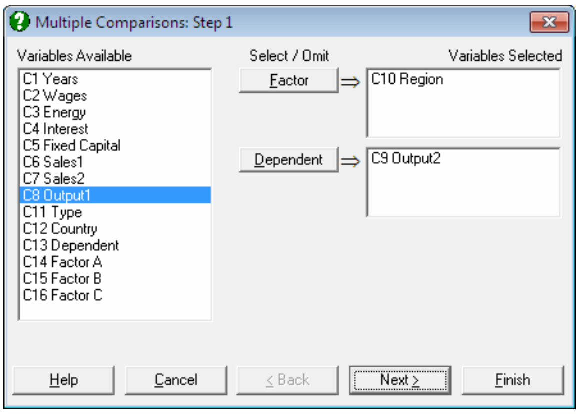
where q(α,k,f) is the range at α significance level, k is the number of subsets, f is the degrees of freedom of the between-groups sum of squares and S_x is the combined standard error. The combined standard error can be weighted by an arithmetic (the default) or harmonic mean of sample sizes. Although the latter is generally used when the sample sizes are not all equal, in such cases UNISTAT will not automatically revert to the harmonic mean. The selection should be made by the user, if the harmonic mean is required.

By default, the test result is reported by comparing a pair's observed q value against the table q for the given significance level and degrees of freedom. However, if required, comparisons can be made according to the probability value or confidence interval. To select one of these options, the following line should be entered and edited in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
ComparisonCriterion=i
```

where:

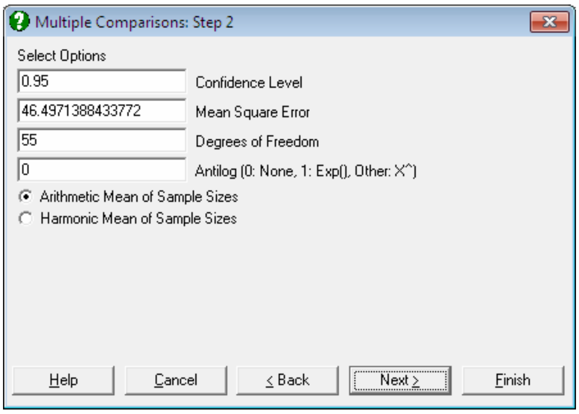
- **i = 1: Critical value:** Significant when the observed q value is greater than the table q (except for the upper-tailed Dunnett test).
- **i = 2: Probability:** Significant when the observed p-value is less than the given significance level (0.05 is the default).
- **i = 3: Confidence interval:** Significant when the confidence interval does not include zero.


Multiple Comparisons: Step 1

Variables Available: C1 Years, C2 Wages, C3 Energy, C4 Interest, C5 Fixed Capital, C6 Sales1, C7 Sales2, C8 Output1, C11 Type, C12 Country, C13 Dependent, C14 Factor A, C15 Factor B, C16 Factor C

Variables Selected:
Factor ⇒ C10 Region
Dependent ⇒ C9 Output2

Select at least one data column by clicking on [Dependent] and at least one factor column by clicking on [Factor], which separates the data column into a number of subgroups (or treatments). The procedure is run separately for each [Factor] / [Dependent] pair. A one-way ANOVA model is estimated and the Mean Square Error and its Degrees of Freedom are displayed in a dialogue.


Multiple Comparisons: Step 2

Select Options

0.95 Confidence Level
46.4971388433772 Mean Square Error
55 Degrees of Freedom
0 Antilog (0: None, 1: Exp(), Other: X^)

● Arithmetic Mean of Sample Sizes
○ Harmonic Mean of Sample Sizes

The ability to edit these values enables the user to make comparisons based on any ANOVA model. For instance, if you wish to perform Multiple Comparisons for a 3-way ANOVA, you can copy the Mean Square Error and its Degrees of Freedom from the 3-way ANOVA table and paste them into the respective text boxes on this dialogue. Here you can also change the significance level and choose between arithmetic and harmonic means for sample sizes.

Back Transformations: Often, the data used in Multiple Comparisons is already transformed with a function like e or 10 based logarithm. In such cases the results need to be transformed back to the original scale and the user is faced with the task of applying back transformations manually. The Antilog box allows the user to specify which back-transformation is to be applied in the output. The output values that are affected by this control are:

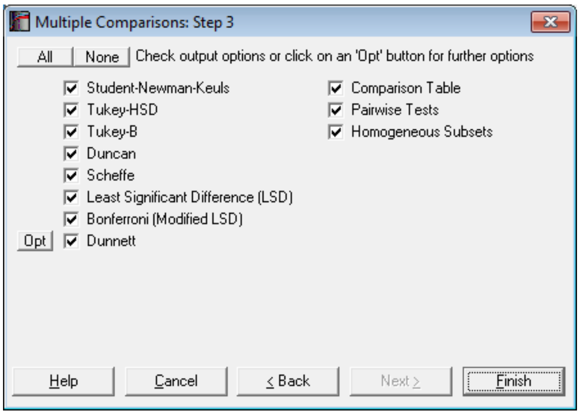
- means,
- difference between means, and
- lower and upper confidence limits for difference between means.

Let X be the value entered into the Antilog box.

- If $X = 0$ then no back-transformation is performed.
- If $X = 1$ then the natural antilog of the output value Y, Exp(Y) is displayed.
- If $1 < X \leq 16$, then X based power of Y, X^Y is displayed.

When a back-transformation base value is specified, the columns affected will be marked by an asterisk in the output.

Next the Output Options Dialogue is displayed. As of this version of UNISTAT, this dialogue features three new check boxes for the three main components of Multiple Comparisons output; Comparison Table, Pairwise Tests and Homogeneous Subsets.



Check output options or click on an 'Opt' button for further options

☑ Student-Newman-Keuls
☑ Tukey-HSD
☑ Tukey-B
☑ Duncan
☑ Scheffe
☑ Least Significant Difference (LSD)
☑ Bonferroni (Modified LSD)
☑ Dunnett

☑ Comparison Table
☑ Pairwise Tests
☑ Homogeneous Subsets

Some or all of the eight comparison tests can be selected for output.

1) Student-Newman-Keuls
2) Tukey-HSD
3) Tukey-B
4) Duncan
5) Scheffe
6) Least Significant Difference (LSD)
7) Bonferroni (Modified LSD)
8) Dunnett

The Dunnett test is different from the others in that it requires selection of a control group and allows selection of two-tailed, lower and upper tail tests. These dialogues are accessible from the [Opt] button situated immediately to the left of the Dunnett check box. By default, the control group is the first group of sorted factor levels and a two-tailed test is performed.

These tests are covered in standard textbooks for experimental design (e.g. Montgomery, D. C. (1991). The present implementation is based on Winer, B. J. (1970).

Multiple Comparisons are also available following a GLM analysis, where the Mean Square Error is based on the error term of the model fitted and therefore there is no need to copy and paste any values (see 7.3.2. General Linear Model).

Comparison Table: The output includes a comparison matrix where significantly different pairs are marked with an asterisk. Where appropriate, the groups are divided into homogenous subsets where their differences are not significant according to the given significance criterion. Vertical bars on the right of the table indicate the homogeneous subsets. Homogeneous subsets are not available for all tests.

Pairwise Tests: A second table displays the pairs compared, their difference, standard error, observed and table q statistics, probability, confidence interval and the test result. Double asterisks are displayed on the last column, if the difference between two means is significantly different from zero. By design, confidence intervals are not available for Student-Newman-Keuls and Duncan methods.

Homogeneous Subsets: The third part of output will list homogenous subsets, their pooled mean and confidence intervals. The critical values are based on t-distribution. Differences between pooled means of homogenous subsets and their confidence intervals are also displayed. Critical values are based on the studentised t-distribution.

Another way to shorten the Multiple Comparisons output is to enter the following line in the [Options] section of *Documents*\Unistat65\Unistat65.ini file:

```
ComparisonShortOutput=1
```

In this case, only a comparison table will be displayed, the contents of which depend on the comparison criterion selected by the `ComparisonCriterion` entry as described above. If `ComparisonCriterion=2`, then the short output will display only the observed probability and the test result based on probability. If `ComparisonCriterion=3`, then the short output will display only the confidence interval and the test result based on whether it contains zero.

UNISTAT also offers a number of nonparametric Multiple Comparisons, comparisons of medians, variances and intercepts and slopes of regression lines. The full list of these procedures and the available Multiple Comparisons is as follows:

Statistics 1 → Nonparametric Tests (Multisample) → Kruskal-Wallis ANOVA
Nonparametric comparisons against a control group:
Dunnett, Dunn.
Nonparametric comparisons:
with rank sums: Tukey, S-N-K
with mean ranks: Tukey, S-N-K
t, Dunn
Statistics 1 → Nonparametric Tests (Multisample) → Multisample Median Test

Multiple comparison of medians: Tukey
Statistics 1 → Nonparametric Tests (Multisample) → Friedman Two-Way ANOVA
Nonparametric Multiple Comparisons: Tukey
Statistics 1 → Nonparametric Tests (Multisample) → Quade Two-Way ANOVA
Nonparametric Multiple Comparisons: Tukey
Statistics 1 → Tests for ANOVA → Homogeneity of Variance Tests
Comparison of variances against a control group: Dunnett
Multiple comparison of variances: Tukey, S-N-K
Statistics 1 → Tests for ANOVA → Heterogeneity of Regression
Comparison of slopes against a control group: Dunnett
Comparison of intercepts against a control group: Dunnett
Multiple comparison of slopes: Tukey
Multiple comparison of intercepts: Tukey

By default, these procedures will display all three main components of the multiple comparisons output; Comparison Table, Pairwise Tests and Homogeneous Subsets. If you wish to make the output from them to obey the Multiple Comparisons Output Options Dialogue check boxes, to display only the selected parts of the output, enter the following line in the [Options] section of *Documents**Unistat65**Unistat65.ini* file:

```
MultiCompOthers=1
```

7.4.3.1. Student-Newman-Keuls

Student-Newman-Keuls procedure is based on the studentised range and has different range values for different size subsets. Comparisons can be made at any confidence level.

The standard error is defined as:

$$SE = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}$$

Confidence intervals are not available for this test due to its multistage character.

Example

Examples 11.1 and 11.4 on p. 228 and p. 234 from Zar, J. H. (2010). A researcher wants to find out which variables have different means at a 95% confidence level.

The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data should be stacked in a single column *Concentration* and a factor column *Water* created to keep track of the group memberships.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Multiple Comparisons, *Water* (*C3*) as [Factor] and *Concentration* (*C4*) as [Dependent], click [Next] to accept default values at the next dialogue and select Student-Newman-Keuls to obtain the following results:

Multiple Comparisons

Student-Newman-Keuls

For Concentration, classified by Water
Mean Square Error: 9.7652, Degrees of Freedom: 25
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	1	2	4	3	5	
1	6	32.0833		**	**	**	**	|
2	6	40.2333	**				**	|
4	6	41.1000	**				**	|
3	6	44.0833	**				**	|
5	6	58.3000	**	**	**	**		|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
5 - 1	26.2167	1.8042	20.5500	4.1534	0.0000
3 - 1	12.0000	1.8042	9.4062	3.8900	0.0000
4 - 1	9.0167	1.8042	7.0677	3.5226	0.0001
2 - 1	8.1500	1.8042	6.3884	2.9126	0.0001
5 - 2	18.0667	1.8042	14.1616	3.8900	0.0000
3 - 2	3.8500	1.8042	3.0178	3.5226	0.1032
4 - 2	0.8667	1.8042	0.6793	2.9126	0.6351
5 - 4	17.2000	1.8042	13.4823	3.5226	0.0000
3 - 4	2.9833	1.8042	2.3385	2.9126	0.1107
5 - 3	14.2167	1.8042	11.1438	2.9126	0.0000

Comparison	Lower 95%	Upper 95%	Result
5 - 1	*	*	**
3 - 1	*	*	**
4 - 1	*	*	**
2 - 1	*	*	**
5 - 2	*	*	**
3 - 2	*	*	
4 - 2	*	*	
5 - 4	*	*	**
3 - 4	*	*	
5 - 3	*	*	**

```
            Homogeneous Subsets:
                        Group 1:    1
                   Pooled mean =    32.0833
       95% Confidence Interval =    29.4559 <> 34.7108
                        Group 2:    2 4 3
                   Pooled mean =    41.8056
       95% Confidence Interval =    40.2886 <> 43.3225
                        Group 3:    5
                   Pooled mean =    58.3000
       95% Confidence Interval =    55.6725 <> 60.9275
 Between Homogeneous subsets:
               Group 2 - Group 1
                    Difference =    9.7222
       95% Confidence Interval =    5.3959 <> 14.0485
               Group 3 - Group 2
                    Difference =    16.4944
       95% Confidence Interval =    12.1681 <> 20.8208
```

7.4.3.2. Tukey-HSD

Tukey-HSD procedure (known as Tukey's honestly significant difference test) is also based on the studentised range though the range value is independent of different subset sizes. The range value used here is the largest range used in Student-Newman-Keuls method. The test can be performed at any confidence level. The standard error is defined as in Student-Newman-Keuls test.

In case sample sizes are not equal, confidence intervals for homogenous subsets and their differences (the third and fourth part of output) are not displayed.

Example 1

Example 11.1 on p. 228 from Zar, J. H. (2010). The data used (which has equal sample sizes) is as in section 7.4.3.1. Student-Newman-Keuls.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Multiple Comparisons, *Water* (*C3*) as [Factor] and *Concentration* (*C4*) as [Dependent] and Tukey-HSD at the next dialogue to obtain the following results:

Multiple Comparisons

Tukey-HSD

For Concentration, classified by Water
Mean Square Error: 9.7652, Degrees of Freedom: 25
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	1	2	4	3	5	
1	6	32.0833		**	**	**	**	│
2	6	40.2333	**				**	│
4	6	41.1000	**				**	│
3	6	44.0833	**				**	│
5	6	58.3000	**	**	**	**		│

Comparison	Difference	Standard Error	q Stat	Table q	Probability
5 - 1	26.2167	1.8042	20.5500	4.1534	0.0000
3 - 1	12.0000	1.8042	9.4062	4.1534	0.0000
4 - 1	9.0167	1.8042	7.0677	4.1534	0.0003
2 - 1	8.1500	1.8042	6.3884	4.1534	0.0011
5 - 2	18.0667	1.8042	14.1616	4.1534	0.0000
3 - 2	3.8500	1.8042	3.0178	4.1534	0.2376
4 - 2	0.8667	1.8042	0.6793	4.1534	0.9885
5 - 4	17.2000	1.8042	13.4823	4.1534	0.0000
3 - 4	2.9833	1.8042	2.3385	4.1534	0.4791
5 - 3	14.2167	1.8042	11.1438	4.1534	0.0000

Comparison	Lower 95%	Upper 95%	Result
5 - 1	20.9180	31.5153	**
3 - 1	6.7014	17.2986	**
4 - 1	3.7180	14.3153	**
2 - 1	2.8514	13.4486	**
5 - 2	12.7680	23.3653	**
3 - 2	-1.4486	9.1486	
4 - 2	-4.4320	6.1653	
5 - 4	11.9014	22.4986	**
3 - 4	-2.3153	8.2820	
5 - 3	8.9180	19.5153	**

```
                   Homogeneous Subsets:
                               Group 1:    1
                          Pooled mean =    32.0833
              95% Confidence Interval =    29.4559 <> 34.7108
                               Group 2:    2 4 3
                          Pooled mean =    41.8056
              95% Confidence Interval =    40.2886 <> 43.3225
                               Group 3:    5
                          Pooled mean =    58.3000
              95% Confidence Interval =    55.6725 <> 60.9275
       Between Homogeneous subsets:
                      Group 2 - Group 1
                           Difference =    9.7222
              95% Confidence Interval =    5.3959 <> 14.0485
                      Group 3 - Group 2
                           Difference =    16.4944
              95% Confidence Interval =    12.1681 <> 20.8208
```

Example 2

Example 11.2 on p. 231 from Zar, J. H. (2010). The data used (which has unequal sample sizes) is as in section 7.4.2.3. Homogeneity of Variance Examples.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Multiple Comparisons, *Feed* (*C1*) as [Factor] and *Weight* (*C2*) as [Dependent] and Tukey-HSD at the next dialogue to obtain the following results:

Multiple Comparisons

Tukey-HSD

For Weight, classified by Feed
Mean Square Error: 9.383333333333, Degrees of Freedom: 15
** denotes significantly different pairs.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	4	1	2	3	
4	5	63.2400			**	**	|
1	5	64.6200			**	**	|
2	5	71.3000	**	**			|
3	4	73.3500	**	**			|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
3 - 4	10.1100	2.0549	6.9580	4.0760	0.0009
2 - 4	8.0600	1.9374	5.8836	4.0760	0.0042
1 - 4	1.3800	1.9374	1.0074	4.0760	0.8907
3 - 1	8.7300	2.0549	6.0082	4.0760	0.0035
2 - 1	6.6800	1.9374	4.8762	4.0760	0.0168
3 - 2	2.0500	2.0549	1.4109	4.0760	0.7530

Comparison	Lower 95%	Upper 95%	Result
3 - 4	4.1876	16.0324	**
2 - 4	2.4763	13.6437	**
1 - 4	-4.2037	6.9637	
3 - 1	2.8076	14.6524	**
2 - 1	1.0963	12.2637	**
3 - 2	-3.8724	7.9724	

7.4.3.3. Tukey-B

Tukey-B procedure (also based on the studentised range) uses at each step the average of range values for Tukey-HSD and Student-Newman-Keuls methods. Therefore, range values are different for different sized pairs. The test can be performed at any confidence level.

7.4.3.4. Duncan

The Duncan method uses different range values for different subsample sizes. The range calculations are based on Duncan's table of significant ranges. Confidence level can be set only at 0.90 or 0.95 or 0.99 and probability values for individual comparisons are not available.

Confidence intervals are not available for this test due to its multistage character.

Example

Example 3-6 on p. 76 from Montgomery, D. C. (1991). Data from Example 3-1 is transformed into a suitable format first, where rows of the table are stacked in one column and a second factor column containing integers from 1 to 5 is created to keep track of the groups.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Multiple Comparisons and select *Cotton percentage* (*C15*) [Factor] and *Tensile strength* (*C16*) as [Dependent]. Then select Duncan, Arithmetic Mean of Sample Sizes from the next two dialogues and accept the default Mean Square Error and Degrees of Freedom values. The following results will be obtained:

Multiple Comparisons

Duncan

For Tensile strength, classified by Cotton percentage
Mean Square Error: 8.06, Degrees of Freedom: 20
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	1	5	2	3	4	
1	5	9.8000			**	**	**	|
5	5	10.8000			**	**	**	|
2	5	15.4000	**	**			**	|
3	5	17.6000	**	**			**	|
4	5	21.6000	**	**	**	**		|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
4 - 1	11.8000	1.7956	9.2939	3.2558	*
3 - 1	7.8000	1.7956	6.1434	3.1960	*
2 - 1	5.6000	1.7956	4.4107	3.0938	*
5 - 1	1.0000	1.7956	0.7876	2.9453	*
4 - 5	10.8000	1.7956	8.5063	3.1960	*
3 - 5	6.8000	1.7956	5.3558	3.0938	*
2 - 5	4.6000	1.7956	3.6231	2.9453	*
4 - 2	6.2000	1.7956	4.8833	3.0938	*
3 - 2	2.2000	1.7956	1.7328	2.9453	*
4 - 3	4.0000	1.7956	3.1505	2.9453	*

Comparison	Lower 95%	Upper 95%	Result
4 - 1	*	*	**
3 - 1	*	*	**
2 - 1	*	*	**
5 - 1	*	*	
4 - 5	*	*	**
3 - 5	*	*	**
2 - 5	*	*	**
4 - 2	*	*	**
3 - 2	*	*	
4 - 3	*	*	**

Homogeneous Subsets:	
Group 1:	1 5
Pooled mean =	10.3000
95% Confidence Interval =	8.4273 <> 12.1727
Group 2:	2 3
Pooled mean =	16.5000
95% Confidence Interval =	14.6273 <> 18.3727
Group 3:	4
Pooled mean =	21.6000
95% Confidence Interval =	18.9516 <> 24.2484

7.4.3.5. Scheffe

This method uses a single range value for all comparisons and is based on the F distribution. Any confidence level can be selected. The Scheffe method is the most conservative range test.

The standard error is defined as in Student-Newman-Keuls test and the critical value for a comparison is obtained from F distribution with k - 1 and n - k degrees of freedom, as follows:

$$S_\alpha = \sqrt{(k-1)F_{1-\alpha,k-1,n-k}}$$

Example

Example 11.5 on p. 238 from Zar, J. H. (2010). The data used (which has equal sample sizes) is as in section Student-Newman-Keuls.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Multiple Comparisons, *Water* (*C3*) as [Factor] and *Concentration* (*C4*) as [Dependent] and Scheffe at the next dialogue to obtain the following results:

Multiple Comparisons

Scheffe

For Concentration, classified by Water
Mean Square Error: 9.7652, Degrees of Freedom: 25
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	1	2	4	3	5	
1	6	32.0833		**	**	**	**	|
2	6	40.2333	**				**	|
4	6	41.1000	**				**	|
3	6	44.0833	**				**	|
5	6	58.3000	**	**	**	**		|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
5 - 1	26.2167	1.8042	14.5311	3.3219	0.0000
3 - 1	12.0000	1.8042	6.6512	3.3219	0.0000
4 - 1	9.0167	1.8042	4.9977	3.3219	0.0013
2 - 1	8.1500	1.8042	4.5173	3.3219	0.0038
5 - 2	18.0667	1.8042	10.0138	3.3219	0.0000
3 - 2	3.8500	1.8042	2.1339	3.3219	0.3613
4 - 2	0.8667	1.8042	0.4804	3.3219	0.9934
5 - 4	17.2000	1.8042	9.5334	3.3219	0.0000
3 - 4	2.9833	1.8042	1.6536	3.3219	0.6100
5 - 3	14.2167	1.8042	7.8798	3.3219	0.0000

Comparison	Lower 95%	Upper 95%	Result
5 - 1	20.2234	32.2099	**
3 - 1	6.0067	17.9933	**
4 - 1	3.0234	15.0099	**
2 - 1	2.1567	14.1433	**
5 - 2	12.0734	24.0599	**
3 - 2	-2.1433	9.8433	
4 - 2	-5.1266	6.8599	
5 - 4	11.2067	23.1933	**
3 - 4	-3.0099	8.9766	
5 - 3	8.2234	20.2099	**

Homogeneous Subsets:	
Group 1:	1
Pooled mean =	32.0833
95% Confidence Interval =	29.4559 <> 34.7108
Group 2:	2 4 3
Pooled mean =	41.8056
95% Confidence Interval =	40.2886 <> 43.3225
Group 3:	5
Pooled mean =	58.3000
95% Confidence Interval =	55.6725 <> 60.9275
Between Homogeneous subsets:	
Group 2 - Group 1	
Difference =	9.7222
95% Confidence Interval =	3.6039 <> 15.8406
Group 3 - Group 2	
Difference =	16.4944
95% Confidence Interval =	10.3761 <> 22.6128

7.4.3.6. Least Significant Difference (LSD)

Least Significant Difference (LSD) method is based on the t-distribution and any confidence level can be selected. The range value is computed as:

$$q = t_{1-\alpha/2,f}$$

UNISTAT Version 5.5 (and in earlier versions) Least Significant Difference (LSD) and Bonferroni (Modified LSD) methods have been based on the F-statistic, using the following equation:

$$q = \sqrt{2F_{1-\alpha,1,f}}$$

The two methods are identical and generate exactly the same output, except for the critical values. If you wish to display the critical values based on the F-distribution, then enter the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
LSDBonferroniBaseT=1
```

Example

Example 3-5 on p. 74 from Montgomery, D. C. (1991). Data from Example 3-1 is transformed into a suitable format first, where rows of the table are stacked in one column and a second factor column containing integers from 1 to 5 is created to keep track of the groups.

Open ANOTESTS and select Statistics 1 → Tests for ANOVA → Multiple Comparisons, *Cotton percentage* (*C15*) [Factor] and *Tensile strength* (*C16*) as [Dependent]. Then select Least Significant Difference, Arithmetic Mean of Sample Sizes from the next two dialogues and accept the default Mean Square Error and Degrees of Freedom values. The following results will be obtained:

Multiple Comparisons

Least Significant Difference

For Tensile strength, classified by Cotton percentage
Mean Square Error: 8.06, Degrees of Freedom: 20
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Mean	1	5	2	3	4	
1	5	9.8000			**	**	**	|
5	5	10.8000			**	**	**	|
2	5	15.4000	**	**			**	|
3	5	17.6000	**	**			**	|
4	5	21.6000	**	**	**	**		|

Comparison	Difference	Standard Error	q Stat	Table q	Probability
4 - 1	11.8000	1.7956	6.5718	2.0860	0.0000
3 - 1	7.8000	1.7956	4.3441	2.0860	0.0003
2 - 1	5.6000	1.7956	3.1188	2.0860	0.0054
5 - 1	1.0000	1.7956	0.5569	2.0860	0.5838
4 - 5	10.8000	1.7956	6.0149	2.0860	0.0000
3 - 5	6.8000	1.7956	3.7871	2.0860	0.0012
2 - 5	4.6000	1.7956	2.5619	2.0860	0.0186
4 - 2	6.2000	1.7956	3.4530	2.0860	0.0025
3 - 2	2.2000	1.7956	1.2253	2.0860	0.2347
4 - 3	4.0000	1.7956	2.2277	2.0860	0.0375

Comparison	Lower 95%	Upper 95%	Result
4 - 1	8.0545	15.5455	**
3 - 1	4.0545	11.5455	**
2 - 1	1.8545	9.3455	**
5 - 1	-2.7455	4.7455	
4 - 5	7.0545	14.5455	**
3 - 5	3.0545	10.5455	**
2 - 5	0.8545	8.3455	**
4 - 2	2.4545	9.9455	**
3 - 2	-1.5455	5.9455	
4 - 3	0.2545	7.7455	**

Homogeneous Subsets:	
Group 1:	1 5
Pooled mean =	10.3000
95% Confidence Interval =	8.4273 <> 12.1727
Group 2:	2 3
Pooled mean =	16.5000
95% Confidence Interval =	14.6273 <> 18.3727
Group 3:	4
Pooled mean =	21.6000
95% Confidence Interval =	18.9516 <> 24.2484

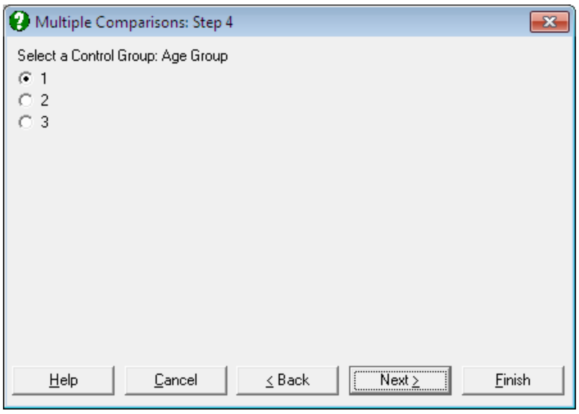
7.4.3.7. Bonferroni (Modified LSD)

The Bonferroni method (also known as Modified Least Significant Difference) is equivalent to the Least Significant Difference (LSD) method except in its definition of the significance level, which is:

$$\alpha' = \alpha(k(k-1)/2)$$

7.4.3.8. Dunnett

Under some circumstances, it may be more desirable to test the means of subgroups against the mean of a control group, rather than testing for all possible pairs. In such cases Dunnett test can be performed with an upper or lower one-tailed or two-tailed null hypothesis.


Multiple Comparisons: Step 4

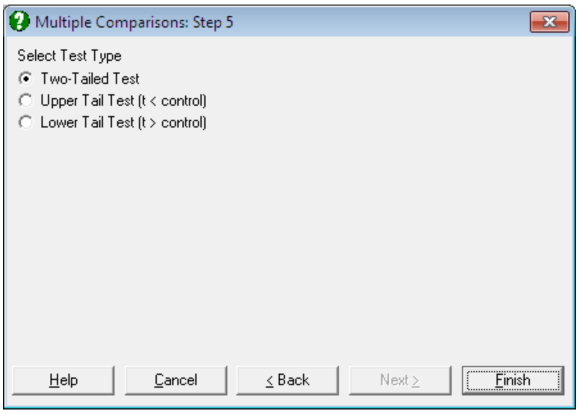
Select a Control Group: Age Group

● 1

○ 2

○ 3

Two further dialogues pop up to select the control subgroup and the type of test.


Multiple Comparisons: Step 5

Select Test Type

● Two-Tailed Test

○ Upper Tail Test (t < control)

○ Lower Tail Test (t > control)

The type of test can be one of the following:

- **Two-tailed test:** The null hypothesis $\mu_i = \mu_{ctrl}$ is tested against $\mu_i \neq \mu_{ctrl}$
- **Upper tail test:** The null hypothesis $\mu_i < \mu_{ctrl}$ is tested against $\mu_i \geq \mu_{ctrl}$
- **Lower tail test:** The null hypothesis $\mu_i > \mu_{ctrl}$ is tested against $\mu_i \leq \mu_{ctrl}$

The standard error is calculated as:

$$SE = \sqrt{s^2\left(\frac{1}{n_{ctrl}} + \frac{1}{n_A}\right)}$$

Critical and p-values are computed using the algorithm developed by Charles Dunnett and the test can be performed at any confidence level. The maximum number groups (factor levels) is limited to 50. Critical and p-values generated by this algorithm are sensitive to difference in sample sizes. When all sample sizes are equal, the results are identical to the published tables for Dunnett's critical values. When the sample sizes are not equal, the correct values may diverge from the published tables, as the latter do not take unequal sample size correction into consideration. By default, UNISTAT will compute the corrected critical values. You may, however, override this and not apply unequal sample size correction by including the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
DunnettSampleSizeCorr=0
```

Example 1

Example 3-7 on p. 80 from Montgomery, D. C. (1991). Data from Example 3-1 is transformed into a suitable format first, where rows of the table are stacked in one column and a second factor column containing integers from 1 to 5 is created to keep track of the groups.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Multiple Comparisons and select *Cotton percentage* (*C15*) [Factor] and *Tensile strength* (*C16*) as [Dependent]. Then select Dunnett, Arithmetic Mean of Sample Sizes from the next two dialogues and select group 5 as the control group from the last dialogue. Then select two-tailed test and accept the default Mean Square Error and Degrees of Freedom values. The following results will be obtained:

Multiple Comparisons

Dunnett

For Tensile strength, classified by Cotton percentage
Control Group: 5, Two-Tailed Test
Mean Square Error: 8.06, Degrees of Freedom: 20
** denotes significantly different pairs. Vertical bars show homogeneous subsets.
A pairwise test result is significant if its q stat value is greater than the table q.

<table>
<tr><th>Group</th><th>Cases</th><th>Mean</th><th>5</th><th></th></tr>
<tr><td>1</td><td>5</td><td>9.8000</td><td></td><td>|</td></tr>
<tr><td>5</td><td>5</td><td>10.8000</td><td></td><td>| |</td></tr>
<tr><td>2</td><td>5</td><td>15.4000</td><td></td><td> |</td></tr>
<tr><td>3</td><td>5</td><td>17.6000</td><td>**</td><td> |</td></tr>
<tr><td>4</td><td>5</td><td>21.6000</td><td>**</td><td> |</td></tr>
</table>

Comparison	Difference	Standard Error	q Stat	Table q	Probability
4 - 5	10.8000	1.7956	6.0149	2.6510	0.0000
3 - 5	6.8000	1.7956	3.7871	2.6510	0.0041
2 - 5	4.6000	1.7956	2.5619	2.6510	0.0600
1 - 5	-1.0000	1.7956	0.5569	2.6510	0.9469

Comparison	Lower 95%	Upper 95%	Result
4 - 5	6.0399	15.5601	**
3 - 5	2.0399	11.5601	**
2 - 5	-0.1601	9.3601	
1 - 5	-5.7601	3.7601	

```
          Homogeneous Subsets:
                      Group 1:   1 5
                Pooled mean =    10.3000
     95% Confidence Interval =   8.4273 <> 12.1727
                      Group 2:   5 2
                Pooled mean =    13.1000
     95% Confidence Interval =   11.2273 <> 14.9727
                      Group 3:   3
                Pooled mean =    17.6000
     95% Confidence Interval =   14.9516 <> 20.2484
                      Group 4:   4
                Pooled mean =    21.6000
     95% Confidence Interval =   18.9516 <> 24.2484
 Between Homogeneous subsets:
             Group 2 - Group 1
                 Difference =    2.8000
     95% Confidence Interval =  -2.5730 <> 8.1730
             Group 3 - Group 2
                 Difference =    4.5000
     95% Confidence Interval =  -2.0805 <> 11.0805
             Group 4 - Group 3
                 Difference =    4.0000
     95% Confidence Interval =  -3.5985 <> 11.5985
```

Example 2

Re-running the above example after including `ComparisonCriterion=3` and `ComparisonShortOutput=1` lines in *Documents\Unistat65\Unistat65.ini* file under the [Options] group, we obtain the following result.

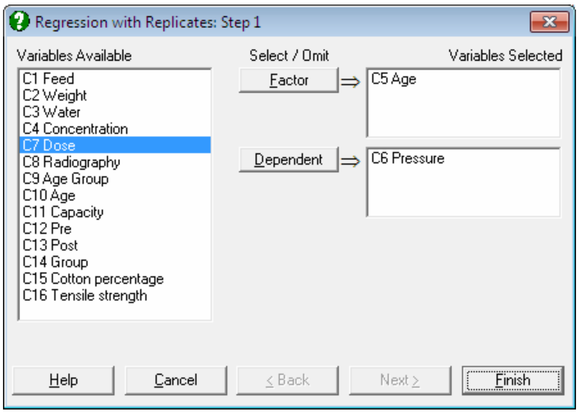
Multiple Comparisons

Dunnett

Comparison	Difference	Lower 95%	Upper 95%	Result
4 - 5	10.8000	6.0399	15.5601	**
3 - 5	6.8000	2.0399	11.5601	**
2 - 5	4.6000	-0.1601	9.3601	
1 - 5	-1.0000	-5.7601	3.7601	

7.4.4. Regression with Replicates

This is a test of linearity for bivariate regressions when the data contains multiple measurements of the dependent variable for each value of the independent variable. The null hypothesis tested is “population regression is linear” against the alternative hypothesis “population regression is not linear”.



Factor ⇒ C5 Age

Dependent ⇒ C6 Pressure

Select at least one variable as [Factor] and at least one data variable as [Dependent]. The procedure is run separately for each [Factor] / [Dependent] pair. Output consists of simple regression results, and ANOVA of regression table (testing the null hypothesis that “the slope of the regression line is zero”), another ANOVA table where the among groups variation is broken down into the Linear Regression and its error sum of squares. The test statistic is the F-test for regression error term, which is defined as:

$$F = \frac{\text{Regression Error MS}}{\text{Within Groups MS}}$$

Example 1

Examples 17.8a and 17.8b on pp. 349, 350 from Zar, J. H. (2010). The null hypothesis that “the population regression is linear” is tested at a 95% confidence level.

The table format given in the book can be transformed into the factor format by using UNISTAT’s Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All data on systolic blood pressure should be stacked in a column *Pressure* (the Y variable) and all data on ages (the X variable) should be expanded to form a column *Age* to keep track of the age groups of pressure measurements.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Regression with Replicates and select *Age* (*C5*) as [Factor] and *Pressure* (*C6*) as [Dependent] to obtain the following results:

Regression with Replicates

Regression results

Constant =	68.7849
Slope =	1.3031
R-squared =	0.9827
Standard Error =	2.5702

ANOVA of regression

Due To	Sum of Squares	DoF	Mean Square	F-stat	Prob
Regression	6750.289	1	6750.289	1021.819	0.0000
Error	118.911	18	6.606		
Total	6869.200	19	361.537		

Test of linearity

Due To	Sum of Squares	DoF	Mean Square	F-stat	Prob
Among groups	6751.933	4	1687.983	215.916	0.0000
Regression	6750.289	1	6750.289	863.454	0.0000
Error	1.644	3	0.548	0.070	0.9750
Within groups	117.267	15	7.818		
Total	6869.200	19	361.537		

Since the probability value for Regression in the ANOVA of regression table is less than 5% reject the null hypothesis that “the regression slope is zero”. The test of linearity is the F-statistic on regression error, which is 0.070 with a 97.5% probability, therefore do not reject the null hypothesis of linearity.

Example 2

Example 11.2 on p. 316 in Armitage & Berry (2002). Data on radiographic assessments of bone healing for three doses of vitamin D are given.

The format of Table 9.3 in the book is not suitable for analysis in UNISTAT. All data should be stacked in a single column *Radiography* and a factor column *Dose* created to keep track of the group memberships.

Open ANOTESTS, select Statistics 1 → Tests for ANOVA → Regression with Replicates and select *Dose* (*C7*) as [Factor] and *Radiography* (*C8*) as [Dependent] to obtain the following results:

Regression with Replicates

For Radiography, classified by Dose

Constant =	1.2195
Slope =	0.7876
R-squared =	0.2399
Standard Error =	1.2408

ANOVA of regression

Due To	Sum of Squares	DoF	Mean Square	F-stat	Prob
Regression	14.089	1	14.089	9.151	0.0052
Error	44.645	29	1.539		
Total	58.734	30	1.958		

Test of linearity

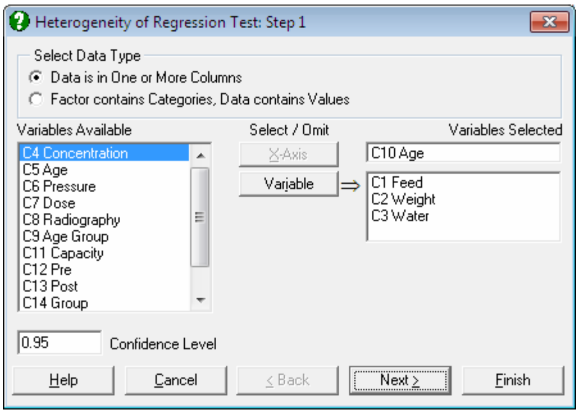
Due To	Sum of Squares	DoF	Mean Square	F-stat	Prob
Among groups	16.992	2	8.496	5.699	0.0084
Regression	14.089	1	14.089	9.451	0.0047
Error	2.903	1	2.903	1.948	0.1738
Within groups	41.742	28	1.491		
Total	58.734	30	1.958		

The test of linearity is the F-statistic on regression error, which is 1.948 with a 17% tail probability. Therefore, do not reject the null hypothesis of linearity.

7.4.5. Heterogeneity of Regression

This procedure (which is also known as analysis of covariance) is used to test whether slopes and / or intercepts of a number of bivariate regression lines are significantly different. These are also known as slope or parallelism tests. Data in two different formats can be analysed:

1) **Data is in One or More Columns:** Select an [X-Axis] variable and any number of Y-Axis variables by clicking [Variable]. Each Y-Axis variable is regressed against the same X-Axis variable.



Heterogeneity of Regression Test: Step 1

Select Data Type

● Data is in One or More Columns

○ Factor contains Categories, Data contains Values

Variables Available: C4 Concentration, C5 Age, C6 Pressure, C7 Dose, C8 Radiography, C9 Age Group, C11 Capacity, C12 Pre, C13 Post, C14 Group

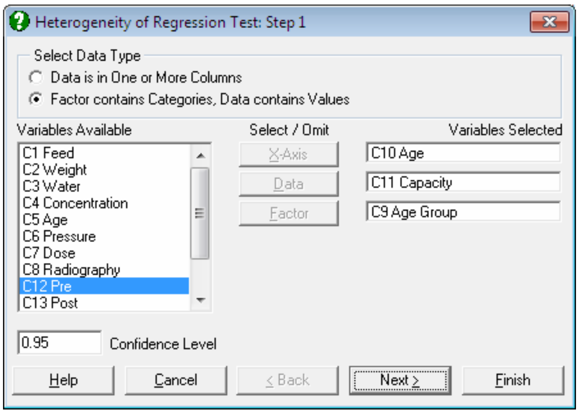
Variables Selected:

C10 Age

C1 Feed, C2 Weight, C3 Water

0.95 Confidence Level

2) **Factor contains Categories, Data contains Values:** Select a [Data] column and a [Factor] column: Subgroups of [Data] defined by the levels of [Factor] are the Y-Axis variables. You are also required to select an [X-Axis] variable which has the same length as the other two. In this case it is possible to regress each Y-Axis variable against different values of the [X-Axis] variable.



Heterogeneity of Regression Test: Step 1

Select Data Type

○ Data is in One or More Columns

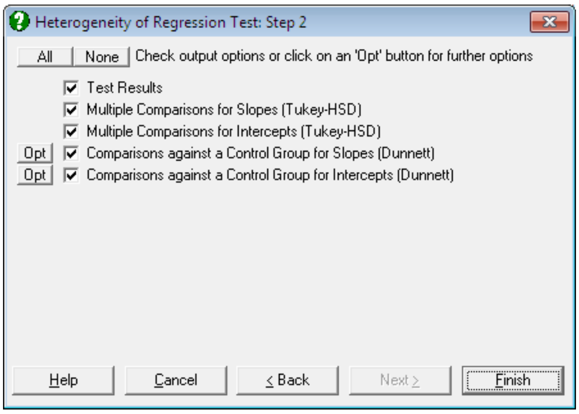
● Factor contains Categories, Data contains Values

Variables Available: C1 Feed, C2 Weight, C3 Water, C4 Concentration, C5 Age, C6 Pressure, C7 Dose, C8 Radiography, C12 Pre, C13 Post

Variables Selected: C10 Age, C11 Capacity, C9 Age Group

0.95 Confidence Level

The output options include the test results and four multiple comparison procedures.



Heterogeneity of Regression Test: Step 2

Check output options or click on an 'Opt' button for further options

☑ Test Results

☑ Multiple Comparisons for Slopes (Tukey-HSD)

☑ Multiple Comparisons for Intercepts (Tukey-HSD)

☑ Comparisons against a Control Group for Slopes (Dunnett)

☑ Comparisons against a Control Group for Intercepts (Dunnett)

7.4.5.1. Heterogeneity of Regression Test Results

The output includes a summary table for each individual regression, as well as the *pooled*, *common* and *total* regressions. For the pooled regression, the residual SS and residual DF are the sums of individual SS and DF values respectively. For the common regression, sum of all individual difference sum of squares are used to compute the residual SS figure. The residual DF is the total number of cases minus number of regressions minus one. For the total regression all Y-Axis variables are regressed on the X-Axis variable. Three null hypotheses are tested:

1) All slopes are equal:

$$F = \frac{\frac{SSc - SSp}{k - 1}}{\frac{SSp}{DFp}}$$

with k - 1 and DFp degrees of freedom.

2)All intercepts are equal:

$$F = \frac{\frac{SSt - SSc}{k - 1}}{\frac{SSc}{DFc}}$$

with k - 1 and DFc degrees of freedom. This test statistic can also be obtained by running an analysis of covariance where X-Axis variable is the covariate, Y-Axis variable is the data, and the factor is the classification variable. The F-statistic on the main effect and its probability are identical to the results obtained using the present method.

3)All regressions are equal:

$$F = \frac{\frac{SSt - SSp}{2(k - 1)}}{\frac{SSp}{DFp}}$$

with 2(k - 1) and DFp degrees of freedom.

7.4.5.2. Heterogeneity of Regression Multiple Comparisons

If one of the three null hypotheses is rejected then multiple comparison tests can be performed to find out which slopes or intercepts are significantly different. Here we provide a Tukey-HSD type test to compare all possible pairs of regressions for their slopes and / or intercepts and a Dunnett type test to compare all regression lines against a control line.

Multiple comparison of intercepts is only meaningful when the equality of slopes is accepted. Accordingly, when the differences between intercepts are calculated, the slopes of individual lines are all assumed to be equal to the slope of the *common* regression. Therefore, the values of the *difference* column in the comparison table do not necessarily correspond to differences between actual intercepts (see Zar (2010), p. 376, equation 18.39).

7.4.5.3. Heterogeneity of Regression Examples

Example 1

Examples 11.3 on p. 325, p. 329 and 11.4. on p. 334 in Armitage & Berry (2002). Ages and vital capacities for three groups of workers in the cadmium industry are given, where x is the age last birthday (years) and y is vital capacity (litres).

The data in Table 9.4. is given in the form of pairs of columns, one for x and one for y in different groups. In order to analyse this data in UNISTAT, all x values should be stacked in one column, y values in another column and a third column (Group) should be created to keep track of group memberships. Therefore, the resulting data matrix should have 84 rows and 3 columns.

Open ANOTESTS and select Statistics 1 → Tests for ANOVA → Heterogeneity of Regression. From the Variable Selection Dialogue select the second data option Factor contains categories Data contains values, assign *Age Group* (*C9*) as [Factor], *Age* (*C10*) as [X-Axis] and *Capacity* (*C11*) as [Data]. The following results are obtained:

Heterogeneity of Regression Test

Test Results

X Axis: Age, Dependent Variable: Capacity, classified by Age Group

Regression	Cases	Intercept	Slope	Residual SS	R-DF
1	12	8.1834	-0.0851	5.1311	10
2	28	6.2300	-0.0465	7.6050	26
3	44	5.6803	-0.0306	14.7991	42
Pooled	*	*	*	27.5352	78
Common	84	6.0048	-0.0398	30.0347	80
Total	84	6.0333	-0.0405	30.1964	82

Null hypothesis: All slopes are equal	
F(2,78) =	3.5402
Right-Tail Probability =	0.0338
Null hypothesis: All intercepts are equal	
F(2,80) =	0.2153
Right-Tail Probability =	0.8067
Null hypothesis: All regressions are identical	
F(4,78) =	1.8846
Right-Tail Probability =	0.1215

The total regression results given by Armitage and Berry on p. 329 correspond to common regression results above. According to the approach adopted here (following Zar (2010), pp. 375-378) the total regression is identical to the one run on all groups. This discrepancy does not affect the test of slopes, but it does affect the tests of intercepts and regressions.

Since the between slopes F-value of 3.54 has a tail probability less than 5% then reject the first null hypothesis that "slopes are the same". It cannot be rejected at a 99% confidence level. The between groups F-value of 0.22 has a tail probability of 0.81, then do not reject the second null hypothesis that "the intercepts are the same". Multiple comparisons will answer the question which slopes are different.

Multiple comparisons for slopes

Method: 95% Tukey-HSD interval.
** denotes significantly different pairs.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	SSQ(x)	Slope	1	2	3	
1	912.2500	-0.0851			**	I
2	2282.7143	-0.0465				I I
3	6197.1591	-0.0306	**			I

Comparison	Difference	Standard Error	q Stat	Table q	Probability
3 - 1	0.0545	0.0146	3.7276	3.3765	0.0269
2 - 1	0.0386	0.0161	2.3890	3.3765	0.2155
3 - 2	0.0159	0.0101	1.5771	3.3765	0.5075

Comparison	Lower 95%	Upper 95%	Result
3 - 1	0.0051	0.1039	**
2 - 1	-0.0159	0.0931	
3 - 2	-0.0182	0.0500	

Multiple comparisons for intercepts

Method: 95% Tukey-HSD interval.
** denotes significantly different pairs.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Intercept	1	2	3	
1	12	5.9630				I
2	28	6.0013				I
3	44	6.0729				I

Comparison	Difference	Standard Error	q Stat	Table q	Probability
3 - 1	0.1099	0.1428	0.7698	3.3765	0.8497
2 - 1	0.0383	0.1669	0.2297	3.3765	0.9856
3 - 2	0.0716	0.1001	0.7156	3.3765	0.8686

Comparison	Lower 95%	Upper 95%	Result
3 - 1	-0.3723	0.5922	
2 - 1	-0.5252	0.6018	
3 - 2	-0.2663	0.4095	

Multiple comparisons for slopes

Method: 95% Dunnett interval.
Control Group: 1, Two-Tailed Test
** denotes significantly different pairs.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	SSQ(x)	Slope	1	
1	912.2500	-0.0851		I
2	2282.7143	-0.0465		I
3	6197.1591	-0.0306	**	I

Comparison	Difference	Standard Error	q Stat	Table q	Probability
3 - 1	0.0545	0.0207	2.6358	2.1922	0.0168
2 - 1	0.0386	0.0228	1.6893	2.1922	0.1442

Comparison	Lower 95%	Upper 95%	Result
3 - 1	0.0092	0.0998	**
2 - 1	-0.0115	0.0886	

Multiple comparisons for intercepts

Method: 95% Dunnett interval.
Control Group: 1, Two-Tailed Test
** denotes significantly different pairs.
A pairwise test result is significant if its q stat value is greater than the table q.

Group	Cases	Intercept	1	
1	12	5.9630		I
2	28	6.0013		I
3	44	6.0729		I

Comparison	Difference	Standard Error	q Stat	Table q	Probability
3 - 1	0.1099	0.2020	0.5443	2.2066	0.7704
2 - 1	0.0383	0.2360	0.1624	2.2066	0.9755

Comparison	Lower 95%	Upper 95%	Result
3 - 1	-0.3357	0.5556	
2 - 1	-0.4825	0.5591	

Example 2

Table 8.1 on p. 326 from Tabachnick, B. G. & L. S. Fidell (1989). A reading test is given to disabled children before and after an experiment where two different teaching methods are applied to two thirds of children and one third are kept as controls. We would like to find out whether the teaching methods have significant effects on test results, having made an adjustment for their pre-test reading abilities.

The table format given in the book can be transformed into the factor format by using UNISTAT's Data → Stack Columns procedure and the **Level()** function (see 3.4.2.5. Statistical Functions). All *Pre* and *Post* data should be stacked in two columns and a factor column *Group* created to keep track of the group memberships. Therefore, the resulting data matrix should have 9 rows and 3 columns.

Open ANOTESTS and select Statistics 1 → Tests for ANOVA → Heterogeneity of Regression. From the Variable Selection Dialogue select the second data option Factor contains categories Data contains values, assign *Group* (*C14*) as [Factor], *Pre* (*C12*) as [X-Axis] and *Post* (*C13*) as [Data]. Selecting only the Test Results output option the following results are obtained:

Heterogeneity of Regression Test

X Axis: Pre, Dependent Variable: Post, classified by Group

Regression	cases	Intercept	Slope	Residual SS	R-DF
1	3	50.3073	0.5917	0.5550	1
2	3	22.8759	0.8759	60.6353	1
3	3	19.1923	0.7821	84.9615	1
Pooled	*****	*****	*****	146.1519	3
Common	9	30.0816	0.7591	149.4387	5
Total	9	17.6851	0.9030	515.6399	7

```
       Null hypothesis: All slopes are equal
                                  F(2,3) =    0.0337
                   Right-Tail Probability =    0.9672
   Null hypothesis: All intercepts are equal
                                  F(2,5) =    6.1263
                   Right-Tail Probability =    0.0452
Null hypothesis: All regressions are identical
                                  F(4,3) =    1.8961
                   Right-Tail Probability =    0.3131
```

Although the null hypotheses “all slopes are equal” is not rejected, “all intercepts are equal” (that the different teaching methods do not have significant effects) should be rejected (since 0.0452 < 0.05).

UNISTAT Statistical Package

Chapter 8
Multivariate Analysis

8.0. Overview

Multivariate analysis is useful when the data consists of various measurements (variables) on the same set of cases. You can determine which cases can be grouped together (Cluster Analysis) or belong to a predetermined group (Discriminant Analysis) or reduce the dimensionality of the data by forming linear combinations of the existing variables (Principal Components Analysis, Factor Analysis and Canonical Correlations). The derived configurations will represent most of the variation in the original data with a smaller number of variables, thus enabling the user to describe the data in a more straightforward manner by graphical or other statistical methods. Therefore, it is important that the user should have immediate access to 2D and 3D graphical representation of results, as well as the ability to save results as data for further analysis.

Central to most multivariate methods is the concept of *proximity*, which can be defined as the relationship between two points in multidimensional space. A proximity measure may reflect *similarity* or *dissimilarity* of the two points. For instance, when we describe the relative positions of two points in terms of the distance between them, we are using a dissimilarity measure. The further apart the two points, the greater their dissimilarity, and when they are identical, the dissimilarity becomes zero. Euclidian distance is just one dissimilarity measure among many others. Using the Multivariate Analysis module you can compute eight proximity measures from the raw data, or enter any square and symmetric matrix for analysis as proximities.

In this chapter an effort will be made to introduce the user to the basic concepts of multivariate analysis. However as it is impossible to *teach* an immense topic such as this in a User's Guide, we recommend that the novice user should consult an introductory text book on the subject. See, for instance, Stevens, J. (1986) or Morrison, D. F. (1990).

Data types: Where relevant, the program will allow you to input raw data or an already formed proximity matrix for a multivariate procedure.

The following table shows which procedures allow what sort of data.

Procedure	Raw Data	Proximities
Cluster Analysis		
Hierarchical	Yes	Yes
K-th Neighbour	Yes	No
K-Means	Yes	No
Discriminant Analysis		
Multiple	Yes	No
K-NN	Yes	No
Multidimensional Scaling	No	Yes
Principal Components	Yes	Yes
Factor Analysis		
Principal Components	Yes	Yes
Principal Axis	Yes	Yes
Canonical Correlations	Yes	No
Reliability Analysis	Yes	No

Table 8.1. Data types accepted by multivariate procedures.

You do not need to take any action to tell the program whether the data to be analysed is a proximity matrix or whether it is raw data. The program will conclude that the data is a proximity matrix if the selected columns form a square and symmetric matrix.

The following proximity matrices can be formed and input for multivariate analysis:

1) Covariance matrix
2) Correlation matrix
3) Euclid
4) Squared Euclid
5) Cosine
6) Chebychev
7) Block
8) Power

The first two matrices can be generated and saved to the Data Processor using the Statistics 1 → Matrix Statistics option. The latter six can be generated using the Cluster Analysis procedure, saved to the Data Processor and then analysed selecting the desired multivariate procedure. It is not necessary to perform a Cluster Analysis to generate proximity matrices.

If the data selected for a multivariate analysis is not a proximity matrix, then the program will permit the formation of a standardised or non standardised proximity matrix from raw data. Standard and non standard proximity matrices are directly proportional to simple (Pearson) correlations matrix and covariance matrix for the same data respectively.

Missing data: If raw data is selected for analysis then any rows containing at least one missing value will be omitted (listwise deletion). Missing data are not allowed in proximity matrices.

Convergence criteria: The first step for almost all multivariate procedures (with the exception of Cluster Analysis) is to compute eigenvalues and eigenvectors of the proximity matrix. The proximity matrix itself is either formed by the procedure from raw data or it is supplied by the user. The core algorithms adopted here are TRED2 which performs a Housholder reduction of the proximity matrix and TQL2 which applies an iterational algorithm (QL) to determine the eigenvalues and eigenvectors (See Smith, B. T. et al, 1976).

Iterations continue until either the reduction in the objective function is less than a given tolerance level, or the maximum number of iterations is reached. A dialogue will allow you to edit these two parameters.

Graphics: All multivariate procedures will offer results in the form of graphical, as well as tabulated numeric output. Each graphics option has its own controls allowing you to annotate and edit the appearance of graphs. The program will display a 2D graph if you select two variables to plot, and a 3D graph if you select three variables. For common graphics controls see 2.3. Graphics Editor.

Saving results for further analysis: In Stand-Alone Mode, it is possible to save all tabulated results to the spreadsheet for further analysis by clicking on the UNISTAT icon situated on the Output Medium Toolbar.

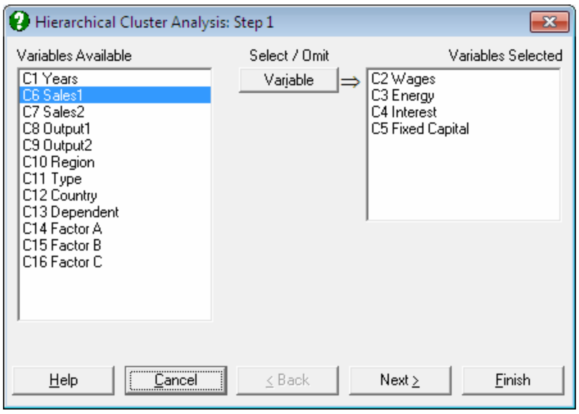
Row labels: In tables and graphs where rows of the data matrix are displayed (e.g. in Cluster Analysis and Discriminant Analysis), rows are referred to by their labels rather than their numbers. Therefore, if the data has no Row Labels, the parts of the table or the graph referring to rows will be left blank. To display row numbers as labels return to the Data Processor, select Edit → Row Labels and then select the Years option. If an initial value of 1 is supplied, then Row Labels will be set to row numbers. Although this may seem extra work on the users part, it makes annotation of output much more flexible.

8.1. Cluster Analysis

Cluster Analysis is used to determine the inherent or natural groupings in data, or provide a convenient summary of data into a given number of groups. There are two main categories of Cluster Analysis, hierarchical and nonhierarchical. The main difference between these two methods is that while the hierarchical method forms clusters sequentially, starting with the most similar pair and forming higher clusters step-by-step, the nonhierarchical methods evaluate the overall distribution of pairs and then classify them into a given number of groups.

Clusters are formed row-wise. If the data is not already in this form, you may use Data Processor's Data → Transpose Matrix utility to obtain the correct format. There is no limitation on the number of cases to be clustered, except for the available memory and hard disk space. But beware: this is an n^3 procedure and you may have to wait a bit (!) if you have thousands of cases. Also, it is not possible to draw character dendrograms for more than 800 cases.

This implementation of Cluster Analysis provides nine hierarchical (Average Between Groups, Average Within Groups, Single Linkage, Complete Linkage, Centroid, Median, Ward, McQuitty, Flexible), one modified hierarchical (K-th neighbour) and one nonhierarchical (K-means) method.



Hierarchical Cluster Analysis: Step 1

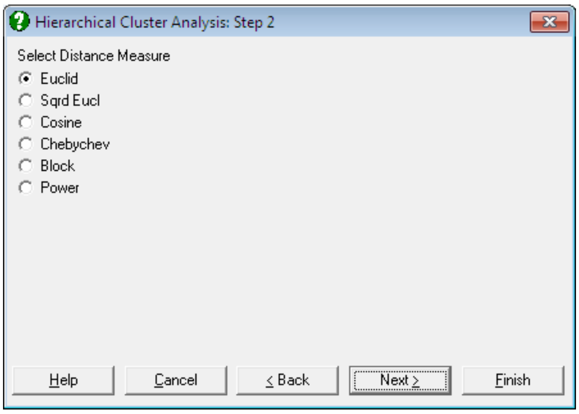
Variables Available:
C1 Years
C6 Sales1
C7 Sales2
C8 Output1
C9 Output2
C10 Region
C11 Type
C12 Country
C13 Dependent
C14 Factor A
C15 Factor B
C16 Factor C

Variables Selected:
C2 Wages
C3 Energy
C4 Interest
C5 Fixed Capital

8.1.1. Hierarchical Cluster Analysis

First, select the data columns to be analysed by clicking on [Variable] from the Variable Selection Dialogue. If the data is not a proximity matrix (if it is not square and symmetric) then another dialogue will appear allowing you to choose from six distance measures. This dialogue will not be available when you input a proximity matrix.

8.1.1.1. Distance Measures


Hierarchical Cluster Analysis: Step 2

Select Distance Measure

● Euclid
○ Sqrd Eucl
○ Cosine
○ Chebychev
○ Block
○ Power

Euclid:

$$\text{Distance}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Squared Euclid:

$$\text{Distance}(x, y) = \sum_i (x_i - y_i)^2$$

Cosine:

$$\text{Distance}(x, y) = \frac{\sum_i (x_i y_i)}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

Chebychev:

$$\text{Distance}(x, y) = \text{Max}_i \mid x_i - y_i \mid$$

Block:

$$\text{Distance}(x,y) = \sum_i | x_i - y_i |$$

Power:

$$\text{Distance}(x,y) = \left(\sum_i | x_i - y_i |^p\right)^{1/r}$$

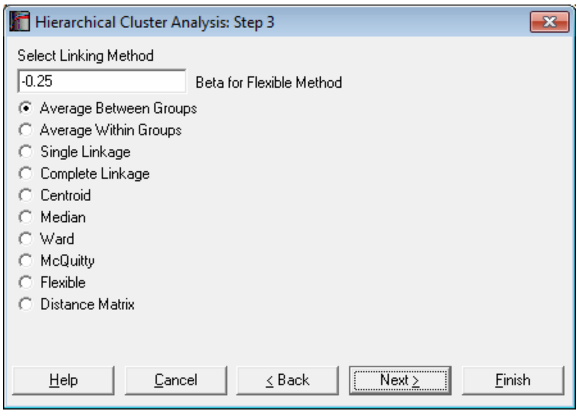
where the power terms p and r are supplied by the user.

8.1.1.2. Distance Matrix

After the distance matrix is computed, a dialogue containing nine hierarchical clustering methods and a Distance Matrix option will appear.

It is possible to select one of the methods and proceed immediately with the analysis, or select the last option to view or save the generated distance matrix. The Distance Matrix option will not be available when you input a proximity matrix for analysis.

8.1.1.3. Hierarchical Methods



Hierarchical Cluster Analysis: Step 3

Select Linking Method

-0.25 Beta for Flexible Method

● Average Between Groups
○ Average Within Groups
○ Single Linkage
○ Complete Linkage
○ Centroid
○ Median
○ Ward
○ McQuitty
○ Flexible
○ Distance Matrix

All hierarchical methods apply the same algorithm. However, they differ in the way they compute the distance between two clusters.

First, the n(n - 1)/2 elements of the proximity matrix are sorted in ascending order. The nearest two points are joined to form the first cluster. At the ith step the remaining points and the existing clusters are considered. Either the next two nearest points, or a cluster and a point, or two clusters are formed into a new cluster. This process is repeated until the number of clusters is reduced to one.

One of the following nine hierarchical clustering methods can be selected, where d_{ij} is the dissimilarity between clusters i and j, $n_i = 1$, $i = 1, ..., n$ is a unity vector, $S_i = 0$, $i = 1, ..., n$ is a zero vector and indices t and r represent a new cluster and all other clusters respectively.

Average Between Groups:

Compute an unweighted average distance between pairs belonging to two clusters. Update:

$$d_{tr} = d_{pr} + d_{qr}$$

$$n_t = n_p + n_q$$

and select the minimum of:

$d_{ij}/(n_i + n_j)$.

Average Within Groups:

Update:

$$d_{tr} = d_{pr} + d_{qr}$$

$$n_t = n_p + n_q$$

$$S_i = S_p + S_q + d_{pq}$$

$$d_{tr} = \frac{S_i + S_j + d_{ij}}{(n_i + n_j)(n_i + n_j - 1)/2}$$

Single Linkage:

Select the smallest distance between pairs of elements in each cluster. Update:

$d_{tr} = \min(d_{pr}, d_{qr})$.

Complete Linkage:

Select the largest distance between pairs of elements in each cluster. Update:

$d_{tr} = \max(d_{pr}, d_{qr})$.

Centroid:

A cluster's location is represented by the centroid of all points within the cluster. Update:

$$d_{tr} = \frac{n_p}{n_p + n_q} d_{pr} + \frac{n_q}{n_p + n_q} d_{qr} - \frac{n_p n_q}{(n_p + n_q)^2} d_{pq}$$

This method should be used only with squared Euclid distance.

Median:

Compute the weighted average distance between pairs belonging to two clusters. Update:

$$d_{tr} = (d_{pr} + d_{qr})/2 - d_{pq}/4$$

This method should be used only with squared Euclid distance.

Ward:

This method is also known as incremental sum of squares. Unlike other methods which minimise the distance between two clusters, the Ward's method minimises the increase in total within-cluster sum of squares of the newly formed cluster. The distance between the two clusters is given as:

$$d_{tr} = \frac{n_p + n_r}{n_T} d_{pr} + \frac{n_q + n_r}{n_T} d_{qr} - \frac{n_r}{n_T} d_{pq}$$

$$n_T = n_p + n_q + n_r$$

where n_r is the number of observations within the current cluster. This method should be used only with squared Euclid distance.

McQuitty:

The distance between the two clusters is calculated as:

$$d_{tr} = (d_{pr} + d_{qr})/2$$

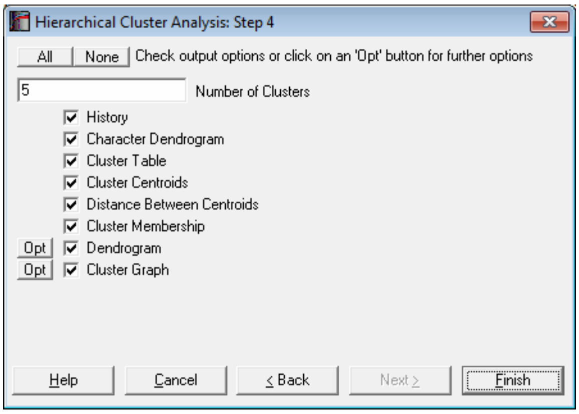
Flexible:

Update:

$$d_{tr} = (d_{pr} + d_{qr})(1 - \beta)/2 + d_{pq}\beta$$

where ß is a constant supplied by the user. The default value for ß is -0.25.

8.1.1.4. Hierarchical Cluster Output Options


Hierarchical Cluster Analysis: Step 4

Check output options or click on an 'Opt' button for further options

5 Number of Clusters

☑ History
☑ Character Dendrogram
☑ Cluster Table
☑ Cluster Centroids
☑ Distance Between Centroids
☑ Cluster Membership
☑ Dendrogram
☑ Cluster Graph

If there are n valid cases in data, the program will start with n clusters and combine them one-by-one until there is only one cluster is left. The History output option will summarise the clustering steps and the two dendrogram diagrams will show this entire process. The remaining output options depend on the Number of Clusters parameter defined by the user.

Number of Clusters: By entering a number between 1 and n, it is possible to display clustering results for any number of clusters. This number can also be changed from within the Cluster Graph, by selecting the Edit → XY Points dialogue.

History: This table shows the two clusters combined at each step, the number of cases in the new cluster and the distance between them. The newly formed cluster is given the label of the cluster in the left hand column.

Character Dendrogram: A dendrogram displays a visual summary of the clustering process, providing you with an understanding of the groups and proximities inherent in data. The order in which clusters are combined does not necessarily coincide with the order they are drawn on a dendrogram. The dendrogram procedure first rearranges the History table to produce an uncluttered tree diagram. The same tree structure can also be output in the form of a graph.

The advantage of this form of output is in its ability to display all Row Labels without any cluttering. However, due its low resolution on the (horizontal) distance axis, some of the clusters which are too close to each other may not be distinguished.

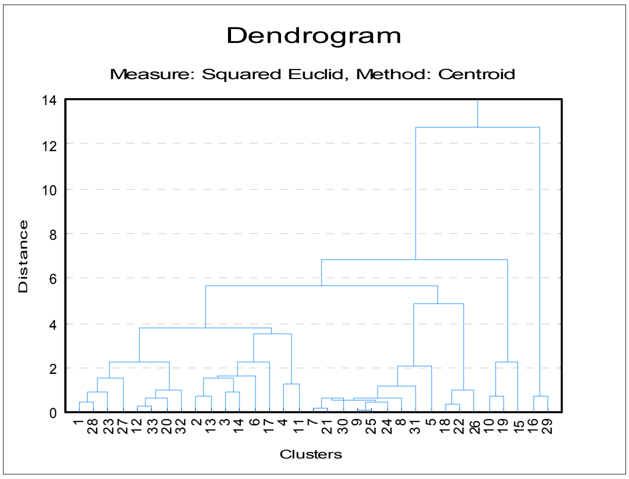
Cluster Table: The number of cases and their percentages are displayed for the number of clusters defined by the user. The within cluster sum of squares, average, minimum and maximum distance of individual cases from their cluster's centroid are also displayed.

Cluster Centroids: For each cluster, the coordinates of the cluster centroid are displayed.

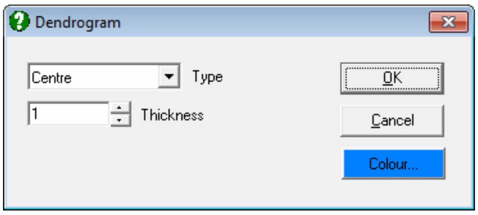
Distance Between Centroids: Distance between each pair of cluster centroids is displayed in a square-symmetric table.

Cluster Membership: A table containing all cases displays which case belongs to which cluster. As in the Cluster Table option, the number of clusters to be formed can be selected by the user.

Hi-res Dendrogram: The high-resolution dendrogram is convenient when the number of rows in the data set does not exceed 100. The vertical axis represents the distance and the horizontal axis represents the clusters combined.



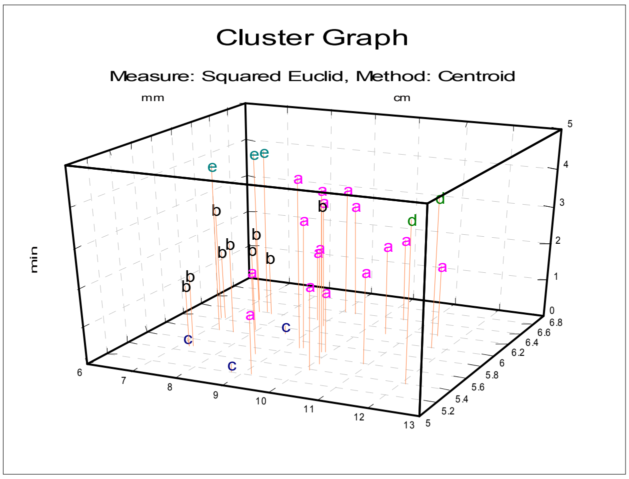
The Edit → XY Points dialogue for the Hi-Res Dendrogram procedure enables you to change the colour and thickness of lines, as well as positions of the stems (the vertical lines representing the newly formed clusters). Stems can be started from the midpoint (the default), the right or the left corner of the line connecting the two old clusters.



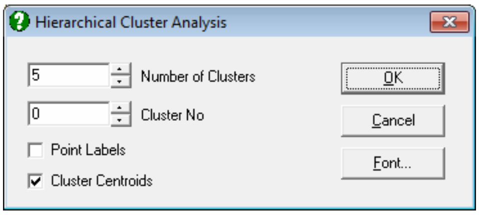
Type: Centre
Thickness: 1

By default, the row numbers are displayed as the X-axis labels. It is also possible to display the Row Labels as X-axis labels, from the Edit → Axes dialogue. If the Row Labels are too long, you can display them up and down or rotate the text by 90º or 270º.

Cluster Graph: Two and three-dimensional scatter diagrams can be displayed showing which data point belongs to which cluster. If you select two variables a 2D graph is displayed and a 3D graph if you select three variables. Different clusters are represented by different letters in different colours.



You can change the number of clusters to be displayed from the Edit → XY Points dialogue, without having to go back to the Output Options Dialogue. It is possible to select the font and the size of the letters and display point labels for them. You can also display cluster centroids in capital letters.



Hierarchical Cluster Analysis

5 Number of Clusters

0 Cluster No

□ Point Labels

☑ Cluster Centroids

If the Cluster No field is zero, all groups will be displayed simultaneously. If this field is set to any other number less than or equal to the Number of Clusters, then only the cases belonging to that cluster will be displayed.

8.1.1.5. Hierarchical Cluster Example

Open MULTIVAR, select Statistics 2 → Cluster Analysis → Hierarchical Cluster Analysis and select *Perf*, *Info*, *Verbexp* and *Age* (*C1* to *C4*) as [Variable]s. Select distance measure as Euclid and linking method as Average Between Groups. Select number of clusters as 3 and all the output options to obtain the following results:

Hierarchical Cluster Analysis

Variables Selected: Perf, Info, Verbexp, Age
Measure: Euclid, Method: Average Between Groups

History

Step	Combined1	Combined2	Cases	Distance
1	1	8	2	4.6915
2	2	9	2	9.4345
3	1	5	3	9.5967
4	1	6	4	10.8672
5	3	4	2	12.5714
6	1	2	6	16.1606
7	3	7	3	19.2953
8	1	3	9	26.9553

Character Dendrogram

```
1+----------+
8+----------+-----------+
5+----------------------+--+
6+-------------------------+-----------+
2+---------------------+               |
9+---------------------+---------------+-------------------------+
3+-----------------------------+                                 |
4+-----------------------------+---------------+                 |
7+---------------------------------------------+-----------------+
```

Cluster Table

	Cases	Percentage	Within SSQ	Average Distance	Minimum	Maximum
Cluster 1	6	66.7%	479.6467	8.2992	3.4567	12.2459
Cluster 2	2	22.2%	79.0200	6.2857	6.2857	6.2857
Cluster 3	1	11.1%	0.0000	0.0000	0.0000	0.0000

Cluster Centroids

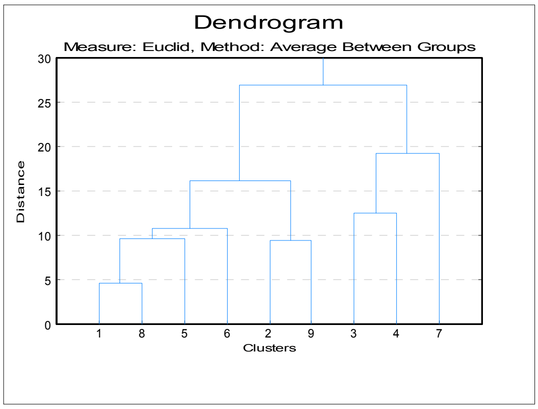
	Cluster 1	Cluster 2	Cluster 3	Overall
Perf	88.1667	107.0000	120.0000	95.8889
Info	8.0000	12.5000	12.0000	9.4444
Verbexp	32.0000	43.5000	30.0000	34.3333
Age	7.1667	7.1000	8.4000	7.2889

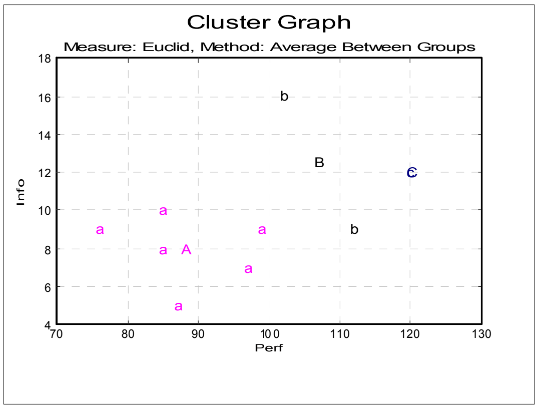
Distance Between Centroids

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0.0000	22.5211	32.1696
Cluster 2	22.5211	0.0000	18.7933
Cluster 3	32.1696	18.7933	0.0000

Cluster Membership

Observation	Cluster
1	1
2	1
3	2
4	2
5	1
6	1
7	3
8	1
9	1


Dendrogram
Measure: Euclid, Method: Average Between Groups
Distance
0
5
10
15
20
25
30
1
8
5
6
2
9
3
4
7
Clusters

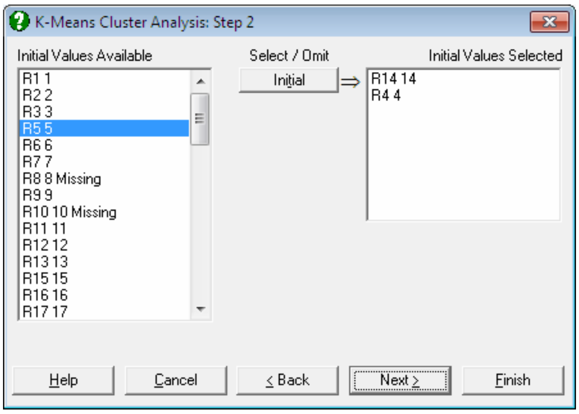

Cluster Graph
Measure: Euclid, Method: Average Between Groups
Info
4
6
8
10
12
14
16
18
70
80
90
100
110
120
130
Perf
b
B
C
a
a
a
b
a
A
a
a

8.1.2. K-th Neighbour Cluster Analysis

K-th Neighbour clustering is also a hierarchical method, however, the basic algorithm is modified to overcome the inherent problem of noise associated with Single Linkage or Complete Linkage methods. See Wong, M A and Lane, T A (1983), pp. 362-8. Instead of linking a point to the nearest (or farthest) point, this method will link a point to its kth neighbour, where k is supplied by the user. However, since by symmetry this principle is also applied to the point to be linked, the distance from each point to its kth neighbour is computed. If the actual distance between the two points is greater than the sum total of their distances to their own kth neighbours, then the distance between the two points is assumed to be infinity (a very large number). Otherwise, the distance is the average of distances to the kth nearest points.

Output options for this procedure are the same as for other hierarchical methods.

8.1.3. K-Means Cluster Analysis



Initial Values Available

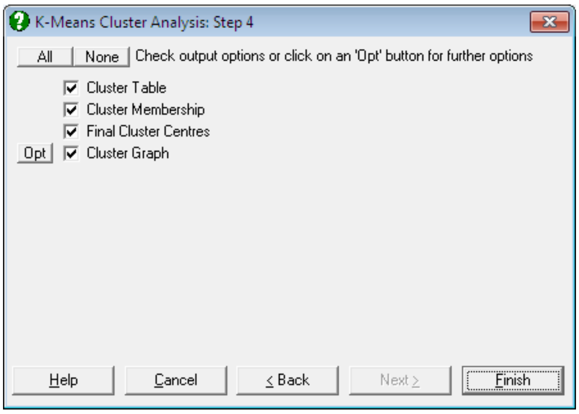
R1 1
R2 2
R3 3
R5 5
R6 6
R7 7
R8 8 Missing
R9 9
R10 10 Missing
R11 11
R12 12
R13 13
R15 15
R16 16
R17 17

Initial Values Selected

R14 14
R4 4

This is a k-means algorithm to divide M points in N dimensions into K clusters. The user selects K initial points from the rows of the data matrix. The procedure applies an iterational algorithm which minimises the within-cluster sum of squares. See Hartigan, J. A. and Wong, M. A. (1979), p. 100.

The following output options are provided:



Check output options or click on an 'Opt' button for further options

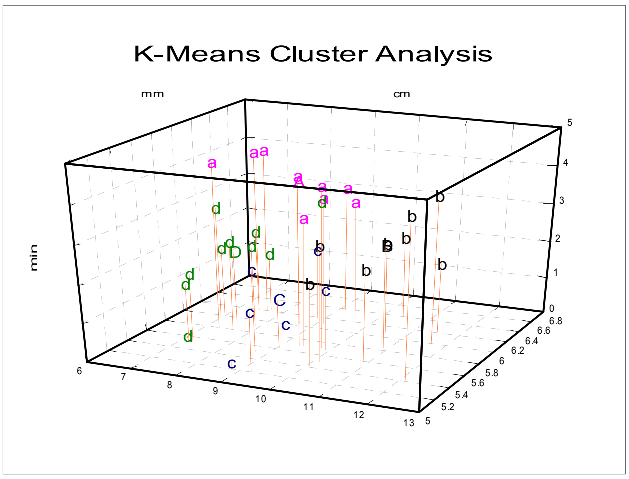
☑ Cluster Table
☑ Cluster Membership
☑ Final Cluster Centres
☑ Cluster Graph

Cluster Table: The number of cases in each cluster, their percentages and the minimised sum of squares are displayed. The number of clusters formed is determined by the number of initial points selected.

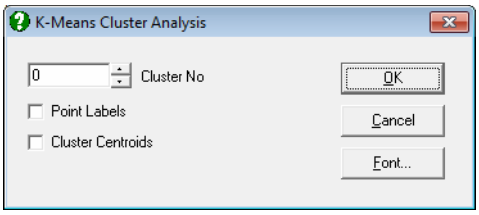
Cluster Membership: This is similar to the membership table for hierarchical methods. The number of the cluster which includes the case is displayed.

Final Cluster Centres: The k-means clustering algorithm computes centroids for each cluster. The final configuration is displayed in a table.

Cluster Graph: This is similar to the Cluster Graph for hierarchical methods.



However, here it is also possible to display the cluster centroids on the same graph, using the Edit → XY Points dialogue. A cluster centroid will be represented by a capital letter. Unlike the hierarchical methods, here the number of clusters cannot be changed, because it is fixed by the number of seeds selected at the start of the analysis.


K-Means Cluster Analysis

0 Cluster No

☐ Point Labels

☐ Cluster Centroids

If the Cluster No field is zero, all groups will be displayed simultaneously. If this field is set to any other number less than or equal to the number of clusters, then only the cases belonging to that cluster will be displayed.

Example

Open MULTIVAR, select Statistics 2 → Cluster Analysis → K-Means Cluster Analysis, and select *Perf*, *Info*, *Verbexp* and *Age* (*C1* to *C4*) as [Variable]s. Select *R2*, *R4* and *R8* as seeds at the next dialogue and accept the default number of maximum iterations to obtain the following results:

K-Means Cluster Analysis

Variables Selected: Perf, Info, Verbexp, Age

Cluster Table

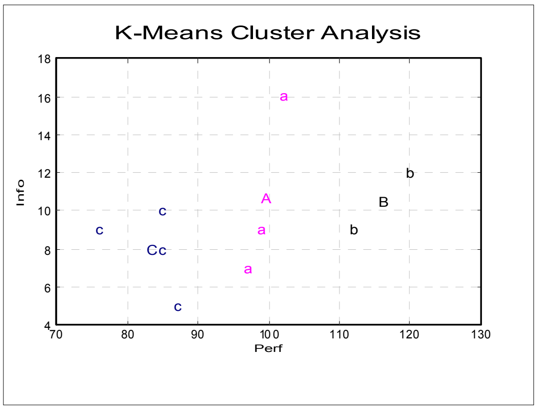
Cluster	Seed	Cases	Percentage	SSQ
1	2	3	33.33%	220.3800
2	4	2	22.22%	109.2200
3	8	4	44.44%	140.7875

Cluster Membership

Observation	Cluster
1	3
2	1
3	2
4	1
5	3
6	3
7	2
8	3
9	1

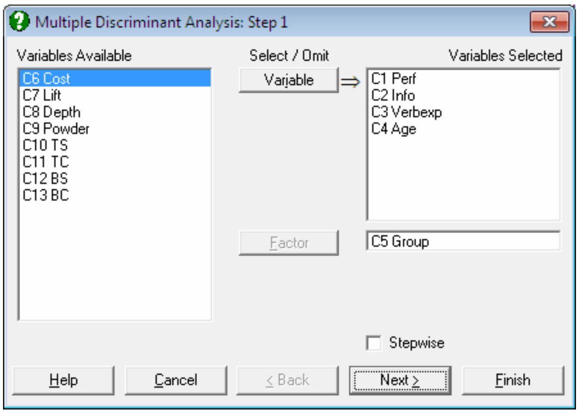
Final Cluster Centres

Seed	Perf	Info	Verbexp	Age
2	99.3333	10.6667	36.0000	7.8333
4	116.0000	10.5000	36.0000	7.8000
8	83.2500	8.0000	32.2500	6.6250


K-Means Cluster Analysis
Info
Perf
18
16
14
12
10
8
6
4
70
80
90
100
110
120
130
a
b
A
B
c
a
b
Cc
a
c

8.2. Discriminant Analysis

Discriminant Analysis is used to determine whether a given classification of cases into a number of groups is an appropriate one. The Discriminant Analysis can be used, for instance, to test whether a particular clustering of cases obtained from a Cluster Analysis is likely. It will report whether the group assignment of a case is true or false, as well as reporting the probability of the case belonging to a particular group.


Multiple Discriminant Analysis: Step 1

Variables Available: C6 Cost, C7 Lift, C8 Depth, C9 Powder, C10 TS, C11 TC, C12 BS, C13 BC

Variables Selected: C1 Perf, C2 Info, C3 Verbexp, C4 Age

Factor: C5 Group

□ Stepwise

The variables to be analysed are selected from the data matrix by clicking on [Variable]. A factor column containing the group assignments of cases must also be selected by clicking on [Factor]. This is typically a string variable or numeric variable containing integers. The program will not proceed unless a factor column is selected.

Two types of Discriminant Analysis are provided: Multiple (including Linear and Canonical Discriminant Functions, which can be stepwise) and nonparametric Kth neighbour (also known as K-NN) discriminant analyses.

Multicollinearity: Existence of a solution depends on the number of degrees of freedom available in data. If there are insufficient degrees of freedom, the program reports Warning: Singular matrix and does not proceed further. The most common cause of a singular matrix is the number of cases (rows) in raw data being less than the number of variables selected for analysis.

Predictions: The estimated discriminant functions can be applied to test cases to predict their group membership. If, for a case (row), all independent variables are non-missing, but only the factor (group) variable is missing, then it is treated as a test case. Such cases are not included in the estimation of discriminant functions, but the estimated coefficients are applied to them. The predicted cases are represented in all plots (by an @ character) and in all relevant tables (by an * character).

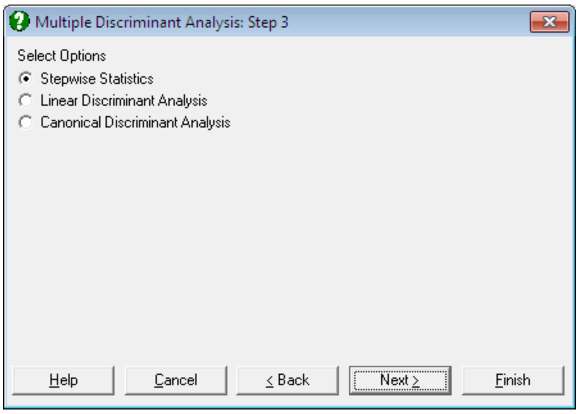
It is possible to use markers other than missing data to designate cases as test cases. Suppose, for instance, you wish the program to interpret cases with –1 in their group variable as test cases. To do this, enter the following line in the [Options] section of *Documents\Unistat65\Unistat65.ini* file:

```
DiscrPredict=-1
```

If the group variable is a string variable, you can use a string value as a test case marker.

8.2.1. Multiple Discriminant Analysis

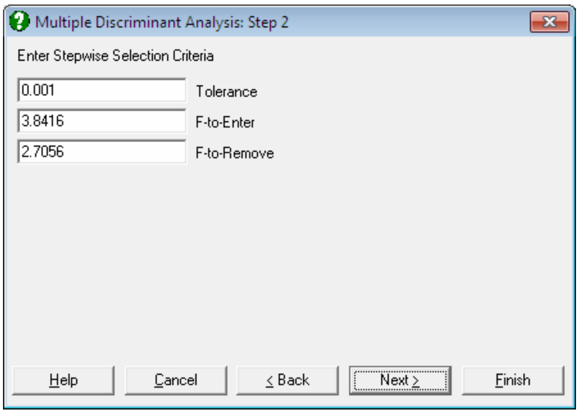
Linear and Canonical discriminant analyses can be performed with or without stepwise selection of variables. A Linear Discriminant Analysis should be performed before a Canonical one. The program will do this automatically, even if only the Canonical option is selected. It is possible to output Stepwise Statistics, Linear and Canonical analysis results separately.


Multiple Discriminant Analysis: Step 3

Select Options
● Stepwise Statistics
○ Linear Discriminant Analysis
○ Canonical Discriminant Analysis

8.2.1.1. Stepwise Discriminant Analysis

The Stepwise check box provided on the Variable Selection Dialogue enables you to select the best subset of variables to run a Discriminant Analysis. When this box is checked, the variables selected for analysis are ranked according to their influence on the final result. They are also tested against a user-defined criterion of eligibility. The variable with the highest influence that passes the test of eligibility is then included in the analysis. At each step, the already selected variables are also tested against an exclusion criterion and they may be excluded from the analysis if they fail to satisfy this criterion. The steps are repeated until there are no variables that can be entered or removed from the analysis.


Multiple Discriminant Analysis: Step 2

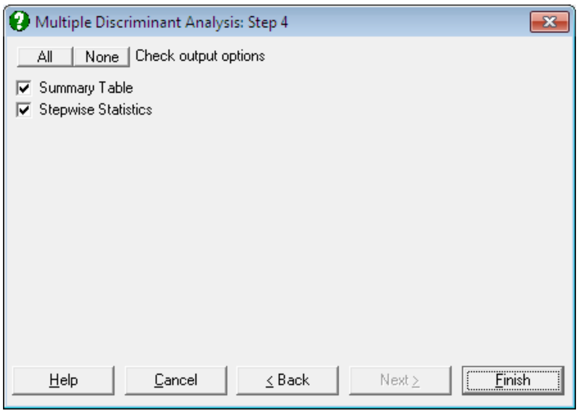
Enter Stepwise Selection Criteria

0.001 Tolerance

3.8416 F-to-Enter

2.7056 F-to-Remove

The stepwise selection method used in Stepwise Discriminant Analysis is almost identical to the one employed in Stepwise Regression. For further information on this method and the interpretation of F-to-enter and F-to-remove statistics see 7.2.3.1. Stepwise Selection Criteria. The following output options are available:


Multiple Discriminant Analysis: Step 4

Check output options

☑ Summary Table

☑ Stepwise Statistics

Summary Table: The variable entered or removed at each step, its F-to-enter or F-to-remove value, its tail probability and Wilks' lambda statistic are displayed.

Stepwise Statistics: First, within group sums of squares and the cross product matrix are computed:

$$w_{ij} = \sum_j^g \sum_k^{mj} x_{ijk} x_{ijk} - \sum_j^g (\sum_k^{mj} x_{ijk})(\sum_j^{mj} x_{ijk})/n_j$$

where g is the number of groups, p is the number of variables, x_{ijk} is the value of variable i for case k in group j and m_j is the number of cases in group j. Define $[w^*_{ij}]$ as the new matrix after a new variable is entered or omitted. Then:

Tolerance$_i$ = 0 if $w_{ij} = 0$

Tolerance$_i$ = w^*_{ii} / w_{ii} if variable i is not in the analysis,

Tolerance$_i$ = $-1/(w^*_{ii} w_{ii})$ if variable i is in the analysis.

$$\text{F-to-remove}_i = \frac{(w^*_{ii} - t^*_{ii})(n-p-g+1)}{t^*_{ii}(g-1)}$$

with degrees of freedom (g – 1) and (n – p – g + 1),

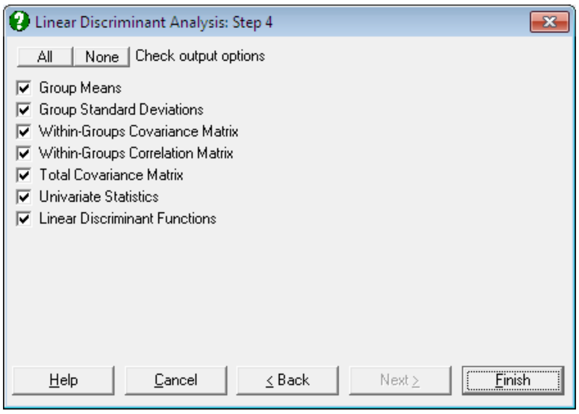
$$\text{F-to-enter}_i = \frac{(t^*_{ii} - w^*_{ii})(n-p-g)}{w^*_{ii}(g-1)}$$

with degrees of freedom (g – 1) and (n – p – g),

and Wilks’ Lambda is:

$$\Lambda = \frac{|W|}{|T|}$$

8.2.1.2. Linear Discriminant Analysis



Linear Discriminant Analysis: Step 4

All | None | Check output options

☑ Group Means
☑ Group Standard Deviations
☑ Within-Groups Covariance Matrix
☑ Within-Groups Correlation Matrix
☑ Total Covariance Matrix
☑ Univariate Statistics
☑ Linear Discriminant Functions

Group Means and Standard Deviations: The means and standard deviations of sub-groups defined by the factor column are tabulated.

Pooled Within-Groups Covariance Matrix:

$[w_{il}/(n\text{-}g)]$

Pooled Within-Groups Correlation Matrix:

$[w_{il}/Sqr(w_{ii}w_{ll})]$.

Total Covariance Matrix: First compute the total sums of squares and cross product matrix:

$$t_{il} = \sum_j^g \sum_k^{mj} x_{ijk} x_{ljk} - (\sum_j^g \sum_k^{mj} x_{ijk})(\sum_j^g \sum_j^{mj} x_{ljk})/n$$

The total covariance matrix is $[t_{il}/(n - 1)]$.

Univariate Statistics: Wilks' lambda is:

$$\Lambda = \frac{w_{ii}}{t_{ii}}$$

and the F-statistic, which has g - 1 and n - g degrees of freedom, is:

$$F_i = \frac{(t_{ii} - w_{ii})(n - g)}{W_{ii}(g - 1)}$$

Linear Discriminant Functions: These are also known as Fisher's Linear Discriminant Functions. The coefficients can be saved to the data matrix and subsequently used to classify cases. Since canonical discrimination is a superior method, classifications are made here in the second level Output Options Dialogue, using the Canonical Discrimination Functions.

Coefficients of the Linear Discriminant Functions are:

$$b_{ij} = (n - g)\sum_t^q w_{il}\overline{x_{ij}}$$

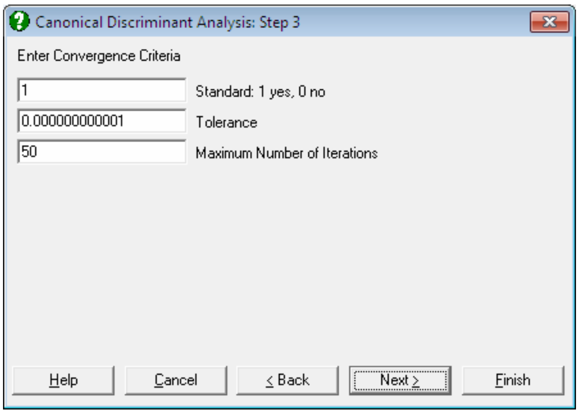
where i = 1, ..., p and j = 1, ..., g and the constant terms are:

$$a_j = Log(P_j) - 1/2\sum_i^q b_{ij}\overline{x_{ij}}$$

where $p_j = n_j/n$.

8.2.1.3. Canonical Discriminant Analysis

The Canonical Discriminant Analysis is based on the eigenvectors and eigenvalues of the proximity matrix and thus it involves an iterational algorithm. Iterations continue until either the reduction in the objective function is less than a given tolerance level, or the maximum number of iterations is reached.


Canonical Discriminant Analysis: Step 3

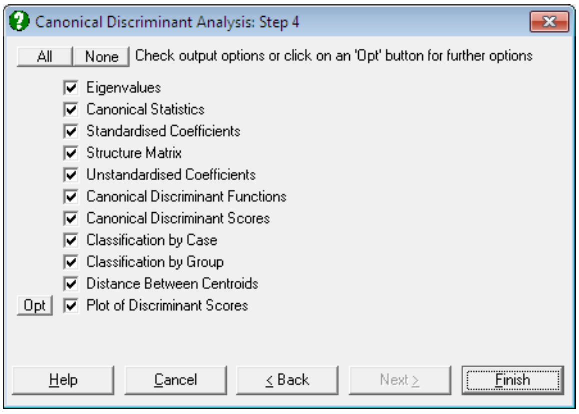
Enter Convergence Criteria

1 Standard: 1 yes, 0 no

0.000000000001 Tolerance

50 Maximum Number of Iterations

A dialogue allows editing two or three parameters, depending on whether the data is raw or it is already formed into a proximity matrix (i.e. it is square and symmetric). In the former case, the program will allow the choice of forming a standardised (the default) or non standardised proximity matrix.


Canonical Discriminant Analysis: Step 4

Check output options or click on an 'Opt' button for further options

☑ Eigenvalues
☑ Canonical Statistics
☑ Standardised Coefficients
☑ Structure Matrix
☑ Unstandardised Coefficients
☑ Canonical Discriminant Functions
☑ Canonical Discriminant Scores
☑ Classification by Case
☑ Classification by Group
☑ Distance Between Centroids
☑ Plot of Discriminant Scores

The eigenvalues and eigenvectors of the system are found using the Cholesky decomposition.

The number of Canonical Discriminant Functions extracted (f) depends on the number of variables and the number of groups:

$$f = \text{Min}(p, g - 1)$$

where p is the number of variables and g is the number of groups. The output options include the following:

Eigenvalues: Canonical Correlations are found as:

$$r_k = \sqrt{\frac{\lambda_k}{(1+\lambda_k)}}$$

Canonical Statistics: Wilks' lambda is used to test the significance of all the discriminating functions after the first k:

$$\Lambda = \prod_{i=k+1}^{m} \frac{1}{1+\lambda_i} \quad k = 0,1,\ldots, m - 1$$

The tail probability for lambda is determined from chi-square distribution:

$$x^2 = -(n - \frac{q+g}{2} - 1)\text{Log}\Lambda_k$$

with (p - k)(g - k - 1) degrees of freedom.

Canonical Discriminant Coefficients: Standardised coefficients matrix **D** is obtained by multiplying the square roots of $[w_{ii}]$ by the corresponding eigenvectors. The unstandardised coefficients are:

$$B = \sqrt{n-g}\, S_{11}^{-1} D$$

and the constants are:

$$a_k = -\sum_i^q b_{ik} \overline{x_i}$$

Canonical Discriminant Scores: The unstandardised coefficients matrix is multiplied by the data matrix. Discriminant scores can be displayed in tabular form and saved to the Data Processor for further analysis. They can also be displayed in 2D and 3D scatter diagrams with group memberships and group centroids. Centroids are the Canonical Discriminant Functions evaluated at the group means.

$$\overline{f_{jk}} = a_k + \sum_1^q b_{ik} \overline{x_{ij}}$$

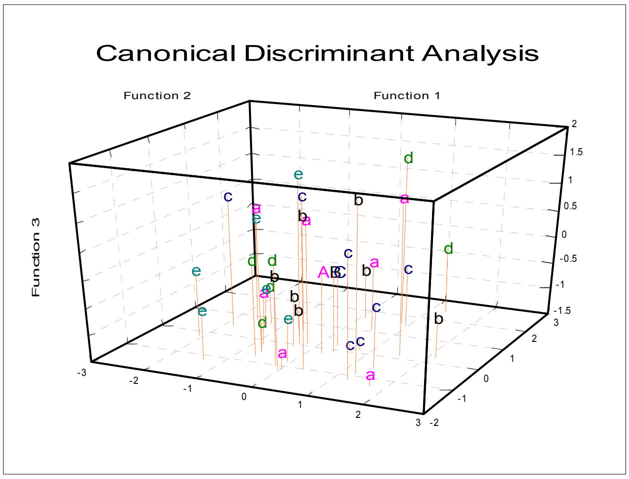
Classification by Case: For each case, the chi-square distances from all centroids are computed. The probability of the case being a member of a group is the tail probability for this chi-square value with m degrees of freedom. A case is classified into the group for which this probability is the highest.

The table displays the given group membership, the highest probability (estimated) group, and the probability of the case belonging to the estimated group. If the estimated and actual groups differ, two stars (representing a misclassification) are printed between the two columns.

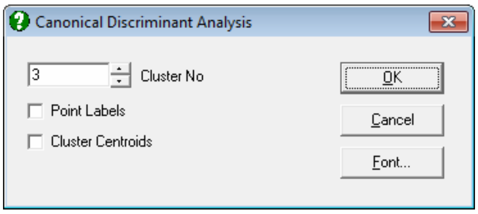
Classification by Group: This is an alternative way of presenting the classification results explained above. A table is formed with rows and columns corresponding to actual and estimated group memberships respectively. Each cell displays the number of elements and their percentage. Diagonal elements are the cases classified correctly and the off-diagonal elements are misclassified cases.

Distance Between Centroids: The distance between every pair of centroids are tabulated in ascending order.

Plot of Discriminant Scores: This provides options to display group centroids (which are represented by capital letters) and groups selectively. If you select two variables a 2D graph is displayed and a 3D graph if you select three variables.



You can choose to display group centroids in capital letters, point labels or change the font of group letters from the Edit → XY Points dialogue. When the Cluster No field is zero all groups will be displayed simultaneously. If this field is set to one, then the first group only, if two, the second group only, etc., will be displayed.



3 Cluster No

□ Point Labels

□ Cluster Centroids

8.2.1.4. Discriminant Examples

Example 1

Table 11.1 on p. 513 from Tabachnick, B. G. & L. S. Fidell (1989). Measurements on four predictors, performance, information, verbal expression and age, are given on three groups of children. In order to keep track of group memberships of children we should add a factor column *Group* to the data matrix. We would like to find out whether the children have been correctly classified into three groups. We would also like to be able to determine the group membership of a child who was not included in the study, but for whom we have measurements on predictor variables.

Open MULTIVAR, select Statistics 2 → Discriminant Analysis → Multiple Discriminant Analysis and select *Perf*, *Info*, *Verbexp*, *Age* (*C1* to *C4*) as [Variable]s and *Group* (*C5*) as [Factor]. Leave the Stepwise box unchecked. The analysis has two stages; first a Linear Discriminant Analysis is performed and its output options are presented alongside the option for the second stage; Canonical Discriminant Analysis. Select all output options in both stages to obtain the following results:

Linear Discriminant Analysis

Group Means

	Perf	Info	Verbexp	Age
Group 1	98.6667	7.0000	36.3333	7.3000
Group 2	87.6667	11.6667	38.3333	6.9333
Group 3	101.3333	9.6667	28.3333	7.6333

Group Standard Deviations

	Perf	Info	Verbexp	Age
Group 1	12.5831	2.0000	5.5076	0.9539
Group 2	13.2035	3.7859	6.5064	0.7024
Group 3	17.6163	2.0817	1.5275	1.1590

Within-Groups Covariance Matrix

	Perf	Info	Verbexp	Age
Perf	214.3333	36.6667	58.0556	8.3333
Info	36.6667	7.5556	12.2778	1.0611
Verbexp	58.0556	12.2778	25.0000	1.6222
Age	8.3333	1.0611	1.6222	0.9156

Within-Groups Correlation Matrix

	Perf	Info	Verbexp	Age
Perf	1.0000	0.9112	0.7931	0.5949
Info	0.9112	1.0000	0.8933	0.4034
Verbexp	0.7931	0.8933	1.0000	0.3391
Age	0.5949	0.4034	0.3391	1.0000

Total Covariance Matrix

	Perf	Info	Verbexp	Age
Perf	200.1111	18.5556	21.0417	8.0611
Info	18.5556	9.7778	10.2083	0.5181
Verbexp	21.0417	10.2083	39.7500	-0.0833
Age	8.0611	0.5181	-0.0833	0.7786

Univariate Statistics

	Lambda	F-statistic	Probability
Perf	0.8033	0.7346	0.5184
Info	0.5795	2.1765	0.1947
Verbexp	0.4717	3.3600	0.1050
Age	0.8819	0.4017	0.6859

Linear Discriminant Functions

	Group 1	Group 2	Group 3
Perf	1.9242	0.5870	1.3655
Info	-17.5622	-8.6992	-10.5870
Verbexp	5.5459	4.1168	2.9728
Age	0.9872	5.0175	2.9114
Constant	-138.9111	-72.3844	-72.3403

Canonical Discriminant Analysis

Eigenvalues

	Eigenvalue	Percent	Cumulative	Correlation
1	13.4859	70.7%	70.7%	0.9649
2	5.5892	29.3%	100.0%	0.9210

Canonical Statistics

	Wilks' lambda	Chi-square	Deg Fre	Probability
0	0.0105	20.5138	8	0.0086
1	0.1518	8.4845	3	0.0370

Standardised Coefficients

	Function 1	Function 2
Perf	-2.5035	-1.4741
Info	3.4896	-0.2838
Verbexp	-1.3247	1.7888
Age	0.5027	0.2362

Structure Matrix

	Function 1	Function 2
Perf	-0.0755	-0.1734
Info	0.2280	0.0664
Verbexp	-0.0223	0.4463
Age	-0.0279	-0.1486

Unstandardised Coefficients

	Function 1	Function 2
Perf	-0.1710	-0.1007
Info	1.2695	-0.1032
Verbexp	-0.2649	0.3578
Age	0.5254	0.2469
Constant	9.6737	-3.4529

Canonical Discriminant Functions

	Function 1	Function 2
Group 1	-4.1023	0.6910
Group 2	2.9807	1.9417
Group 3	1.1217	-2.6327

Canonical Discriminant Scores

	Function 1	Function 2
1	-3.7063	-0.0581
2	-3.2036	0.9864
3	-5.3972	1.1446
4	4.2997	2.4526
5	1.7594	2.4276
6	2.8829	0.9448
7	0.8531	-3.9675
8	1.1866	-1.2645
9	1.3253	-2.6660

Classification by Case

	ActGroup	EstGroup	Probability
1	1	1	0.6984
2	1	1	0.6392
3	1	1	0.3902
4	2	2	0.3677
5	2	2	0.4215
6	2	2	0.6055
7	3	3	0.3958
8	3	3	0.3914
9	3	3	0.9789

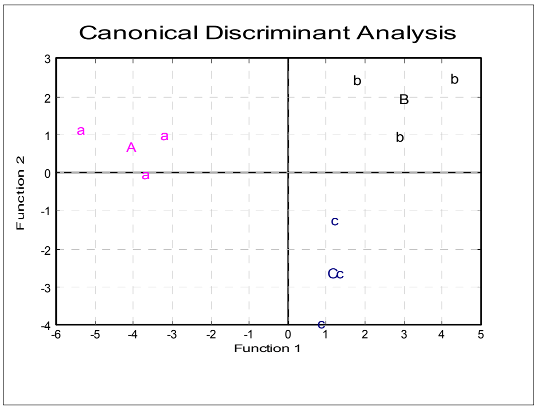
Classification by Group

	Group 1	Group 2	Group 3
Group 1	3	0	0
Group 2	0	3	0
Group 3	0	0	3

Correctly classified: 100.00%

Distance Between Centroids

Clusters	Distance
B - C	4.9377
A - C	6.1917
A - B	7.1926



Example 2

Open STEPDSCR, select Statistics 2 → Discriminant Analysis → Multiple Discriminant Analysis and select *Var1*, to *Var7* (*C1* to *C7*) as [Variable]s, *Groups* (*C8*) as [Factor] and check the Stepwise box. On the next dialogue, accept the default values of Tolerance, F-to-Enter and F-to-Remove. Next select Stepwise Statistics and leave both output options checked. The output below is abbreviated:

Multiple Discriminant Analysis: Stepwise Statistics

Summary Table

Tolerance: 0.001
F-to-Enter: 3.8416 (5.0%)
F-to-Remove: 2.7056 (10.0%)

Variable	Entered at Step	F-to-Enter	Probability	Wilks' Lambda
Var3	1	42.2648	0.0000	0.6167
Var5	2	9.1029	0.0036	0.5429
Var7	3	7.7673	0.0070	0.4858
Var6	4	10.7627	0.0017	0.4168

Step 0

Variable	Entered at Step	Tolerance	F-to-Enter/Remove	Wilks' Lambda
Var1		1.0000	4.0471	0.9438
Var2		1.0000	0.7221	0.9895
Var3		1.0000	42.2648	0.6167
Var4		1.0000	0.1175	0.9983
Var5		1.0000	24.2906	0.7368
Var6		1.0000	0.1060	0.9984
Var7		1.0000	2.0781	0.9703

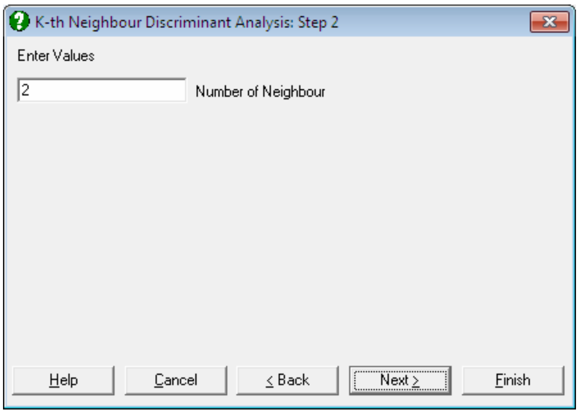
…

Step 4

Variable	Entered at Step	Tolerance	F-to-Enter/Remove	Wilks' Lambda
Var1		0.9923	0.8803	0.4111
Var2		0.9777	3.1368	0.3973
Var3	1	-1.0000	35.3364	0.6433
Var4		0.9910	2.3970	0.4017
Var5	2	-1.0000	7.9404	0.4677
Var6	4	-1.0000	10.7627	0.4858
Var7	3	-1.0000	19.3810	0.5410

8.2.2. K-th Neighbour Discriminant Analysis

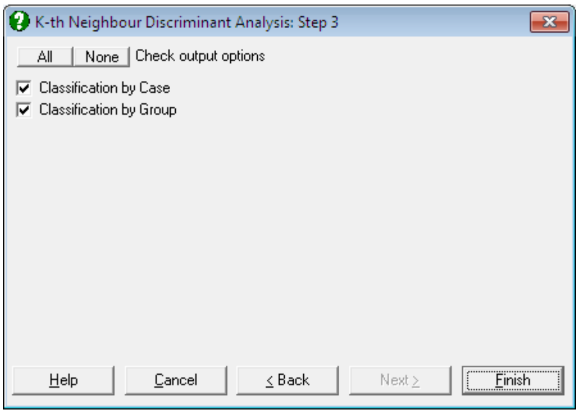
This is a nonparametric discriminant method, which is especially useful when the sample sizes are large. The probability that a point falls within the neighbourhood of a class is based on the distance of the kth nearest point to the centroid. In this way, the effects of outliers can be minimised. The number of neighbour k is provided by the user.


K-th Neighbour Discriminant Analysis: Step 2

Enter Values

2 Number of Neighbour

Since the algorithm does not involve computing covariances etc., the output consists only of the classification tables. See Hand, D J (1981).


K-th Neighbour Discriminant Analysis: Step 3

Check output options

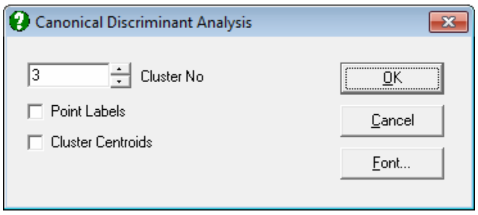
☑ Classification by Case
☑ Classification by Group

Classification by Case: The form of this output is identical to the Classification by Case output in section 8.2.1.3. Canonical Discriminant Analysis.

Classification by Group: The form of this output is identical to the Classification by Group output in section 8.2.1.3. Canonical Discriminant Analysis.

Example

Let us solve the example for the Multiple Discriminant Analysis using the Kth neighbour method (see You can choose to display group centroids in capital letters, point labels or change the font of group letters from the Edit → XY Points dialogue. When the Cluster No field is zero all groups will be displayed simultaneously. If this field is set to one, then the first group only, if two, the second group only, etc., will be displayed. 8.2.1.4. Discriminant Examples).



3 Cluster No

□ Point Labels

□ Cluster Centroids

K-th Neighbour Discriminant Analysis

Variables Selected: Perf, Info, Verbexp, Age

Classification by Case

	ActGroup	Miscls	EstGroup	Probability
1	1		1	0.5000
2	1		1	0.5000
3	1		1	0.5000
4	2	*	1	0.5000
5	2	*	1	0.5000
6	2		2	0.5000
7	3	*	1	0.5000
8	3	*	1	0.5000
9	3	*	1	0.5000

Classification by Group

	1	2	3
1	3 100.00%	0 0.00%	0 0.00%
2	2 66.67%	1 33.33%	0 0.00%
3	3 100.00%	0 0.00%	0 0.00%

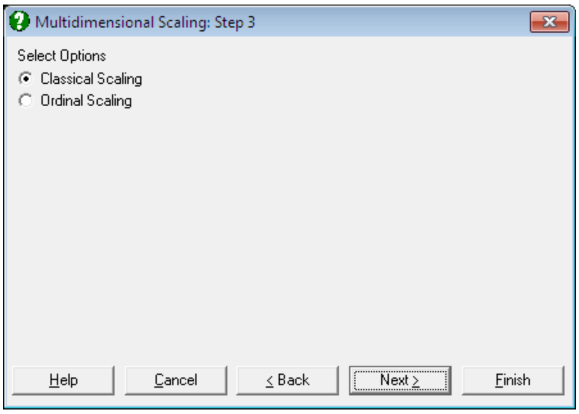
Correctly Classified: 44.44%

8.3. Multidimensional Scaling

This is, in a sense, the reverse of other multivariate methods. Instead of computing dissimilarities from raw data, Multidimensional Scaling (or MDS analysis) attempts to reconstruct the coordinates of points from a given dissimilarity matrix. A typical example is to reconstruct the locations of cities in a country from a given mileage chart. In general, Multidimensional Scaling methods will provide a satisfactory reconstruction of a map, except their orientation (rotation and reflection) and scaling.

The two general categories of Multidimensional Scaling are metric (or classical) and nonmetric (or ordinal) analyses, where the former attempts to reconstruct a map with correct distances, the latter concentrates on the relative positions of points, rather than their distances. The present implementation of Multidimensional Scaling analysis brings together these two methods in a single procedure. First a classical scaling is performed on the dissimilarity matrix, and then the derived configuration is input as initial values for the ordinal Multidimensional Scaling analysis.

A linear classical analysis should be performed before an ordinal one. The program will do this automatically, even if only the ordinal option is selected. It is, however, possible to obtain the classical and ordinal analysis results separately in two different Output Options Dialogues.



Multidimensional Scaling: Step 3

Select Options

● Classical Scaling

○ Ordinal Scaling

8.3.1. Classical Multidimensional Scaling

The Classical Multidimensional Scaling method was developed by Torgerson, W. S. (1958). It is a slight variation of principle components analysis and is also referred to as principle coordinates analysis.

First the proximity data is normalised and the normalised elements are squared:

$$d_{ij}^2 = \frac{d_{ij}^2}{\frac{\sum_i \sum_j d_{ij}^2}{n(n-1)}}$$

Then a scalar products matrix is formed:

$$b_{ij} = -.5[d_{ij}^2 - d_i^2 \text{ - } d_j^2 + d^2]$$

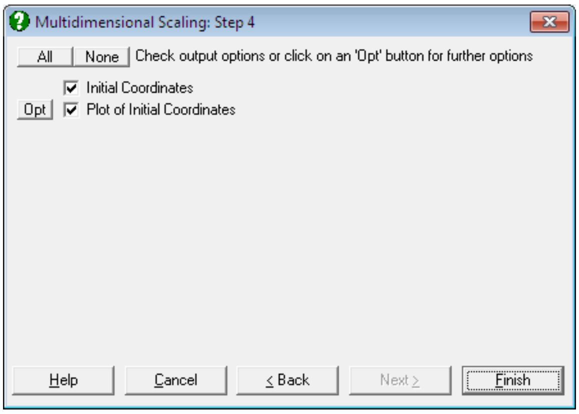
where:

$$d_i^2 = \frac{1}{n} \sum_j d_{ij}^2$$

$$d_j^2 = \frac{1}{n} \sum_i d_{ij}^2$$

$$d^2 = \frac{1}{n^2} \sum_i \sum_j d_{ij}^2$$

The classical scaling Output Options Dialogue displays a table and a graph.



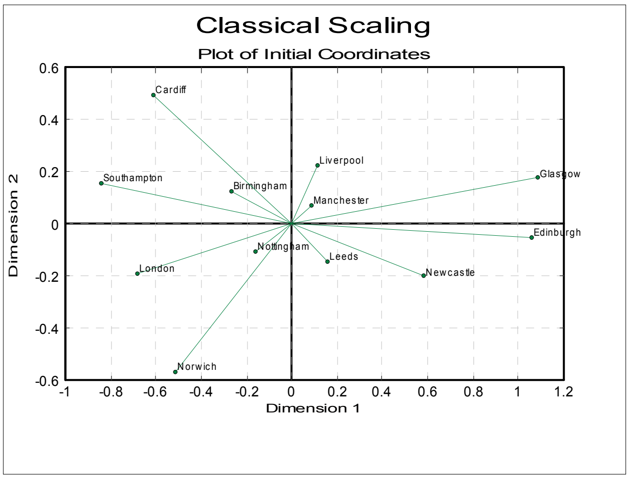
Check output options or click on an 'Opt' button for further options

☑ Initial Coordinates

☑ Plot of Initial Coordinates

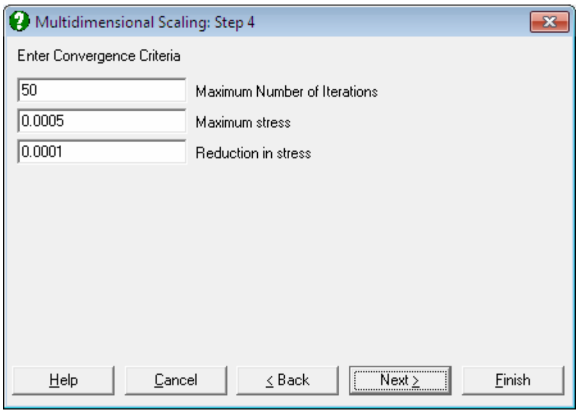
Initial Coordinates: Eigenvectors of $[b_{ij}]$, multiplied by the square root of the corresponding eigenvalues constitute the initial coordinates.

Plot of Initial Coordinates: A 2D plot of the initial coordinates is displayed.



8.3.2. Ordinal Multidimensional Scaling

In reality, a dissimilarity matrix is never an exact measure of the Euclidian distance between the points it represents. The general problem is better described in terms of the relative positions of points rather than their exact distances. For this purpose, the ordinal scaling methods rank distances in the original dissimilarity matrix and aim at obtaining coordinates of points whose ranked distances are the same as that of the original data.


Multidimensional Scaling: Step 4

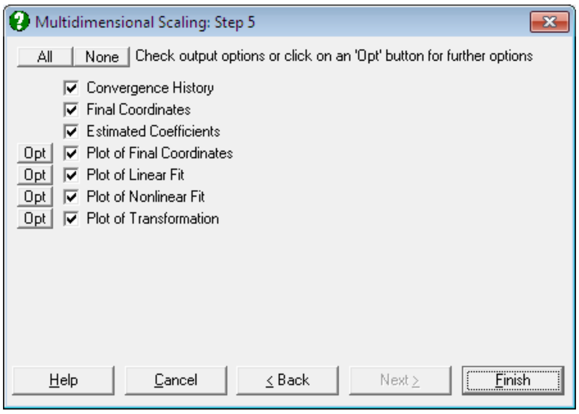
Enter Convergence Criteria

50 Maximum Number of Iterations

0.0005 Maximum stress

0.0001 Reduction in stress

The ordinal multidimensional scaling procedure adopted here uses an iterational procedure known as monotonic least squares (see Kruskal, J. B. and Wish, M. 1978). This is a least squares (in this context also called *stress*) minimisation method, subject to the constraint that the solution should be in the same rank order as the original data. The convergence parameters for the monotonic least squares procedure can be controlled by the user.


Multidimensional Scaling: Step 5

Check output options or click on an 'Opt' button for further options

☑ Convergence History
☑ Final Coordinates
☑ Estimated Coefficients
☑ Plot of Final Coordinates
☑ Plot of Linear Fit
☑ Plot of Nonlinear Fit
☑ Plot of Transformation

Convergence History: The stress and reduction in stress values are displayed for each iteration. The reason for stopping iterations is also displayed.

Final Coordinates: The coordinates, when the iterations have been stopped, are displayed

Estimated Coefficients: n(n - 1)/2 coefficients for the least squares coefficients are displayed.

Plot of Final Coordinates: Final coordinates are plotted in 2D.

Plots of Linear Fit, Nonlinear Fit and Transformation: A scatter graph is plotted for each pair of the last three columns of the Estimated Coefficients table.

Example

The file UKCITIES contains a matrix which gives the distance (in miles) between twelve cities in the UK. This data is used to calculate a map showing the position of these cities relative to each other.

Open UKCITIES, select Statistics 2 → Multidimensional Scaling and select from *Birmingham* to *Southampton* (*C1* to *C12*) as [Variable]s. Leave the default options on the second dialogue and then select the Classical Scaling option. Then select Plot of Initial Coordinates from the output dialogue and Dimension 2 as [X-Axis] and Dimension 1 as [Y-Axis].

The resulting map shows the calculated position of the twelve cities. This is actually a reflection (see 8.3. Multidimensional Scaling) down the y-axis of the true position of these cities, since *Manchester* is actually to the west of *Leeds* in the UK.

Continuing from the last example, click on the [Last Procedure Dialogue] button. Then click [Back] and select Ordinal Scaling. Accept the default values on the next two dialogues. The Estimated Coefficients table below is truncated to save space.

Ordinal Scaling

Convergence History

Terminated at iteration 1, since stress < 0.0005

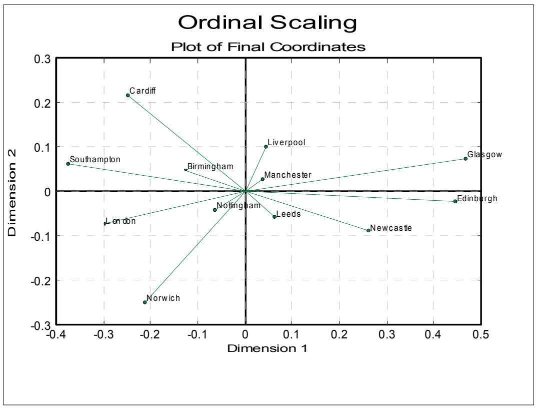
Iteration	Stress	Reduction	%Reduction
1	0.0003	0.0002	48.18%

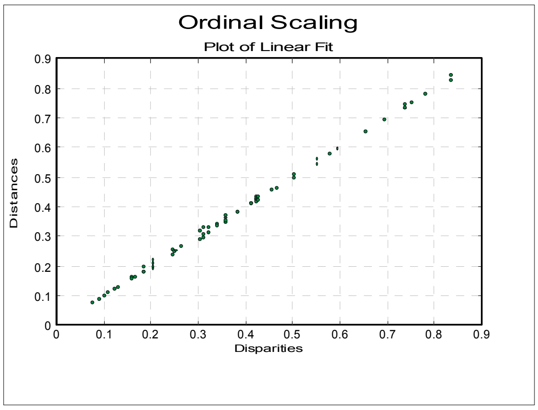
Final Coordinates

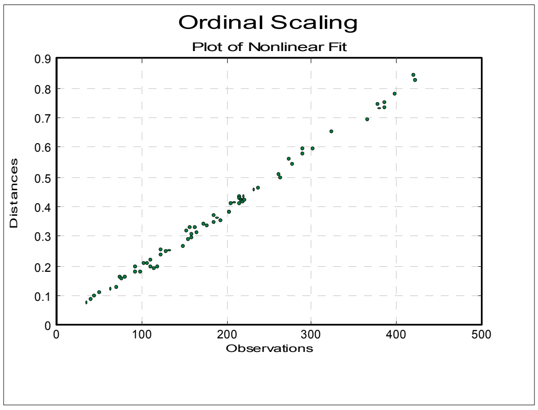
	Dimension 1	Dimension 2
Birmingham	-0.1263	0.0486
Cardiff	-0.2484	0.2156
Edinburgh	0.4462	-0.0230
Glasgow	0.4691	0.0754
Leeds	0.0641	-0.0582
Liverpool	0.0448	0.1018
London	-0.2972	-0.0738
Manchester	0.0372	0.0279
Newcastle	0.2607	-0.0888
Norwich	-0.2118	-0.2477
Nottingham	-0.0632	-0.0404
Southampton	-0.3750	0.0626

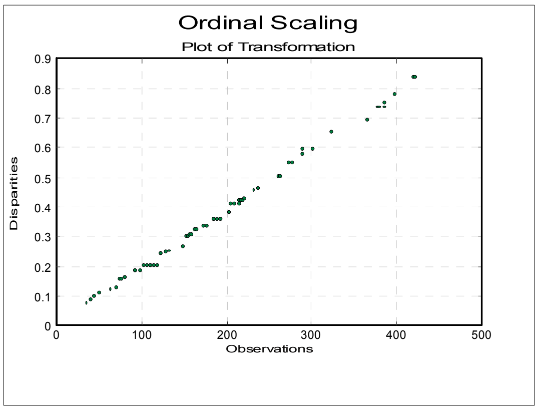
Estimated Coefficients

	Observed	Distances	Disparities	Differences
1	35.0000	0.0743	0.0743	0.0000
2	40.0000	0.0902	0.0902	0.0000
3	44.0000	0.1011	0.1011	0.0000
4	50.0000	0.1091	0.1091	0.0000
5	63.0000	0.1215	0.1215	0.0000
...	...	...	...	...




Ordinal Scaling
Plot of Linear Fit
Distances
Disparities
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

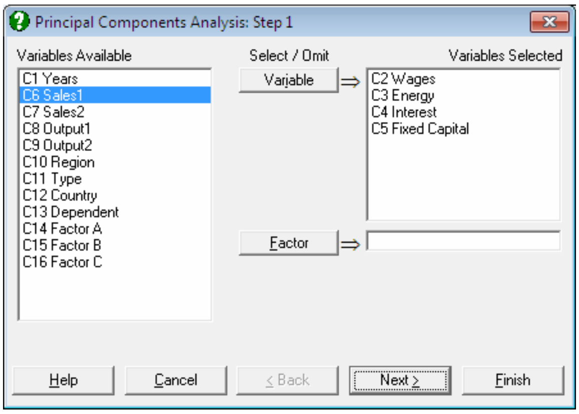

Ordinal Scaling
Plot of Nonlinear Fit
Distances
Observations
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
100
200
300
400
500


Ordinal Scaling
Plot of Transformation
Disparities
Observations
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
100
200
300
400
500

8.4. Principal Components Analysis

This is the core multivariate analysis procedure. All other multivariate methods (except for Cluster Analysis) can be considered as variations of Principal Components Analysis (PCA). The basic idea behind PCA is to redraw the axis system for n dimensional data such that points lie as close as possible to the axes. The derived variables, also called *principal components*, can express a large proportion of total variance of the data with a smaller number of variables.

From a mathematical point of view, the problem of PCA consists of finding eigenvectors of the standardised or non standardised sum of squared products (SSP) matrix for the raw data. The standard and non standard SSP matrices are directly proportional to simple correlations and covariance matrices for the same data respectively.



Variables Available:
C1 Years
C6 Sales1
C7 Sales2
C8 Output1
C9 Output2
C10 Region
C11 Type
C12 Country
C13 Dependent
C14 Factor A
C15 Factor B
C16 Factor C

Variables Selected:
C2 Wages
C3 Energy
C4 Interest
C5 Fixed Capital

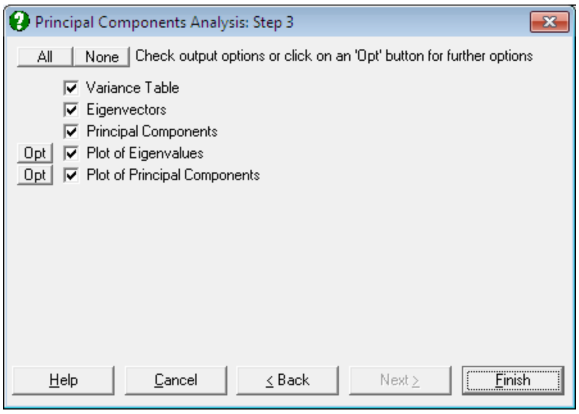
Select raw data columns to analyse by clicking on [Variable]. There is also an optional [Factor] button available to run predictions.

Predictions: The rows of the factor column containing a missing value will not be included in calculations. However, if no other variable contains a missing value, then the PCA transformation will be applied to them. They will be indicated in all plots by an @ character and in case-wise tables by an asterisk (*). In this way, it is possible to obtain transformations on a set of observed cases and simultaneously apply the transformation to a number of test cases.

It is possible to use markers other than missing data to designate cases as test cases. Suppose, for instance, you wish the program to interpret cases with –1 in their group variable as test cases. To do this, enter the following line in the [Options] section of *Documents**Unistat65**Unistat65.ini* file:

```
DiscrPredict=-1
```

The following output options are available:



Principal Components Analysis: Step 3

Check output options or click on an 'Opt' button for further options

☑ Variance Table
☑ Eigenvectors
☑ Principal Components
Opt ☑ Plot of Eigenvalues
Opt ☑ Plot of Principal Components

Variance Table: Eigenvalues are scaled such that their total variance is equal to the total number of variables. It is often concluded that a principle component with an eigenvalue greater than one makes a significant contribution to the total variance.

Eigenvectors: These are the coefficients which transform original data into the new coordinates. Each eigenvector is scaled such that the sum of squares of its elements is unity.

Principal Components: These are the transformed variables obtained by multiplying the original data matrix with the matrix of eigenvectors. When the analysis is carried out on a correlation or covariance matrix, the Principal Components table and plot options will not be available.

The Principal Components have the following properties:

1) They are uncorrelated. The Pearson's correlation between any two Principal Components is zero.
2) Their variances are equal to their corresponding eigenvalues.
3) They are sorted in decreasing order according to their variances.

Therefore, you may examine the Variance Table (the eigenvalues), decide on the first r eigenvalues according to the percentage of variation you want to retain, then save the Principal Components to data and then retain only those first r Principal Components for further analysis.

Plot of Eigenvalues (Scree Plot): This is also called the scree plot. Eigenvalues and their corresponding eigenvectors are sorted in decreasing order. Typically, this plot will fall sharply with the first few eigenvalues and then get less and less steep.

Plot of Principal Components: This is the plot of transformed variables displayed in the Principal Components table. The Edit → XY Points menu option will provide the possibility to display the transformed data points alongside the original variables.

Example

Table 12.2 on p. 607. Tabachnick, B. G. & L. S. Fidell (1989).

Open MULTIVAR, select Statistics 2 → Principal Components Analysis and select *Cost*, *Lift*, *Depth*, *Powder* (*C6* to *C9*) as [Variable]s. Select Output and All to obtain the following results:

Principal Components Analysis

Variance Table

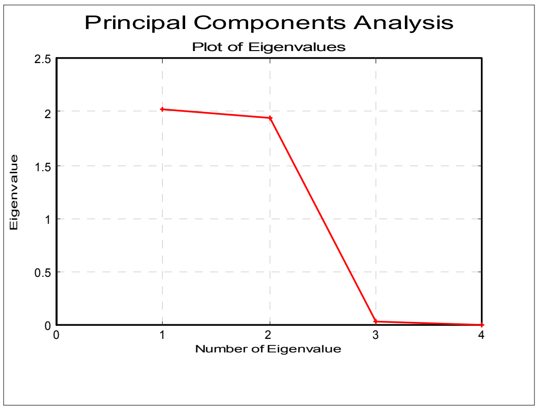
Component No	Eigenvalue	Cumulative Variance	Percent	Cumulative
1	2.0163	2.0163	0.5041	0.5041
2	1.9415	3.9578	0.4854	0.9895
3	0.0378	3.9956	0.0095	0.9989
4	0.0044	4.0000	0.0011	1.0000

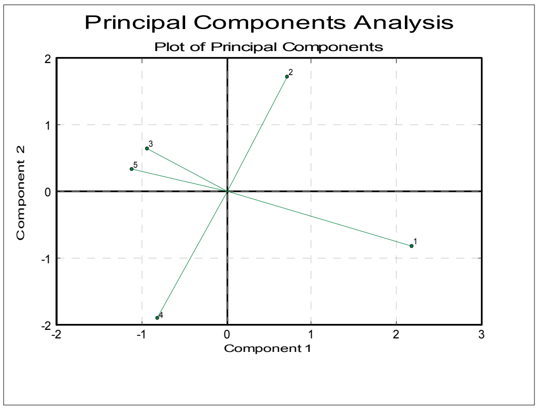
Eigenvectors

	Dimension 1	Dimension 2	Dimension 3	Dimension 4
Cost	-0.3524	0.6143	0.6625	0.2439
Lift	0.2511	-0.6638	0.6759	0.1988
Depth	0.6274	0.3222	0.2755	-0.6532
Powder	0.6474	0.2796	-0.1685	0.6887

Principal Components

	Dimension 1	Dimension 2	Dimension 3	Dimension 4
1	2.1766	-0.8161	0.0820	-0.0379
2	0.7102	1.7180	-0.1123	0.0692
3	-0.9445	0.6479	-0.1456	-0.0930
4	-0.8213	-1.8991	-0.1302	0.0494
5	-1.1210	0.3494	0.3062	0.0123





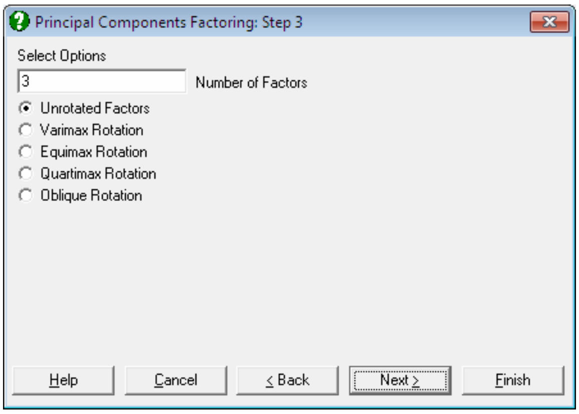
8.5. Factor Analysis

While the aim of Principal Components Analysis is simply to transform the original variables into a new set of variables, Factor Analysis attempts to construct a mathematical model explaining the correlations between a large set of variables. The Factor Analysis does this by deriving some variables (factors) that cannot be observed directly from the raw data.

8.5.1. Factoring Methods

The Factor Analysis is based on the Principal Components Analysis (see Mardia, K. V., Kent, J. T. and Bibby, J. M., 1979). UNISTAT provides two variations of this method called Principal Components Factoring and Principal Axis Factoring. Central to both methods is the concept of *communality*, which is the squared multiple correlations for each factor with all other factors and represent the proportion of variance explained by the common factors. When the number of factors is set to its maximum (the number of variables), communalities become unity and the Factor Analysis is reduced to a Principal Components Analysis.

Principal Components Factoring: This procedure will first perform a Principal Components Analysis and determine the number of components with an eigenvalue greater than unity. This number will be displayed in a dialogue as the number of factors to be extracted, and it can be changed to any value between one and the total number of variables. Then you can either display the unrotated factors or proceed with one of the four rotation options available.

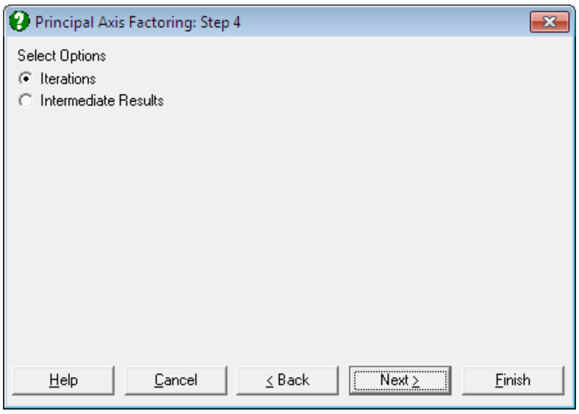

Principal Components Factoring: Step 3

Select Options

3 Number of Factors

● Unrotated Factors
○ Varimax Rotation
○ Equimax Rotation
○ Quartimax Rotation
○ Oblique Rotation

Principal Axis Factoring: First a Principal Components Analysis is performed and the communalities are computed for the selected number of factors. At this stage, it is possible to display the eigenvalues and eigenvectors or to proceed with the next stage of the analysis.

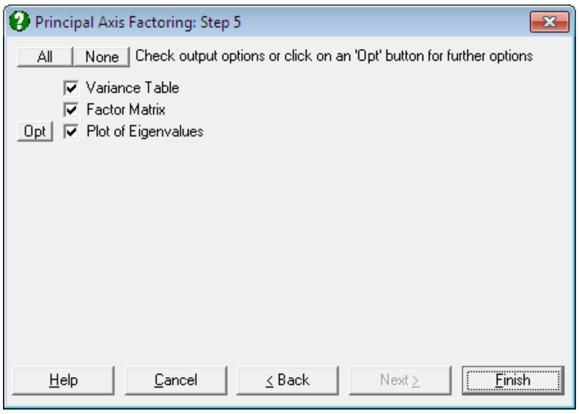

Principal Axis Factoring: Step 4

Select Options

● Iterations

○ Intermediate Results

These initial communalities are then substituted into the diagonal of the correlation matrix replacing ones, and then another PCA is performed and the new communalities are computed. Iterations continue until the maximum number of iterations is reached or the change in the communality estimates is less than the given tolerance level.


Principal Axis Factoring: Step 5

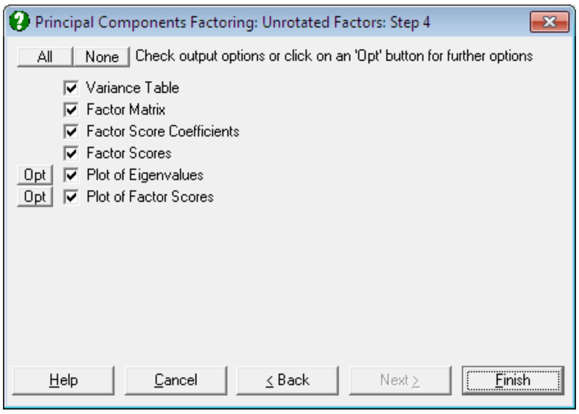
Check output options or click on an 'Opt' button for further options

☑ Variance Table

☑ Factor Matrix

☑ Plot of Eigenvalues

8.5.2. Unrotated Output Options



Principal Components Factoring: Unrotated Factors: Step 4

Check output options or click on an 'Opt' button for further options

☑ Variance Table
☑ Factor Matrix
☑ Factor Score Coefficients
☑ Factor Scores
Opt ☑ Plot of Eigenvalues
Opt ☑ Plot of Factor Scores

The Principal Components Factoring and Principal Axis Factoring methods offer the same initial results menu, also providing access to various rotation options.

Variance Table: This table is similar to the table of eigenvalues discussed in Principal Components Analysis (also see plot of eigenvalues or scree plot), except that only the selected number of factors are displayed and there is a new column called communalities - the squared multiple correlations for each selected factor with all other factors.

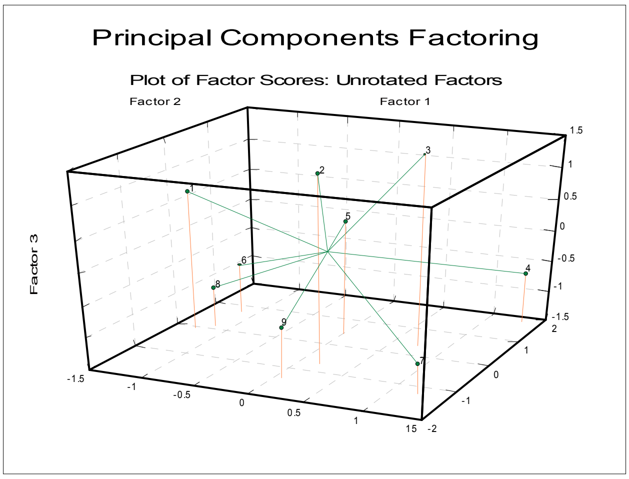
As we have seen above, the program will select by default only those factors with an eigenvalue greater than unity. By examining this table, however, you may wish to adopt a different criterion and choose, say, the first *r* factors which explain 90% of the variation in data. In this case, press <Escape/Back> twice to edit the number of factors suggested by the program, and display the Variance Table again. This process does not involve recomputing the principal components.

Factor Matrix: Elements of this matrix are commonly referred to as *factor loadings* since they indicate how much weight is assigned to each factor. Factor loadings are the correlations between the derived factors and the original variables and are used to establish the link between the two sets. The factors can then be interpreted as representing those characteristics of the original data with which they have the highest correlations. For instance, if the principal component number 1's correlation with the original variable called *intelligence* is the highest, it may then be said to represent *intelligence*. Of course, other original variables will also be correlated with the first principal component, representing other influences on *intelligence*.

Factor Score Coefficients: When multiplied by the original data matrix, these coefficients will transform the original data to a smaller set representing the values of factors. This is the matrix of unrotated factor score coefficients. After a rotation is performed, the rotated factor score coefficients will also be given. These coefficients may be saved to the Data Processor to transform various related data sets outside the procedure. If the data input is a raw data matrix, then the transformation will be carried out by the procedure and the results will be displayed in tabular and graphical forms.

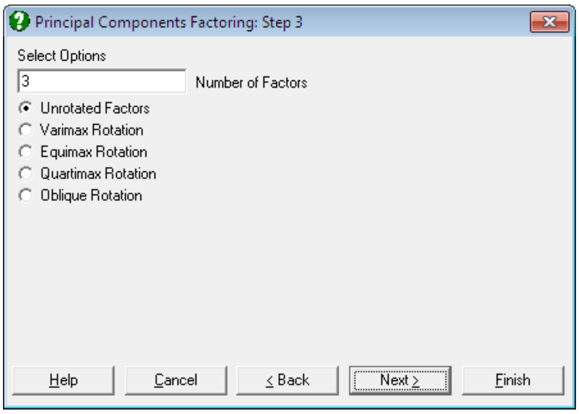
Plot of Eigenvalues (Scree Plot): This is also called the scree plot. Eigenvalues are sorted in decreasing order. Typically, this plot will fall sharply with the first few eigenvalues and then get less and less steep.

Table and Plot of Factor Scores: These are the derived variables representing the most of the variation in the original data, and as such are analogous to principal components in the Principal Components Analysis. Factor scores may be saved for further analysis.



When the input data for the analysis is a correlation or covariance matrix, the table and plot of factor scores options will not be available.

8.5.3. Rotations


Principal Components Factoring: Step 3

Select Options

Number of Factors: 3

● Unrotated Factors
○ Varimax Rotation
○ Equimax Rotation
○ Quartimax Rotation
○ Oblique Rotation

Often, the first phase of the analysis will produce factor loadings showing correlations between all factors and all original variables, making it difficult to assign meaningful attributes to factors. It is possible to improve interpretation of factors by rotating the factor matrix.

There are two major types of rotation; *orthogonal*, keeping the axes at right angles, and *oblique*, where angles between axes are not necessarily 90º. UNISTAT supports three orthogonal (varimax, equimax and quartimax) and one oblique (oblimin) rotation. All these are iterational procedures and may take a long time to compute. The oblimin method also requires the input of a parameter (*Delta*) which determines the degree of obliqueness.

The outputs for orthogonal and oblique rotations are also slightly different. However, all rotation options will display the rotated factor matrix, plot of factors and factor score coefficients. If the input data is not a correlation or covariance matrix, then the factor scores will also be displayed in tabular and graphical forms.

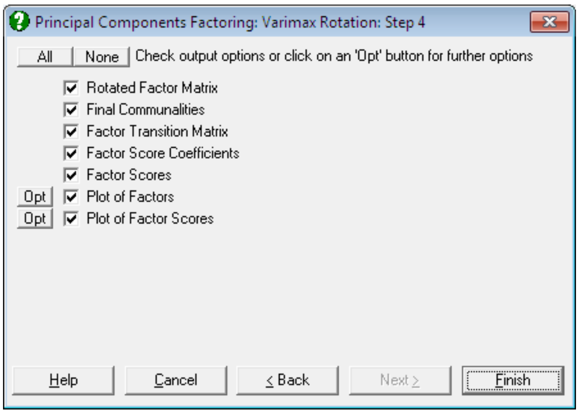
Orthogonal Rotations: The most popular rotation method is varimax, which minimises the number of variables that have high loadings. Quartimax method, on the other hand, minimises the number of factors needed to explain a variable. Equimax is a combination of varimax and quartimax methods.

The orthogonal rotation output options include the final communalities and the factor transition matrix. The latter is the matrix which affects rotations when multiplied by the original factors.

Oblimin Rotation: A parameter determining the degree of obliqueness (*Delta*) is entered by the user. This usually varies from 0 to negative numbers, 0 giving the most oblique solution. Oblimin is an iterational method involving highly demanding calculations, including determining the roots of a third degree polynomial at each iteration.

In addition to the output options of the orthogonal rotation, the structure matrix and the factor correlations matrix options are also available for oblimin rotation.

8.5.4. Rotated Output Options



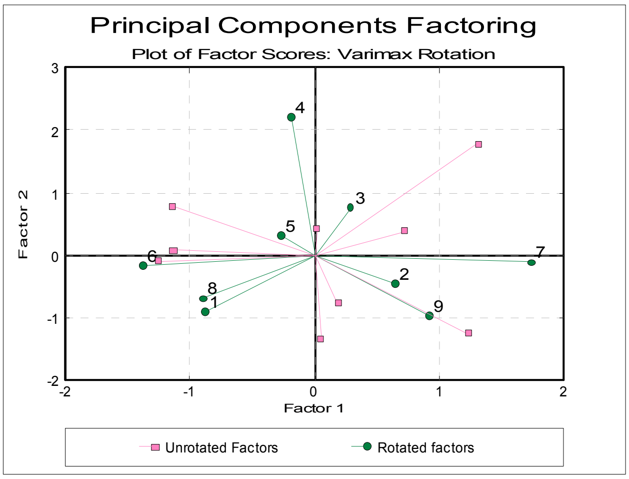
Principal Components Factoring: Varimax Rotation: Step 4

Check output options or click on an 'Opt' button for further options

☑ Rotated Factor Matrix
☑ Final Communalities
☑ Factor Transition Matrix
☑ Factor Score Coefficients
☑ Factor Scores
Opt ☑ Plot of Factors
Opt ☑ Plot of Factor Scores

Rotated factors, factor score coefficients and factor scores can be displayed in tabular or graphical form. The factor scores options will not be available when a correlation or covariance matrix is input.

The plot of factors option allows display of both unrotated and rotated factors on the same graph, enabling you to compare the effects of rotation. The rotated factors can be greater than unity after an oblique rotation. The factor scores graph may take a long time to draw due to the amount of computing necessary, particularly when the number of factors is relatively large.



8.5.5. Factor Analysis Example

Table 12.2 on p. 607 from Tabachnick, B. G. & L. S. Fidell (1989).

Open MULTIVAR, select Statistics 2 → Factor Analysis → Principal Axis Factoring and select *Cost*, *Lift*, *Depth* and *Powder* (*C6* to *C9*) as [Variable]s. The output contains three stages. First, initial communalities, eigenvalues and factor matrix are reported. Then, the final communalities and eigenvalues and thirdly, the results after Varimax rotation.

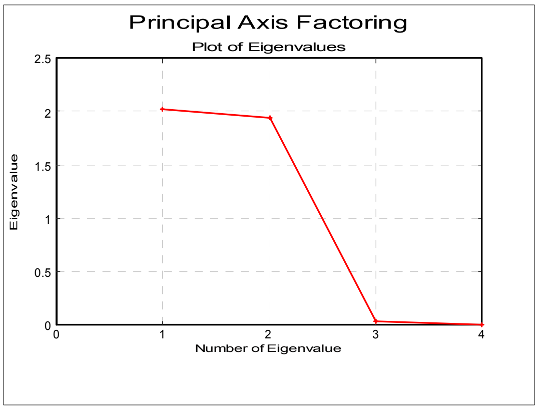
Principal Axis Factoring

Variance Table

	Communality	Factor	Eigenvalue	Percent	Cumulative
Cost	0.9608	1	2.0163	50.4%	50.4%
Lift	0.9532	2	1.9415	48.5%	98.9%
Depth	0.9900				
Powder	0.9909				

Factor Matrix

	Factor 1	Factor 2
Cost	-0.5004	0.8560
Lift	0.3566	-0.9249
Depth	0.8909	0.4490
Powder	0.9193	0.3896



Principal Axis Factoring: Unrotated Factors

Variance Table

	Communality	Factor	Eigenvalue	Percent	Cumulative
Cost	0.9719	1	2.0052	50.1%	50.1%
Lift	0.9587	2	1.9101	47.8%	97.9%
Depth	0.9852				
Powder	0.9995				

Factor Matrix

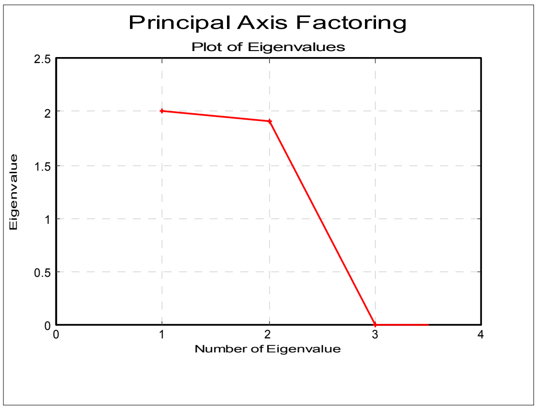
	Factor 1	Factor 2
Cost	-0.4049	0.8989
Lift	0.2548	-0.9454
Depth	0.9283	0.3515
Powder	0.9564	0.2913

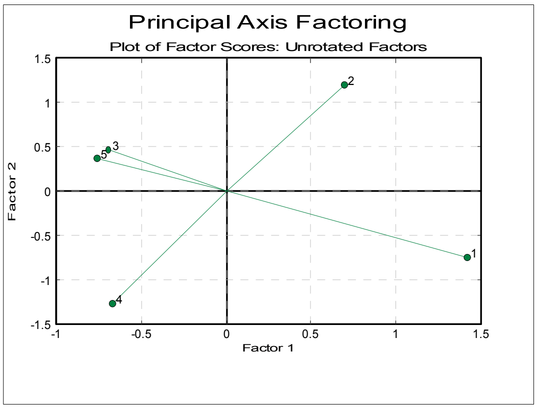
Factor Score Coefficients

	Factor 1	Factor 2
Cost	0.0121	0.6664
Lift	0.2914	-0.3046
Depth	-0.1813	-0.1372
Powder	1.1480	0.5027

Factor Scores

	Factor 1	Factor 2
1	1.4177	-0.7455
2	0.6990	1.1896
3	-0.6946	0.4601
4	-0.6661	-1.2735
5	-0.7560	0.3693





Principal Axis Factoring: Varimax rotation

Rotated Factor Matrix

	Factor 1	Factor 2
Cost	-0.0860	-0.9821
Lift	-0.0710	0.9765
Depth	0.9922	-0.0259
Powder	0.9990	0.0403

Final Communalities

	Communality
Cost	0.9719
Lift	0.9587
Depth	0.9852
Powder	0.9995

Factor Transition Matrix

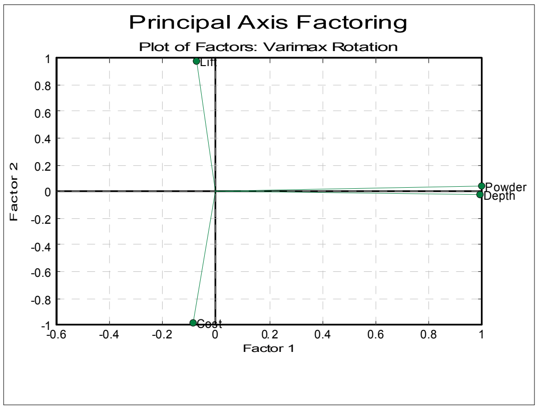
	Factor 1	Factor 2
Factor 1	0.9441	-0.3296
Factor 2	0.3296	0.9441

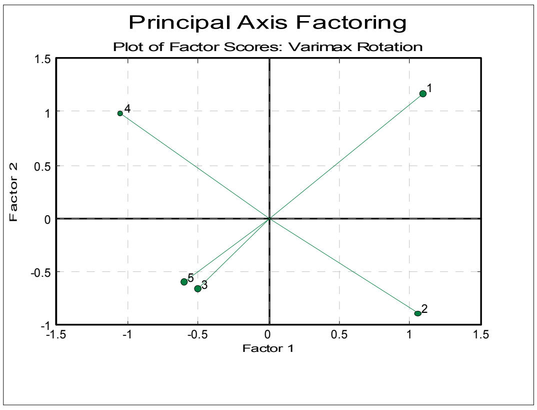
Factor Score Coefficients

	Factor 1	Factor 2
Cost	0.2310	-0.6252
Lift	0.1747	0.3837
Depth	-0.2164	0.0698
Powder	1.2496	-0.0963

Factor Scores

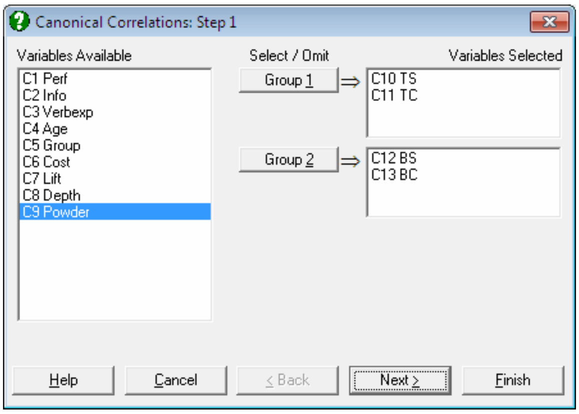
	Factor 1	Factor 2
1	1.0928	1.1712
2	1.0521	-0.8927
3	-0.5042	-0.6634
4	-1.0486	0.9827
5	-0.5920	-0.5978




Principal Axis Factoring
Plot of Factor Scores: Varimax Rotation
Factor 2
Factor 1
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
0.5
1
1.5
1
2
3
4
5

8.6. Canonical Correlations

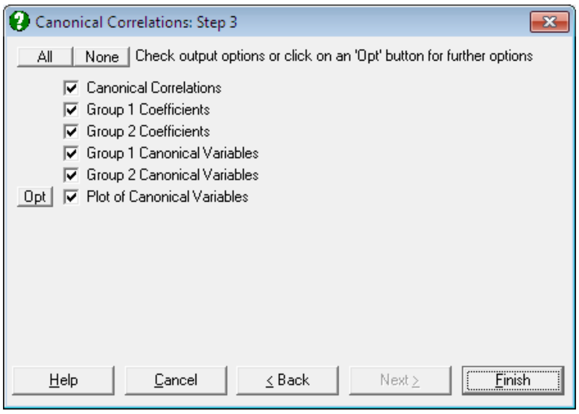
Like the Principal Components Analysis, the Canonical Correlations procedure forms a linear combination of variables that explain most of the variation in data. However, while in Principal Components Analysis the relationships within a single set of variables are sought, in Canonical Correlations relationships between two groups of variables are analysed. Often, an analogy is drawn between Canonical Correlations and Regression Analysis. In a way, the former can be considered as a regression with multiple dependent variables and with no asymmetry between the two groups of variables.



Group 1
C10 TS
C11 TC

Group 2
C12 BS
C13 BC

Select the data columns belonging to two groups by clicking on [Group 1] and [Group 2] from the Variable Selection Dialogue. Although the number of variables in two groups need not necessarily be the same, the number of correlations computed will be the smaller of the two group sizes.



Check output options or click on an 'Opt' button for further options

☑ Canonical Correlations
☑ Group 1 Coefficients
☑ Group 2 Coefficients
☑ Group 1 Canonical Variables
☑ Group 2 Canonical Variables
☑ Plot of Canonical Variables

The following output options are provided.

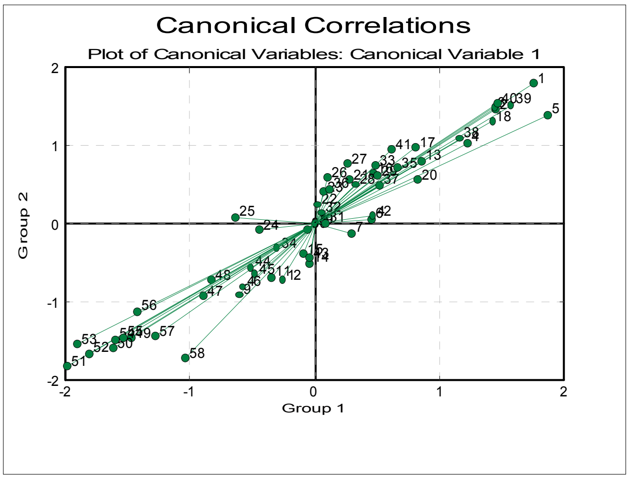
Canonical Correlations: These are simply the square roots of eigenvalues. Wilks' Lambda is computed as follows.

$$L_i = -[(n-1.5)-(p+q)/2]\sum_{i=2}^{m} \mathrm{Log}(1-R_i^2)$$

where i = 2, ..., n and is chi-square distributed with (p - i)(q - i) degrees of freedom.

Transformation Coefficients: These are the multipliers transforming the original variables into two sets of linear combinations (Canonical Variables).

Canonical Variables: Canonical Variables are the linear combinations of the original variables in groups 1 and 2. Correlations between these variables help to determine whether groups 1 and 2 are correlated.



Example

Table 6.1 on p. 197 from Tabachnick, B. G. & L. S. Fidell (1989).

Open MULTIVAR, select **Statistics 2** → Canonical Correlations and select *TS* and *TC* (*C10* and *C11*) as [Group 1] and *BS* and *BC* (*C12* and *C13*) as [Group 2].

Canonical Correlations

Group 1: TS, TC
Group 2: BS, BC

	Eigenvalue	Correlation	Wilks' lambda	Chi-Square	DoF	Probability
1	0.9982	0.9991	0.0001	14.6209	4	0.0056
2	0.9675	0.9836	0.0325	5.1415	1	0.0234

Group 1 Coefficients

Group 1	Grp1CanVarCo1	Grp1CanVarCo2
TS	-0.0469	1.0707
TC	1.0159	-0.3414

Group 2 Coefficients

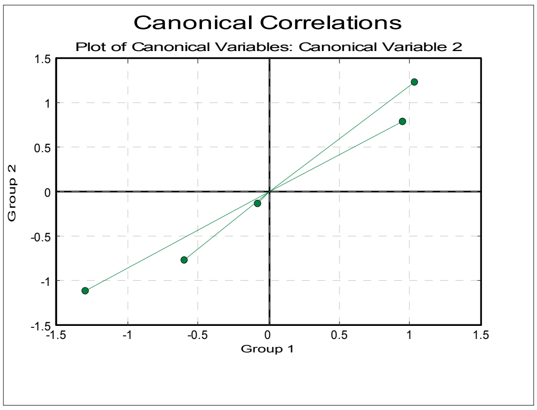
Group 2	Grp2CanVarCo1	Grp2CanVarCo2
BS	0.0529	1.0380
BC	0.9843	-0.3338

Group 1 Canonical Variables

Group 1	Grp1CanVar1	Grp1CanVar2
1	1.2619	0.9462
2	0.8540	-1.2966
3	-0.4013	-0.0815
4	-0.7031	1.0295
5	-1.0114	-0.5976

Group 2 Canonical Variables

Group 2	Grp2CanVar1	Grp2CanVar2
1	1.2492	0.7865
2	0.8460	-1.1180
3	-0.3284	-0.1365
4	-0.7176	1.2348
5	-1.0492	-0.7669

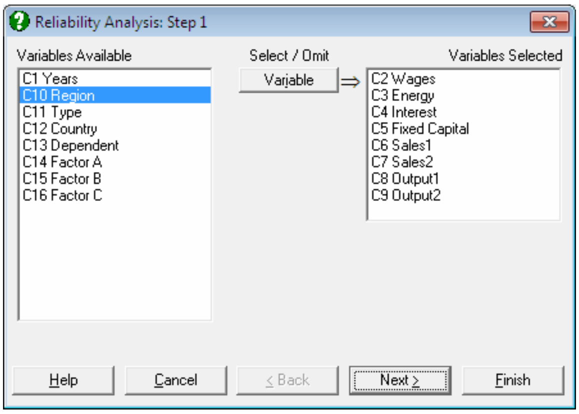


8.7. Reliability Analysis

Reliability Analysis is used to create a measurement scale for a number of variables none of which is representative of the complete group on their own. A number of reliability coefficients are computed to construct the sum scales.

Reliability Analysis is an attempt to find the true score using a set of items. These items are believed to reflect the true score and would involve some random error. The sum of these items (called the sum scale) will give an indication of the true score. The mean value of the error terms across the items will be zero. The true score component remains the same when summing across items. Therefore, the more items that are added, the more true score (relative to the error score) will be reflected in the sum scale.

The assessment of scale reliability is based on the correlation between the individual items or measurements that make up the scale, relative to the variances of the items.

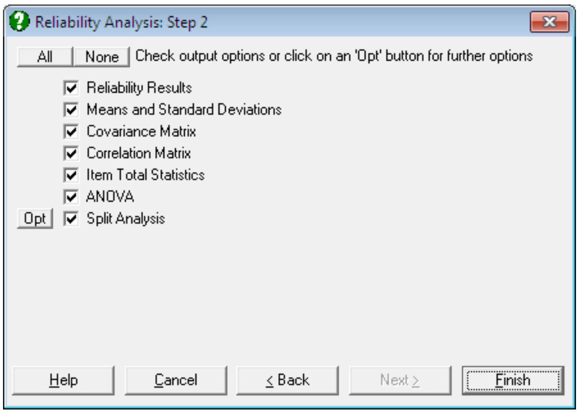


Reliability Analysis: Step 1

Variables Available: C1 Years, C10 Region, C11 Type, C12 Country, C13 Dependent, C14 Factor A, C15 Factor B, C16 Factor C

Variables Selected: C2 Wages, C3 Energy, C4 Interest, C5 Fixed Capital, C6 Sales1, C7 Sales2, C8 Output1, C9 Output2

The columns containing the items are selected by clicking on [Variable]. Any rows containing one or more missing values are removed from the analysis.



Check output options or click on an 'Opt' button for further options

☑ Reliability Results
☑ Means and Standard Deviations
☑ Covariance Matrix
☑ Correlation Matrix
☑ Item Total Statistics
☑ ANOVA
☑ Split Analysis

Reliability Results: The output will display summary information about the sum scale and the items. It will also display the Cronbach's alpha statistic.

The variance of the sum scale will be smaller than the sum of item variances if the items measure the same variability. We can estimate the proportion of true score variance in the sum scale by comparing the sum of item variance with the variance of the sum scale. If the items have no error and measure the true score then alpha will equal 1. If the items are unrelated then alpha will equal 0. Standardised alpha is the value of alpha if the items had been standardised before the Reliability Analysis.

Means and Standard Deviations: The mean and standard deviation of each item is displayed.

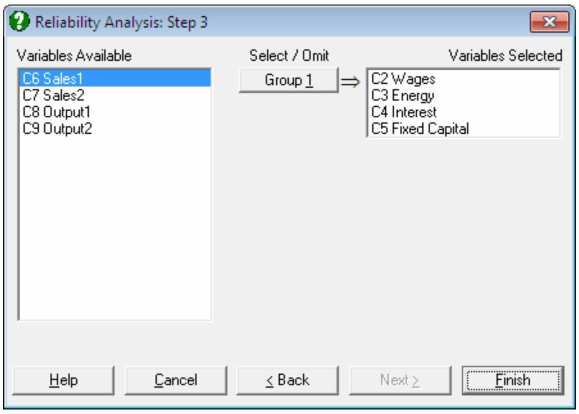
Covariance Matrix: The covariance between the items is displayed.

Correlation Matrix: The correlation between the items is displayed.

Item Total Statistics: Sum statistics relating to each item are displayed. The first two columns contain the mean and variance that the sum scale would take if the item was removed from the scale. The third column shows the correlation between the item and the sum scale with the item removed. The final column shows the value of Cronbach's alpha if the item was removed from the scale.

ANOVA: A one way repeated measures ANOVA table is displayed where the items (columns) are the levels of the factor and the rows are the repeated measures (subjects).

Split Analysis: The items are split into two groups. Select the items for group one by clicking on [Group 1]. The remaining items are placed in group two. If the sum scale is reliable, the two groups should be highly correlated. Split analysis calculates sum scale statistics for both groups and the correlation between the two groups.



Variables Available: C6 Sales1, C7 Sales2, C8 Output1, C9 Output2

Variables Selected: C2 Wages, C3 Energy, C4 Interest, C5 Fixed Capital

Example

Open DEMODATA, select Statistics 2 → Reliability Analysis and select *Wages* to *Fixed Capital* (*C2* to *C5*), *Output 1* and *Output 2* (*C8* and *C9*) as [Variable]s. Select all the output options to obtain the following results.

Reliability Analysis

Variables Selected
Wages, Energy, Interest, Fixed Capital, Output1, Output2

Reliability Results

Statistics for	Mean	Variance	Standard Deviation	Variables
Scale	582.5786	3733.7172	61.1042	6

Statistics for	Mean	Minimum	Maximum	Variance
Item Mean	97.0964	84.1014	107.3121	105.2151
Item Variance	128.4060	44.8069	214.6677	4778.0683
Covariance	98.7760	23.3530	186.6305	2706.6123
Correlation	0.7996	0.5141	0.9653	0.0176

Number of Cases =	56
Cronbach's Alpha =	0.9524
Standardised Alpha =	0.9599

Means and Standard Deviations

Item	Mean	Standard Deviation
Wages	101.9893	13.1962
Energy	100.9995	14.6515
Interest	84.1995	12.1420
Fixed Capital	84.1014	11.9729
Output1	103.9768	6.6938
Output2	107.3121	6.7855

Covariance Matrix

	Wages	Energy	Interest	Fixed Capital	Output1	Output2
Wages	174.1406	186.6305	149.4251	145.7093	60.4203	67.2986
Energy	186.6305	214.6677	162.2643	165.7807	76.1006	65.4243
Interest	149.4251	162.2643	147.4287	135.4810	61.1438	59.3716
Fixed Capital	145.7093	165.7807	135.4810	143.3492	65.5625	57.6750
Output1	60.4203	76.1006	61.1438	65.5625	44.8069	23.3530
Output2	67.2986	65.4243	59.3716	57.6750	23.3530	46.0431

Correlation Matrix

	Wages	Energy	Interest	Fixed Capital	Output1	Output2
Wages	1.0000	0.9653	0.9326	0.9222	0.6840	0.7516
Energy	0.9653	1.0000	0.9121	0.9450	0.7759	0.6581
Interest	0.9326	0.9121	1.0000	0.9319	0.7523	0.7206
Fixed Capital	0.9222	0.9450	0.9319	1.0000	0.8181	0.7099
Output1	0.6840	0.7759	0.7523	0.8181	1.0000	0.5141
Output2	0.7516	0.6581	0.7206	0.7099	0.5141	1.0000

Item Total Statistics

	MeanDel	VarDel	CorrDel	AlphaDel
Wages	480.5893	2340.6090	0.9547	0.9315
Energy	481.5791	2206.6489	0.9534	0.9352
Interest	498.3791	2450.9170	0.9444	0.9323
Fixed Capital	498.4771	2449.9511	0.9622	0.9301
Output1	478.6018	3115.7501	0.7670	0.9589
Output2	475.2664	3141.4291	0.7181	0.9618

Scale mean, scale variance and Cronbach’s alpha denote the values if the item is deleted from the scale.
Correlation denotes the corrected item-total correlation

ANOVA

Due To	Sum of Squares	DoF	Mean Square	F-stat	Prob
Between Subjects	34225.741	55	622.286		
Within Subjects	37608.484	280	134.316		
Between Measures	29460.233	5	5892.047	198.854	0.0000
Error	8148.251	275	29.630		
Total	71834.225	335	214.431		

Split Analysis

Group 1
Wages, Energy, Interest

Group 2
Fixed Capital, Output1, Output2

	Number of Items	Mean	Sum	Standard Deviation	Variance	Cronbach's Alpha
Group 1	3	287.1882	16082.5400	39.1520	1532.8768	0.9753
Group 2	3	295.3904	16541.8600	22.9648	527.3802	0.8339

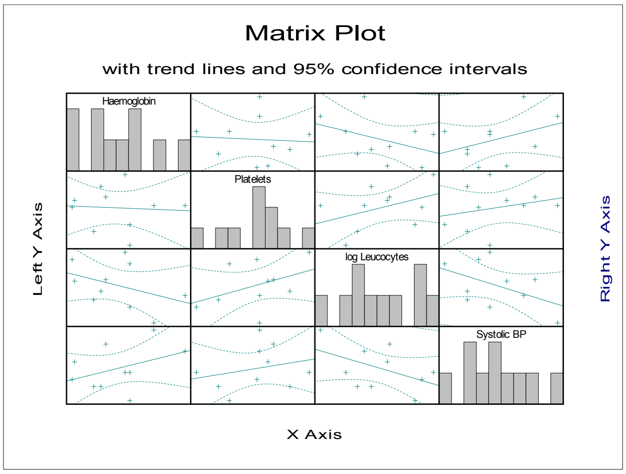
Correlation Between Part 1 and Part 2 =	0.9306
Spearman Brown Split Half Reliability =	0.9641
Guttman Split Half Reliability =	0.8964

8.8. Multivariate Plots

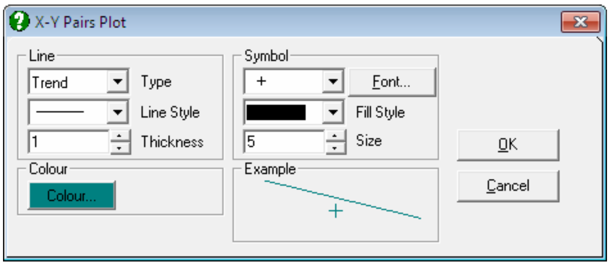
Multivariate plotting procedures are used to inspect and discover the relationships inherent within the matrix (multivariate) data.

The first two options (Matrix Plot and Rectangular Plot) will produce pairwise scatter plots of two or more variables, which are useful in visualising the column-wise characteristics of data. There are also eight plot types collected under Graph → Multivariate Plots → Icon Plots, which display the row-wise characteristics.

8.8.1. Matrix Plot



The Matrix Plot procedure draws n(n - 1) scatter plots and n histograms for n selected variables. The variables are selected from the Variable Selection Dialogue by clicking on [Variable]. All combinations of the selected variables are plotted against each other. Down the diagonal of the matrix, where each variable is matched with itself, a histogram of the variable is drawn and its label is printed. Where each variable is matched with a different variable then a scatter plot of the two variables is drawn.



X-Y Pairs Plot

Line

Trend Type

Line Style

1 Thickness

Symbol

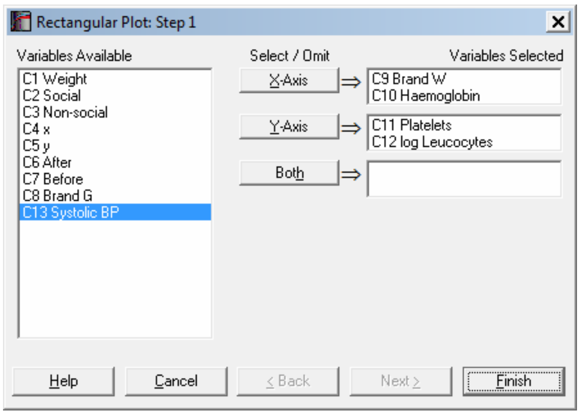
+

Fill Style

5 Size

The Edit → Lines dialogue can be used to connect x-y points with lines or to fit a trend line on data with optional confidence intervals. The usual Symbol panel can be used to select any symbols in any colour. The Edit → Bars dialogue can be used to change the colour or filling patterns of the histogram bars.

8.8.2. Rectangular Plot



Rectangular Plot: Step 1

Variables Available

C1 Weight
C2 Social
C3 Non-social
C4 x
C5 y
C6 After
C7 Before
C8 Brand G
C13 Systolic BP

Select / Omit

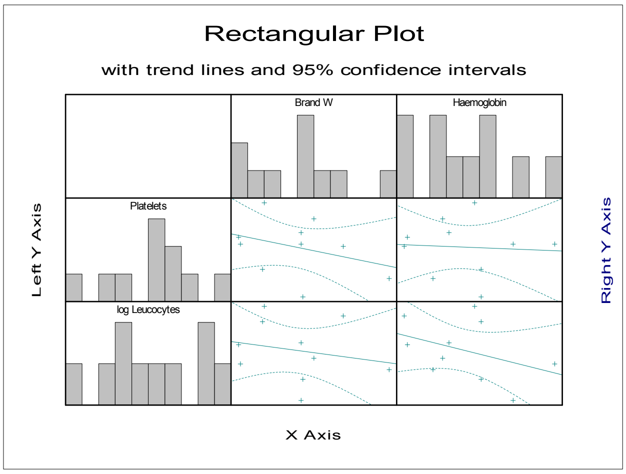
Variables Selected

X-Axis ⇒ C9 Brand W, C10 Haemoglobin

Y-Axis ⇒ C11 Platelets, C12 log Leucocytes

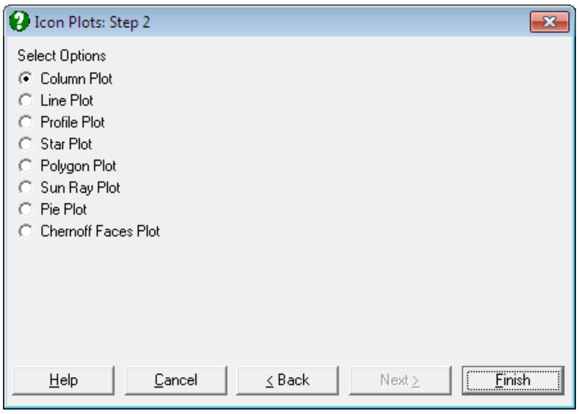
Both ⇒

This draws multiple scatter plots in a matrix format, which resembles the Matrix Plot above. However in Rectangular Plot, variables are selected for the [X Axis], [Y Axis] or [Both]. Variables selected for both axes appear on both the X-axis and the Y-axis. All the variables on the X-axis are plotted against all the variables on the Y-axis. A histogram is drawn in the first column and the first row for each variable. In the remaining cells of the matrix, a scatter plot of the two variables is drawn. If a variable is selected as [Both], it will appear on both axes and will also be plotted against itself. When this happens, all the points will lie on a straight line.



The Edit → Lines dialogue can be used to connect x-y points with lines or to fit a trend line on data with optional confidence intervals. The usual Symbol panel can be used to select any symbols in any colour. The Edit → Bars dialogue can be used to change the colour or filling patterns of the histogram bars.

8.8.3. Icon Plots

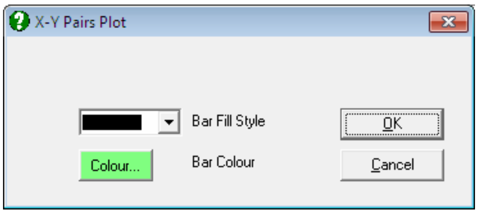

Icon Plots: Step 2

Select Options

● Column Plot
○ Line Plot
○ Profile Plot
○ Star Plot
○ Polygon Plot
○ Sun Ray Plot
○ Pie Plot
○ Chernoff Faces Plot

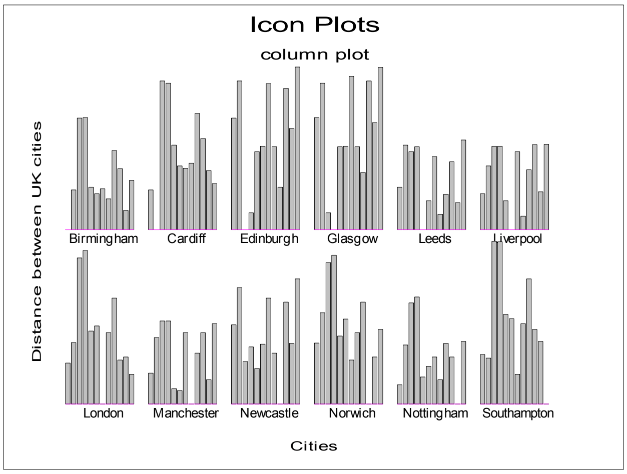
Icon Plots are used to visually inspect the patterns or relationships present within multidimensional (matrix) data. Each icon represents one row of the same set of variables. Each attribute of the icon represents a separate variable. Examining the icons can help to visually cluster cases or identify relationships between the variables.

Select the columns to be included in the plot by clicking on [Variable] in the Variable Selection Dialogue.

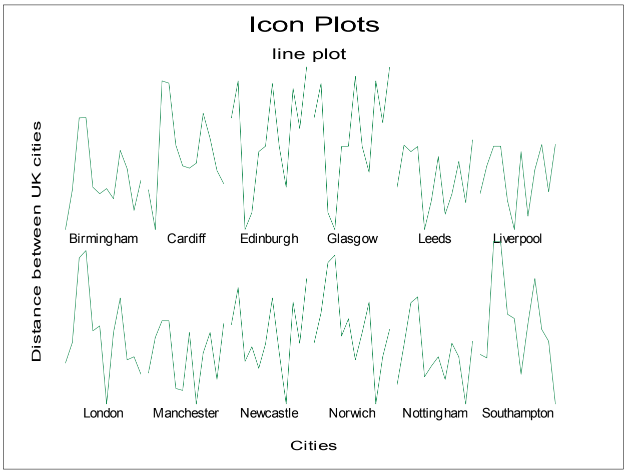


Once the plot is displayed, you will be able to change the number of icons displayed per row or per column from Edit → Axes menu option. The Edit → Bars dialogue can be used to change the colour or filling patterns of the histogram bars.

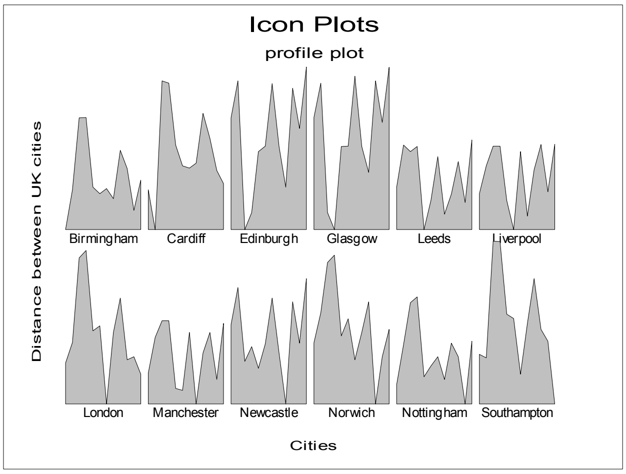
Column Plot: In column icon plot, a small bar chart is drawn for each row of the matrix. Each column in the chart represents the relative value of each selected variable.



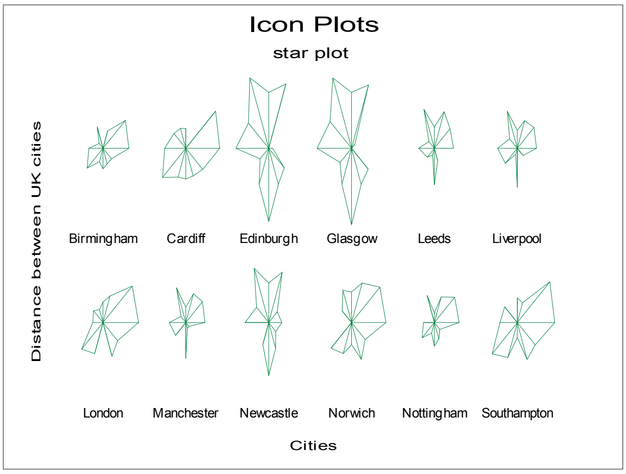
Line Plot: In line icon plot, a line is drawn for each row of the matrix. The height of the line, as it moves from left to right, represents the relative value of each selected variable.



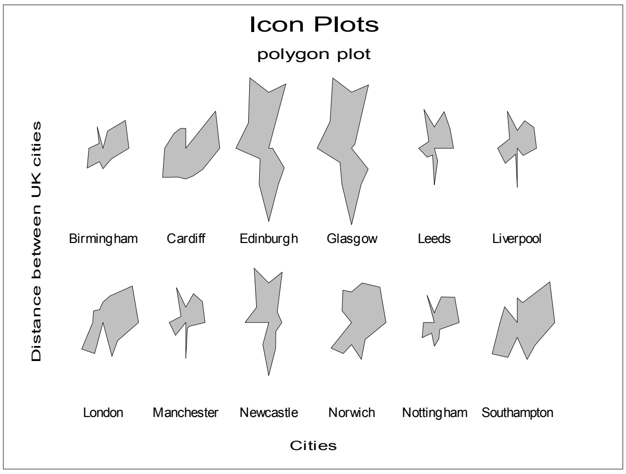
Profile Plot: In profile icon plot, the area enclosed under a line plot is drawn for each row of the matrix. The height of the area, as it moves from left to right, represents the relative value of each selected variable.



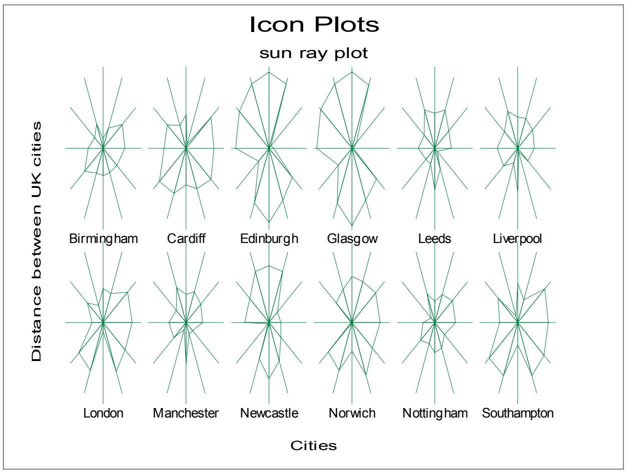
Star Plot: In star icon plot, a star like shape is drawn for each row of the matrix. The shape is constructed by connecting the rays which extend from the centre of the shape together. Each ray (plotted clockwise from the twelve o'clock position) represents a different variable. The length of each ray represents the relative value of a particular variable.



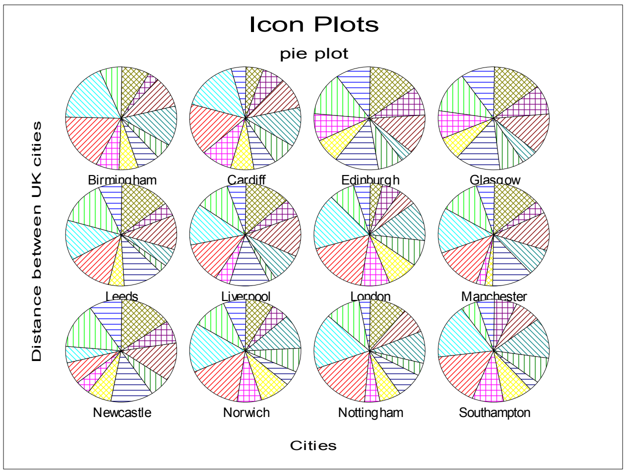
Polygon Plot: In polygon icon plot, a polygon is drawn for each row of the matrix. The distance of each vertex from the centre, as it rotates clockwise, represents the relative value of a particular variable.



Sun Ray Plot: In sun ray icon plot, a star like shape is drawn for each case. Each ray (plotted clockwise from the twelve o'clock position) represents a different variable. The length of the ray represents 4 standard deviations for each variable, the middle of the ray the mean value of the variable, the end of the ray the mean + 2 standard deviations and the centre of the shape the mean - 2 standard deviations. Values for each parameter are connected by a cord.

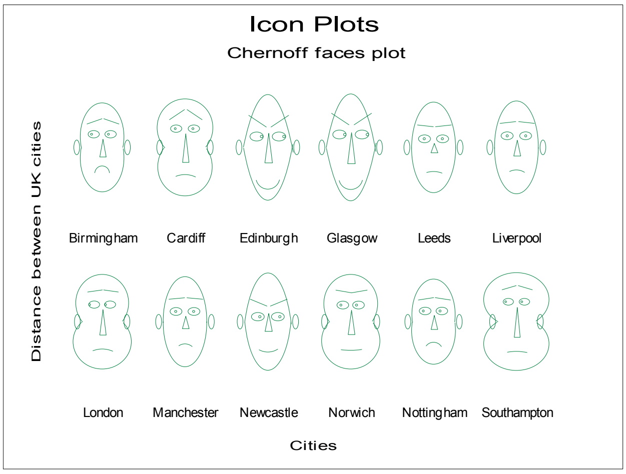


Pie Plot: In pie icon plot, a small pie chart is drawn for each row of the matrix. Each piece of the pie (plotted clockwise from the twelve o'clock position) represents a different variable.



The size of each piece represents the relative value of each variable. This means that if all variables are proportional to each other, then all pie segments will be identical. This is not the case with the other Icon Plots.

Chernoff Faces Plot: In Chernoff faces icon plot, a face is drawn for each row of the matrix. Features on the face represent the relative value of variables. This plot will only work with up to 18 variables.



The features controlled are:

Curvature of mouth
Eccentricity of upper face
Length of nose
Position of pupils
Height of eyes
Face height
Nose width
Length of eyes
Length of eyebrows
Angle of eyebrow
Eccentricity of lower face
Height of centre of mouth
Length of mouth
Separation of eyes
Face width
Eccentricity of eyes
Radius of ears
Height of eyebrows

UNISTAT Statistical Package

Chapter 9
Time Series Analysis

9.1. Box-Jenkins ARIMA

9.1.0. Overview

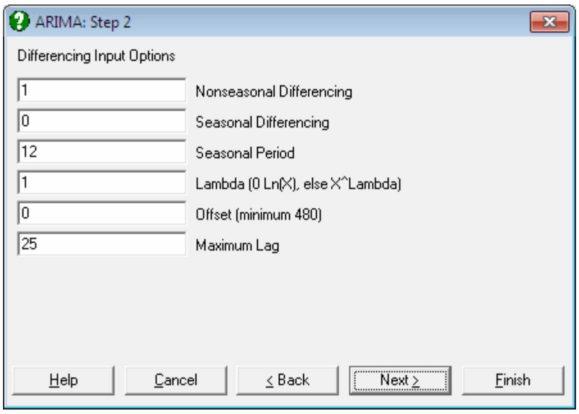
ARIMA stands for *Auto Regressive Integrated Moving Average* model. So called, because the model fits autoregressive and moving average parameters to a transformed (differenced) time series and integrates back to the original scale before forecasts are generated. The differencing transformation makes use of B, the backshift operator, which shifts the subscript of a time series observation backwards in time by one period.

$$Bx_t = x_{t-1}$$

$$B^2 x_t = x_{t-2}$$

Select the column to analyse by clicking on [Dependent]. This column should not contain any missing values.

9.1.1. Differencing Input Options



ARIMA: Step 2

Differencing Input Options

1 Nonseasonal Differencing

0 Seasonal Differencing

12 Seasonal Period

1 Lambda (0 Ln(X), else X^Lambda)

0 Offset (minimum 480)

25 Maximum Lag

The Differencing Input Options dialogue is for entering the parameter values used in transforming the original data. The data can be transformed by differencing, taking logs, raising to a power and adding an offset to it.

$$w_t = (1-B)^d (1-B^s)^D y_t$$

Nonseasonal Differencing: The degree of differencing across whole series (d).

Seasonal Differencing: The degree of differencing between points with seasonal period units apart in the series (D).

Seasonal Period: The number of time units per season (s).

Lambda: This is a coded value which determines any logarithmic or power transformation. This is performed before any differencing:

= 1 No Effect $y_t = x_t + \text{offset}$

= 0 Log of series $y_t = \text{Log}(x_t + \text{offset})$

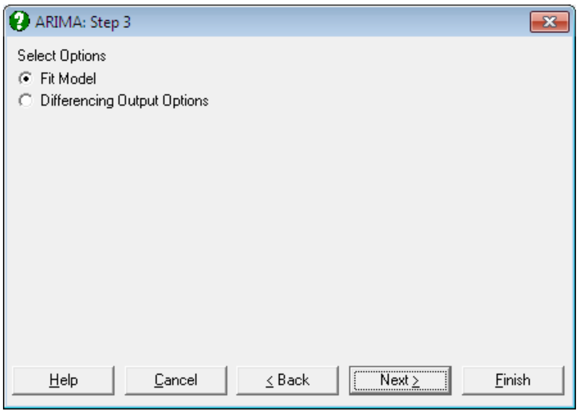
else $y_t = (x_t + \text{offset})^{\lambda}$

Offset: The value of offset is added to every value in the time series. This is used to allow taking logarithms of a series including negative values.

Maximum Lag: This is the maximum lag to calculate in correlation displays.

You can apply any transformation to the series during the data preparation phase. In this case, forecasts will be generated in the transformed scale. Transformations made within Box-Jenkins ARIMA are reversed to give forecasts in the original scale.

The program then displays a dialogue with two options. The Fit Model option proceeds with the next step in the analysis (where the Box-Jenkins ARIMA model is selected.) and the Differencing Output Options gives access to the intermediate results.



ARIMA: Step 3

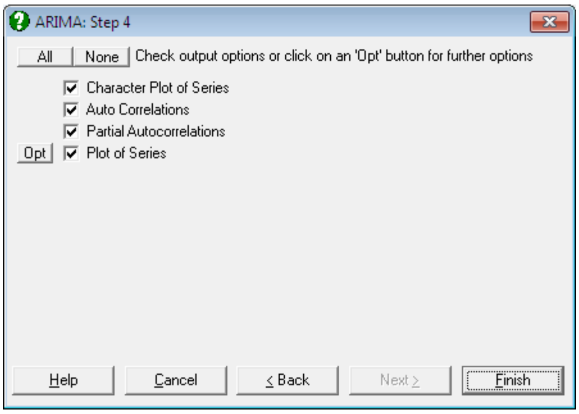
Select Options

● Fit Model

○ Differencing Output Options

The Fit Model option should not be selected until the data series has been transformed into a stationary series.

9.1.2. Differencing Output Options


ARIMA: Step 4
All
None
Check output options or click on an 'Opt' button for further options
☑ Character Plot of Series
☑ Auto Correlations
☑ Partial Autocorrelations
Opt
☑ Plot of Series
Help
Cancel
< Back
Next >
Finish

Differencing Output Options should be used to help set the transformation values before the model is estimated. You may move between this dialogue and the second dialogue a number of times before the series is transformed appropriately. The input series for the Box-Jenkins ARIMA model must be stationary. A stationary series has a constant mean, variance and autocorrelation. Also, the Autocorrelation Function and Partial Autocorrelation Function should give an idea about the number of parameters to fit to the model.

Character Plot of Series: The transformed series is displayed in the form of a character plot together with the values.

Autocorrelation Function: The Autocorrelation Function (ACF) displays the autocorrelations of the transformed series. The autocorrelation represents the correlation between points in the series at displacement lag. The number of autocorrelations displayed is controlled in the Differencing Input Options dialogue.

In Box-Jenkins ARIMA modelling the series is required to be stationary. If the ACF either cuts off fairly quickly or dies down fairly quickly, then the time series values should be considered stationary. If the ACF dies down extremely slowly, then the time series values should be considered non stationary and some differencing will be required.

The ACF should be examined to decide which model to fit in the final stages.

Partial Autocorrelation Function: The Partial Autocorrelation Function (PACF) displays the partial autocorrelations of the transformed series. The partial autocorrelations represent the correlation between points in the series at displacement lag, with the effects of the intervening observations eliminated. Hence the partial autocorrelation at lag 1 is equivalent to the autocorrelation at lag 1. The number of autocorrelations displayed is controlled in the Differencing Input Options dialogue.

The PACF should be examined to decide which model to fit in the final stages.

Hi-Res Plot of Series: This displays a graphical view of the transformed series.

Example

Table 3.1 on p. 83 from Bowerman, Bruce L. & Richard T. O'Connell (1987). Data on monthly Hotel Room Averages for 1973-1986 are given.

Open TIMESER, select Statistics 2 → Box-Jenkins ARIMA and *Room Averages* (*C1*) as [Variable]. At the Differencing Input Options dialogue enter:

- **0** Nonseasonal Differencing
- **1** Seasonal Differencing
- **12** Seasonal Period
- **0** Lambda (0 Log(X), else X^Lambda)
- **0** Offset (minimum 480)
- **25** Maximum Lag

Check all differencing output option boxes to obtain the following results:

ARIMA

Character Plot of Series: Room Averages

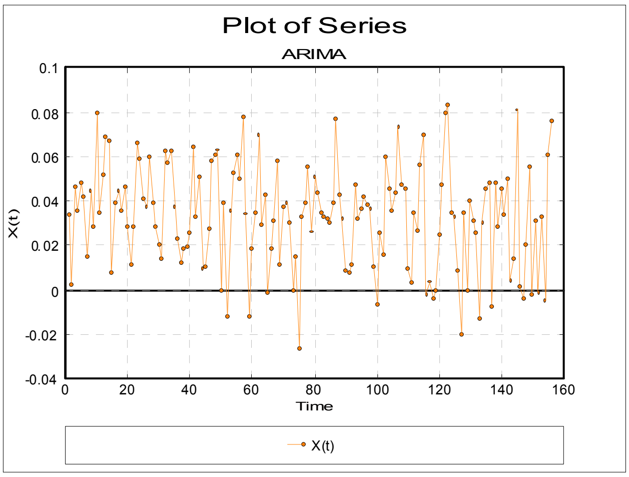
```
Row |   X(t) | -0.0263                                               0.0831
  1 | 0.0334 |                                 *
  2 | 0.0020 |                 *
  3 | 0.0465 |                                        *
  4 | 0.0357 |                                  *
  5 | 0.0484 |                                         *
... |    ... | ...
```

Autocorrelations: Room Averages

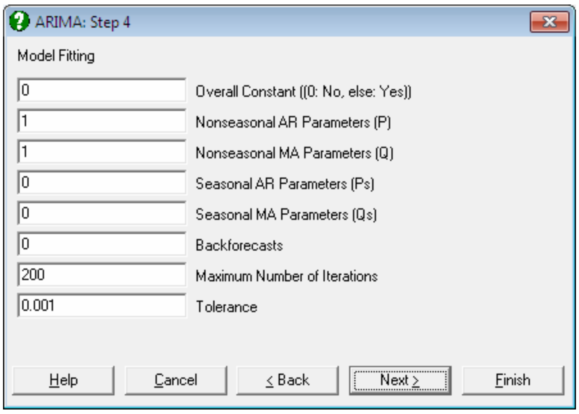
Lag	Correlation	Standard	-1.0000 1.0000
		Error	
1	0.1933	0.0801	(*****)*
2	0.0244	0.0830	(**)
3	-0.2442	0.0830	***(****)
4	-0.1515	0.0875	(*****)
5	-0.2119	0.0892	*(*****)
...	...	...	...

Partial Autocorrelations: Room Averages

Lag	Correlation	Standard	-1.0000 1.0000
		Error	
1	0.1933	0.0801	(*****)*
2	-0.0135	0.0801	(*)
3	-0.2560	0.0801	***(****)
4	-0.0622	0.0801	(**)
5	-0.1760	0.0801	*(****)
...	...	...	...



9.1.3. Model Fitting


ARIMA: Step 4

Model Fitting

0 Overall Constant ((0: No, else: Yes))
1 Nonseasonal AR Parameters (P)
1 Nonseasonal MA Parameters (Q)
0 Seasonal AR Parameters (Ps)
0 Seasonal MA Parameters (Qs)
0 Backforecasts
200 Maximum Number of Iterations
0.001 Tolerance

The Model Fitting dialogue should only be used when the time series is considered stationary. The following guidelines can be used to help choose an Box-Jenkins ARIMA model to fit. It is often possible to try different models on the same data.

9.1.3.1. Seasonal and Nonseasonal Operators

Nonseasonal Operators (p and q)

The ACF has spikes at lags 1, 2, ..., r and cuts off after lag r, and the PACF dies down; use q = r and p = 0.

The ACF dies down and the PACF has spikes at lags 1, 2, ..., r and cuts off after lag r; use q = 0 and p = r.

The ACF has spikes at lags 1, 2, ...,r and cuts off after lag r, and the PACF has spikes at lags 1, 2, ... ,s and cuts off after lag s; use q = r and p = s.

The ACF contains small autocorrelations at all lags and the PACF contains small autocorrelations at all lags; use q = 0 and p = 0.

The ACF dies down and the PACF dies down; use p = 1 and q = 1.

Seasonal Operators (P and Q)

The previous guidelines apply to P and Q, but only consider autocorrelations at s, 2s, 3s, ... where s is the seasonal period.

9.1.3.2. Model Fitting Parameters

The Model Fitting dialogue requires inputting the following parameters:

Overall Constant: If the overall constant value is non zero, then θ_0 is included in the model, otherwise it is not.

Nonseasonal AR Parameters: The nonseasonal AR parameter determines the number of nonseasonal autoregressive parameters (p) to include in the model. This value is normally not larger than 2.

Nonseasonal MA Parameters: The nonseasonal MA parameter determines the number of nonseasonal moving average parameters (q) to include in the model. This value is normally not larger than 2.

Seasonal AR Parameters: The seasonal AR parameter determines the number of seasonal autoregressive parameters (P) to include in the model. This value is normally not larger than 2.

Seasonal MA Parameters: The seasonal MA parameter determines the number of seasonal moving average parameters (Q) to include in the model. This value is not normally larger than 2.

Backforecasts: This is the number of backforecasts generated before the model is fitted. If this value is zero then no backforecasts are generated.

Maximum Number of Iterations: The maximum number of iterations is the number of iterations allowed before the model declares non convergence.

The model fitted is given by the following equations:

$$\varphi(B)w_t = \theta_0 + \theta(B)\alpha_t$$

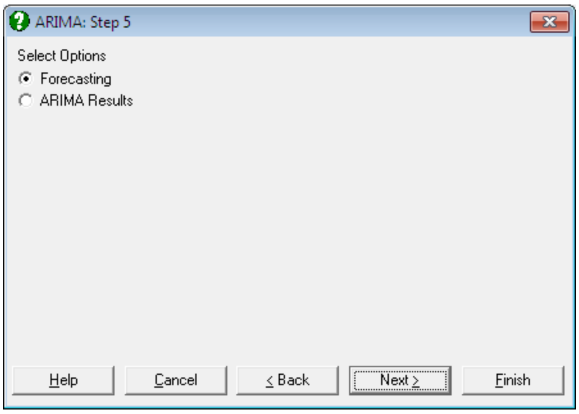
$$\Phi(B^s)\alpha_t = \Theta(B^s)a_t$$

where:

- θ_0 is the overall constant.
- $\varphi(B) = 1 - \varphi_1 B - \ldots - \varphi_p B^p$ is the AR operator.

- $\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$ is the MA operator.
- $\Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{Ps}$ is the seasonal AR operator.
- $\Theta(B^s) = 1 - \Theta_1 B^s - \ldots - \Theta_Q B^{Qs}$ is the seasonal MA operator.
- α_t is the seasonal white noise.
- a_t is the white noise and assumed $a_t \approx Z(0,\sigma^2)$

The model is fitted by an iterative least squares method. The output options are accessed by selecting the ARIMA Results from the following dialogue.

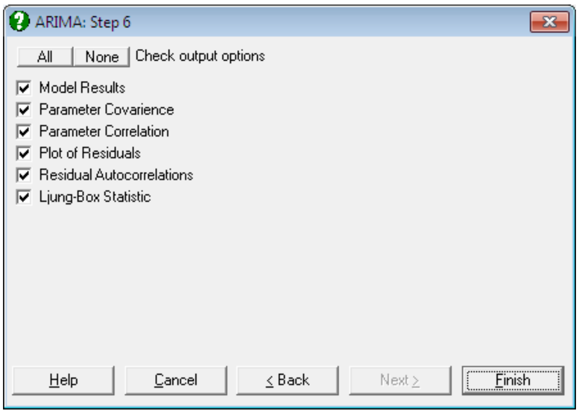

ARIMA: Step 5

Select Options

● Forecasting

○ ARIMA Results

9.1.4. Model Output Options



Check output options

☑ Model Results
☑ Parameter Covarience
☑ Parameter Correlation
☑ Plot of Residuals
☑ Residual Autocorrelations
☑ Ljung-Box Statistic

When a model has been fitted, you will have the following output options:

Model Results: The number of iterations made and the transformation used are displayed. For each fitted parameter the estimated value, the standard error and the t-value are displayed.

Parameter Covariance Matrix: A table of the covariance between each fitted parameter is displayed.

Parameter Correlation Matrix: A table of the correlation between each fitted parameter is displayed.

Plot of Residuals: This displays the residual values in a table and allows them to be saved back to the data matrix.

Residual Autocorrelation: This displays the Autocorrelation Function of the residuals. The residuals should be unrelated because the model should account for the relationship in the time series data. If the residuals are unrelated then the autocorrelations of the residuals should be small. The Ljung-Box statistic (see below) is a test of the residual autocorrelations.

Ljung-Box Statistic: This displays the Ljung-Box statistic, the degrees of freedom and the associated chi-square probabilities at various values up to the lag. The Ljung-Box statistic is a test of the relationship between the residuals. A large value shows the residuals to be related, and hence the model being inadequate.

Example

Following the example in Differencing Output Options, select Fit Model to select an Box-Jenkins ARIMA model. On the Model Fitting dialogue enter:

- **1** Overall Constant (0 No, Else Yes)
- **3** Nonseasonal AR Parameters (P)
- **0** Nonseasonal MA Parameters (Q)
- **0** Seasonal AR Parameters (Ps)
- **1** Seasonal MA Parameters (Qs)
- **0** Backforecasts
- **200** Maximum Number of Iterations
- **0.0001** Tolerance

On the Model Output Options dialogue check only the Model Results box to obtain the following results:

ARIMA: Fit Model

Model Results

Transformation: X(t) = (1-B^12) log(Room Averages)

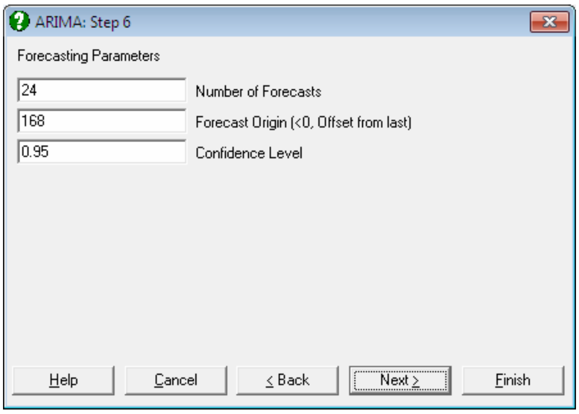
Parameter	Estimate	Std error	t ratio
Overall Constant	0.02699	0.01690	1.5976
(AR) P(1)	0.26089	0.06798	3.8380
(AR) P(2)	0.15688	0.06293	2.4929
(AR) P(3)	-0.23467	0.07074	-3.3175
(SMA) Qs(1)	0.51389	0.07311	7.0293

Number of Iterations =	148 (Converged)
Seasonal Period =	12

The numbers obtained here are not identical to the ones given in the book, though the general characteristics of the fitted models are similar. This is due to the highly iterative nature of the estimation process and the results may differ from one implementation to the other.

9.1.5. Forecasting

9.1.5.1. Forecasting Input Options



Forecasting Parameters

24 Number of Forecasts

168 Forecast Origin (<0, Offset from last)

0.95 Confidence Level

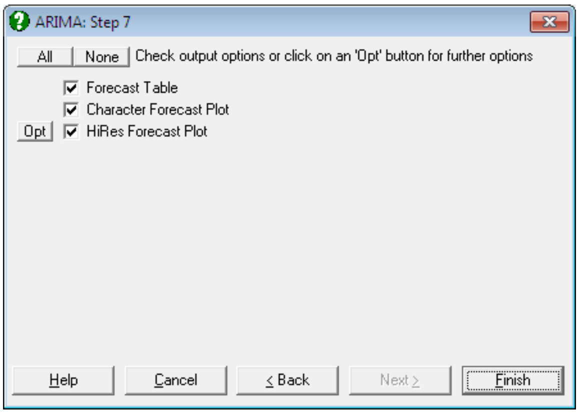
This dialogue requests the following parameters:

Number of Forecasts: This is the number of forecasts to be generated.

Forecast Origin (<0, Offset): The forecast origin determines the location in the series at which the forecasts will start. If you enter a positive value, this is used as the forecast origin. If 0 is entered, then the last point in the series is used as the origin. If a negative value is entered, then this value is used as an offset from the last point. For instance, -1 represents the penultimate point as the forecast origin.

Confidence Level: The forecasts will be given with confidence intervals at this level. The value must be greater than 0 and less than 1. Typical values are 0.95, 0.99 or 0.9.

9.1.5.2. Forecasting Output Options



Check output options or click on an 'Opt' button for further options

☑ Forecast Table
☑ Character Forecast Plot
☑ HiRes Forecast Plot

When the above parameters have been specified, the following output options will be available:

Forecast Table: A table of forecasts and confidence intervals from the given forecast origin is displayed.

Character Forecast Plot: A character plot of the original data, lead 1 forecasts and forecasts from the forecast origin are displayed.

Plot of Forecast: A graphical display of the original data, lead 1 forecasts and forecasts from the forecast origin is generated. These are the same values as the Character Forecast Plot.

Example

Following the example in Model Output Options dialogue select Forecasting. On the Forecasting Input Options dialogue enter:

- **24** Number of Forecasts
- **168** Forecast Origin (<0, Offset from last)
- **0.95** Confidence Level

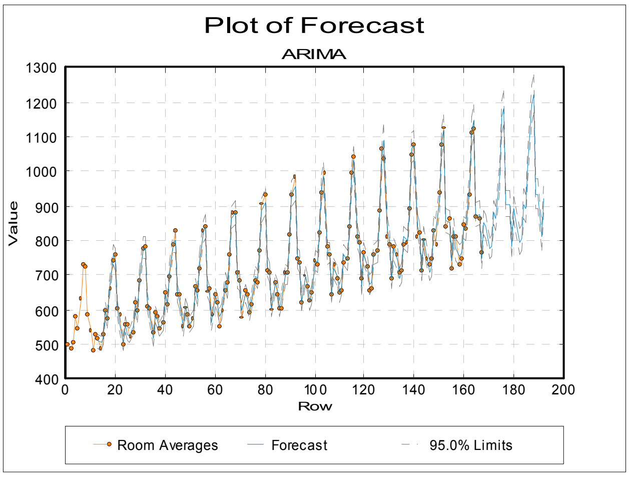
Select the Forecast Table output option to obtain the following results:

ARIMA: Forecasting

Forecast Table: Room Averages

Forecasts with Origin at 168

Row	Forecast	Lower 95%	Upper 95%
169	840.0521	808.0632	873.3075
170	771.1056	741.7421	801.6315
171	777.0039	747.4158	807.7633
172	872.1992	838.9860	906.7271
173	858.4281	825.7393	892.4109
174	982.0313	944.6357	1020.9071
175	1154.9294	1110.9500	1200.6498
176	1181.2955	1136.3121	1228.0598
177	902.9520	868.5678	938.6974
178	903.5714	869.1636	939.3412
179	783.1983	753.3743	814.2029
180	892.2127	858.2375	927.5330
181	860.6177	823.8597	899.0157
182	794.2411	760.3182	829.6776
183	804.2494	769.8990	840.1324
184	903.2180	864.6406	943.5167
185	888.6314	850.6769	928.2792
186	1015.3943	972.0257	1060.6980
187	1193.5981	1142.6182	1246.8526
188	1220.5764	1168.4441	1275.0345
189	933.1099	893.2557	974.7422
190	933.8563	893.9703	975.5220
191	809.5330	774.9569	845.6517
192	922.2238	882.8345	963.3704

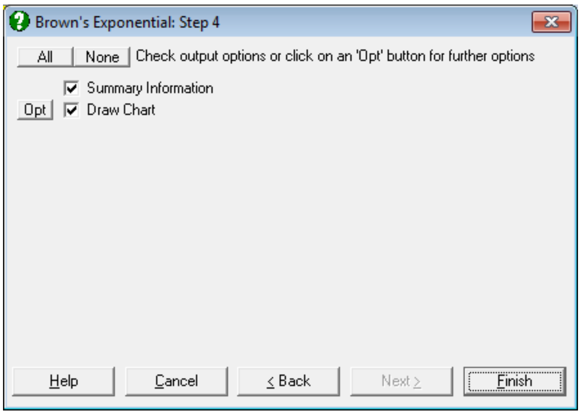


9.2. Forecasting and Smoothing

This section brings together a collection of procedures based on exponential weights moving averages. This technique attempts to track changes in a time series by using newly observed values to update the estimates of the parameters describing the time series. These procedures will work with missing values, by simply using the current forecast as the missing value.

One column is selected for analysis by clicking on [Variable]. Then you will need to select the number of values to forecast and the values of the smoothing constants used in the model. The smoothing constants should usually lie between 0.05 and 0.3. The output from these procedures takes the form of a graphical display of the forecasts and the smoothed values or a table of these values. You will have the choice of editing the initial values which UNISTAT calculates for the exponential weights moving averages.

All forecasting procedures feature an Output Options Dialogue giving access to numeric output tables and graphics. Clicking the [Opt] button situated to the left of the Draw Chart option will place the chart in UNISTAT's Graphics Editor. The chart can be further customised and annotated using the tools available under the graphics window's Edit menu.



Brown's Exponential: Step 4

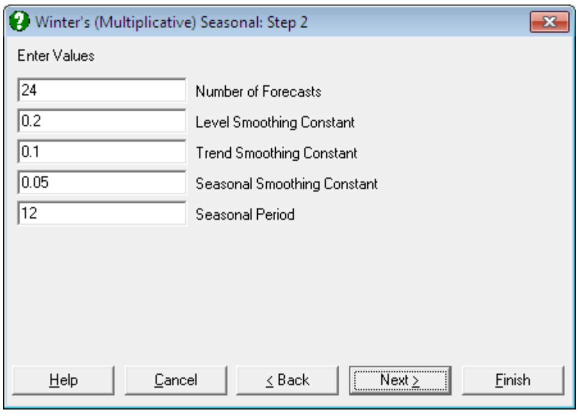
Check output options or click on an 'Opt' button for further options

☑ Summary Information

☑ Draw Chart

9.2.0. Exponential Weights Moving Average

An exponential weights moving average is an average that weights the observed time series values unequally, with more recent observations being weighted more heavily than older observations. This unequal weighting is achieved through smoothing constants that determine how much weight is given to each observation.


Winter's (Multiplicative) Seasonal: Step 2

Enter Values

24 Number of Forecasts
0.2 Level Smoothing Constant
0.1 Trend Smoothing Constant
0.05 Seasonal Smoothing Constant
12 Seasonal Period

If m_{t-1} is the moving average calculated for the first t - 1 points in the series x_t then, given the value x_t, the new moving average is found as:

$$m_t = \lambda x_t + (1 - \lambda) m_{t-1}$$

where:

- λ is the smoothing constant ($0 \leq \lambda \leq 1$).

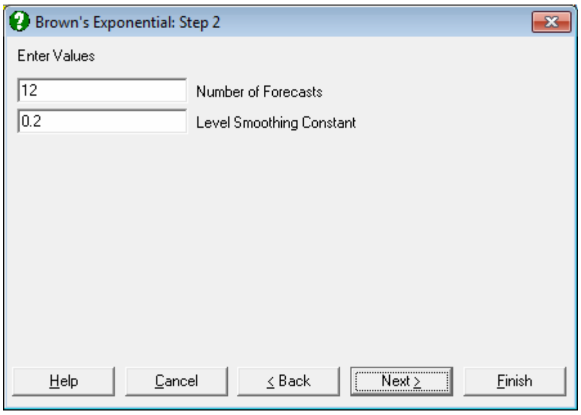
9.2.1. Brown’s Exponential

This simply calculates the exponential weighted average in time. For a data series x_t forecasts are given by:

$$x_{t|t-1} = m_{t-1}$$

where:

- $m_t = \lambda x_t + (1-\lambda)m_{t-1}$ is the level at time t.
- λ is the level smoothing constant.

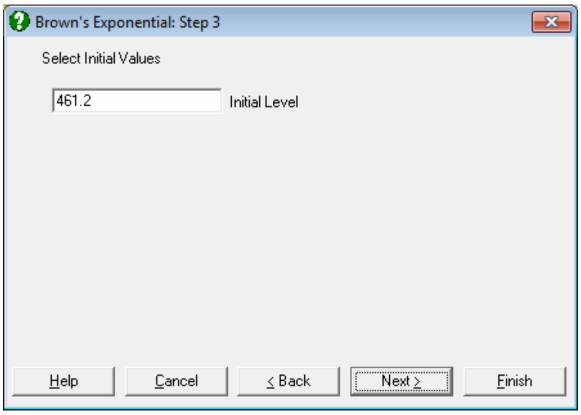

Brown's Exponential: Step 2

Enter Values

12 Number of Forecasts

0.2 Level Smoothing Constant

The initial value m_0 is calculated as the average level in the first quarter of the series.


Brown's Exponential: Step 3

Select Initial Values

461.2 Initial Level

Example

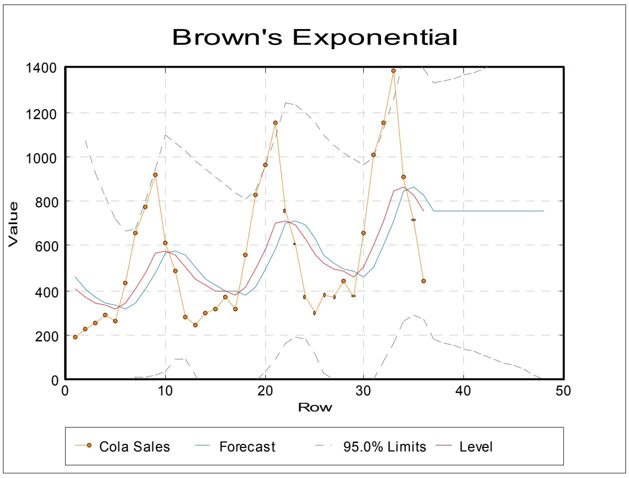
Open TIMESER and select Statistics 2 → Forecasting → Brown's Exponential and select *Cola Sales* (*C2*) as [Variable]. On the following dialogues accept the program's suggestions:

Brown's Exponential

Level Smoothing Constant =	0.2000
Sum of Squares =	3029058.7929

Summary Table

Row	Cola Sales	Forecast	Lower 95%	Upper 95%	Level
1	189.0000	461.2000	*	*	406.7600
2	229.0000	406.7600	-260.1178	1073.6378	371.2080
3	249.0000	371.2080	-179.9829	922.3989	346.7664
…	…	…	…	…	…
34	904.0000	847.5220	265.3873	1429.6567	858.8176
35	715.0000	858.8176	289.7348	1427.9004	830.0541
36	441.0000	830.0541	267.1638	1392.9444	752.2433
37		752.2433	178.5120	1325.9746	
38		752.2433	166.5580	1337.9286	
39		752.2433	153.6842	1350.8023	
…		…	…	…	
46		752.2433	42.4548	1462.0317	
47		752.2433	24.1409	1480.3456	
48		752.2433	5.3481	1499.1385	



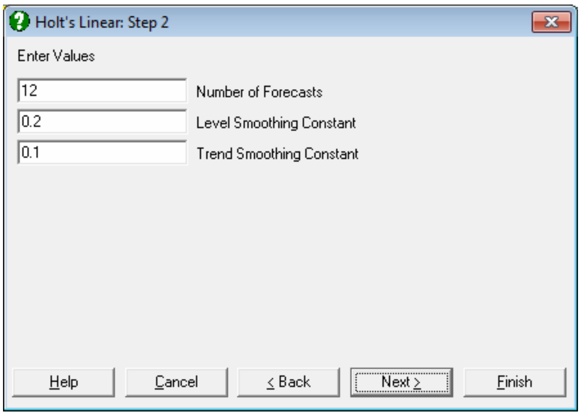
9.2.2. Holt's Linear

This adds a trend component to Brown's Exponential method. For a data series x_t forecasts are given by:

$$\hat{x}_{t|t-1} = m_{t-1} + b_{t-1}$$

where:

- $m_t = \lambda_0 x_t + (1-\lambda_0)(m_{t-1} + b_{t-1})$ is the level at time t.
- $b_t = \lambda_1(m_t - m_{t-1}) + (1-\lambda_1)b_{t-1}$ is the trend at time t.
- λ_0 is the level smoothing constant.
- λ_1 is the trend smoothing constant.



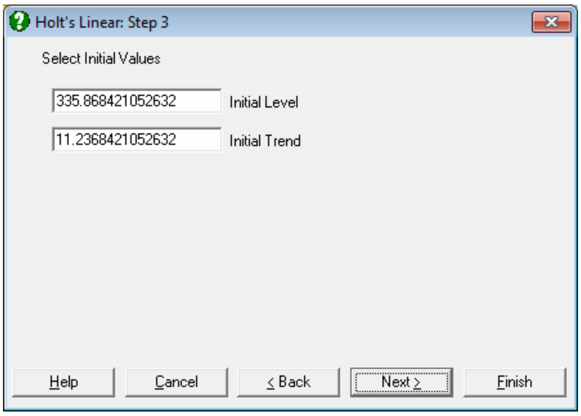
Enter Values

12 Number of Forecasts

0.2 Level Smoothing Constant

0.1 Trend Smoothing Constant

The initial values m_0 and b_0 are calculated by a linear regression on the first half of the series. Double exponential smoothing is a special case of Holt's Linear method with $\lambda_0 = 1-\omega^2$ and $\lambda_1 = (1-\omega)/(1+\omega)$, where $0 \le \omega \le 1$ is the discount factor.



Select Initial Values

335.868421052632 Initial Level

11.2368421052632 Initial Trend

Example

Open TIMESER and select Statistics 2 → Forecasting → Holt's Linear and select *Cola Sales* (*C2*) as [Variable]. On the following dialogues accept the program's suggestions:

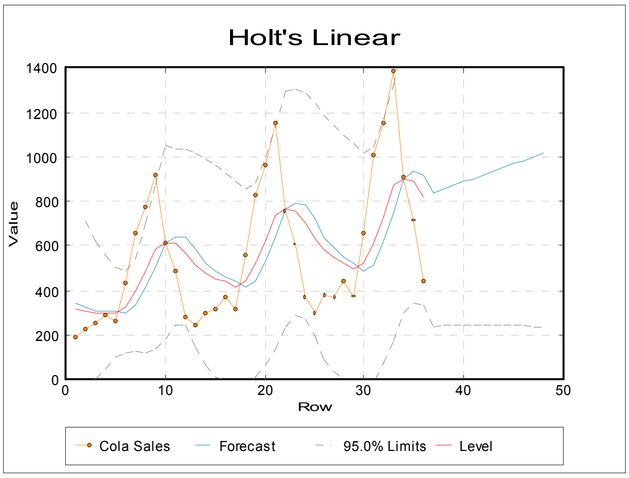
Holt's Linear

Level Smoothing Constant =	0.2000
Trend Smoothing Constant =	0.1000
Sum of Squares =	3224162.8564

Summary Table

Row	Cola Sales	Forecast	Lower 95%	Upper 95%	Level	Trend
1	189.0000	347.1053	*	*	315.4842	8.0747
2	229.0000	323.5589	-63.7919	710.9098	304.6472	6.1836
3	249.0000	310.8307	1.3227	620.3387	298.4646	4.9469
4	289.0000	303.4115	46.5787	560.2443	300.5292	4.6587
5	260.0000	305.1879	103.7364	506.6394	296.1503	3.7550
6	431.0000	299.9053	116.6024	483.2082	326.1242	6.3768
7	660.0000	332.5011	126.2193	538.7829	398.0009	12.9268
8	777.0000	410.9277	119.4922	702.3632	484.1422	20.2483
9	915.0000	504.3904	137.2768	871.5041	586.5123	28.4605
10	613.0000	614.9728	176.8746	1053.0710	614.5782	28.4210
11	485.0000	642.9993	248.2275	1037.7710	611.3994	25.2610
12	277.0000	636.6604	242.5869	1030.7339	564.7283	18.0678
13	244.0000	582.7962	148.1328	1017.4595	515.0369	11.2919
14	296.0000	526.3288	61.2522	991.4054	480.2631	6.6853
15	319.0000	486.9484	14.7847	959.1120	453.3587	3.3263
16	370.0000	456.6850	-11.4321	924.8022	439.3480	1.5926
...	...	...	...	...	...	...

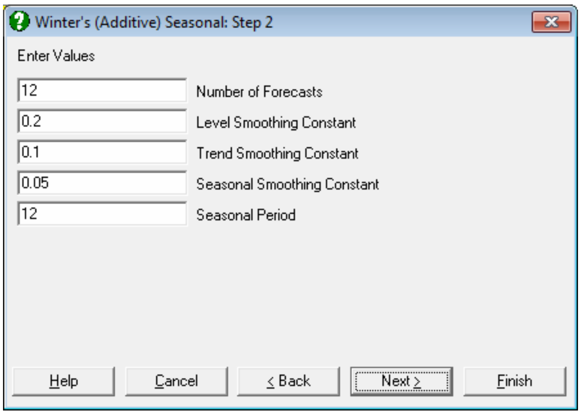
Row	Cola Sales	Forecast	Lower 95%	Upper 95%	Level	Trend
30	660.0000	485.3266	-47.3881	1018.0413	520.2613	-3.0428
31	1004.0000	517.2184	-12.0038	1046.4407	614.5747	6.6928
32	1153.0000	621.2675	70.6462	1171.8888	727.6140	17.3274
33	1388.0000	744.9415	170.8170	1319.0659	873.5532	30.1886
34	904.0000	903.7418	299.2737	1508.2098	903.7934	30.1938
35	715.0000	933.9872	347.2790	1520.6954	890.1897	25.8140
36	441.0000	916.0038	330.7298	1501.2777	821.0030	16.3139
37		837.3170	235.9746	1438.6593		
38		853.6309	239.7592	1467.5026		
39		869.9449	242.5799	1497.3098		
40		886.2588	244.4973	1528.0203		
41		902.5728	245.5710	1559.5745		
42		918.8867	245.8581	1591.9153		
43		935.2006	245.4136	1624.9877		
44		951.5146	244.2895	1658.7397		
45		967.8285	242.5346	1693.1225		
46		984.1425	240.1951	1728.0899		
47		1000.4564	237.3138	1763.5991		
48		1016.7704	233.9305	1799.6103		



9.2.3. Winters' Additive Seasonal

This adds an additive seasonal component to Holt's Linear method. For a data series x_t forecasts are given by:

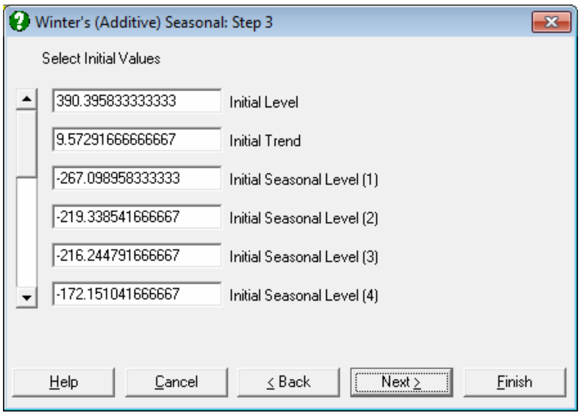
$$\hat{x}_{t|t-1} = m_{t-1} + b_{t-1} + c_{t-s}$$



12 Number of Forecasts
0.2 Level Smoothing Constant
0.1 Trend Smoothing Constant
0.05 Seasonal Smoothing Constant
12 Seasonal Period

where:

- $m_t = \lambda_0(x_t - c_{t-s}) + (1-\lambda_0)(m_{t-1} + b_{t-1})$ is the level at time t.
- $c_t = \lambda_2(y_t - m_t) + (1-\lambda_2)c_{t-s}$ is the trend at time t.
- $c_t = \lambda_2(y_t - m_t) + (1-\lambda_2)c_{t-s}$ is the seasonal component at time t.
- λ_0 is the level smoothing constant.
- λ_1 is the trend smoothing constant.
- λ_2 is the seasonal smoothing constant.
- s is the season period.



Select Initial Values

390.395833333333 Initial Level

9.57291666666667 Initial Trend

-267.098958333333 Initial Seasonal Level (1)

-219.338541666667 Initial Seasonal Level (2)

-216.244791666667 Initial Seasonal Level (3)

-172.151041666667 Initial Seasonal Level (4)

The initial values are based on the complete data series. Suppose the series includes complete data for L seasons. The initial trend and level values are given by:

$$b_0 = \frac{\bar{y}_L - \bar{y}_1}{(L-1)s}$$

$$m_0 = \bar{y}_1 - \frac{s}{2} b_0$$

where $\bar{y}_i$ is the average level in season i.

The initial seasonal components $c_{1-s},\ldots,c_0$ are given by the average value of S_t, the observed value minus the expected value if no seasonal component is used. That is:

$$S_t = x_t - \left(\bar{y}_i - \left(\frac{s+1}{2} - j \right) b_0 \right)$$

where:

- i is the year in which t falls.
- j is season with in t falls.

Example

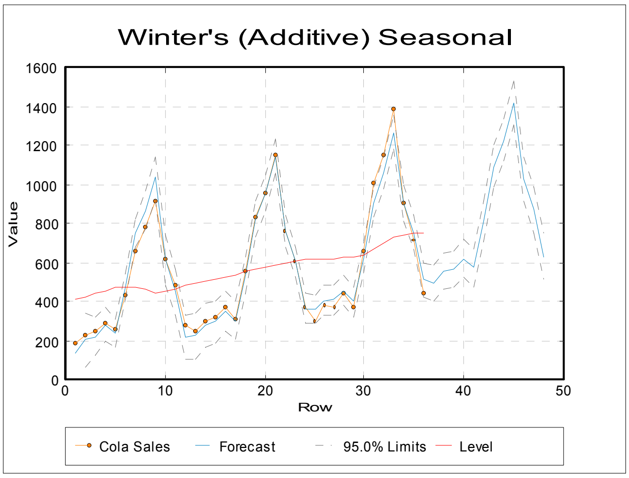
Open TIMESER and select Statistics 2 → Forecasting → Winters’ Additive Seasonal and select *Cola Sales* (*C2*) as [Variable].

Winters' Additive Seasonal

Level Smoothing Constant =	0.2000
Trend Smoothing Constant =	0.1000
Seasonal Smoothing Constant =	0.0500
Seasonal Period =	12
Sum of Squares =	92856.6862

Summary Table

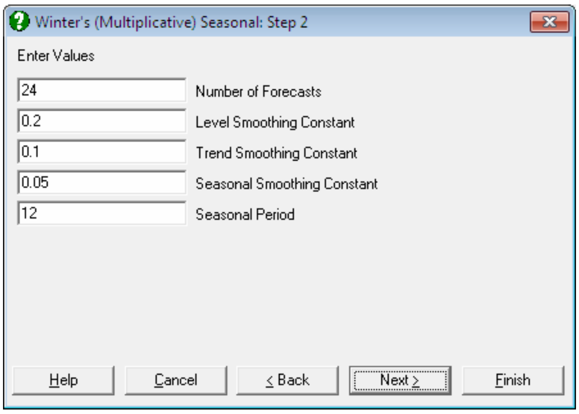
Row	Cola Sales	Forecast	Lower 95%	Upper 95%	Level	Trend	Seasonal
1	189.0000	132.8698	*	*	411.1948	10.6955	-264.8537
2	229.0000	202.5518	65.0353	340.0683	427.1800	11.2245	-218.2806
3	249.0000	222.1597	121.0029	323.3164	443.7725	11.7613	-215.1712
...	…	…	…	…	…	…	…
34	904.0000	906.4846	814.8228	998.1464	744.1618	13.7218	161.7265
35	715.0000	755.1910	666.0461	844.3359	749.8454	12.9180	-4.3002
36	441.0000	512.3320	422.9208	601.7432	748.4969	11.4913	-253.2846
37		493.3129	401.5309	585.0949			
38		552.6437	458.9494	646.3381			
39		567.1610	471.4072	662.9148			
40		622.7754	524.8243	720.7266			
41		572.9176	472.6404	673.1948			
42		808.3282	705.6048	911.0516			
43		1092.9881	987.7069	1198.2693			
44		1226.4534	1118.5106	1334.3961			
45		1416.6969	1305.9963	1527.3975			
46		1025.1368	911.5891	1138.6844			
47		870.6014	754.1240	987.0787			
48		633.1083	513.6246	752.5920			



9.2.4. Winters' Multiplicative Seasonal

This adds a mulitiplicative seasonal component to Holt's Linear method. For a data series x_t forecasts are given by:

$$\hat{x}_{t|t-1} = (m_{t-1} + b_{t-1})c_{t-s}$$



Enter Values

24 Number of Forecasts

0.2 Level Smoothing Constant

0.1 Trend Smoothing Constant

0.05 Seasonal Smoothing Constant

12 Seasonal Period

where:

- $m_t = \lambda_0 \dfrac{x_t}{c_{t-s}} + (1-\lambda_0)(m_{t-1} + b_{t-1})$ is the level at time t,
- $b_t = \lambda_1(m_t - m_{t-1}) + (1-\lambda_1)b_{t-1}$ is the trend at time t,
- $c_t = \lambda_2 \dfrac{y_t}{m_t} + (1-\lambda_2)c_{t-s}$ is the seasonal component at time t,
- λ_0 is the level smoothing constant,
- λ_1 is the trend smoothing constant,
- λ_2 is the seasonal smoothing constant,
- s is the season period.

The initial values are based on the complete data series. Suppose the series includes complete data for L seasons. The initial trend and level values are given by:

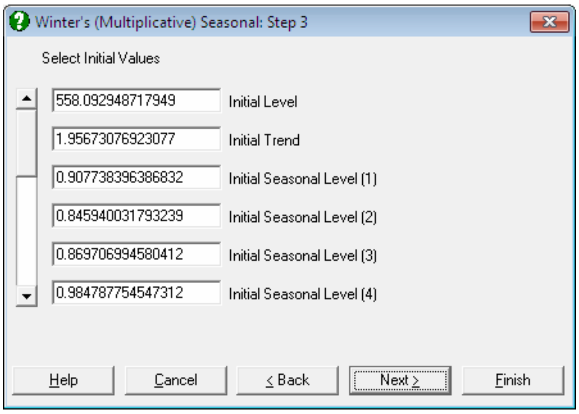
$$b_0 = \frac{\bar{y}_L - \bar{y}_1}{(L-1)s}$$

$$m_0 = \bar{y}_1 - \frac{s}{2} b_0$$

where:

- $\bar{y}_i$ is the average level in season i.

The initial seasonal components c_{1-s}, ..., c_0 are given by the average value of S_t, the observed value divided by the expected value if no seasonal component is used. That is:



Winter's (Multiplicative) Seasonal: Step 3

Select Initial Values

558.092948717949 Initial Level

1.95673076923077 Initial Trend

0.907738396386832 Initial Seasonal Level (1)

0.845940031793239 Initial Seasonal Level (2)

0.869706994580412 Initial Seasonal Level (3)

0.984787754547312 Initial Seasonal Level (4)

$$S_t = \frac{x_t}{\bar{y}_i - [(s+1)/2 - j]b_0}$$

where:

- i is the year in which t falls.
- j is season with in t falls.

Example

Consider the Tasty Cola Sales given in Table 6.4 Bowerman, Bruce L. & Richard T. O'Connell (1987). Open TIMESER and select Statistics 2 → Forecasting → Winters' Multiplicative Seasonal and select *Cola Sales* (*C2*) as [Variable]. On the second dialogue enter:

- **12** Number of Forecasts.
- **.2** Level Smoothing Constant.
- **.15** Trend Smoothing Constant.
- **.05** Seasonal Smoothing Constant.
- **12** Seasonal Period.

Select the default values on the third dialogue. Select the Summary Table on the Output Options Dialogue to obtain the following results. The table is shortened here for space considerations.

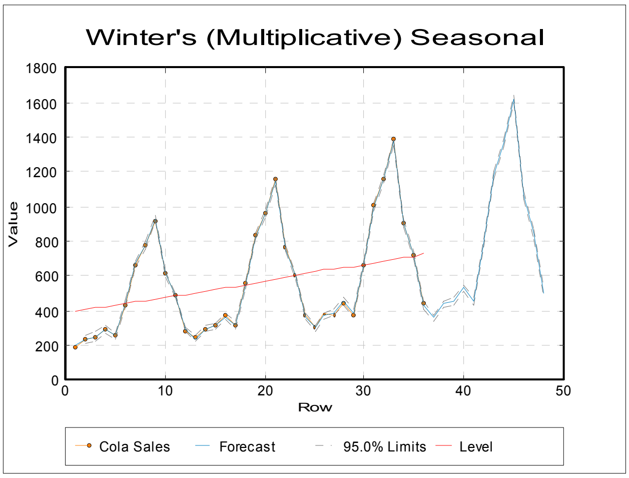
Winters' Multiplicative Seasonal

Level Smoothing Constant =	0.2000
Trend Smoothing Constant =	0.1500
Seasonal Smoothing Constant =	0.0500
Seasonal Period =	12
Sum of Squares =	2254.8254

Summary Table

Row	Cola Sales	Forecast	Lower 95%	Upper 95%	Level	Trend	Seasonal
1	189.0000	193.6418	*	*	398.0512	9.2853	0.4837
2	229.0000	238.1693	214.6797	261.6590	404.2000	8.8148	0.5838
3	249.0000	248.7039	217.7488	279.6590	413.1132	8.8296	0.6022
4	289.0000	291.9661	271.1169	312.8153	421.5858	8.9831	0.6909
5	260.0000	252.2613	233.9958	270.5268	433.2106	9.2472	0.5866
6	431.0000	440.9282	419.8437	462.0127	440.4653	9.0480	0.9956
7	660.0000	667.1568	645.5183	688.7952	448.5488	8.9515	1.4835
8	777.0000	774.4063	754.1714	794.6412	457.8068	8.9822	1.6929
9	915.0000	927.4468	909.2720	945.6216	465.5361	8.8569	1.9858
10	613.0000	611.8198	593.9592	629.6805	474.5760	8.8752	1.2898
11	485.0000	487.0187	470.7199	503.3175	483.0504	8.8351	1.0072
12	277.0000	292.4919	277.2284	307.7553	486.6750	8.3141	0.5934
13	244.0000	239.4143	220.1039	258.7248	496.8852	8.5037	0.4840
14	296.0000	295.0392	275.4274	314.6510	505.7181	8.5366	0.5839
15	319.0000	309.6720	291.1731	328.1710	517.3527	8.8464	0.6029
16	370.0000	363.5268	343.7310	383.3225	528.0731	9.0338	0.6913
17	313.0000	315.0635	295.0703	335.0568	536.4034	8.9634	0.5864
18	556.0000	542.9897	523.6656	562.3139	547.9802	9.2248	0.9966
19	831.0000	826.6347	806.6055	846.6638	557.7935	9.2836	1.4839
20	960.0000	960.0143	940.6599	979.3688	567.0755	9.2835	1.6929
21	1152.0000	1144.5310	1126.1433	1162.9188	577.1112	9.3587	1.9863
22	759.0000	756.4226	738.4717	774.3735	586.8695	9.3987	1.2900
23	607.0000	600.5685	583.2110	617.9260	597.5453	9.5264	1.0076
24	371.0000	360.2125	342.9294	377.4955	610.7077	9.8900	0.5941
25	298.0000	300.3973	281.9785	318.8161	619.6072	9.7909	0.4839
26	378.0000	367.4819	349.3146	385.6493	633.0010	10.1512	0.5845
27	373.0000	387.7550	368.5888	406.9211	638.2576	9.6617	0.6020

Row	Cola Sales	Forecast	Lower 95%	Upper 95%	Level	Trend	Seasonal
28	443.0000	447.9353	427.2584	468.6123	646.4916	9.5190	0.6910
29	374.0000	385.4918	364.6530	406.3307	653.3489	9.0469	0.5858
30	660.0000	660.1998	638.4242	681.9754	662.3557	9.0409	0.9967
31	1004.0000	996.3279	975.2618	1017.3940	672.4306	9.1960	1.4844
32	1153.0000	1154.0077	1133.2125	1174.8028	681.5076	9.1781	1.6930
33	1388.0000	1371.9822	1351.7913	1392.1730	692.2985	9.4201	1.9873
34	904.0000	905.2261	885.0484	925.4038	701.5284	9.3915	1.2899
35	715.0000	716.3717	696.7190	736.0244	710.6477	9.3507	1.0076
36	441.0000	427.7323	408.5459	446.9188	724.4651	10.0207	0.5948
37		355.4162	335.2428	375.5896			
38		435.1913	414.5975	455.7850			
39		454.2166	433.1702	475.2630			
40		528.3472	506.8178	549.8765			
41		453.7447	431.7040	475.7853			
42		781.9787	759.4004	804.5570			
43		1179.5346	1156.3941	1202.6751			
44		1362.2084	1338.4830	1385.9339			
45		1618.9822	1594.6506	1643.3139			
46		1063.7802	1038.8227	1088.7376			
47		841.0292	815.4278	866.6306			
48		502.4411	476.1790	528.7033			



9.2.5. Dixon-Grubbs-Neumann Outlier Tests

A single column is selected as [Variable]. For $3 \le n \le 25$ the Dixon statistic, Q is given. The data is considered in an ordered sequence and Q is calculated as:

$$Q = \begin{cases} \left|\dfrac{x_1 - x_2}{x_1 - x_n}\right|, & 3 \le n \le 7 \\ \left|\dfrac{x_1 - x_2}{x_1 - x_{n-1}}\right|, & 8 \le n \le 10 \\ \left|\dfrac{x_1 - x_3}{x_1 - x_{n-1}}\right|, & 11 \le n \le 13 \\ \left|\dfrac{x_1 - x_3}{x_1 - x_{n-2}}\right|, & 14 \le n \le 25 \end{cases}$$

UNISTAT reports the maximum Q and minimum Q. These two values can be considered as the two different ways of sorting the data. The maximum Q tests the largest value in the column and the minimum Q tests the smallest value in the column. For probability values of the Dixon statistic refer to tables.

For $n > 25$ the Grubbs statistic is given instead of the Dixon statistic. The test statistic Q is calculated as follows:

$$Q = \frac{|x_m - \bar{x}|}{s}$$

where x_m is the maximum observation. The Grubbs test requires that the data is approximately normally distributed. The 1-tail probability is computed according to the following inequality:

$$Q > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{(\alpha/(2n)),n-2)}}{n - 2 + t^2_{(\alpha/(2n)),n-2)}}}$$

The Neumann trend statistic T is calculated as follows:

$$T = \frac{\sum (x_i - x_{i+1})^2}{\sum (x_i - \bar{x})^2}$$

This can be thought of as the mean square successive difference divided by the variance. The approximate tail probability for the Neumann trend statistic is calculated from the following Z transformation, which is assumed to follow a N(0,1) distribution.

$$Z = \left(\frac{2-T}{2}\right)\sqrt{\frac{(n-1)(n+1)}{n-2}}$$

See Sachs, L. (1984). For Dixon and Grubbs tests see pp. 277-79 and for Neumann test pp. 373-75.

Example

Open TIMESER and select Statistics 2 → Forecasting → 9.2.5. Dixon-Grubbs-Neumann and select *Cola Sales* (*C2*) as [Variable].

Dixon-Grubbs-Neumann Outlier Tests

```
Data variable: Cola Sales
Number of Cases = 36
Grubbs outlier test is selected.

  Maximum deviation from mean / Standard Deviation:
                                        Q =    2.6584
                       1-Tail Probability =    0.1879
                            Neumann trend =    0.3995
                   Approximate Probability =    0.0000
```

Open TIMESER and select Statistics 2 → Forecasting → 9.2.5. Dixon-Grubbs-Neumann and select *Failure time* (*C7*) as [Variable].

Dixon-Grubbs-Neumann Outlier Tests

```
Data variable: Failure time
Number of Cases = 8
Dixon outlier test is selected.

                   Q(min) =    0.1173 (X(1) - X(2)) / (X(1) - X(N-1))
                   Q(max) =    0.2582 (X(N) - X(N-1)) / (X(N) - X(2))
            Neumann trend =    0.2101
  Approximate Probability =    0.0019
```

9.3. Quality Control

9.3.0. Overview

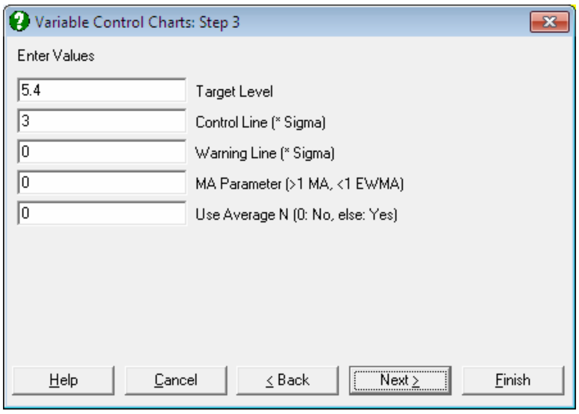
9.3.0.1. Control Charts

A control chart consists of a lower control limit (LCL), an upper control limit (UCL) and a centre line (CL). Samples are taken in time series order and plotted on the control chart. The centre line represents no deviation from the sample characteristic. The control limits are selected so that if the process is in control, almost all of the points will fall within the control limits. A point falling outside the control limits is an evidence of the process being out of control. Even if all the points lie within the control limits, the process can still be out of control. This condition exists when the points fall in a systematic manner, e.g. if the last 10 samples were above (or below) the centre line.

A control chart is a visual hypothesis test. The null hypothesis is that “the variable’s true value is the target level”. The control lines are set such that the sample would not take values outside the control lines. The typical value for the control lines is target level ±3 standard deviations, so if the null hypothesis holds then only 0.2% of the sample values should lie outside the control lines.

Selection of data columns to be analysed is specific to types of control charts (see 9.3.1. Variable Control Charts and 9.3.2. Attribute Control Charts). The following sections summarise the aspects common to most control charts.

9.3.0.2. Control Chart Inputs



Variable Control Charts: Step 3

Enter Values

5.4 Target Level

3 Control Line (* Sigma)

0 Warning Line (* Sigma)

0 MA Parameter (>1 MA, <1 EWMA)

0 Use Average N (0: No, else: Yes)

UNISTAT will suggest values for these parameters based on the data.

Target Level: The centre line for the chart.

Sigma: The standard error of the variable. On some charts this is related to the target level and hence not requested.

Control Line: The control lines will be at target level ± control line times sigma.

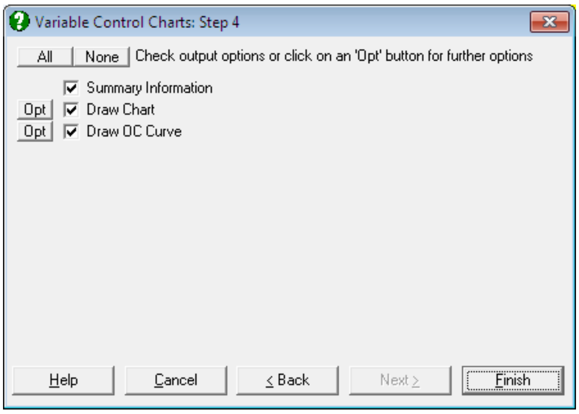
Warning Line: Warning lines will be added at target level ± warning line times sigma. If this value is zero then warning lines will not be drawn. A typical level for warning lines is 2 times sigma then you would expect 5% of points to lie outside the warning lines.

MA Parameter: UNISTAT will add either the standard moving average or the exponential weights moving average of the data to the chart. The following ranges of values can be entered:

0 No moving average line,
0-1 Exponential weights moving average is added to the chart,
1- Standard moving average of data is added to the chart.

Use Average N: The control / warning lines will be calculated using the average value of sample size. This will result in the control / warning lines being straight.

9.3.0.3. Control Chart Output Options


Variable Control Charts: Step 4

Check output options or click on an 'Opt' button for further options

☑ Summary Information
☑ Draw Chart
☑ Draw OC Curve

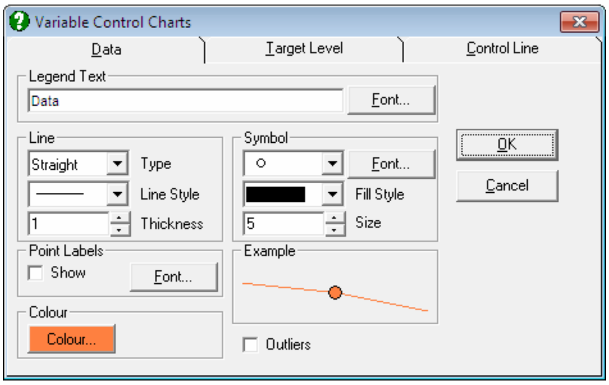
After the calculations for a chart have been made, two or three output options become available. An OC Curve option is not available for all control charts:

Draw Chart: Draws the quality control chart (see 9.3.0.4. Control Chart Options).

Summary Information: For each sample, the control variable, the LCL and the UCL are displayed in a table. Also, any samples outside the control limits are marked.

Draw OC Curve: Draw the Operating Characteristics Curve for the data (see 9.3.0.5. Operating Characteristic Curve).

9.3.0.4. Control Chart Options



Variable Control Charts

Legend Text: Data

Line: Type: Straight; Line Style; Thickness: 1

Symbol: ○; Fill Style; Size: 5

Point Labels: □ Show

□ Outliers

In the Edit → Data Series dialogue, the Point Labels and Outliers check boxes work in the following way:

Data Values: If Outliers and Point Labels are both checked then only outlying points will be labelled. If Point Labels alone is checked then each point is labelled.

Target Level: If Point Labels is checked then the target level will be shown on the right Y-axis. The Outliers check box has no effect.

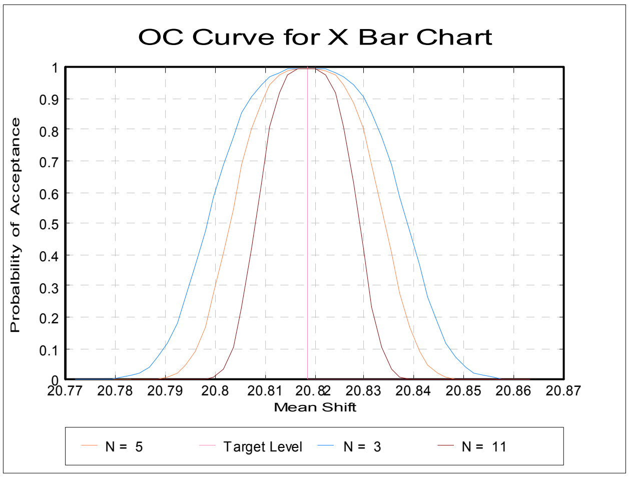
Control Lines: If Point Labels is checked then the final control level is shown on the right Y-axis. This is most useful when average N values are used. If Outliers is checked then a box is drawn around the points which lie outside the control lines.

Warning Lines: If Point Labels is checked then the final warning level is shown on the right Y-axis. This is most useful when average N values are used. If

Outliers is checked then a box is drawn around the points which lie outside the warning lines.

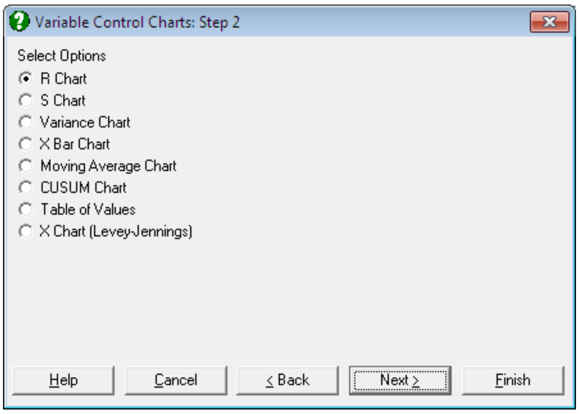
9.3.0.5. Operating Characteristic Curve

The Operating Characteristic Curve (or OC curve) shows how sensitive the current control chart is. It is a graph of the probability of accepting a sample with a true level as shown on the x-axis. This is usually referred to as the Type II or beta error probability.



The OC chart is generated on an X Bar Chart. This shows that with the current average sample size of 5, there is a 70% chance of a sample (with an actual mean of 20.80) lying outside the control lines. If the sample size was increased to 11 then there would be 98% chance of the point lying outside the control limits.

9.3.1. Variable Control Charts



Variable Control Charts: Step 2

Select Options

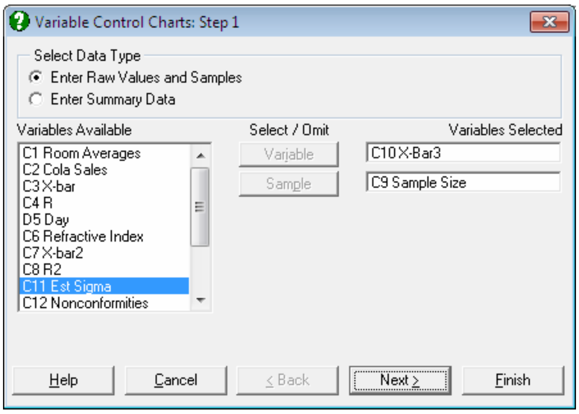
● R Chart
○ S Chart
○ Variance Chart
○ X Bar Chart
○ Moving Average Chart
○ CUSUM Chart
○ Table of Values
○ X Chart (Levey-Jennings)

These methods (which are also known as Shewhart charts) attempt to control quality characteristics that can be expressed as numerical measures. Examples include the diameter of a piston, the weight of beverage placed in a bottle, etc. The general approach is straightforward: simply extract samples of a certain size from the process, produce charts for the variability of those samples and their closeness to target specifications. In an application, it is general practice to ensure that the variability is in control before examining the central tendency. The variability is controlled with the R Chart, S Chart and Variance Chart in UNISTAT. The central tendency is controlled with the X Bar Chart, Moving Average Charts and CUSUM Chart.

9.3.1.0. Overview

Two types of data can be analysed.

9.3.1.0.1. Raw Values and Samples



Variable Control Charts: Step 1

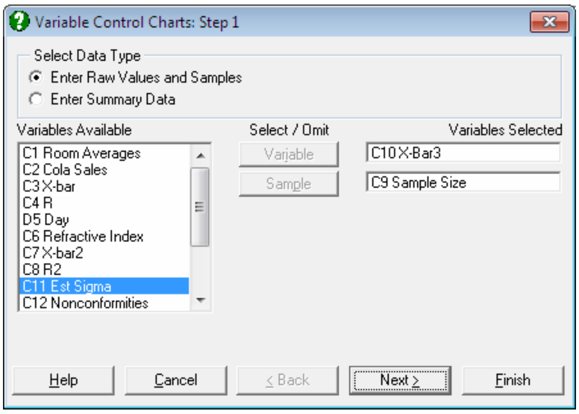
Select Data Type
● Enter Raw Values and Samples
○ Enter Summary Data

Variables Selected
Variable: C10 X-Bar3
Sample: C9 Sample Size

When this method is used, you will need to select one column that contains the data values by clicking on [Variable] and a second column that specifies which sample the value is associated with by clicking on [Sample]. The sample column is like a factor column, in that the levels within the column are used to classify the data into groups, but there is also an order implied in these classes. It is assumed that sample numbers with higher values are taken later in time, but the sample numbers need not be strictly sequential. It is also possible to use a date column as the sample column. In this case each measurement taken on the same day will be in the same sample.

It must be noted here that all Variable Control Charts involve *grouping* of the original raw data and the points on the plot will correspond to *samples* rather than the individual observations. For instance, if you have a [Variable] column with 200 observations and a [Sample] column which also has 200 rows but only 5 distinct values, then Variable Control Charts will produce a graph with 5 points only, corresponding to each sample. One exception to this is X Chart (Levey-Jennings), which does not involve a grouping of observations. It will plot all 200 observations and the [Sample] column will not play a role in the appearance of the curve. However, we assume that there is an underlying *sample* here, in parallel with other Variable Control Charts. In X Chart, the [Sample] column can be used for labelling purposes, particularly in marking which outlying points belong to which sample.

9.3.1.0.2. Summary Data



Variable Control Charts: Step 1

Select Data Type

● Enter Raw Values and Samples

○ Enter Summary Data

Variables Available: C1 Room Averages, C2 Cola Sales, C3 X-bar, C4 R, D5 Day, C6 Refractive Index, C7 X-bar2, C8 R2, C11 Est Sigma, C12 Nonconformities

Variables Selected: C10 X-Bar3, C9 Sample Size

When this data option is selected, each row in the data matrix is assumed to represent a sample. You may select up to four columns containing sample sizes, means, standard errors and the sample ranges by clicking on [Size], [Mean], [Std Dev] and [Range] respectively.

Only one of standard error and range columns must be selected. If all the variables are not selected from the Variables Available list, output options will be reduced in the following way:

Sample size not specified: UNISTAT will prompt for a common sample size to apply to all samples.

Mean not specified: The X Bar Chart, Moving Average Charts and CUSUM Chart will not be available.

Standard error not specified: The S Chart and Variance Chart will not be available.

Range not specified: The R Chart will not be available.

If the columns have been selected but within these columns values are missing, UNISTAT behaves in the following way. If the sample size is missing then the average sample size is used. If the mean value is missing then this row will be missing on X Bar Chart, Moving Average Charts and CUSUM Chart. If the standard error is missing then the row will be missing on S Chart and Variance Chart. If the range is missing then the row will be missing on an R Chart. If both standard error and range are missing then the row is considered to be missing.

9.3.1.0.3. Standard Error Estimation

In X Bar Chart, Moving Average Charts and CUSUM Chart the standard error for each sample is required. If the standard error for a sample is not given then it is estimated from the range within the given sample.

9.3.1.1. R Chart

In R Chart, the sample ranges are plotted in order to control the variability of a variable. The sample sizes should be small and fairly consistent. This chart only uses the two most extreme values from a sample and so is weaker than the S Chart. Also the true target level varies with the sample size. The R Chart does not account for this, and use the average sample size to estimate the target level. So samples should have sizes near this value.

$$UCL_R = T + C \times \frac{d_3}{d_2} T$$

$$CL_R = T$$

$$LCL_R = T - C \times \frac{d_3}{d_2} T$$

where:

- T is the target level. UNISTAT gives the mean range of the samples as the default value.
- C is the control line parameter.
- d_2, d_3 are known parameters dependent on sample size such that:

$$R \approx d_2 \sigma$$

$$\sigma_R \approx d_3 \sigma$$

Example

Table 7.1 on p. 189 from Banks, Jerry (1989). X-bar and R values for Socket Dimensions are given in centimetres.

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. Select the data option 2 Enter Summary Data and *X-Bar* (*C3*) as [Mean] and *R* (*C4*) as [Range]. From the next dialogue select the sample size as 5 and from the next one R Chart as the output option. At the next dialogue leave the default values as follows:

- **2.7136E-02** Target Level
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0** MA Parameter (>1 MA, 1< EWMA)
- **0** Use Average N (0 No, Else Yes)

And finally, from the Output Options Dialogue click [Finish].

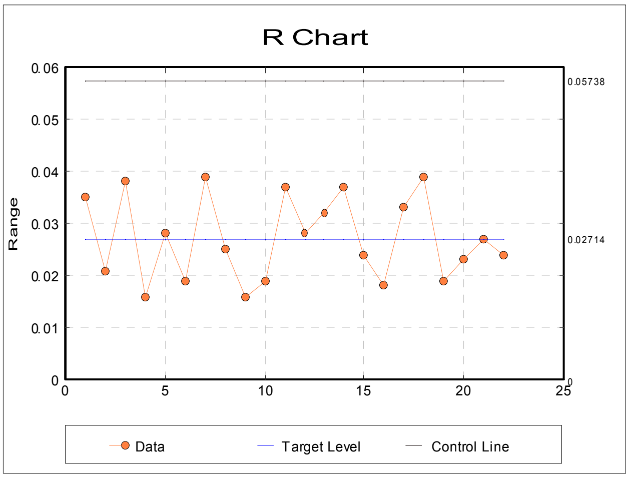
Variable Control Charts

R Chart

Size = 5.0000
Mean: X-bar
Range: R
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
1	0.0350	0.0000	0.0574
2	0.0210	0.0000	0.0574
3	0.0380	0.0000	0.0574
4	0.0160	0.0000	0.0574
5	0.0280	0.0000	0.0574
6	0.0190	0.0000	0.0574
7	0.0390	0.0000	0.0574
8	0.0250	0.0000	0.0574
9	0.0160	0.0000	0.0574
10	0.0190	0.0000	0.0574
11	0.0370	0.0000	0.0574
12	0.0280	0.0000	0.0574
13	0.0320	0.0000	0.0574
14	0.0370	0.0000	0.0574
15	0.0240	0.0000	0.0574
16	0.0180	0.0000	0.0574
17	0.0330	0.0000	0.0574
18	0.0390	0.0000	0.0574
19	0.0190	0.0000	0.0574
20	0.0230	0.0000	0.0574
21	0.0270	0.0000	0.0574
22	0.0240	0.0000	0.0574

Target Level =	0.0271
Average Sample Size =	5.0000
Control Limits =	3.0000 × Sigma
Sigma =	0.0117



This shows that the range (and hence the variability) is under control and we can proceed to examine the central tendency.

9.3.1.2. S Chart

In S Chart the sample standard errors are plotted in order to control the variability of a variable. For large sample sizes this chart is more powerful than the R Chart.

$$UCL_S = T + C \times \frac{T}{c_4}\sqrt{1-c_4^2}$$

$$CL_S = T$$

$$LCL_S = T - C \times \frac{T}{c_4}\sqrt{1-c_4^2}$$

where:

- T is the target level. UNISTAT gives the average standard deviation of the samples as the default value,
- C is the control line parameter.
- c_4 is a known parameter dependent on sample size, n.

$$c_4 = \sqrt{\frac{2}{n-1}}\frac{\left[(n-2)/2\right]!}{\left[(n-3)/2\right]!}$$

Example

Table 7.5 on p. 214 from Banks, Jerry (1989). Refractive Index of Fiber Optic Cable is given.

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. Select the data option 1 Enter Raw Values and Samples and *Day* (*C5*) as [Sample] and *Refractive Index* (*C6*) as [Variable]. From the next dialogue select S Chart and leave the default values as follows:

- **1.2715** Target Level
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0** MA Parameter (>1 MA, 1< EWMA)
- **0** Use Average N (0 No, Else Yes)

From the final output dialogue click [Finish] to obtain the following graph, which shows that the variability is under control.

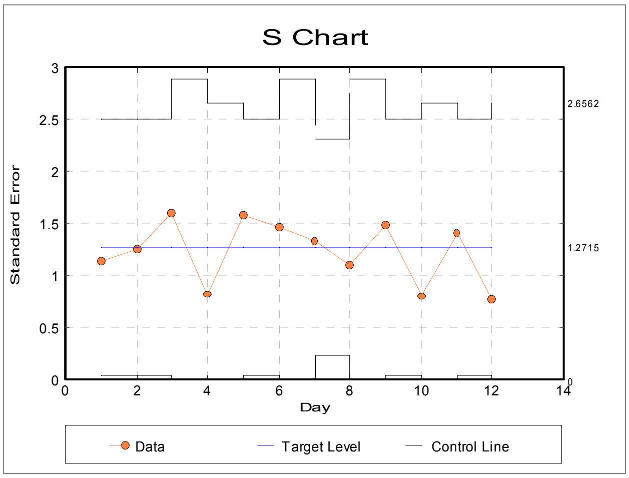
Variable Control Charts

S Chart

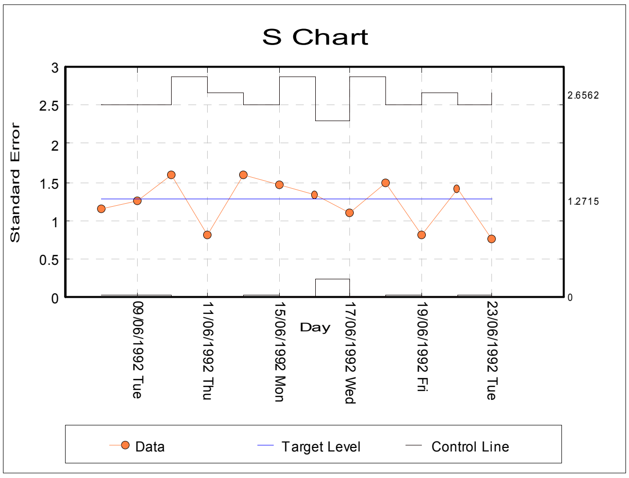
Sample: Day
Variable: Refractive Index
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
08/06/1992 Mon	1.1466	0.0386	2.5044
09/06/1992 Tue	1.2501	0.0386	2.5044
10/06/1992 Wed	1.5948	0.0000	2.8813
11/06/1992 Thu	0.8204	0.0000	2.6562
12/06/1992 Fri	1.5870	0.0386	2.5044
15/06/1992 Mon	1.4617	0.0000	2.8813
16/06/1992 Tue	1.3307	0.2353	2.3077
17/06/1992 Wed	1.0996	0.0000	2.8813
18/06/1992 Thu	1.4828	0.0386	2.5044
19/06/1992 Fri	0.8019	0.0000	2.6562
22/06/1992 Mon	1.4166	0.0386	2.5044
23/06/1992 Tue	0.7668	0.0000	2.6562

Target Level =	1.2715
Average Sample Size =	5.4167
Control Limits =	3.0000 × Sigma
Sigma =	1.2715



It is possible to edit the graph settings to display dates on the x-axis, instead row numbers.



9.3.1.3. Variance Chart

This is simply the S Chart with all the values squared. Keeping the selections for the previous example we obtain:

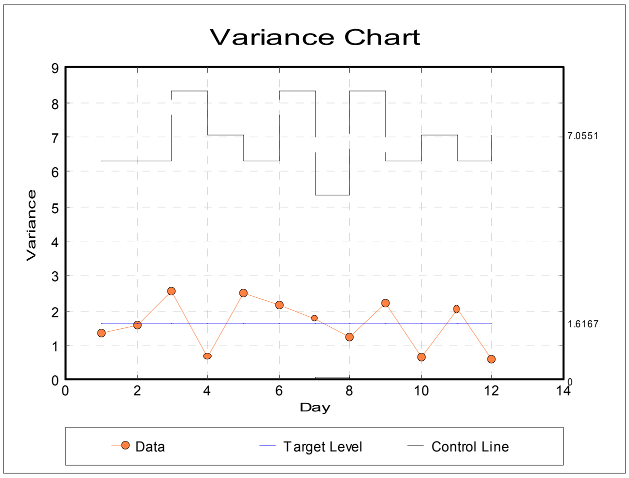
Variable Control Charts

Variance Chart

Sample: Day
Variable: Refractive Index
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
08/06/1992 Mon	1.3147	0.0015	6.2719
09/06/1992 Tue	1.5627	0.0015	6.2719
10/06/1992 Wed	2.5433	0.0000	8.3017
11/06/1992 Thu	0.6730	0.0000	7.0551
12/06/1992 Fri	2.5187	0.0015	6.2719
15/06/1992 Mon	2.1367	0.0000	8.3017
16/06/1992 Tue	1.7707	0.0554	5.3253
17/06/1992 Wed	1.2092	0.0000	8.3017
18/06/1992 Thu	2.1987	0.0015	6.2719
19/06/1992 Fri	0.6430	0.0000	7.0551
22/06/1992 Mon	2.0067	0.0015	6.2719
23/06/1992 Tue	0.5880	0.0000	7.0551

Target Level =	1.6167
Average Sample Size =	5.4167
Control Limits =	3.0000 × Sigma
Sigma =	1.6167



9.3.1.4. X Bar Chart

In this chart the sample means are plotted to control the central tendency of a variable. X Bar Chart assumes that the target level and sigma values have been specified as targets to achieve.

$$UCL_{\overline{X}} = T + C \times \frac{S}{\sqrt{n}}$$

$$CL_{\overline{X}} = T$$

$$LCL_{\overline{X}} = T - C \times \frac{S}{\sqrt{n}}$$

where:

- T is the target level. UNISTAT gives the overall mean of the samples as the default value. The weighted mean of the sample means.
- S is the sigma value. UNISTAT gives the overall sigma of the samples as the default value.
- C is control line parameter,
- n is the sample size.

Example

Table 7.2 on p. 192 from Banks, Jerry (1989). This is the continuation of the power socket manufacturing example in section 9.3.1.1. R Chart.

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. Select the data option 2 Enter Summary Data. Select *X-Bar2* (*C7*) as [Mean] and *R2* (*C8*) as [Range]. From the next dialogue select the sample size as 5. The variability is known to be under control, so select X Bar Chart at the next dialogue. The above data is continuation data from a process which has been under control previously, with calculated target levels (see example in 9.3.1.1. R Chart). So, at the next dialogue overwrite the default values with the following values which were obtained when the process was under control:

- **20.8185** Target Level
- **0.0116** Sigma
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0** MA Parameter (>1 MA, <1 EWMA)
- **0** Use Average N (0 No, Else Yes)

and from the final output dialogue click [Finish] to obtain the following chart:

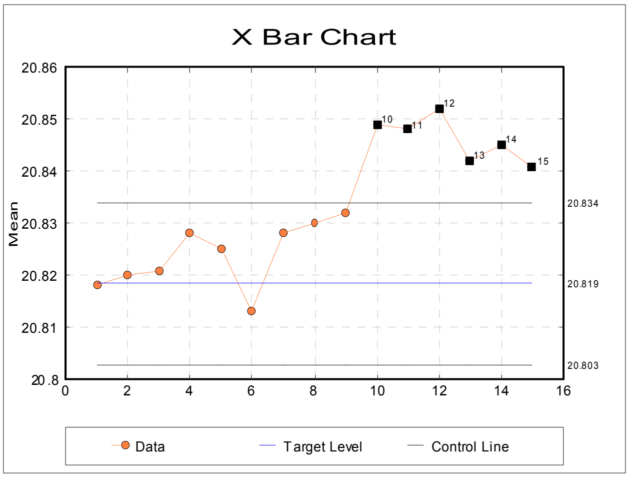
Variable Control Charts

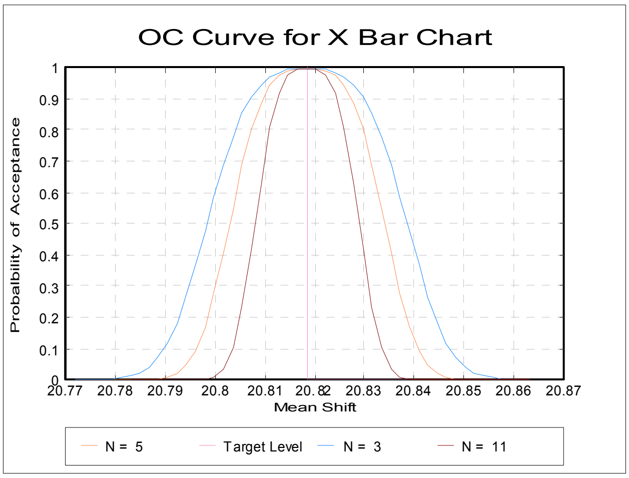
X Bar Chart

Size = 5.0000
Mean: X-bar2
Range: R2
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
31/12/1899 Sun	20.8180	20.8029	20.8341
01/01/1900 Mon	20.8200	20.8029	20.8341
02/01/1900 Tue	20.8210	20.8029	20.8341
03/01/1900 Wed	20.8280	20.8029	20.8341
04/01/1900 Thu	20.8250	20.8029	20.8341
05/01/1900 Fri	20.8130	20.8029	20.8341
06/01/1900 Sat	20.8280	20.8029	20.8341
07/01/1900 Sun	20.8300	20.8029	20.8341
08/01/1900 Mon	20.8320	20.8029	20.8341
*** 09/01/1900 Tue**	20.8490	20.8029	20.8341
*** 10/01/1900 Wed**	20.8480	20.8029	20.8341
*** 11/01/1900 Thu**	20.8520	20.8029	20.8341
*** 12/01/1900 Fri**	20.8420	20.8029	20.8341
*** 13/01/1900 Sat**	20.8450	20.8029	20.8341
*** 14/01/1900 Sun**	20.8410	20.8029	20.8341

Target Level =	20.8185
Average Sample Size =	5.0000
Control Limits =	3.0000 × Sigma
Sigma =	0.0116





This shows that the process means are out of control. Since from the tenth sample onwards the points are above the UCL, the process must have gone out of control.

9.3.1.5. Moving Average Charts

Even if all the points lie inside the control lines on an X Bar Chart, this does not mean that the central tendency is under control. If a number of consecutive sample means all lie above the target level but under the control level, this may be an indication of a problem. The Moving Average Charts can be more sensitive than the standard X Bar Chart. The moving average of the sample means is plotted with the appropriate control lines. In UNISTAT two types of Moving Average Charts are available: (i) the standard moving average and (ii) the exponential weights moving average. The control lines for the two charts are quite different.

In these charts the moving average values of the sample means are plotted to control the central tendency of a variable. They will plot either standard moving average values or exponential weights moving average values. Both these charts assume the average sample size for all samples.

9.3.1.5.1. Standard Moving Average Charts

In this chart the standard moving average of the sample means are plotted to control the central tendency of a variable.

$$\mathrm{UCL} = \mathrm{T} + \mathrm{C} \times \frac{\mathrm{S}}{\sqrt{\mathrm{nw}}}$$

$$\mathrm{CL} = \mathrm{T}$$

$$\mathrm{LCL} = \mathrm{T} - \mathrm{C} \times \frac{\mathrm{S}}{\sqrt{\mathrm{nw}}}$$

where:

- T is the target level. UNISTAT gives the overall mean of the samples as the default value. The weighted mean of the sample means.
- S is the sigma value. UNISTAT gives the overall sigma of the samples as the default value.
- C is the control line parameter,
- n is the sample size,
- w is the span of the moving average.

Example

Table 8.5 on p. 245 from Banks, Jerry (1989).

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. Select the data option 2 Enter Summary Data and *Sample Size* (*C9*) as [Size], *X-Bar3* (*C10*) as [Mean] and *Est Sigma* (*C11*) as [Std Dev]. The variability is known to be under control, so select Moving Average Chart at the next dialogue. Next enter the following parameters:

- **10.1704545** Target Level
- **2.28** Sigma
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **6** MA Parameter (>1 MA, <1 EWMA)

At the final output dialogue click [Next] to obtain the following chart:

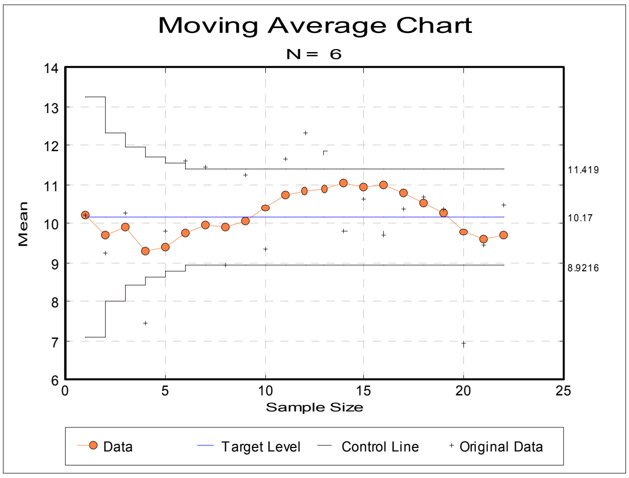
Variable Control Charts

Moving Average Chart

```
Size: Sample Size
Mean: X-Bar3
Standard Deviation: Est Sigma
'*' denotes sample outside control limits
```

Sample	Value	LCL	UCL
31/12/1899 Sun	10.2300	7.1115	13.2294
01/01/1900 Mon	9.7250	8.0075	12.3335
02/01/1900 Tue	9.9000	8.4044	11.9365
03/01/1900 Wed	9.2900	8.6410	11.6999
04/01/1900 Thu	9.3920	8.8025	11.5385
05/01/1900 Fri	9.7600	8.9216	11.4193
06/01/1900 Sat	9.9600	8.9216	11.4193
07/01/1900 Sun	9.9117	8.9216	11.4193
08/01/1900 Mon	10.0783	8.9216	11.4193
09/01/1900 Tue	10.3950	8.9216	11.4193
10/01/1900 Wed	10.7067	8.9216	11.4193
11/01/1900 Thu	10.8250	8.9216	11.4193
12/01/1900 Fri	10.8917	8.9216	11.4193
13/01/1900 Sat	11.0383	8.9216	11.4193
14/01/1900 Sun	10.9350	8.9216	11.4193
15/01/1900 Mon	10.9967	8.9216	11.4193
16/01/1900 Tue	10.7767	8.9216	11.4193
17/01/1900 Wed	10.5050	8.9216	11.4193
18/01/1900 Thu	10.2633	8.9216	11.4193
19/01/1900 Fri	9.7800	8.9216	11.4193
20/01/1900 Sat	9.5833	8.9216	11.4193
21/01/1900 Sun	9.7067	8.9216	11.4193

Moving Average, N =	6
Target Level =	10.1705
Average Sample Size =	5.0000
Control Limits =	3.0000 × Sigma
Sigma =	2.2800



This chart shows that the process is in control.

9.3.1.5.2. Exponential Weights Moving Average Chart

This is sometimes called the geometric moving average control chart.

$$\mathrm{UCL} = \mathrm{T} + \mathrm{C} \times \mathrm{Sx}\sqrt{\frac{\mathrm{r}}{\mathrm{n}(2-\mathrm{r})}\left(1-(1-\mathrm{r})^{2\mathrm{t}}\right)}$$

$$\mathrm{CL} = \mathrm{T}$$

$$\mathrm{LCL} = \mathrm{T} - \mathrm{C} \times \mathrm{Sx}\sqrt{\frac{\mathrm{r}}{\mathrm{n}(2-\mathrm{r})}\left(1-(1-\mathrm{r})^{2\mathrm{t}}\right)}$$

where:

- T is the target level. The default value is the overall mean of the samples, which is the weighted mean of the sample means.
- S is the sigma value. The default value is the overall sigma of the samples.
- C is the Control line parameter.
- n is the average sample size.
- r is the exponential weight parameter ($1 \le r \le 0$).
- t is the number of sample. This sequentially counts through the samples. It represents the number of samples which have contributed to the current moving average value.

Example

Table 8.6 on p. 249 from Banks, Jerry (1989).

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. Select the data option 2 Enter Summary Data and *Sample Size* (*C9*) as [Size], *X-Bar3* (*C10*) as [Mean] and *Est Sigma* (*C11*) as [Std Dev]. The variability is known to be under control, so select Moving Average Chart at the next dialogue. Next enter the following parameters:

- **10.1704545** Target Level
- **2.28** Sigma
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0.25** Parameter (>1 MA, <1 EWMA)

and finally click [Finish] to obtain the following results:

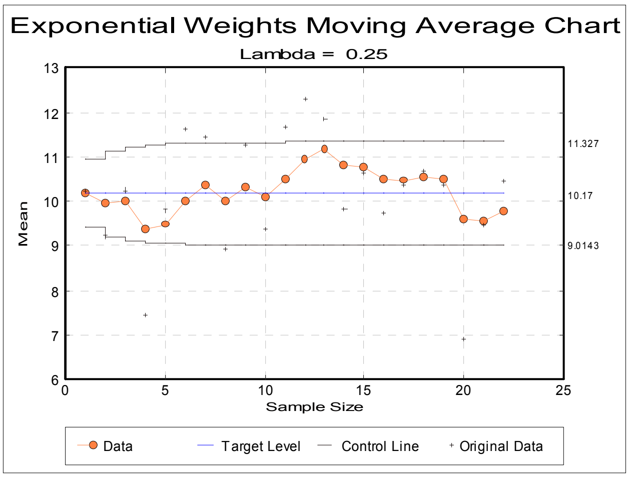
Variable Control Charts

Moving Average Chart

Size: Sample Size
Mean: X-Bar3
Standard Deviation: Est Sigma
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
31/12/1899 Sun	10.1853	9.4057	10.9352
01/01/1900 Mon	9.9440	9.2145	11.1264
02/01/1900 Tue	10.0205	9.1222	11.2187
...	…	…	…
14/01/1900 Sun	10.7737	9.0144	11.3265
15/01/1900 Mon	10.5128	9.0143	11.3266
16/01/1900 Tue	10.4721	9.0143	11.3266
17/01/1900 Wed	10.5241	9.0143	11.3266
18/01/1900 Thu	10.4881	9.0143	11.3266
19/01/1900 Fri	9.5935	9.0143	11.3266
20/01/1900 Sat	9.5577	9.0143	11.3266
21/01/1900 Sun	9.7857	9.0143	11.3266

Moving Average, Lambda =	0.2500
Target Level =	10.1705
Average Sample Size =	5.0000
Control Limits =	3.0000 × Sigma
Sigma =	2.2800



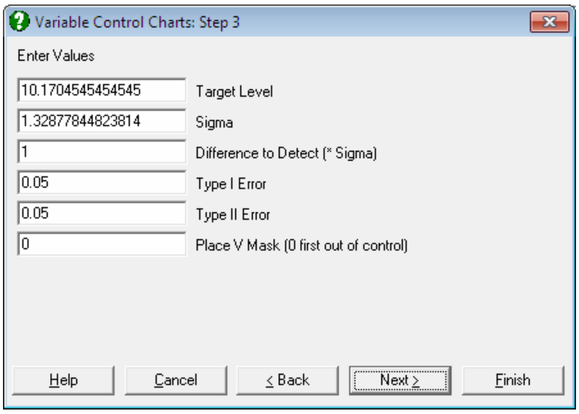
This chart shows that the process is in control.

9.3.1.6. CUSUM Chart

The CUSUM Chart is used to control the central tendency, where the cumulative sum of differences from the target level is plotted. This means that even minor permanent shifts in the process will eventually lead to a sizeable cumulative sum of deviations.

The CUSUM Chart is completely different from other control charts because it has no control lines. Instead, a *V-mask* is used to control the variable, which is considered at each point in the series. If a previous point lies outside the V-mask the process is declared to be out of control.

Missing values are handled differently in CUSUM Charts, compared to the other control charts. To leave an empty column on the X-axis would affect the test of the V-mask. So, missing values are not drawn on the chart at all. This often results in the numbering on the X-axis looking somewhat strange.


Variable Control Charts: Step 3

Enter Values

10.1704545454545 Target Level
1.32877844823814 Sigma
1 Difference to Detect (* Sigma)
0.05 Type I Error
0.05 Type II Error
0 Place V Mask (0 first out of control)

The CUSUM Chart has the following input parameters:

Target Level: The target level is the centre line for the chart.

Sigma: The sigma value is the standard error of the variable.

Difference to Detect: This is the shift in the process mean to detect. Shifts less than this value are considered inconsequential. The value given is in terms of the standard deviation of each sample (sigma / average sample size).

Type I Error: The Type I error (alpha error) is the probability of declaring a process out of control when it is not.

Type II Error: The Type II error (beta error) is the probability of accepting a process is in control when it is out of control. A process is considered out of control when the mean shifts beyond the difference to detect.

Place V-Mask: This value allows the user to place the V-mask on any particular point in the data. If 0 is entered then the V-mask is placed on the first point where the chart is out of control. If the chart is in control then the V-mask is drawn on the last point in the series.

Example

Table 8.6 on p. 249 from Banks, Jerry (1989).

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. Select the data option 2 Enter Summary Data and *Sample Size* (*C9*) as [Size], *X-Bar3* (*C10*) as [Mean] and *Est Sigma* (*C11*) as [Std Dev]. The variability is known to be under control, so select Cusum Chart at the next dialogue. Next enter the following parameters:

- **10.1704545** Target Level
- **2.28** Sigma
- **1** Difference to Detect (* Sigma)
- **0.05** Type I Error
- **0.05** Type II Error
- **0** Place V Mask (0 first out of control)

and finally click [Finish] to obtain the following results:

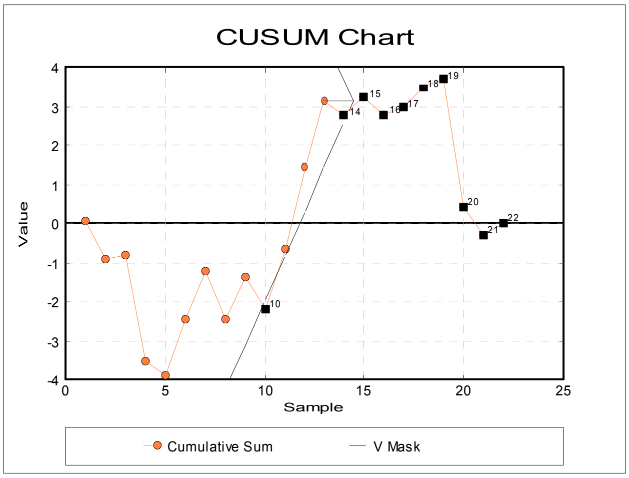
Variable Control Charts

CUSUM Chart

Size: Sample Size
Mean: X-Bar3
Standard Deviation: Est Sigma
'*' denotes sample outside V Mask at 13

Sample	Value	V Mask Upper	V Mask Lower
31/12/1899 Sun	0.0595	-12.2146	18.4628
01/01/1900 Mon	-0.8909	-11.0746	17.3228
02/01/1900 Tue	-0.8114	-9.9346	16.1828
03/01/1900 Wed	-3.5218	-8.7946	15.0428
04/01/1900 Thu	-3.8923	-7.6546	13.9028
05/01/1900 Fri	-2.4627	-6.5146	12.7628
06/01/1900 Sat	-1.2032	-5.3746	11.6228
07/01/1900 Sun	-2.4436	-4.2346	10.4828
08/01/1900 Mon	-1.3641	-3.0946	9.3428
* 09/01/1900 Tue	-2.1745	-1.9546	8.2028
10/01/1900 Wed	-0.6750	-0.8146	7.0628
11/01/1900 Thu	1.4645	0.3254	5.9228
12/01/1900 Fri	3.1241	1.4654	4.7828
* 13/01/1900 Sat	2.7636	-1.0000e+030	1.0000e+030
* 14/01/1900 Sun	3.2232	-1.0000e+030	1.0000e+030
* 15/01/1900 Mon	2.7827	-1.0000e+030	1.0000e+030
* 16/01/1900 Tue	2.9623	-1.0000e+030	1.0000e+030
* 17/01/1900 Wed	3.4718	-1.0000e+030	1.0000e+030
* 18/01/1900 Thu	3.6814	-1.0000e+030	1.0000e+030
* 19/01/1900 Fri	0.4209	-1.0000e+030	1.0000e+030
* 20/01/1900 Sat	-0.2995	-1.0000e+030	1.0000e+030
* 21/01/1900 Sun	0.0000	-1.0000e+030	1.0000e+030

Target Level =	10.1705
Average Sample Size =	5.0000
Sigma =	2.2800
Difference to Detect =	1.0000 × Sigma
Type I Error =	0.0500
Type II Error =	0.0500



9.3.1.7. Table of Values

For each sample, the sample sizes, means, standard deviations and ranges are displayed in a table. These values can be saved to the data matrix and used as input to Variable Control Charts.

Example

Table 7.5 on p. 214 from Banks, Jerry (1989). Refractive Index of Fiber Optic Cable is given.

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. Select the data option 1 Enter Raw Values and Samples and *Day* (*C5*) as [Sample] and *Refractive Index* (*C6*) as [Variable]. From the next dialogue select Table of Values to obtain the following results:

Variable Control Charts

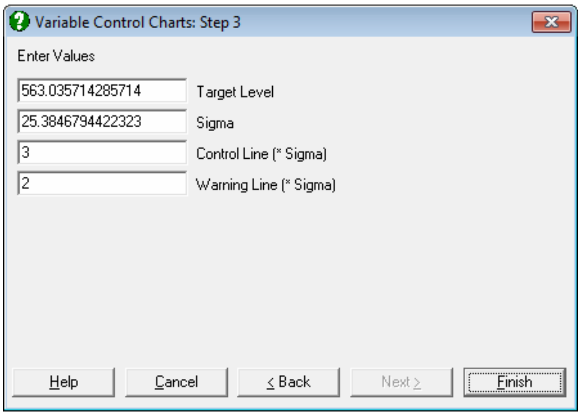
Table of Values

Sample	Size	Mean	Standard Deviation	Range
08/06/1992 Mon	6	95.7333	1.1466	3.2000
09/06/1992 Tue	6	95.4667	1.2501	3.7000
10/06/1992 Wed	4	96.6500	1.5948	3.2000
11/06/1992 Thu	5	97.4600	0.8204	1.9000
12/06/1992 Fri	6	96.8667	1.5870	3.8000
15/06/1992 Mon	4	96.8500	1.4617	3.5000
16/06/1992 Tue	8	96.5250	1.3307	3.4000
17/06/1992 Wed	4	96.0750	1.0996	2.3000
18/06/1992 Thu	6	97.2333	1.4828	4.4000
19/06/1992 Fri	5	96.5400	0.8019	1.9000
22/06/1992 Mon	6	96.6333	1.4166	4.0000
23/06/1992 Tue	5	96.4600	0.7668	1.8000

9.3.1.8. X Chart (Levey-Jennings)

This is a simple X-Y plot with the ability to display target and control lines and mark and annotate outlying points.

Variable selection is similar to that of other Variable Control Charts (see 9.3.0.2. Control Chart Inputs). A data column containing control measurements is selected clicking on [Variable] and a categorical data column (containing string or numeric values) is selected by clicking on [Sample]. The latter will not affect the appearance of the curve, but it can be used to label the group membership of outlying points (see 9.3.1.0.1. Raw Values and Samples).



Variable Control Charts: Step 3

Enter Values

563.035714285714 Target Level

25.3846794422323 Sigma

3 Control Line (* Sigma)

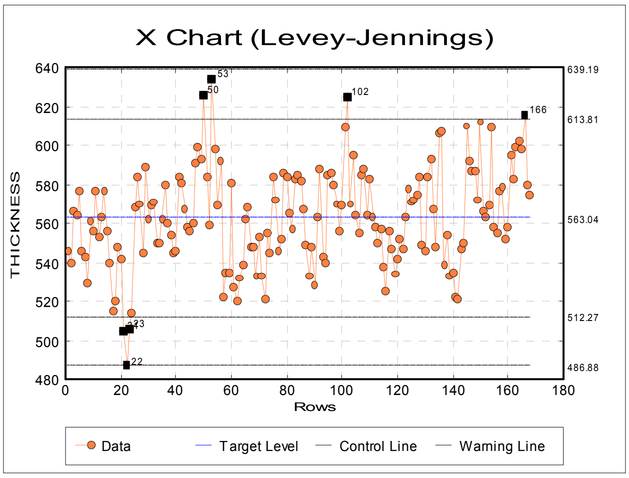
2 Warning Line (* Sigma)

The parameter input dialogue displays the mean and standard deviation as calculated from data. Also displayed are the levels of warning and control lines as multiples of sigma. Here, you can either accept these values or enter other values.

By default, all points are plotted without annotation. It is possible, however, to select Edit → Data Series and display case labels for outlying values.

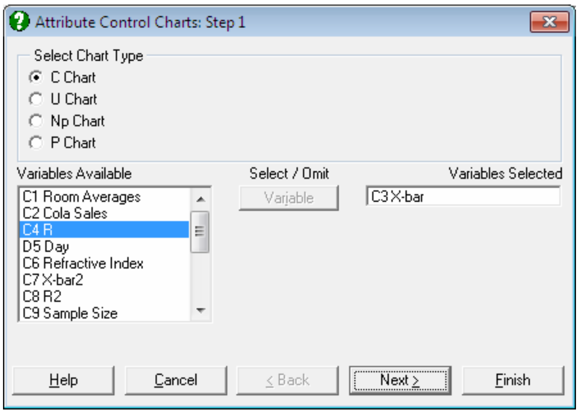
Example

Open TIMESER and select Statistics 2 → Quality Control → Variable Control Charts. From the Variable Selection Dialogue select *THICKNESS* (*C21*) as [Variable] and *ZONE* (*C19*) as [Factor]. On Step 2 select X Chart (Levey-Jennings) and on the next step enter 2 for Warning Line (* Sigma).



9.3.2. Attribute Control Charts

An attribute, as used in quality control, refers to a characteristic that does or does not conform to specifications. For example, in a computer assembly operation, computers are switched on after they have been assembled. They either work (conform) and undergo further tests or they do not switch on (non conform) in which case they are sent for repair.



Attribute Control Charts: Step 1

Select Chart Type

● C Chart
○ U Chart
○ Np Chart
○ P Chart

Variables Available: C1 Room Averages, C2 Cola Sales, C4 R, D5 Day, C6 Refractive Index, C7 X-bar2, C8 R2, C9 Sample Size

Variables Selected: C3 X-bar

First select the type of attribute chart. Next, select one column for the number of non conforming cases in each sample by clicking on [Variable] and one column for the sample size by clicking on [Size]. Sample size is not required for the C Chart. Each row represents a sample.

9.3.2.1. C Chart

In this chart, the number of non conforming cases are plotted and the control limits are based on the Poisson distribution. This chart should be used when the non conforming cases are rare. The program does not ask for the sample size, but all samples should have similar sizes.

$$\mathrm{UCL} = \mathrm{T} + \mathrm{C} \times \sqrt{\mathrm{T}}$$

$$\mathrm{CL} = \mathrm{T}$$

$$\mathrm{LCL} = \mathrm{T} - \mathrm{C} \times \sqrt{\mathrm{T}}$$

If LCL < 0 then set LCL = 0.

where:

- T is the target level. The default value is the mean of the sample values.
- C is control line parameter.

Example

Table 6.6 on p. 168 from Banks, Jerry (1989). Number of nonconformities on 22 samples of 50 EGA cards is given.

Open TIMESER and select Statistics 2 → Quality Control → Attribute Control Charts. Select C Chart and *Nonconformities* (*C12*) as [Variable]. On the next dialogue enter:

- **12.8636** Target Level
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0** Parameter (>1 MA, <1 EWMA)

and from the Output Options Dialogue select [Finish] to obtain the following output:

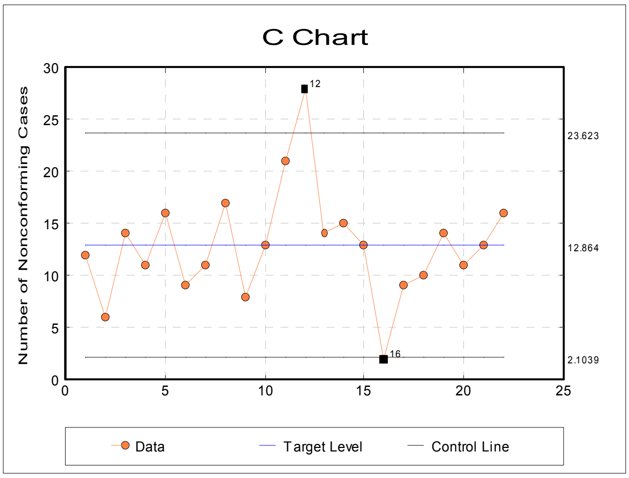
Attribute Control Charts

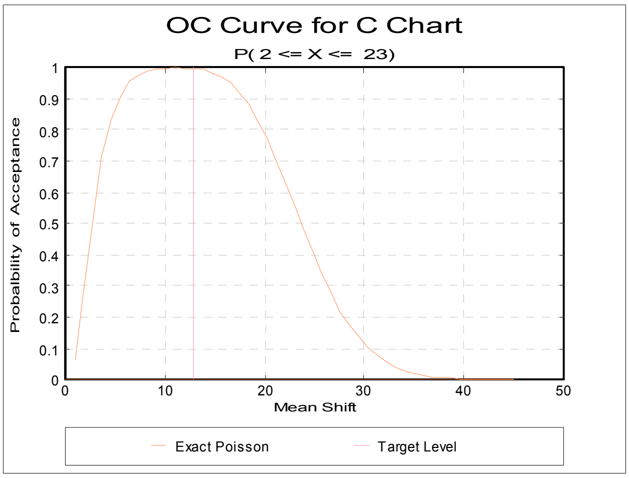
C Chart

Variable: Nonconformities
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
1	12.0000	2.1039	23.6234
2	6.0000	2.1039	23.6234
3	14.0000	2.1039	23.6234
...	...	...	...
11	21.0000	2.1039	23.6234
*** 12**	28.0000	2.1039	23.6234
13	14.0000	2.1039	23.6234
14	15.0000	2.1039	23.6234
15	13.0000	2.1039	23.6234
*** 16**	2.0000	2.1039	23.6234
17	9.0000	2.1039	23.6234
18	10.0000	2.1039	23.6234
19	14.0000	2.1039	23.6234
20	11.0000	2.1039	23.6234
21	13.0000	2.1039	23.6234
22	16.0000	2.1039	23.6234

Target Level =	12.8636
Control Limits =	3.0000 × Sigma
Sigma =	3.5866





This chart shows that the process is out of control. It might be instructive to investigate any changes that occurred in the process when samples 12 and 16 were taken.

9.3.2.2. U Chart

In U Chart, the rate of non conforming cases is plotted. The control limits are based on the Poisson distribution. Therefore, non conforming cases should be rare. Unlike the Np Chart and the P Chart, the U Chart does not require that the number of non conforming cases is less than the sample size. Each item can have more than 1 non conforming feature; e.g. number of scratches per car door.

$$\mathrm{UCL} = \mathrm{T} + \mathrm{C} \times \sqrt{\frac{\mathrm{T}}{\mathrm{n_i}}}$$

$$\mathrm{CL} = \mathrm{T}$$

$$\mathrm{LCL} = \mathrm{T} - \mathrm{C} \times \sqrt{\frac{\mathrm{T}}{\mathrm{n_i}}}$$

If LCL < 0 then set LCL = 0.

where:

- T is the target level. The default value is the total sample value divided by the total sample size, which is the weighted mean of the sample means.
- C is the control line parameter,
- n_i is the size of each sample.

Example

Table 6.8 on p. 173 from Banks, Jerry (1989). Paint Blemishes on Left Front Doors are given.

Open TIMESER and select Statistics 2 → Quality Control → Attribute Control Charts. Select U Chart and *Surface Area* (*C13*) as [Size] and *Blemishes* (*C14*) as [Variable]. On the next dialogue enter:

- **5.86766** Target Level
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0** Parameter (>1 MA, <1 EWMA)
- **0** Use Average N (0 No, Else Yes)

and from the Output Options Dialogue click [Finish] to obtain the following output:

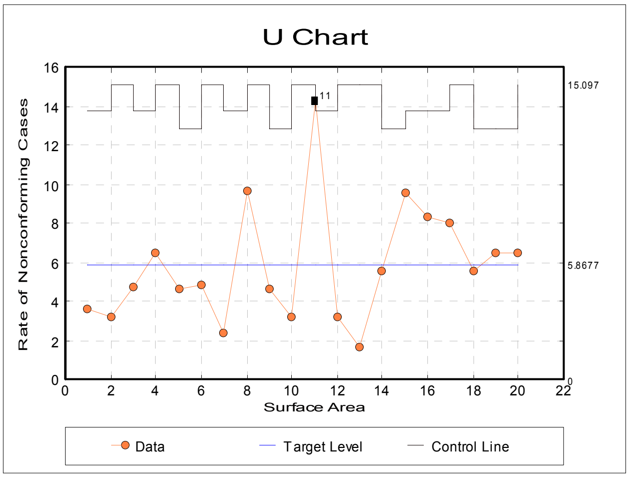
Attribute Control Charts

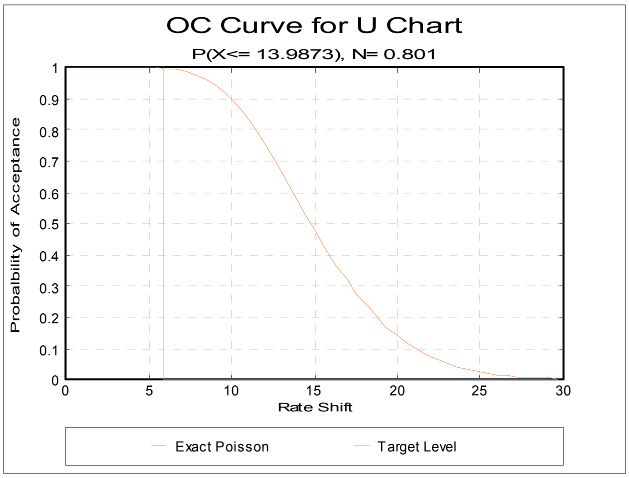
U Chart

Variable: Blemishes
Size: Surface Area
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
1	3.5714	0.0000	13.7966
2	3.2258	0.0000	15.0967
3	4.7619	0.0000	13.7966
…	…	…	…
10	3.2258	0.0000	15.0967
*** 11**	14.2857	0.0000	13.7966
12	3.2258	0.0000	15.0967
13	1.6129	0.0000	15.0967
14	5.5556	0.0000	12.8603
15	9.5238	0.0000	13.7966
16	8.3333	0.0000	13.7966
17	8.0645	0.0000	15.0967
18	5.5556	0.0000	12.8603
19	6.4815	0.0000	12.8603
20	6.4516	0.0000	15.0967

Target Level =	5.8677
Average Sample Size =	0.8010
Control Limits =	3.0000 × Sigma
Sigma =	2.4223





This chart shows that the process is out of control.

9.3.2.3. Np Chart

In this chart, the number of non conforming cases is plotted. The control limits are based on the Binomial distribution, so non conforming cases need not be rare. This chart requires that the number of non conforming cases is less than the sample size. The target level should change depending on each particular sample size. However, this is not the case in the Np Chart. So the sample sizes should be reasonably consistent. If sample sizes vary greatly then P Chart should be used instead.

$$UCL = T + C \times \sqrt{T\left(1 - \frac{T}{n_i}\right)}$$

$$CL = T$$

$$LCL = T - C \times \sqrt{T\left(1 - \frac{T}{n_i}\right)}$$

If LCL < 0 then set LCL = 0.

where:

- T is the target level. The default value is the total number of non conforming cases divided by the total sample size, which is the weighted mean of non conforming cases.

- C is the control line parameter,
- n_i is the size of each sample.

Example

Open TIMESER and select Statistics 2 → Quality Control → Attribute Control Charts. Select Np Chart and *Number Produced* (*C15*) as [Size] and *Nonconforming* (*C16*) as [Variable]. On the next dialogue enter:

- **13.35** Target Level
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0** Parameter (>1 MA, <1 EWMA)
- **0** Use Average N (0 No, Else Yes)

and from the Output Options Dialogue click [Finish] to obtain the following output:

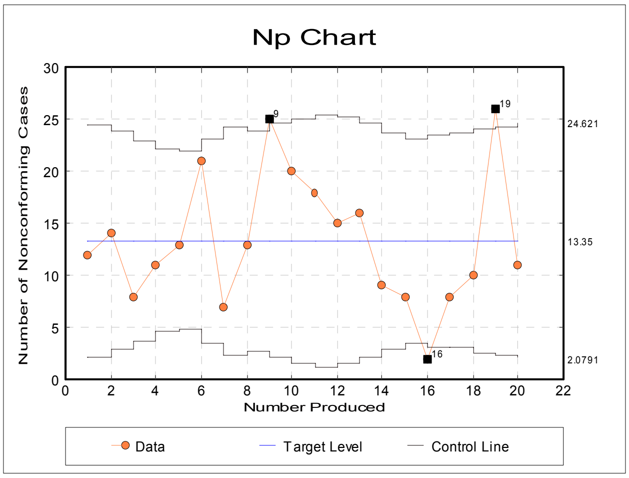
Attribute Control Charts

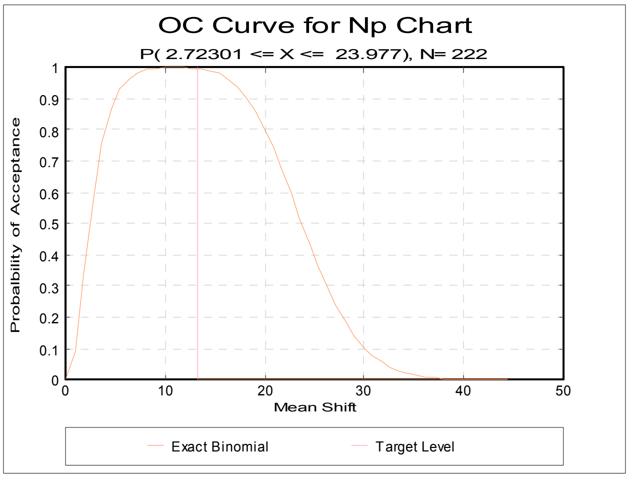
Np Chart

Variable: Nonconforming
Size: Number Produced
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
1	12.0000	2.1924	24.5076
2	14.0000	2.8735	23.8265
3	8.0000	3.7069	22.9931
...	...	...	...
*** 9**	25.0000	2.1243	24.5757
10	20.0000	1.5936	25.1064
11	18.0000	1.1899	25.5101
12	15.0000	1.5290	25.1710
13	16.0000	2.1243	24.5757
14	9.0000	2.9710	23.7290
15	8.0000	3.5501	23.1499
*** 16**	2.0000	3.1937	23.5063
17	8.0000	3.0447	23.6553
18	10.0000	2.6100	24.0900
*** 19**	26.0000	2.4224	24.2776
20	11.0000	2.0791	24.6209

Target Level =	13.3500
Average Sample Size =	222.2500
Control Limits =	3.0000 × Sigma
Sigma =	0.2376





9.3.2.4. P Chart

In this chart, the rate of non conforming cases is plotted. The control limits are based on the Binomial distribution, so non conforming cases need not be rare.

This chart requires that the number of non conforming cases is less than the sample size.

$$UCL = T + C \times \sqrt{\frac{T(1-T)}{n_i}}$$

$$CL = T$$

$$LCL = T - C \times \sqrt{\frac{T(1-T)}{n_i}}$$

If LCL < 0 then set LCL = 0.

where:

- T is the target level. UNISTAT gives the total sample values divided by the total sample sizes. This is the weighted mean of the sample values.
- C is the control line parameter,
- n_i is the size of the i^{th} sample.

Example

Table 6.3 on p.156 from Banks, Jerry (1989). PC Production for July is given.

Open TIMESER and select Statistics 2 → Quality Control → Attribute Control Charts. Select P Chart and *Number Produced* (*C15*) as [Size] and *Nonconforming* (*C16*) as [Variable]. On the next dialogue enter:

- **0.0601** Target Level
- **3** Control Line (* Sigma)
- **0** Warning Line (* Sigma)
- **0** Parameter (>1 MA, <1 EWMA)
- **0** Use Average N (0 No, Else Yes)

and from the Output Options Dialogue click [Finish] to obtain the following output:

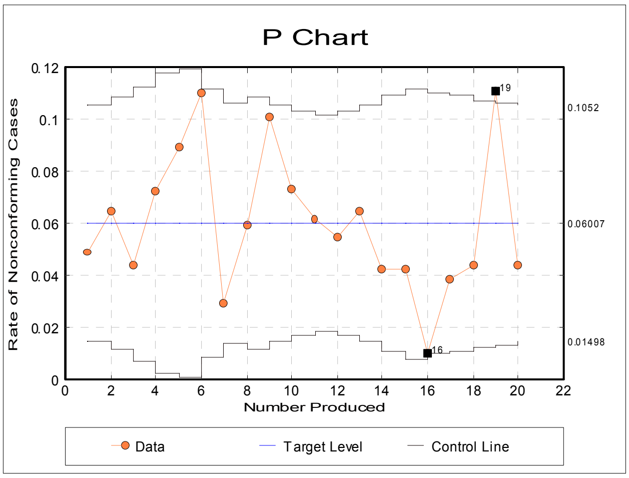
Attribute Control Charts

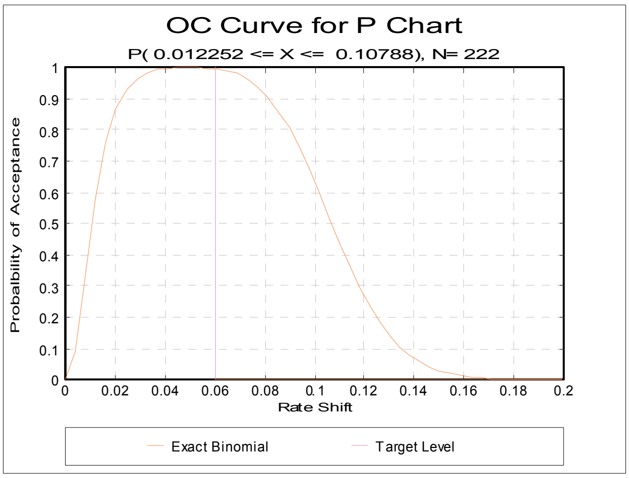
P Chart

Variable: Nonconforming
Size: Number Produced
'*' denotes sample outside control limits

Sample	Value	LCL	UCL
1	0.0490	0.0145	0.1056
2	0.0648	0.0116	0.1086
3	0.0437	0.0074	0.1128
4	0.0724	0.0022	0.1179
5	0.0897	0.0009	0.1193
6	0.1105	0.0084	0.1118
7	0.0295	0.0138	0.1064
8	0.0596	0.0118	0.1083
9	0.1008	0.0148	0.1053
10	0.0735	0.0168	0.1033
11	0.0619	0.0183	0.1019
12	0.0545	0.0171	0.1031
13	0.0645	0.0148	0.1053
14	0.0425	0.0111	0.1090
15	0.0423	0.0082	0.1119
*** 16**	0.0099	0.0100	0.1101
17	0.0383	0.0108	0.1094
18	0.0441	0.0128	0.1074
*** 19**	0.1106	0.0136	0.1066
20	0.0440	0.0150	0.1052

Target Level =	0.0601
Average Sample Size =	222.2500
Control Limits =	3.0000 × Sigma
Sigma =	0.2376





This shows that the process is out of control.

9.3.3. Pareto Chart

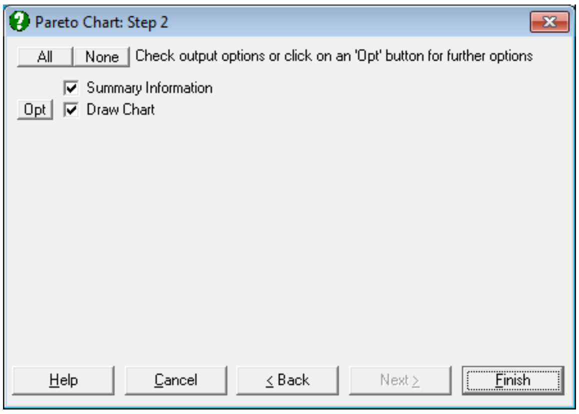
A Pareto Chart is a bar chart sorted in descending order with the cumulative totals shown above the bars. The chart is based on the Pareto principle, which states that quality losses are distributed in such a way that a small percentage of causes are responsible for the majority of the quality problems. This chart allows you to identify which areas are causing most problems. The Pareto chart is useful in comparing the process under investigation before and after the corrections are made.

Each row represents a different category. Select one column by clicking [Variable], which contains the number of defectives associated with each category. If a value is negative, zero or missing, then it is ignored in the analysis.

9.3.3.1. Pareto Output Options

The following output options are available:

Summary Information: This displays the row number, value, percentage, cumulative value and cumulative percentage in a table.

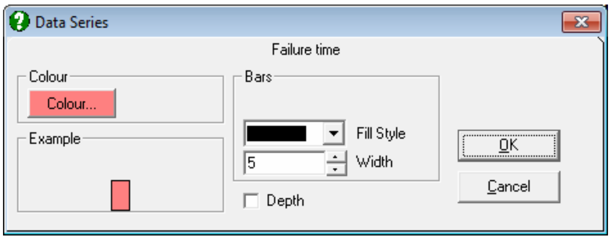

Pareto Chart: Step 2

Check output options or click on an 'Opt' button for further options

☑ Summary Information

☑ Draw Chart

Pareto Chart: Clicking the [Opt] button situated to the left of the Draw Chart option will place the graph in UNISTAT's Graphics Editor.



Data Series

Failure time

Colour

Example

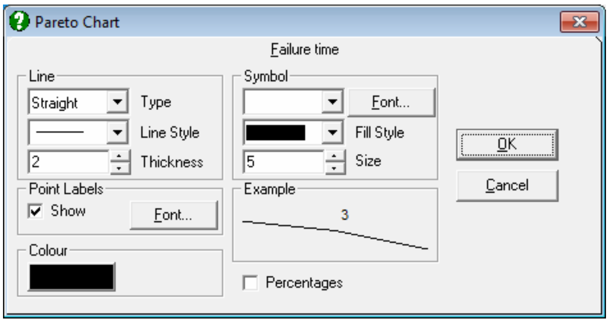
Bars

Fill Style

5 Width

□ Depth

It is possible to control the appearance of bars using the Edit → Data Series dialogue.



Pareto Chart

Failure time

Line

Straight Type

Line Style

2 Thickness

Symbol

Fill Style

5 Size

Point Labels

☑ Show

Colour

□ Percentages

In the Edit → Line dialogue, the Point Labels / Show check box controls the labels displayed along the cumulative total line. If the Percentages box is checked then the cumulative percentages will be displayed. Otherwise, the cumulative values are displayed.

9.3.3.2. Example

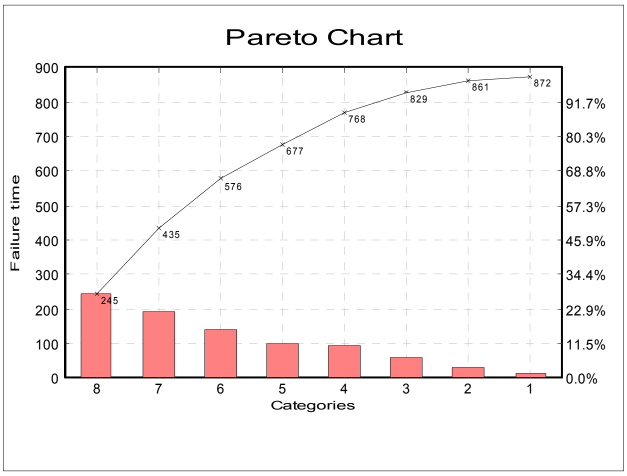
Example 16.15 on p. 580 from Banks, Jerry (1989). Failure times of 8 components are given.

Open TIMESER and select Statistics 2 → Quality Control → Pareto Chart. Select *Failure time* (*C17*) as [Variable] and click [Finish].

Pareto Chart

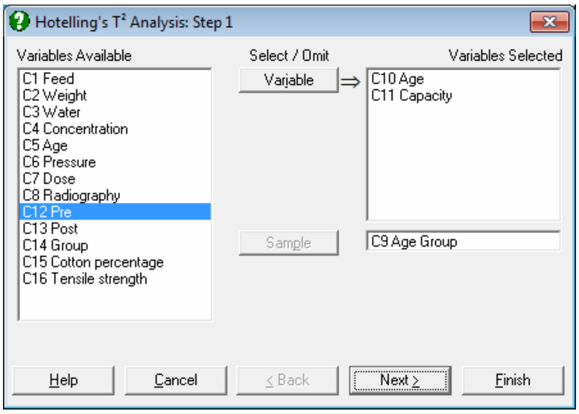
Variable: Failure time

Label	Row	Value	Percentage	Cumulative Value	Cumulative Percentage
8	8	245.0000	28.1%	245.0000	28.1%
7	7	190.0000	21.8%	435.0000	49.9%
6	6	141.0000	16.2%	576.0000	66.1%
5	5	101.0000	11.6%	677.0000	77.6%
4	4	91.0000	10.4%	768.0000	88.1%
3	3	61.0000	7.0%	829.0000	95.1%
2	2	32.0000	3.7%	861.0000	98.7%
1	1	11.0000	1.3%	872.0000	100.0%



9.3.4. Hotelling’s T-Squared Analysis

This is also known as the multivariate control analysis where multiple variables can be controlled in a single chart using the Hotelling’s T-Squared statistic. The Hotelling’s T-Squared statistic is the multivariate equivalent of the t-statistic (see 6.1.1. t- and F-Tests and 6.1.4. Hotelling’s T-Squared Test). In this case, the Hotelling T-Squared statistic is calculated using the average within sample covariance matrix as an estimate of the population covariance matrix, rather than using each sample covariance matrix to calculate T-Squared for that sample only.

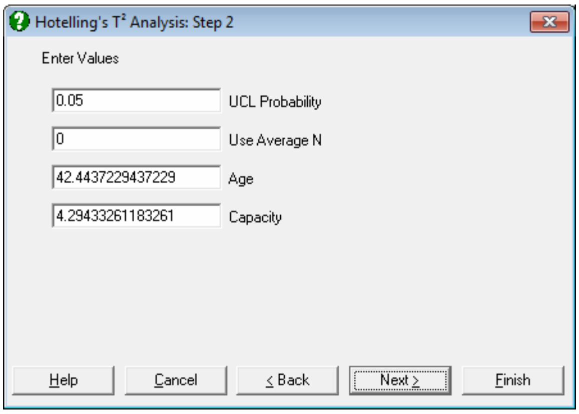


Variables Selected: C10 Age, C11 Capacity

Sample: C9 Age Group

To run this procedure, you will need to select two or more variables by clicking on [Variable] and one [Sample] column for subgroups.

9.3.4.1. Hotelling’s T-Squared Inputs



UCL Probability: 0.05

Use Average N: 0

Age: 42.4437229437229

Capacity: 4.29433261183261

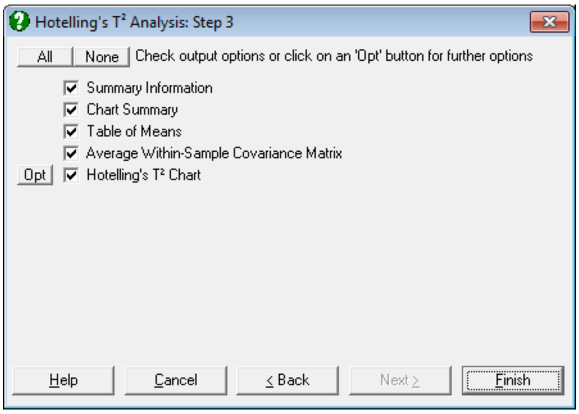
The parameter input dialogue asks for the following:

UCL Probability: This is critical probability of the test and is defined as the probability of declaring the process out of control when it is not out of control. This is equivalent to the Type I error probability.

Use Average N: If this is 1 then the average sample size will be used to determine the critical value for each sample. This will mean that the UCL will be a straight line. Otherwise, if the samples have different sample sizes, then the UCL will vary from sample to sample.

Variable n: You can enter a target level for each variable selected. The default value suggested is the sample mean of each variable.

9.3.4.2. Hotelling's T-Squared Output Options



Check output options or click on an 'Opt' button for further options

☑ Summary Information
☑ Chart Summary
☑ Table of Means
☑ Average Within-Sample Covariance Matrix
☑ Hotelling's T² Chart

The Output Options Dialogue offers the following choices:

Summary Information: For each sample, the sample size, Hotelling's T-Squared statistic, the estimated F- and its associated probability are displayed in a table. The estimated F-value is a transformation from the T-Squared statistic to a variable which follows the F distribution.

Chart Summary: For each sample, the sample size, Hotelling's T-Squared statistic and the UCL are displayed in a table.

Table of Means: For each sample, the sample means of all variables are displayed in a table.

Average Within-sample Covariance Matrix: This is the average of each sample covariance matrix. It is used in calculating the Hotelling's T-Squared statistic, instead of taking a different sample covariance matrix for each sample.

Hotelling's T-Squared Chart: T-Squared and UCL values are plotted against the group numbers. Clicking the [Opt] button situated to the left of the Hotelling's T-Squared Chart option will place the plot in UNISTAT's Graphics Editor. The plot can be further customised and annotated using the tools available under the graphics window's Edit menu.

9.3.4.3. Hotelling's T-Squared Example

Open ANOTESTS and select Statistics 2 → Quality Control → Hotelling's T-Squared Analysis. Select *Age* and *Capacity* (*C10 - C11*) as [Variable]s and *Age Group* (*C9*) as [Sample]. In the next dialogue accept the default values.

Hotelling's T-Squared Analysis

Summary Information

Sample: Age Group
F-Statistic with (2, N-2) Degrees of Freedom

Sample	N	T-Squared	F-Statistic	Probability
1	12	6.1748	2.5728	0.1255
2	28	6.0366	2.8027	0.0790
3	44	3.1140	1.4862	0.2379

Chart Summary

Sample	N	T-Squared	UCL
1	12	6.1748	9.8468
2	28	6.0366	7.2563
3	44	3.1140	6.7465

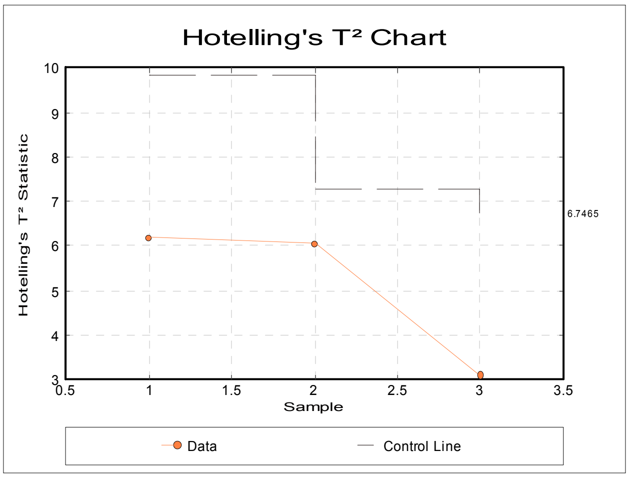
Beta Probability =	0.0500

Table of Means

Sample	Age	Capacity
1	49.7500	3.9492
2	37.7857	4.4718
3	39.7955	4.4620
Overall	42.4437	4.2943

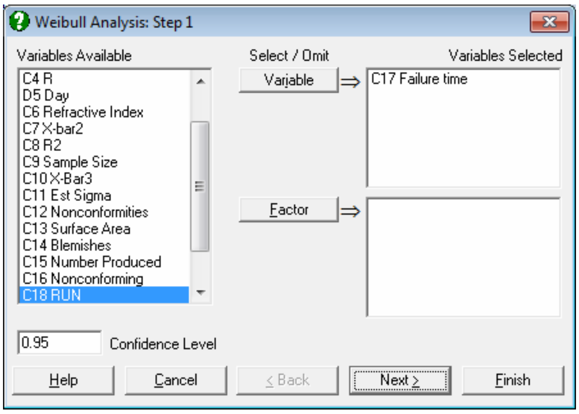
Average Within-Sample Covariance Matrix

	Age	Capacity
Age	103.8656	-5.1348
Capacity	-5.1348	0.6704



9.3.5. Weibull Analysis

Two or three parameter Weibull models can be estimated employing maximum likelihood or ordinary least squares (OLS) regression methods. The program reports the estimated parameters, their confidence intervals and covariance matrix and features an interpolation facility.



Variables Available: C4 R, D5 Day, C6 Refractive Index, C7 X-bar2, C8 R2, C9 Sample Size, C10 X-Bar3, C11 Est Sigma, C12 Nonconformities, C13 Surface Area, C14 Blemishes, C15 Number Produced, C16 Nonconforming, C18 RUN

Variables Selected: C17 Failure time

0.95 Confidence Level

It is sufficient to select one data column to run a Weibull Analysis. Multisample data can be selected either in the form of multiple columns (not necessarily of equal length) or data columns classified by one or more factor columns (see 6.0.4. Multisample Tests). If at least one factor column is selected, then a further dialogue will pop up asking for the combination of factor levels to be included.

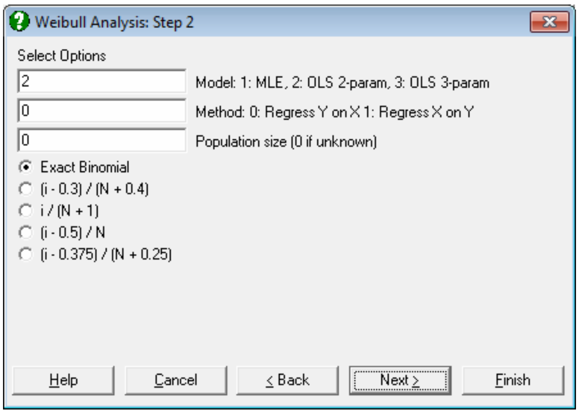
9.3.5.1. Weibull Inputs

An Intermediate Inputs dialogue enables you to control the following parameters:

Model: This can be one of two-parameter maximum likelihood or two- or three-parameter OLS regressions.

Method: If the method is OLS, then you can choose to regress Y on X (the default) or X on Y.

Population size: If the population size is known, you can enter it here. This is the number N used in generating the median ranks as described below. If this box is left as 0, then the program assumes that N is equal to the number of cases in the sample n.



Weibull Analysis: Step 2

Select Options

2 Model: 1: MLE, 2: OLS 2-param, 3: OLS 3-param

0 Method: 0: Regress Y on X 1: Regress X on Y

0 Population size (0 if unknown)

● Exact Binomial
○ (i - 0.3) / (N + 0.4)
○ i / (N + 1)
○ (i - 0.5) / N
○ (i - 0.375) / (N + 0.25)

Median Ranks: In order to run a regression and to plot a chart, Weibull Analysis needs to generate median ranks as Y-axis values. Median ranks can be computed in one of the following ways.

- Exact Binomial
- (i - 0.3) / (N + 0.4)
- i / (N + 1)
- (i - 0.5) / N
- (i - 0.375) / (N + 0.25)

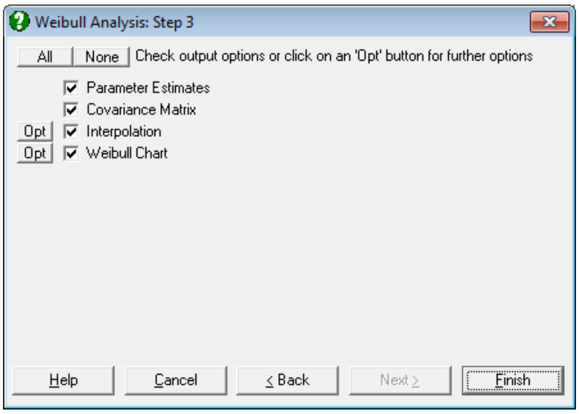
where N is the population size supplied by the user as described above. If N is not known, it is assumed to be equal to the sample size n. The default method adopted by UNISTAT is Exact Binomial, where the exact median ranks are determined from the binomial distribution as:

$$\sum_{k=i}^{N}\binom{N}{k}R_i^k(1-R_i)^{N-k}=0.5$$

where R_i is the median rank of the i^th observation. Other options in this list are various approximations to median ranks. For large values of N not all exact median ranks can be computed. When this is the case, the program cannot proceed with the analysis. However, you can still run your model by selecting one of the approximation formulas.

In Stand-Alone Mode, you can generate a data column containing median ranks using the Data Processor's **MdRk(*N*)** function (see 3.4.2.5. Statistical Functions).

9.3.5.2. Weibull Output Options


Weibull Analysis: Step 3

Check output options or click on an 'Opt' button for further options

☑ Parameter Estimates
☑ Covariance Matrix
☑ Interpolation
☑ Weibull Chart

9.3.5.2.1. Weibull Parameter Estimates

Parameters of the Weibull distribution can be estimated by one of the following three methods.

Two-parameter maximum likelihood estimation: The probability density function for the two-parameter Weibull distribution is given as:

$$f(T) = \frac{\beta}{\eta}(T/\eta)^{\beta-1} \mathrm{Exp}(-(T/\eta)^{\beta})$$

and the log likelihood function is:

$$L = \prod_{i=1}^{n} f(T_i)$$

where n is the sample size. Differentiating L with respect to β and η and setting equal to zero, we obtain the following two equations used in estimation:

$$\frac{\partial L}{\partial \beta} = \frac{\eta}{\beta} + \sum_{i=1}^{n} \mathrm{Log}(T_i) - \frac{1}{\eta}\sum_{i=1}^{n} T_i^{\beta} \mathrm{Log}(T_i) = 0$$

$$\frac{\partial L}{\partial \eta} = -\frac{n}{\eta} + \frac{1}{\eta^2}\sum_{i=1}^{n} T_i^{\beta} = 0$$

Two-parameter OLS estimation: The cumulative distribution function for the two-parameter Weibull distribution is given as:

$$F(T) = 1 - \mathrm{Exp}(-(T/\eta)^{\beta})$$

where T is the failure time, β is the shape parameter, η is the scale parameter. Taking the natural logarithm of both sides and rearranging we obtain:

$$\mathrm{Log}(-\mathrm{Log}(1 - F(T))) = -\beta \mathrm{Log}(\eta) + \beta \mathrm{Log}(T)$$

This is clearly a line equation of the form:

$$y = a + bx$$

where, conveniently, the left hand side is the gompit (or cloglog) function:

$$y = \mathrm{Log}(-\mathrm{Log}(1 - F(T)))$$

and F(T) is the median rank function, with the slope:

$$b = \beta$$

and the intercept:

$$a = -\beta \mathrm{Log}(\eta)$$

Three-parameter OLS estimation: The cumulative distribution function for the three-parameter Weibull distribution is given as:

$$F(T) = 1 - \mathrm{Exp}\left(-\left(\frac{T-\gamma}{\eta}\right)^{\beta}\right)$$

where β is the shape parameter, η is the scale parameter, and γ is the location parameter. This equation is linearised as above and all three parameters are estimated simultaneously using the iterative least squares method.

Mean failure time: This is found using the gamma function:

$$\mu = \eta\Gamma(1/\beta + 1)$$

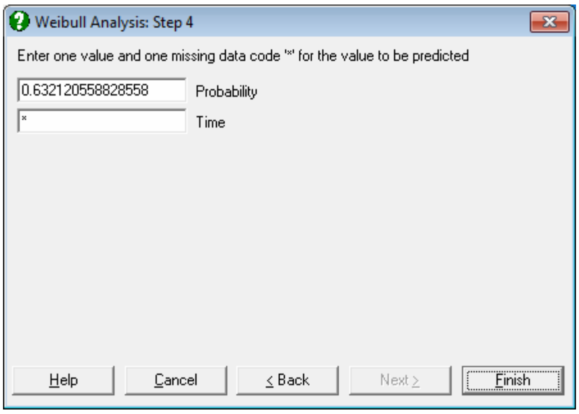
9.3.5.2.2. Weibull Covariance Matrix

For all estimation methods, the variance-covariance matrix is defined as the inverse of the second partial derivatives matrix of the log likelihood function:

$$\mathrm{Cov} = \begin{bmatrix} -\dfrac{\partial^2 L}{\partial \beta^2} & -\dfrac{\partial^2 L}{\partial \beta \partial \eta} \\ -\dfrac{\partial^2 L}{\partial \beta \partial \eta} & -\dfrac{\partial^2 L}{\partial \eta^2} \end{bmatrix}^{-1}$$

9.3.5.2.3. Weibull Interpolation

In the Output Options Dialogue, clicking on the [Opt] button situated to the left of the Interpolation check box, the interpolation dialogue can be accessed.


Weibull Analysis: Step 4

Enter one value and one missing data code "*" for the value to be predicted

0.632120558828558 Probability

* Time

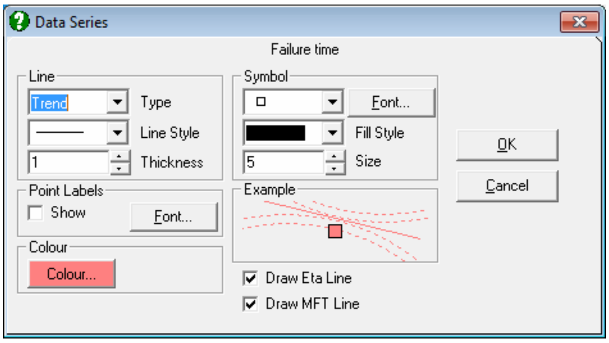
You can enter a failure probability to estimate the failure time or enter a failure time to estimate the probability of failure. Each time you will need to enter an asterisk for the parameter to be estimated. The program will display the estimated parameter together with its confidence interval.

9.3.5.2.4. Weibull Chart

If you use UNISTAT's X-Y Plots procedure and set the Y-axis scale as gompit and the X-axis scale as logarithmic, you instantly have a Weibull probability paper. If you select median ranks as the Y-axis and failure times as the X-axis variable and fit a trend line, you would have already estimated the two-parameter Weibull model. This is precisely the way we generate the Weibull chart here.

In Stand-Alone Mode, you can generate a data column containing median ranks using the Data Processor's **MdRk(*N*)** function (see 3.4.2.5. Statistical Functions).

If the data lies on a near-straight line, then it is said to conform to the Weibull distribution.


Data Series

Failure time

Line: Trend Type; Line Style; 1 Thickness

Point Labels: □ Show

Symbol: □; Fill Style; 5 Size

☑ Draw Eta Line

☑ Draw MFT Line

Clicking the [Opt] button situated to the left of the Weibull Chart option will place the graph in UNISTAT's Graphics Editor. In the Edit → Data Series dialogue, two check boxes allow you to show or hide the eta and mean failure time lines on the chart.

9.3.5.3. Examples

Example 1

Example 16.15 on p. 580 from Banks, Jerry (1989). Failure times of 8 components are given. The population size is known to be 100.

Open TIMESER and select Statistics 2 → Quality Control → Weibull Analysis. From the Variable Selection Dialogue select *Failure time* (*C17*) as [Variable]. In the next dialogue, enter 100 for the population size and select (i – 0.5) / N for median rank approximation. The following output is obtained:

Weibull Analysis

Parameter Estimates

Data variable: Failure time
Number of Cases: 8
Population Size: 100
Median Rank Method: (i - 0.5) / N
2-Parameter OLS Estimation
Regress Y on X

	Value	Lower 95%	Upper 95%
Beta	0.8962	0.5791	1.3870
Eta	3689.4178	2098.0830	6487.7337
Mean Failure Time	3890.9191		

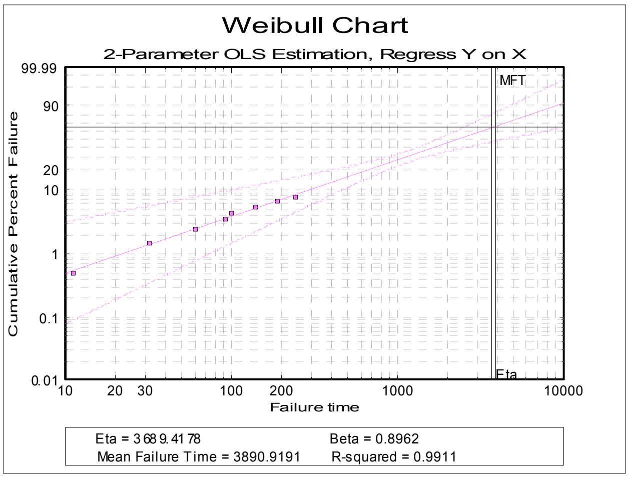
Correlation Coefficient =	0.9956
R-squared =	0.9911

Covariance Matrix

	Beta	Eta
Beta	0.0399	-192.7873
Eta	-192.7873	1128917.0915

Interpolation

Probability	Time	Lower 95%	Upper 95%
0.632120558828558	3689.4178	2098.0830	6487.7337



Example 2

Continuing from the last example, this time select 3-parameter model with population size 0 (unknown) and Exact Binomial median ranks:

Weibull Analysis

Parameter Estimates

Data variable: Failure time
Number of Cases: 8
Median Rank Method: Exact Binomial
3-Parameter OLS Estimation
Regress Y on X

	Value	Lower 95%	Upper 95%
Beta	1.4466	0.8149	2.5682
Eta	143.7028	85.1939	242.3940
Gamma	-15.9995		
Mean Failure Time	130.3404		

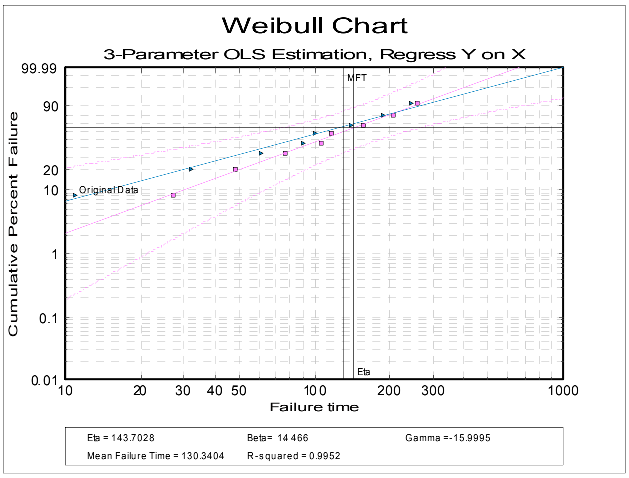
Correlation Coefficient =	0.9976
R-squared =	0.9952

Covariance Matrix

	Beta	Eta
Beta	0.1795	1.7235
Eta	1.7235	1469.3800

Interpolation

Probability	Time	Lower 95%	Upper 95%
0.632120558828558	143.7028	85.1939	242.3940

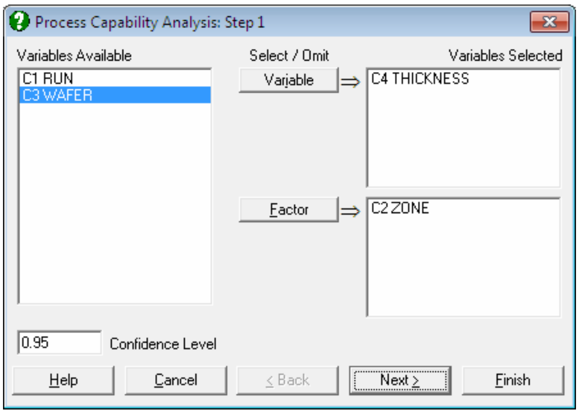


9.3.6. Process Capability Analysis

Process Capability Analysis is used to study the variation within a process. The aim is to determine how well a process meets its specification limits, by comparing the distribution of output to the desired specification limits. A process is said to be capable when a desired (high) number of output values fall within the specification limits.

The standard Process Capability Analysis is based on two important assumptions; (i) process data is normally distributed and (ii) process is in control. When the data is not normally distributed, capability analysis can still produce useful results by using nonparametric indices, or by transforming the data so that it conforms better to normal distribution than its original form. In order to find whether a process is in control, quality control charts like R and X Bar charts can be used (see 9.3.1.1. R Chart and 9.3.1.4. X Bar Chart).

9.3.6.1. Process Capability Analysis Input Options



Process Capability Analysis: Step 1

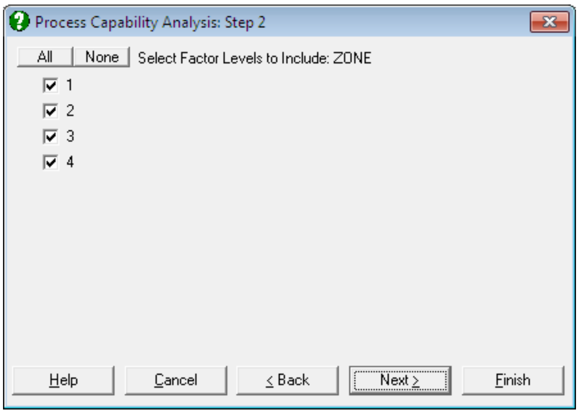
Variables Available: C1 RUN, C3 WAFER

Variables Selected (Variable): C4 THICKNESS

Factor: C2 ZONE

0.95 Confidence Level

The Variable Selection Dialogue is of the multisample type (see 6.0.4. Multisample Tests), allowing selection of multiple data and factor columns. If only one data variable and no factors are selected, then only an overall analysis is performed. Otherwise, all selected subgroups of data are pooled to form one continuous variable and an overall and a pooled sample analyses are performed.



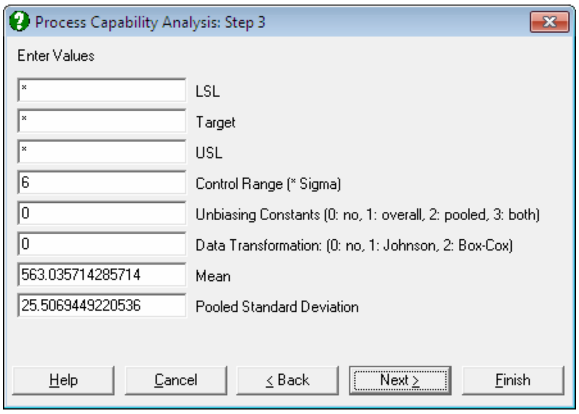
Select Factor Levels to Include: ZONE

☑ 1
☑ 2
☑ 3
☑ 4

If a factor column is selected, then a further dialogue will pop up allowing you to include only the desired subgroups in the analysis. If more than one factor is selected, then all combinations of all factor levels will be listed.

9.3.6.2. Process Capability Analysis Intermediary Input Options

The next dialogue allows you to enter or edit various input parameters.



Enter Values

*	LSL
*	Target
*	USL
6	Control Range (* Sigma)
0	Unbiasing Constants (0: no, 1: overall, 2: pooled, 3: both)
0	Data Transformation: (0: no, 1: Johnson, 2: Box-Cox)
563.035714285714	Mean
25.5069449220536	Pooled Standard Deviation

LSL: Lower Specification Limit: The smallest value below which a process is deemed not to perform satisfactorily. This value and its counterpart USL, Upper Specification Limit, are supplied by the user. If a one-sided lower specification analysis is to be carried out, then enter a single asterisk in the

USL field. In this case the output will not include some indices. The program will not proceed until an LSL or USL or both are entered.

Target: Enter a value if you wish to measure the process capability performance against a target. If a value is entered, the Cpm and related indices are computed. If there is no target, enter a single asterisk in this field.

USL: Upper Specification Limit: The largest value above which a process is deemed not to perform satisfactorily. This value and its counterpart LSL, Lower Specification Limit, are supplied by the user. If a one-sided upper specification analysis is to be carried out, then enter a single asterisk in the LSL field. In this case the output will not include some indices. The program will not proceed until an LSL or USL or both are entered.

Control Range: The observed process variation in terms of sample standard deviation. This is usually 6, ± 3 sigma around the centre. If the sample is from a normally distributed population, then approximately 99% of all data points would fall within this range.

Unbiasing Constants: You can choose to apply unbiasing constants in calculations. It is possible to apply them to either or both overall and pooled standard deviations, as defined in Wheeler D. J. and Chambers D. S. (1992). The magnitude of correction gets larger as the sample size decreases.

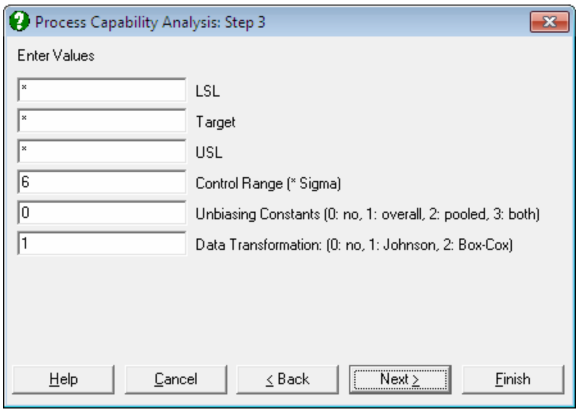
Data Transformation: The options are 0: no transformation, 1: Johnson Transformation and 2: Box-Cox Transformation. Box-Cox Transformation will only work with positive data, whereas Johnson Transformation has no such restrictions. In most cases Johnson Transformation is more powerful than Box-Cox Transformation and it will generate transformed variables conforming better to the normal distribution. These two transformations are also available under the Statistics 2 → Quality Control → Data Transformation procedure.

The number of boxes displayed after this depends on the selection made in the Data Transformation box.

When the no transformation option is selected, one or more boxes will be displayed.

Mean: This field is available only when the no transformation option is selected. Here you can override the default mean computed from the sample and enter your own (i.e. the historical mean).

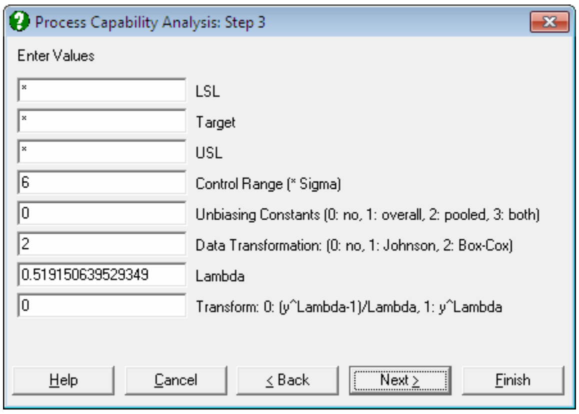
Pooled Standard Deviation: This field is available only when the no transformation option is selected and the data consists of multiple samples (i.e. if more than one variable or factor column(s) are selected). By default, the pooled sample standard deviation calculated from data is displayed. However, you can override this value and enter your own (i.e. the historical) standard deviation. If your data does not have subgroups, you can create a constant factor column (e.g. consisting of 1s) and still be able to enter a historical standard deviation.


Process Capability Analysis: Step 3

Enter Values

*	LSL
*	Target
*	USL
6	Control Range (* Sigma)
0	Unbiasing Constants (0: no, 1: overall, 2: pooled, 3: both)
1	Data Transformation: (0: no, 1: Johnson, 2: Box-Cox)

When the Johnson Transformation option is selected, there will be no further input fields.


Process Capability Analysis: Step 3

Enter Values

*	LSL
*	Target
*	USL
6	Control Range (* Sigma)
0	Unbiasing Constants (0: no, 1: overall, 2: pooled, 3: both)
2	Data Transformation: (0: no, 1: Johnson, 2: Box-Cox)
0.519150639529349	Lambda
0	Transform: 0: (y^Lambda-1)/Lambda, 1: y^Lambda

When the Box-Cox Transformation option is selected, two further input fields will be displayed, similar to ones displayed in Box-Cox Transformation (see 9.3.7.2.1 Box-Cox Transformation Intermediary Inputs).

Lambda: You can override the estimated lambda and enter your own value here. You may wish to do this to use a round power value (like -1, -0.5, 0.5, 2). If the estimated lambda is changed, confidence intervals and chi-squared tests for lambda will not be available.

Transform: Once the optimal lambda is estimated using the standard Box-Cox Transformation, you will have a chance to generate the transformed variable using, (i) the same transformation:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0$$

$$y^{(\lambda)} = \text{Ln}(y) \text{ if } \lambda = 0$$

or, (ii) the simple power transformation:

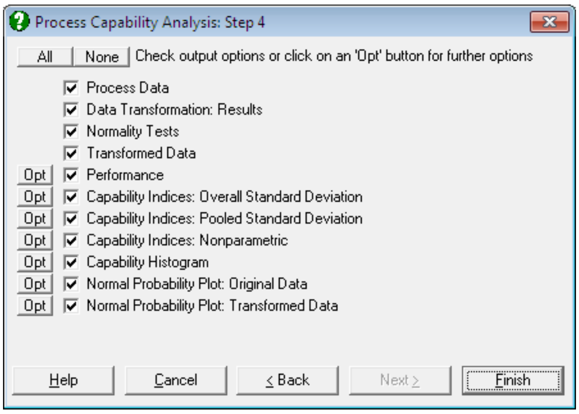
$$y^{(\lambda)} = y^{\lambda}$$

In some cases the second formula may be preferable to the first, since it will not generate nonpositive values. The choice made here will not affect normality of the transformed variable.

Remember that during the maximum likelihood estimation of lambda the original variable is always transformed using the first set of equations.

9.3.6.3. Process Capability Analysis Output Options

Some of the options displayed on this dialogue may appear disabled, depending on the choices made in previous dialogues.



Process Capability Analysis: Step 4

Check output options or click on an 'Opt' button for further options

☑ Process Data
☑ Data Transformation: Results
☑ Normality Tests
☑ Transformed Data
☑ Performance
☑ Capability Indices: Overall Standard Deviation
☑ Capability Indices: Pooled Standard Deviation
☑ Capability Indices: Nonparametric
☑ Capability Histogram
☑ Normal Probability Plot: Original Data
☑ Normal Probability Plot: Transformed Data

Process Data: Variables selected, summary statistics (size, mean, overall and pooled standard deviations) and parameters supplied by the user (control range, LSL target, USL) are displayed. If a Data Transformation option was selected, summary statistics for the transformed data will also be displayed.

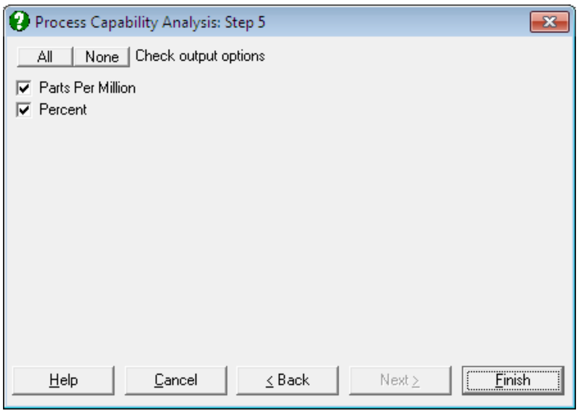
Data Transformation Results: This option is enabled if a transformation has been selected in the previous dialogue. Optimal parameter values estimated by the program and the equation applied in transforming the dependent variable is displayed. The same equation is also printed on a separate line with estimated parameter values, in a format suitable for cell calculations in Excel. You can simply copy this equation, replace the variable x with a cell reference and run interpolations.

Normality Tests: This option is enabled if a transformation has been selected in the previous dialogue. The Anderson-Darling Test of normality is performed on the original and the transformed dependent variable thus allowing the user to judge whether the transformation was useful. No or a small increase in the tail probability indicates that the Box-Cox Transformation was not useful.

Transformed Data: This option is enabled if a transformation has been selected in the previous dialogue. The original and transformed dependent variable values and their group membership (if any) are sorted and displayed in a table.

If you are using UNISTAT in Stand-Alone Mode, click on the UNISTAT icon on the Output Medium Toolbar to send all output to UNISTAT spreadsheet. In Excel Add-In Mode select the output matrix as data for further calculations.

Performance: The number of cases that are expected to fall outside the specification limits is calculated and this is displayed as Parts Per Million or Percentage points.

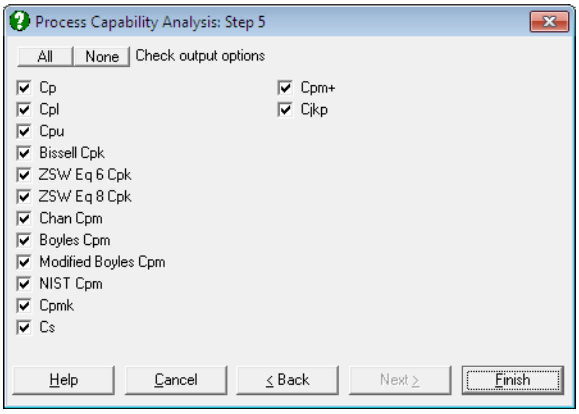


Check output options

☑ Parts Per Million
☑ Percent

The expected value is calculated from the cumulative normal distribution using the sample standard deviation and, in case of subgroup analysis, using both overall and pooled standard deviations.

Capability Indices: Overall Standard Deviation: Capability indices are indifferent to data transformations, since they are unitless ratios. However, when interpretation of results involves parameters in actual data units (such as standard deviation or confidence intervals), they should be transformed back into the original scale.



Check output options

☑ Cp
☑ Cpl
☑ Cpu
☑ Bissell Cpk
☑ ZSW Eq 6 Cpk
☑ ZSW Eq 8 Cpk
☑ Chan Cpm
☑ Boyles Cpm
☑ Modified Boyles Cpm
☑ NIST Cpm
☑ Cpmk
☑ Cs
☑ Cpm+
☑ Cjkp

For most practical applications, many of the indices available here will be irrelevant. When this is the case, simply uncheck the unwanted options and UNISTAT will remember your choices when you run this procedure next time. You can also remove the authors' names from the output by entering the following line in *Documents\Unistat65\Unistat65.ini* file under the [QC] section:

```
WithNames=0
```

Cp: Capability indices are used to compare the variability in the output of a process which is in control, to the desired specification limits. Cp is defined as the ratio of the difference between the specification limits to the control range (i.e. process variation):

$$\mathrm{Cp} = \frac{\mathrm{USL} - \mathrm{LSL}}{6\sigma}$$

It can be seen that when Cp is greater than one, the process specification covers almost all observations, as 6 sigma (the default control range) would cover approximately 99% of all data points from a normally distributed population.

The following interpretation of Cp values is widely accepted:

Cp < 1: not adequate
1 ≤ Cp ≤ 1.33: adequate
Cp > 1.33: satisfactory for existing processes
Cp > 1.50: for critical variables
Cp > 1.67: for new processes with a critical variable.

Confidence intervals for Cp are calculated as:

$$\mathrm{LB} = \mathrm{Cp}\sqrt{\frac{\chi^2_{\alpha/2,n-k}}{n-k}}$$

$$\mathrm{UB} = \mathrm{Cp}\sqrt{\frac{\chi^2_{1-\alpha/2,n-k}}{n-k}}$$

where n is the number of observations and k is the number of subgroups.

The Cp index is only available when both LSL and USL are supplied.

Cpl: This gives the capability of the lower half of the process. When USL is not available, it is still possible to determine the one-sided capability of the process:

$$\mathrm{Cpl} = \frac{\overline{\mathrm{X}} - \mathrm{LSL}}{3\sigma}$$

The following interpretation of Cpl values is widely accepted:

Cpl > 1.25: satisfactory for existing processes
Cpl > 1.45: for critical variables or new processes
Cpl > 1.60: for new processes with a critical variable

The 95% confidence intervals for Cpl are calculated using the noncentral t-distribution as:

$$\Pr\left\{T_{n-k,\delta} \leq 3\mathrm{Cpl}\sqrt{n}\right\} = 0.95$$

where the noncentrality parameter is:

$$\delta = 3\mathrm{LB}\sqrt{n}$$

$$\Pr\left\{T_{n-k,\delta} \leq 3\mathrm{Cpl}\sqrt{n}\right\} = 0.05$$

where the noncentrality parameter is:

$$\delta = 3\mathrm{UB}\sqrt{n}$$

WARNING: Due to the iterational nature of this confidence interval algorithm, the results are accurate to about three significant digits.

Cpu: This gives the capability of the upper half of the process. When LSL is not available, it is still possible to determine the one-sided capability of the process:

$$\mathrm{Cpu} = \frac{\mathrm{USL} - \overline{\mathrm{X}}}{3\sigma}$$

The confidence intervals for Cpu are calculated in the same way as for Cpl.

Cpk: The Cp index may be misleading when the centres of the specification range and process mean are significantly different. In such cases the Cpk index, which is defined as the minimum of Cpl and Cpu, will provide a better measure:

$$\mathrm{Cpk} = \mathrm{Min}\left(\mathrm{Cpl}, \mathrm{Cpu}\right) = \mathrm{Min}\left(\frac{\mathrm{USL} - \overline{\mathrm{X}}}{3\sigma}, \frac{\overline{\mathrm{X}} - \mathrm{LSL}}{3\sigma}\right)$$

This is the minimum distance between specification limits and the process mean divided by half of the control range.

The confidence intervals for Cpk are calculated using three different methods:

(i) Normal approximation suggested by Bissell, A. F. (1990):

$$\mathrm{LU}, \mathrm{LB} = \mathrm{Cpk} \pm \mathrm{Z}_{1-\alpha/2}\sqrt{\frac{1}{9\mathrm{n}} + \frac{\mathrm{Cpk}^2}{2(\mathrm{n}-\mathrm{k})}}$$

(ii) Equation 6 from Zhang, N. F., Stenback, G. A., Wardrop, D. M. (1990):

$$\mathrm{LU}, \mathrm{LB} = \mathrm{Cpk}\left(1 \pm \mathrm{Z}_{1-\alpha/2}\sqrt{\mathrm{a}}\right)$$

where:

$$\mathrm{a} = \frac{\mathrm{n}-1}{\mathrm{n}-3} + \mathrm{Exp}\left(\mathrm{Ln}\left(\frac{(\mathrm{n}-1)}{2}\right) + \left(2\left(\mathrm{Ln}\left(\Gamma\left(\frac{\mathrm{n}}{2} - 1\right)\right) - \mathrm{Ln}\left(\Gamma\left(\frac{\mathrm{n}-1}{2}\right)\right)\right)\right)\right)$$

(iii) Equation 8 from Zhang, N. F., Stenback, G. A., Wardrop, D. M. (1990):

$$\mathrm{LB} = \mathrm{Cpk}\left(1 - \mathrm{Z}_{\alpha/2}\sqrt{\mathrm{a}}\right)$$

$$\mathrm{UB} = \mathrm{Cpk}\left(1 + \mathrm{Z}_{1-\alpha/2}\sqrt{\mathrm{a}}\right)$$

where:

$$\mathrm{a} = \frac{\mathrm{n}-1}{\mathrm{n}-3} - \frac{(\mathrm{n}-1)\Gamma^2\left(\frac{\mathrm{n}-2}{2}\right)}{2\Gamma^2\left(\frac{\mathrm{n}-1}{2}\right)}$$

Cpm: This index is available only when a target value is specified. It is used to measure the variability of process data around a target value. Two different methods of calculating Cpm and its confidence intervals are provided.

(i) Chan L.J., Cheng S.K., and Spiring, F.A. (1989) suggest replacing the standard deviation around the mean with deviation around the target.

$$\sigma_T = \sqrt{\sum_{i=1}^{n} \frac{(x_i - T)^2}{n-1}}$$

When both LSL and USL are given and target level is equal to the sample mean (i.e. $\overline{X} = T$), then:

$$Cpm = \frac{USL - LSL}{6\sigma_T}$$

When both LSL and USL are given and target level is not equal to the sample mean (i.e. $\overline{X} \neq T$), then:

$$Cpm = Min\left(\frac{USL - \overline{X}}{3\sigma_T}, \frac{\overline{X} - LSL}{3\sigma_T}\right)$$

When LSL is given but USL is not available:

$$Cpm = \frac{\overline{X} - LSL}{3\sigma_T}$$

and when USL is given but LSL is not available:

$$Cpm = \frac{USL - \overline{X}}{3\sigma_T}$$

The confidence intervals are given as:

$$LB = Cpm\sqrt{\frac{\chi^2_{\alpha,\nu}}{\nu}}$$

$$UB = Cpm\sqrt{\frac{\chi^2_{1-\alpha,\nu}}{\nu}}$$

where the degrees of freedom is:

$$\nu = n\left(\frac{(1 + a^2)^2}{(1 + 2a^2)}\right)$$

and:

$$a = \frac{\overline{X} - T}{\sigma_O}$$

When both LSL and USL are given and target level is equal to the sample mean (i.e. $\overline{X} = T$), then the one-sided alpha is used:

$$LB = Cpm\sqrt{\frac{\chi^2_{\alpha/2,\nu}}{\nu}}$$

$$UB = Cpm\sqrt{\frac{\chi^2_{1-\alpha/2,\nu}}{\nu}}$$

(ii) Boyles, R. A. (1991) suggests the use of following term to replace the standard deviation around the mean:

$$\sigma'_T = \sqrt{\left(\frac{n-1}{n}\right)\sigma_O^2 + \left(\overline{X} - T\right)^2}$$

and calculations are as above.

(iii) Modified Boyles: In the calculation of Cpm, a standard deviation around the target similar to Boyles' is used, but without the term correcting for the degrees of freedom:

$$\sigma''_T = \sqrt{\sigma_O^2 + \left(\overline{X} - T\right)^2}$$

$$Cpm = \frac{USL - LSL}{6\sigma''_T}$$

This is the Cpm reported by SAS. However, when SAS calculates the confidence intervals, it re-calculates the Cpm, this time using Boyles' definition of standard deviation around the target as:

$$\sigma'_T = \sqrt{\left(\frac{n-1}{n}\right)\sigma_O^2 + \left(\overline{X} - T\right)^2}$$

$$Cpm = \frac{USL - LSL}{6\sigma'_T}$$

and calculates the confidence intervals using this value and the two-tailed chi-square distribution as:

$$LB = Cpm\sqrt{\frac{\chi^2_{\alpha/2,\nu}}{\nu}}$$

$$UB = Cpm\sqrt{\frac{\chi^2_{1-\alpha/2,\nu}}{\nu}}$$

where, as above, the degrees of freedom is:

$$\nu = n\left(\frac{\left(1+a^2\right)^2}{\left(1+2a^2\right)}\right)$$

and:

$$a = \frac{\overline{X}-T}{\sigma_O}$$

(iv) NIST also employs a standard deviation around the target without the term correcting the degrees of freedom:

$$\sigma''_T = \sqrt{\sigma^2_O + \left(\overline{X}-T\right)^2}$$

$$Cpm = \frac{USL-LSL}{6\sigma''_T}$$

This is the Cpm reported by NIST, which is also used in calculating the confidence intervals. NIST also reports confidence intervals based on the two-tailed chi-square distribution:

$$LB = Cpm\sqrt{\frac{\chi^2_{\alpha/2,\nu}}{\nu}}$$

$$UB = Cpm\sqrt{\frac{\chi^2_{1-\alpha/2,\nu}}{\nu}}$$

Cpmk: This index is a useful variation of Cpm, which not only warns against process mean deviating from the target value, but also process variation getting larger.

$$\text{Cmpk} = \frac{d - \left|\overline{X} - m\right|}{3\sqrt{s_n + \left(\overline{X} - T\right)^2}}$$

where:

$$d = \frac{USL - LSL}{2}$$

$$m = \frac{USL + LSL}{2}$$

$$s_n = \left(\frac{n-1}{n}\right)\sigma_O^2$$

and the confidence intervals are given as:

$$LB, LU = \text{Cpmk} \pm Z_{1-\alpha/2}\frac{\sigma_{pmk}}{\sqrt{n}}$$

where:

$$\sigma_{pmk}^2 = \left(\frac{1}{9\left(1+\delta^2\right)} + \frac{2\delta}{3\left(1+\delta^2\right)^{1.5}}\right)C_{pmk} + \frac{72\delta^2 + d\left(\frac{m_4}{s_n^4} - 1\right)}{72\left(1+\delta^2\right)^2}C_{pmk}^2$$

and:

$$\delta = \frac{\overline{X} - T}{s_n}$$

Cs: Also known as *Wright's index*, this index is a variation of Cpmk that performs well even when the data is skewed. Wright, P. A. (1995):

$$Cs = \frac{d - \left|\overline{X} - T\right|}{3\sqrt{\frac{n-1}{n}\sigma_T^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)}\frac{c_4}{\sigma_O}\right|}}$$

where, as before:

$$\sigma_T = \sqrt{\sum_{i=1}^{n}\frac{(x_i - T)^2}{n-1}}$$

$$m_3 = \sum_{i=1}^{n}\frac{\left(x_i - \overline{X}\right)^3}{n}$$

and c4 is the unbiasing constant.

Cpm+: Boyles, R. A. (1992) proposed this index for use when both LSL and USL are given and $LSL \neq T$, $USL \neq T$:

$$Cpm+ = \frac{1}{3n}\sqrt{\frac{\sum_{Xi<T}(Xi - T)^2}{(T - LSL)^2} + \frac{\sum_{Xi>T}(Xi - T)^2}{(USL - T)^2}}$$

Cjkp: Also called the *flexible index*, this is for use when both LSL and USL are given, Kotz, S. and Johnson, N. L. (1993):

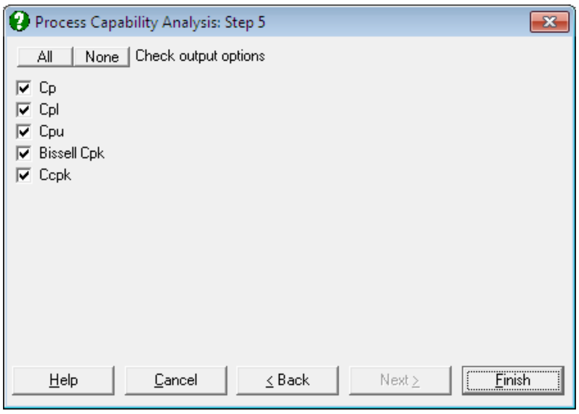
$$Cjkp = \frac{1}{3\sqrt{2}}\mathrm{Min}\left(\frac{T - LSL}{\sqrt{\sum_{Xi<T}(Xi - T)^2 / n}} + \frac{USL - T}{\sqrt{\sum_{Xi>T}(Xi - T)^2 / n}}\right)$$

Capability Indices: Pooled Standard Deviation: If more than one data variable and / or one or more factor variables are selected, then capability indices are also calculated based on the pooled standard deviation and degrees of freedom.

$$\sigma_P = \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1)\sigma_i^2}{\sum_{i=1}^{k}(n_i - 1)}}$$

$$df = \sum_{i=1}^{k}(n_i - 1) = n - k$$

where k is the number of subgroups.



Check output options

☑ Cp
☑ Cpl
☑ Cpu
☑ Bissell Cpk
☑ Ccpk

Cp, Cpl, Cpu, Cpk: These indices are calculated as above, but the pooled standard deviation and the pooled degrees of freedom are used.

Ccpk: This index is reported for subgroup analysis only. It is similar to Cpk, but is centred at the target, when a target value is provided.

When both LSL and USL are given, then:

$$\mathrm{Ccpk} = \mathrm{Min}\left(\frac{\mathrm{USL} - \mathrm{t}}{3\sigma_\mathrm{P}}, \frac{\mathrm{t} - \mathrm{LSL}}{3\sigma_\mathrm{P}}\right)$$

where σ_P is the pooled standard deviation and:

$$\mathrm{t} = \mathrm{T}$$

when a target value is given, and:

$$\mathrm{t} = \frac{\mathrm{USL} + \mathrm{LSL}}{2}$$

otherwise.

When LSL is given but USL is not available:

$$\mathrm{Ccpk} = \frac{\mathrm{t} - \mathrm{LSL}}{3\sigma_\mathrm{P}}$$

and when USL is given but LSL is not available:

$$\mathrm{Ccpk} = \frac{\mathrm{USL} - \mathrm{t}}{3\sigma_{\mathrm{P}}}$$

In both cases,

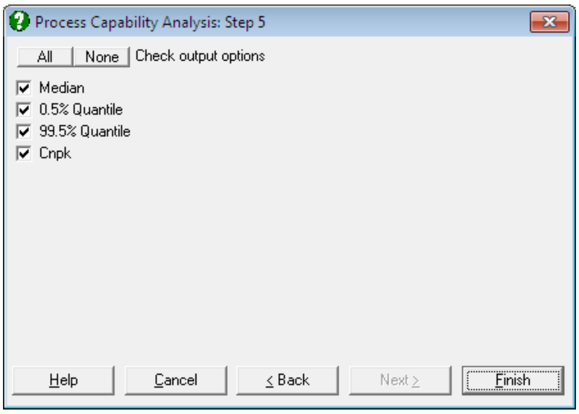
$$\mathrm{t} = \mathrm{T}$$

when a target value is given, and:

$$\mathrm{t} = \overline{\mathrm{X}}$$

otherwise.

Capability Indices: Nonparametric: The standard Process Capability Analysis assumes normal distribution of data. When this is not the case, a nonparametric (i.e. distribution-free) capability index will be useful. The methods used in computing the quantiles and their confidence limits are reported in the header. These methods can be changed using the dialogues of the Quantiles (Percentiles) procedure (see sections 5.1.3.1. Quantile Methods and 5.1.3.2. Quantile Interval Methods).



Check output options

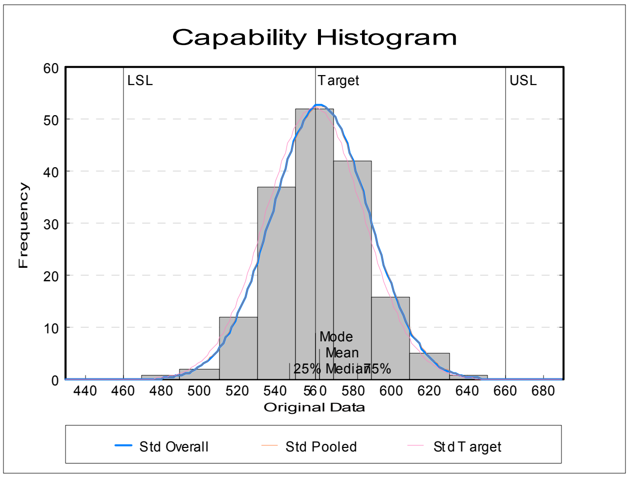
☑ Median
☑ 0.5% Quantile
☑ 99.5% Quantile
☑ Cnpk

Cnpk is a variant of Cpk used for non-normal data and is defined as:

$$\mathrm{Cnpk} = \mathrm{Min}\left(\frac{\mathrm{m} - \mathrm{LSL}}{\mathrm{m} - \mathrm{P}_{(0.005)}}, \frac{\mathrm{USL} - \mathrm{m}}{\mathrm{P}_{(0.995)} - \mathrm{m}} \right)$$

where m is the median and P(0.005) and P(0.995) are the 0.5^{th} and 99.5^{th} percentiles respectively.

Capability Histogram: A histogram of data is displayed to help visualise its distribution. LCL, target, UCL values are indicated as well as mean median mode and quartiles (also see Histogram).



Up to six distributions can also be fitted and displayed on the histogram. The first three of these are reserved by the program to display normal curves with overall and pooled (if data has subgroups) standard deviations and with deviation around the target (if a target has been specified, see definition of σ_T above). The remaining three distributions are set by default to Weibull, lognormal and gamma distributions, but these can be changed by selecting Edit → Distributions dialogue from the graphics menu, after clicking on the [Opt] button situated to the left of the Capability Histogram check box on the Output Options Dialogue.

Normal Probability Plot: Original Data: A Normal Probability Plot of the original data is displayed together with Anderson-Darling Test results in the legend. You can compare this graph with the next one to visualise the improvement provided by the transformation – if there is one.

Normal Probability Plot: Transformed Data: This option is enabled if a transformation has been selected in the previous dialogue. A Normal Probability Plot of the transformed data is displayed together with Anderson-Darling Test results in the legend. You can compare this graph with the previous one to visualise the improvement provided by the transformation.

9.3.6.4. Process Capability Analysis Example

Open TIMESER and select Statistics 2 → Quality Control → Process Capability Analysis. From the Variable Selection Dialogue select *THICKNESS* (*C21*) as [Variable] and *ZONE* (*C19*) as [Factor]. On Step 2 select all levels of *ZONE* enter on the next dialogue 460, 560 and 660 for LSL, Target and USL respectively. Select Johnson Transformation.

Process Capability Analysis

Process Data

Variables Selected: THICKNESS
Subsample selected by: ZONE = 1, 2, 3, 4

Valid Cases =	168
Number of Subgroups =	4
Control Range =	6.0000
Original Process Data:	
LSL =	460.0000
Target =	560.0000
USL =	660.0000
Mean =	563.0357
Overall Standard Deviation =	25.3847
Pooled Standard Deviation =	25.5069
Transformed Process Data:	
LSL =	-3.7597
Target =	-0.1694
USL =	3.3276
Mean =	-0.0588
Overall Standard Deviation =	0.9600
Pooled Standard Deviation =	0.9644

Data Transformation: Results

Z-statistic for best fit =	0.7100
Gamma =	-0.3867
Delta =	4.8986
Xi =	126.7984
Lambda =	554.3748

Transformation selected: Johnson Unbounded System (SU)
z = Gamma + Delta * ASINH((x - Xi) / Lambda), Xi < x
z = -0.386691047854535 + 4.89859585514262 * ASINH((x - 554.374813248812) / 126.798395166468)

Normality Tests

Smaller probabilities indicate non-normality.

	A-D Stat	Probability
Original Data	0.2495	0.7427
Transformed Data	0.2299	0.8044

Transformed Data

	Original Data	Transformed Data	Group
1	487.0000	-2.8805	3
2	505.0000	-2.2490	3
3	506.0000	-2.2130	4
…	…	…	…
166	625.0000	2.2174	3
167	626.0000	2.2511	1
168	634.0000	2.5165	3

Performance: Parts Per Million

	PPM < LSL	PPM > USL	PPM Total
Observed	0.0000	0.0000	0.0000
Overall	57.7873	209.5823	267.3697
Pooled	62.1538	222.8973	285.0511

Performance: Percent

	% < LSL	% > USL	% Total
Observed	0.0000	0.0000	0.0000
Overall	0.0058	0.0210	0.0267
Pooled	0.0062	0.0223	0.0285

Capability Indices: Overall Standard Deviation

	Value	Lower 95%	Upper 95%
Cp	1.2305	1.0986	1.3623
Cpl	1.2851	0.9906	0.9919
Cpu	1.1759	0.9906	0.9919
Bissell Cpk	1.1759	1.0401	1.3117
ZSW Eq 6 Cpk	1.1759	1.0484	1.3034
ZSW Eq 8 Cpk	1.1759	1.0392	1.3126
Chang Cpm	1.2063	1.0974	1.3137
Boyles Cpm	1.2099	1.1006	1.3176
SAS Cpm	1.2224	1.0950	1.3569
NIST Cpm	1.2224	1.0918	1.3529
Cpmk	1.1716	1.0752	1.2680
Cs	1.1547		
Cpm+	0.0070		
Cjkp	1.1054		

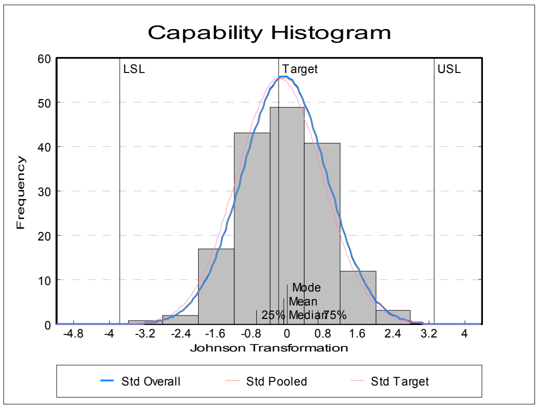
Capability Indices: Pooled Standard Deviation

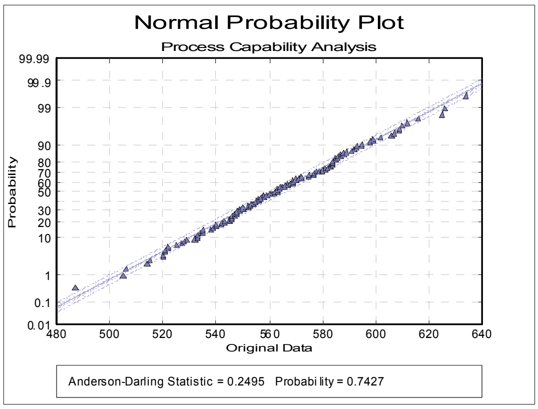
	Value	Lower 95%	Upper 95%
Cp	1.2248	1.0923	1.3571
Cpl	1.2792	0.9906	0.9919
Cpu	1.1705	0.9906	0.9919
Bissell Cpk	1.1705	1.0341	1.3068
Ccpk	1.2087		

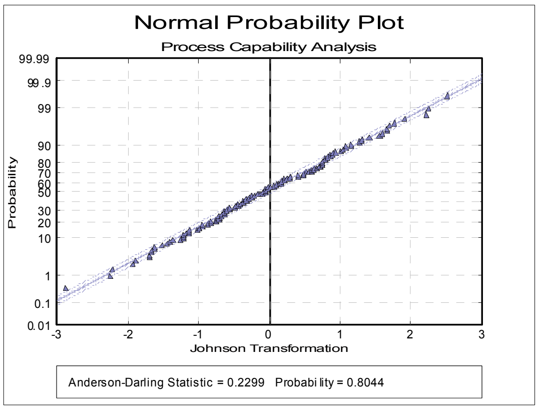
Capability Indices: Nonparametric

Quantile Method: Simple Average
Interval Method: Normal Approximation

	Value	Lower 95%	Upper 95%
Median	-0.0730	-0.3239	0.1003
0.5% Quantile	-2.8805	*	-1.9213
99.5% Quantile	2.5165	1.9091	*
Cnpk	1.3132		


Capability Histogram
Frequency
60
50
40
30
20
10
0
LSL
Target
USL
Mode
Mean
25%
Median
75%
-4.8
-4
-3.2
-2.4
-1.6
-0.8
0
0.8
1.6
2.4
3.2
4
Johnson Transformation
Std Overall
Std Pooled
Std Target

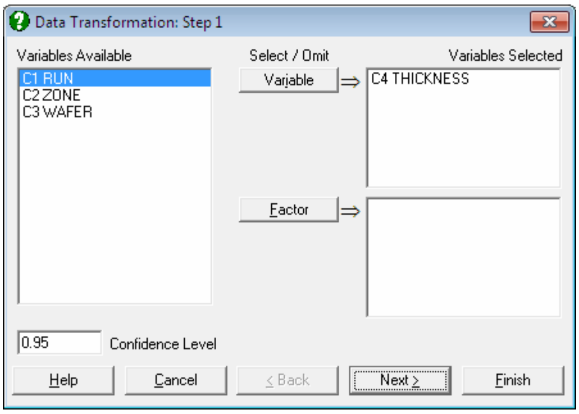

Normal Probability Plot
Process Capability Analysis
Probability
99.99
99.9
99
90
80
70
60
50
30
20
10
1
0.1
0.01
480
500
520
540
560
580
600
620
640
Original Data
Anderson-Darling Statistic = 0.2495 Probability = 0.7427


Normal Probability Plot
Process Capability Analysis
Probability
99.99
99.9
99
90
80
70
60
50
30
20
10
1
0.1
0.01
-3
-2
-1
0
1
2
3
Johnson Transformation
Anderson-Darling Statistic = 0.2299 Probability = 0.8044

9.3.7. Data Transformation

The standard Process Capability Analysis is one of many statistical procedures that assume normal distribution of data. When this cannot be assumed, either capability indices should be computed based on distributions other than normal, or the data should be transformed so that it conforms better to the normal distribution. This procedure provides two Data Transformation options; a family of Johnson transformations (see Kotz, S. and Johnson, N.L. 1993) and the Box-Cox Transformation (see Box, G. E. P. and Cox, D. R. 1964). These two methods are also available as Data Transformation options within the Process Capability Analysis procedure. You can also use the Box-Cox Regression procedure to transform a dependent variable when predictor (independent) variables exist.

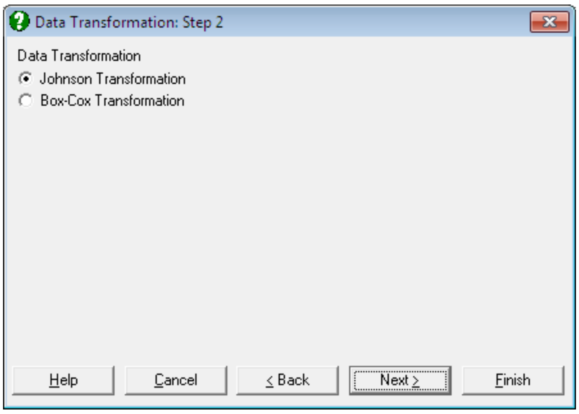
The Variable Selection Dialogue is of the multisample type (see 6.0.4. Multisample Tests), allowing selection of multiple data and factor columns. All selected data is pooled and sorted to form one continuous variable before it is transformed.



Variables Selected: C4 THICKNESS

0.95 Confidence Level

The next dialogue will ask for the type of transformation. Box-Cox Transformation will only work with positive data, whereas Johnson Transformation has no such restrictions. It must also be noted that in most cases Johnson Transformation is more powerful than Box-Cox Transformation and it will generate transformed variables more normally distributed.



Data Transformation: Step 2

Data Transformation

● Johnson Transformation

○ Box-Cox Transformation

9.3.7.1. Johnson Transformation

Johnson Transformation system consists of three types of curves:

Bounded system (SB):

$$y = \gamma + \eta \ \mathrm{Ln}\left(\frac{x - \varepsilon}{\lambda + \varepsilon - \chi}\right)$$

Log-normal system (SL):

$$y = \gamma + \eta \ \mathrm{Ln}\left(\frac{x - \varepsilon}{\lambda}\right)$$

Unbounded system (SU):

$$y = \gamma + \eta \ \mathrm{Sinh}^{-1}\left(\frac{x - \varepsilon}{\lambda}\right)$$

where:

$\mathrm{Sinh}^{-1}(z) = \mathrm{Ln}\left(z + \mathrm{Sqr}(1 + z^2)\right)$

y is the transformed value

γ is the shape 1 parameter

η is the shape 2 parameter

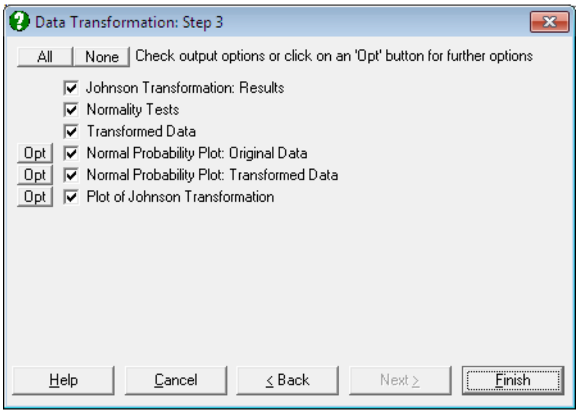
ε is the location parameter

λ is the scale parameter.

The program evaluates all three functions with the current estimates of four parameters, transforms the data and runs a normality test on the transformed data. The four parameters are optimised until one of the three transformation functions produces the best normality test result. The algorithm is based on Polansky, A. M., Chou, Y.-M., and Mason, R. L., (1999), but the Shapiro-Wilk normality test is replaced with the more accurate Anderson-Darling Test.

Johnson Transformation may not always provide a solution. The best way to find out whether a solution has been found is to ensure that the transformed data produces a higher Anderson-Darling Test probability than the original data.

9.3.7.1.1. Johnson Transformation Output Options



Data Transformation: Step 3

Check output options or click on an 'Opt' button for further options

☑ Johnson Transformation: Results
☑ Normality Tests
☑ Transformed Data
☑ Normal Probability Plot: Original Data
☑ Normal Probability Plot: Transformed Data
☑ Plot of Johnson Transformation

Johnson Transformation Results:

> **Parameter estimates:** The optimum levels for the four parameters are displayed.
>
> **Transformation Selected:** The selected Johnson function is displayed together with constraints on parameters. The same equation is also printed on a separate line with estimated parameter values, in a format suitable for cell calculations in Excel. You can simply copy this equation, replace the variable x with a cell reference and run interpolations.

Normality Tests: The Anderson-Darling Test of normality results are displayed for the original and transformed data. Higher probability values indicate better conformity to normal distribution.

Transformed Data: The sorted original data, the transformed data and their group membership (if any) are displayed in a table. If you are using UNISTAT in Stand-Alone Mode, click on the UNISTAT icon on the Output Medium Toolbar to send all output to UNISTAT spreadsheet. In Excel Add-In Mode select the output matrix as data for further calculations.

Normal Probability Plot of Original Data: A Normal Probability Plot of original data is displayed, together with its Anderson-Darling Test statistic and probability. You can compare this graph with the next one to visualise the improvement provided by the transformation.

Normal Probability Plot of Transformed Data: A Normal Probability Plot of transformed data is displayed, together with its Anderson-Darling Test statistic and probability. You can compare this graph with the previous one to visualise the improvement provided by the transformation.

Plot of Johnson Transformation: Probabilities for Anderson-Darling Test on the transformed data are plotted against the z-values. The maximum probability is indicated on the graph. The curve generated may not always be continuous.

9.3.7.1.2. Johnson Transformation Example

Open REGRESS and select Statistics 2 → Quality Control → Data Transformation. From the Variable Selection Dialogue select *cm* (*C2*) as [Variable]. On Step 2 leave convergence parameters unchanged. On the Output Options Dialogue check all options to obtain the following output.

Data Transformation

Johnson Transformation: Results

Variables Selected: cm

Z-statistic for best fit =	0.7200
Gamma =	0.5500
Delta =	0.6075
Xi =	5.6319
Lambda =	6.8365

```
Transformation selected: Johnson Bounded System (SB)
z = Gamma + Delta * LN((x - Xi) / (Xi + Lambda - x)), Xi < x < Xi + Lambda
z = 0.549976426928764 + 0.607452018062996 * LN((x - 6.83647857110731) /
(5.63190833413029 + 6.83647857110731 - x))
```

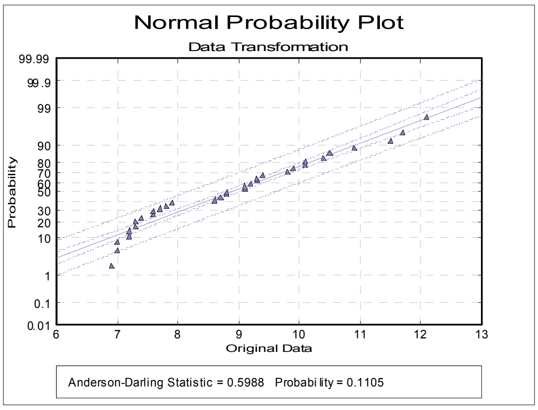
Normality Tests

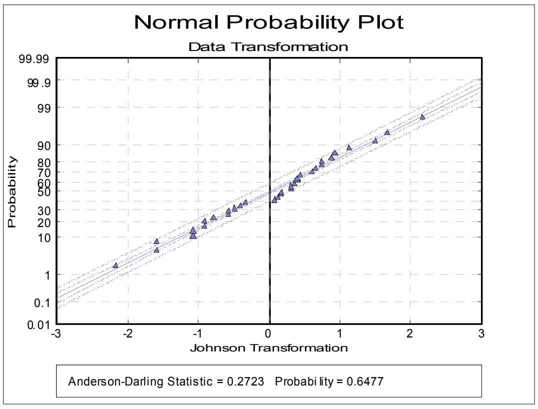
Smaller probabilities indicate non-normality.

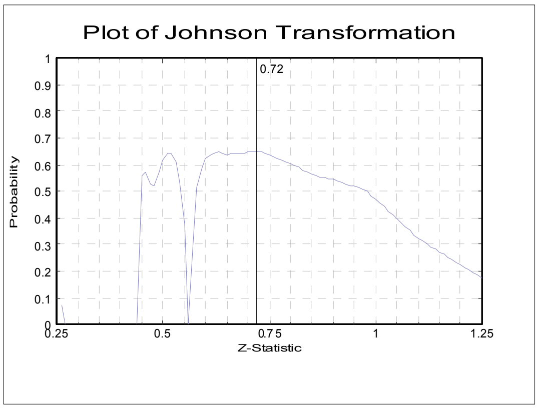
	A-D Stat	Probability
Original Data	0.5988	0.1202
Transformed Data	0.2723	0.6936

Transformed Data

	Original Data	Transformed Data
1	6.9000	-2.1675
2	7.0000	-1.5821
3	7.0000	-1.5821
…	…	…
31	11.5000	1.5048
32	11.7000	1.6709
33	12.1000	2.1654







9.3.7.2. Box-Cox Transformation

Box-Cox Transformation is a power transformation of the type:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0$$

$$y^{(\lambda)} = \text{Ln}(y) \text{ if } \lambda = 0$$

The optimal value of lambda is determined by maximising the following log-likelihood function:

$$L^{(\lambda)} = -\frac{N}{2} \text{Ln}(\hat{\sigma}^2_{(\lambda)}) + (\lambda - 1) \sum_{i=1}^{N} \text{Ln}(y_i)$$

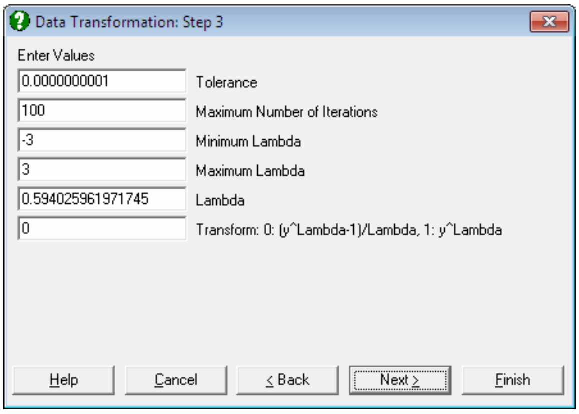
where $\hat{\sigma}^2_{(\lambda)}$ is the estimate of the variance of the transformed y variable.

The negative of the log likelihood function is minimised within a range defined by the user. The default range is $-3 \leq \lambda \leq 3$.

Box-Cox Transformation may not always provide a solution. The best way to find out whether a solution has been found is to ensure that the transformed data produces a higher Anderson-Darling Test probability than the original. Also, you will notice that in most cases Johnson Transformation provides a better transformation than Box-Cox Transformation.

9.3.7.2.1 Box-Cox Transformation Intermediary Inputs

This dialogue is similar to the Intermediate Inputs dialogue for Box-Cox Regression, except for the last item (see 7.2.9. Box-Cox Regression).


Data Transformation: Step 3

Enter Values

0.0000000001 Tolerance

100 Maximum Number of Iterations

-3 Minimum Lambda

3 Maximum Lambda

0.594025961971745 Lambda

0 Transform: 0: (y^Lambda-1)/Lambda, 1: y^Lambda

Tolerance: This value is used to control the sensitivity of minimisation procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Minimum Lambda: Limits for the range where the optimum lambda will be search can be set. Change this value if the optimal lambda cannot be found within the specified range. If the lambda displayed is the same or very near to this minimum, change it to a smaller value. When the limit is changed, a re-calculation is forced and lambda is estimated again.

Maximum Lambda: Change this value if the optimal lambda cannot be found within the specified range. If the lambda displayed is the same or very near to this maximum, change it to a higher value. When the limit is changed, a re-calculation is forced and lambda is estimated again.

Lambda: You can override the estimated lambda and enter your own value here. You may wish to do this to use a round power value (like -1, -0.5, 0.5, 2). If the estimated lambda is changed, confidence intervals and chi-squared tests for lambda will not be available.

Transform: Once the optimal lambda is estimated using the standard Box-Cox Transformation, you will have a chance to generate the transformed variable using, (i) the same transformation:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0$$

$$y^{(\lambda)} = \mathrm{Ln}(y) \text{ if } \lambda = 0$$

or, (ii) the simple power transformation:

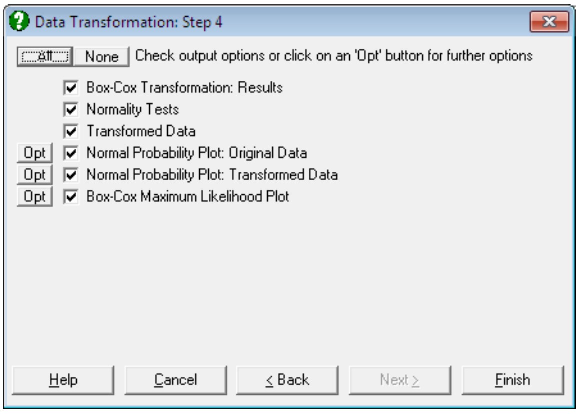
$$y^{(\lambda)} = y^{\lambda}$$

In some cases the second formula may be preferable to the first, since it will not generate nonpositive values. The choice made here will not affect normality of the transformed variable.

Remember that during the maximum likelihood estimation of lambda the original variable is always transformed using the first set of equations.

9.3.7.2.2 Box-Cox Transformation Output Options

This dialogue is similar to the maximum likelihood output dialogue for Box-Cox Regression (see 7.2.9.4. Box-Cox Regression Maximum Likelihood Output Options).



Data Transformation: Step 4

Check output options or click on an 'Opt' button for further options

☑ Box-Cox Transformation: Results
☑ Normality Tests
☑ Transformed Data
☑ Normal Probability Plot: Original Data
☑ Normal Probability Plot: Transformed Data
☑ Box-Cox Maximum Likelihood Plot

Box-Cox Transformation Results: This part of the output contains results for the maximum likelihood estimation.

Lambda with Confidence Intervals: The confidence interval for optimum lambda is based on the likelihood ratio statistic and defined as:

$$f(y,\hat{\lambda}) \geq f(y,\hat{\lambda}) - \frac{\chi^2_{\alpha,1}}{2}$$

Values corresponding to lower and upper bound of lambda are computed separately using an iterational procedure.

Likelihood Ratio Test: This test performed by evaluating the regression equation for lambda fixed at $\lambda_1 = -1$, 0 and 1.

$$L = 2\left[f(y,\hat{\lambda}) - f(y,\lambda_1)\right]$$

which is chi-square distributed with one degree of freedom.

Transformation Selected: The selected Box-Cox function is displayed. The two possibilities are:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0$$

and:

$$y^{(\lambda)} = \text{Ln}(y) \text{ if } \lambda = 0$$

The equation is also printed on a separate line with estimated parameter values, in a format suitable for cell calculations in Excel. You can copy this equation, replace the variable x with a cell reference and run interpolations.

Normality Tests: The Anderson-Darling Test of normality is performed on the original and the transformed data thus allowing you to judge whether the transformation was useful. No or little increase in the tail probability indicates that the Box-Cox Transformation was not useful.

Transformed Data: The sorted original data, the transformed data and their group membership (if any) are displayed in a table. If you are using UNISTAT in Stand-Alone Mode, click on the UNISTAT icon on the Output Medium Toolbar to send all output to UNISTAT spreadsheet. In Excel Add-In Mode select the output matrix as data for further calculations.

Normal Probability Plot of Original Data: A Normal Probability Plot of original data is displayed, together with its Anderson-Darling Test statistic and probability. You can compare this graph with the next one to visualise the improvement provided by the transformation.

Normal Probability Plot of Transformed Data: A Normal Probability Plot of transformed data is displayed, together with its Anderson-Darling Test statistic and probability. You can compare this graph with the previous one to visualise the improvement provided by the transformation.

Box-Cox Maximum Likelihood Plot: Values of the log-likelihood function are plotted against lambda. The estimated lambda and its confidence intervals are also indicated.

9.3.7.2.3 Box-Cox Transformation Example

Open REGRESS and select Statistics 2 → Quality Control → Data Transformation. From the Variable Selection Dialogue select *cm* (*C2*) as [Variable]. On Step 2 leaving the convergence parameters unchanged produces an invalid lower bound for lambda. Change the minimum value for lambda from -3 to -4 and on the Output Options Dialogue check all options to obtain the following output.

Data Transformation

Box-Cox Transformation: Results

Variables Selected: cm

	Value	Lower 95%	Upper 95%
Lambda	-0.7406	-3.1442	1.5260

Box-Cox Transformation:
y = (y ^ Lambda - 1) / Lambda
y = (POWER(y, -0.740557931816275) - 1) / -0.740557931816275

Lambda	Chi-Square	DoF	Probability
-1	0.0477	1	0.8272
0	0.3982	1	0.5280
1	2.2454	1	0.1340

Log of Likelihood = -11.9389

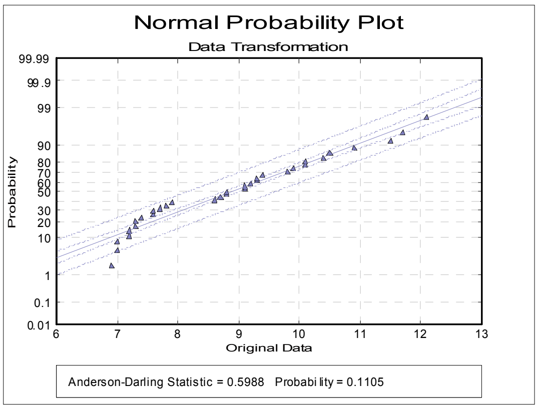
Normality Tests

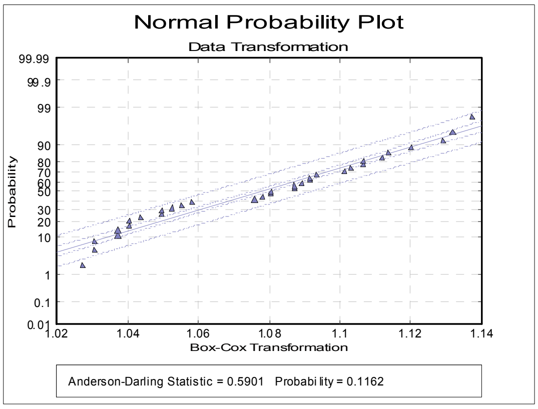
Smaller probabilities indicate non-normality.

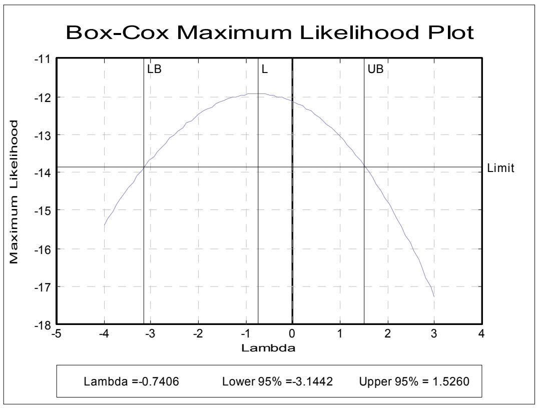
	A-D Stat	Probability
Original Data	0.5988	0.1202
Transformed Data	0.5901	0.1264

Transformed Data

	Original Data	Transformed Data
1	6.9000	1.0273
2	7.0000	1.0307
3	7.0000	1.0307
...	...	...
31	11.5000	1.1291
32	11.7000	1.1319
33	12.1000	1.1372







Box-Cox Maximum Likelihood Plot
Maximum Likelihood
-11
-12
-13
-14
-15
-16
-17
-18
LB
L
UB
Limit
-5
-4
-3
-2
-1
0
1
2
3
4
Lambda
Lambda =-0.7406
Lower 95% =-3.1442
Upper 95% = 1.5260

9.3.8. Gauge R&R Analysis

Gauge repeatability and reproducibility analysis is designed to ensure stability and consistency of measurements made on an instrument by one or more operators. *Repeatability* refers to variation in measurements due to the nature of the equipment while *reproducibility* refers to variation introduced by the operators.

UNISTAT can perform Gauge R&R Analysis employing or Average and Range or Analysis of Variance (ANOVA) methods. Output includes ANOVA table, gauge variances, standard deviations and confidence intervals. Range Chart and Average Chart can be displayed in Stacked or Unstacked form.

9.3.8.1. Gauge R&R Data Preparation

If data is given in the form of a table where measurements corresponding to different parts are in separate columns, it must be transformed into a more convenient format for analysis.

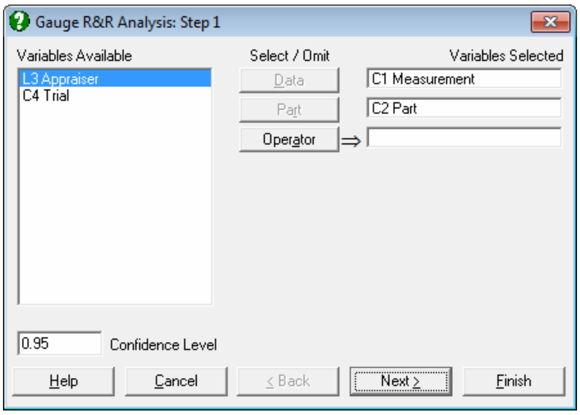
Consider the data given in Figure 12, p. 101 of AIAG (2002), with 3 operators, 3 trials and 10 parts.

Operator		Part									
	Trials	1	2	3	4	5	6	7	8	9	10
A	1	0.29	-0.56	1.34	0.47	-0.80	0.02	0.59	-0.31	2.26	-1.36
	2	0.41	-0.68	1.17	0.50	-0.92	-0.11	0.75	-0.20	1.99	-1.25
	3	0.64	-0.58	1.27	0.64	-0.84	-0.21	0.66	-0.17	2.01	-1.31
B	1	0.08	-0.47	1.19	0.01	-0.56	-0.20	0.47	-0.63	1.80	-1.68
	2	0.25	-1.22	0.94	1.03	-1.20	0.22	0.55	0.08	2.12	-1.62
	3	0.07	-0.68	1.34	0.20	-1.28	0.06	0.83	-0.34	2.19	-1.50
C	1	0.04	-1.38	0.88	0.14	-1.46	-0.29	0.02	-0.46	1.77	-1.49
	2	-0.11	-1.13	1.09	0.20	-1.07	-0.67	0.01	-0.56	1.45	-1.77
	3	-0.15	-0.96	0.67	0.11	-1.45	-0.49	0.21	-0.49	1.87	-2.16

We need to stack all measurements in a single column and create three categorical data columns (or *factors*) to keep track of which measurement belongs to which operator, which trial and which part. For a fully specified table we need to end up with four columns in the worksheet, one data and three factor columns, though the last factor column *Trial* is not used in Gauge R&R Analysis.

In Stand-Alone Mode, *Part*, *Operator* and *Trial* factors can be generated automatically using the UNISTAT spreadsheet functions **Level(10)**, **Level(30);B** and **Level(10);B** respectively (see 3.4.2.5. Statistical Functions).

Data	Pt	Op	Tr
0.29	1	A	1
-0.56	2	A	1
1.34	3	A	1
0.47	4	A	1
-0.80	5	A	1
0.02	6	A	1
0.59	7	A	1
-0.31	8	A	1
2.26	9	A	1
-1.36	10	A	1
0.41	1	A	2
-0.68	2	A	2
1.17	3	A	2
0.50	4	A	2
-0.92	5	A	2
-0.11	6	A	2
0.75	7	A	2
-0.20	8	A	2
1.99	9	A	2
-1.25	10	A	2
0.64	1	A	3
-0.58	2	A	3
1.27	3	A	3
0.64	4	A	3
-0.84	5	A	3
-0.21	6	A	3
0.66	7	A	3
-0.17	8	A	3
2.01	9	A	3
-1.31	10	A	3
0.08	1	B	1
-0.47	2	B	1
1.19	3	B	1
0.01	4	B	1
-0.56	5	B	1
-0.20	6	B	1
0.47	7	B	1
-0.63	8	B	1
1.80	9	B	1
-1.68	10	B	1
0.25	1	B	2
-1.22	2	B	2
0.94	3	B	2
1.03	4	B	2
-1.20	5	B	2
0.22	6	B	2
0.55	7	B	2
0.08	8	B	2
2.12	9	B	2
-1.62	10	B	2
0.07	1	B	3
-0.68	2	B	3
1.34	3	B	3
0.20	4	B	3
-1.28	5	B	3
0.06	6	B	3
0.83	7	B	3
-0.34	8	B	3
2.19	9	B	3
-1.50	10	B	3
0.04	1	C	1
-1.38	2	C	1
0.88	3	C	1
0.14	4	C	1
-1.46	5	C	1
-0.29	6	C	1
0.02	7	C	1
-0.46	8	C	1
1.77	9	C	1
-1.49	10	C	1
-0.11	1	C	2
-1.13	2	C	2
1.09	3	C	2
0.20	4	C	2
-1.07	5	C	2
-0.67	6	C	2
0.01	7	C	2
-0.56	8	C	2
1.45	9	C	2
-1.77	10	C	2
-0.15	1	C	3
-0.96	2	C	3
0.67	3	C	3
0.11	4	C	3
-1.45	5	C	3
-0.49	6	C	3
0.21	7	C	3
-0.49	8	C	3
1.87	9	C	3
-2.16	10	C	3



Gauge R&R Analysis: Step 1

Variables Available: L3 Appraiser, C4 Trial

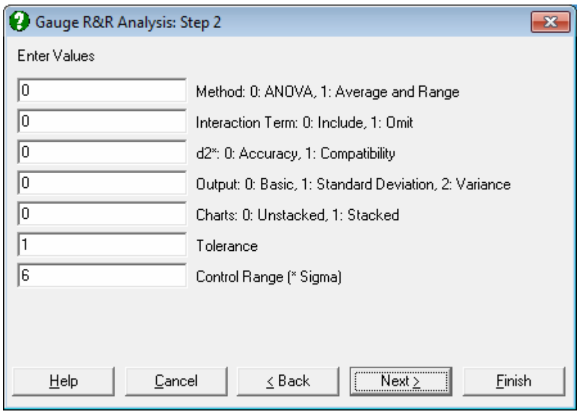
Variables Selected: Data: C1 Measurement; Part: C2 Part; Operator:

0.95 Confidence Level

In the Variable Selection Dialogue, the column containing all measurements is selected as [Data] and the factor column containing part numbers is selected as [Part]. These are the two compulsory variables without which an analysis is not possible. If there are two or more operators taking measurements, this information should be provided in the factor column [Operator].

9.3.8.2. Gauge R&R Intermediary Input Options

In step 2 you can define parameters for the analysis and its charts.



0 Method: 0: ANOVA, 1: Average and Range
0 Interaction Term: 0: Include, 1: Omit
0 d2*: 0: Accuracy, 1: Compatibility
0 Output: 0: Basic, 1: Standard Deviation, 2: Variance
0 Charts: 0: Unstacked, 1: Stacked
1 Tolerance
6 Control Range (* Sigma)

Method: This can be Analysis of Variance (ANOVA) or Average and Range. The ANOVA method is more powerful than Average and Range and it includes confidence intervals.

Interaction Term: When the ANOVA method is selected, you can choose to omit the interaction term between Part and Operator from the analysis. It is advisable to perform the analysis including the interaction term first. If this term in the ANOVA table is highly insignificant (i.e. if it has a high probability), then the analysis can be repeated excluding the interaction term.

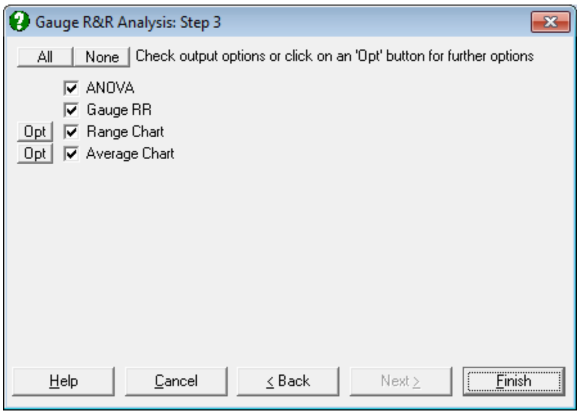
d2*: When the Average and Range method is selected, UNISTAT uses accurate unbiasing constants with four significant digits (Appendix C, p. 195, AIAG, 2002). These may, however, generate slightly different results compared with other applications which use less accurate constants with only two significant digits (Table D3, Duncan 1974). You can select here the use of accurate or compatible constants.

Output: When this is Basic, for the Average and Range method, only the study variation is displayed; for the ANOVA method, output also includes the confidence intervals of study variation. When Standard Deviation is selected, output contains standard deviations, percentages and, if the method is ANOVA, confidence intervals for standard deviations. When Variance is selected, standard deviation values are squared.

Charts: Range Chart (R Chart) and Average Chart (X Bar Chart) can be displayed. By default (Unstacked), trials by different operators are represented separately along the X-axis. When the Stacked option is selected, each operator's data is plotted for the same trial value on the X-axis.

Tolerance: When this value is other than the default value of unity, a % tolerance column is added to the output. % tolerance values are obtained by dividing the study variation by the supplied tolerance value (times 100).

Control Range: The observed process variation in terms of sample standard deviation. This is usually 6, ± 3 sigma around the centre. If the sample is from a normally distributed population, then approximately 99.73% of all data points would fall within this range. Other commonly used values are 5.15 for 99% and 4 for 95% coverage respectively.


Gauge R&R Analysis: Step 3

Check output options or click on an 'Opt' button for further options

☑ ANOVA
☑ Gauge RR
☑ Range Chart
☑ Average Chart

9.3.8.3. Gauge R&R Average and Range Method

This method is supported for historical reasons, because it is possible to perform calculations by hand, with the help of some published tables. Where possible, the use of ANOVA method should be preferred.

For Average and Range method and charts we need to find the mean range. The difference between maximum and minimum measurements for each trial and each operator are computed:

$$R_{ij} = \lfloor \mathrm{Max}(X_{ijk}), k = 1,\ldots,t \rfloor - \lfloor \mathrm{Min}(X_{ijk}), k = 1,\ldots,t \rfloor, i = 1,\ldots,o, j = 1,\ldots,p$$

where t is the number of trials. The mean range is then defined as:

$$\overline{\overline{R}} = \frac{1}{op}\sum_{i=1}^{o}\sum_{j=1}^{p} R_{ij}$$

where o is the number of operators and p is the number of parts. Also define the range of operator averages and the range of part averages respectively as:

$$R_o = \lfloor \mathrm{Max}(\overline{X}_j), j = 1,\ldots,p \rfloor - \lfloor \mathrm{Min}(\overline{X}_j), j = 1,\ldots,p \rfloor$$

$$R_p = [\mathrm{Max}(\overline{X}_i), i = 1,\ldots,o] - [\mathrm{Min}(\overline{X}_i), i = 1,\ldots,o]$$

where:

$$\overline{X}_j = \frac{1}{ot}\sum_{i=1}^{o}\sum_{k=1}^{t} X_{ijk}, j = 1,\ldots,p$$

$$\overline{X}_i = \frac{1}{pt}\sum_{j=1}^{p}\sum_{k=1}^{t} X_{ijk}, i = 1,\ldots,o$$

Then the output parameters for Average and Range method is calculated as follows.

Repeatability (Equipment Variation):

$$EV = \frac{\overline{\overline{R}}}{d_2^*(t, op)}$$

Reproducibility (Appraiser Variation):

$$AV = \sqrt{\left(\frac{R_o}{d_2^*(o,1)}\right)^2 - \left(\frac{EV^2}{op}\right)}$$

Repeatability & Reproducibility (Gauge R&R):

$$RR = \sqrt{EV^2 + AV^2}$$

Part Variation:

$$PV = \frac{R_P}{d_2^*(p,1)}$$

Total Variation:

$$TV = \sqrt{RR^2 + PV^2}$$

The percentage variation is calculated as 100 x Variation / TV and the number of distinct categories as Int(1.41(PV/RR)).

9.3.8.4. Gauge R&R ANOVA Method

The default model is a balanced two-way ANOVA with interaction. Balanced means each operator measures the same number of parts and the same number of trials. It is possible to omit the interaction term or run a balanced one-way ANOVA when there is one operator only. By default, the following table is displayed.

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Part	SS_P	DF_P	MS_P	F_P	P_P
Operator	SS_O	DF_O	MS_O	F_O	P_O
Part x Operator	SS_{PO}	DF_{PO}	MS_{PO}	F_{PO}	P_{PO}
Repeatability	SS_R	DF_R	MS_R		
Total	SS_T	DF_T	MS_T		

The Mean Square (MS) values from the ANOVA table are used to construct study variation values as follows.

Repeatability (Equipment Variation):

$$EV = \sqrt{MS_R}$$

Reproducibility (Appraiser Variation):

$$AV = \sqrt{\frac{MS_O \text{ - } MS_{OP}}{pt}}$$

Interaction:

$$IV = \sqrt{\frac{MS_{OP} \text{ - } EV^2}{t}}$$

Repeatability & Reproducibility (Gauge R&R):

$$RR = \sqrt{EV^2 + AV^2 + IV^2}$$

Part Variation:

$$PV = \sqrt{\frac{MS_P \text{ - } MS_{OP}}{ot}}$$

Total Variation:

$$TV = \sqrt{RR^2 + PV^2}$$

The confidence intervals are computed employing the modified large sample method given in Montgomery, D. C. (2009).

The percentage variation is calculated as 100 x Variation / TV and the number of distinct categories as Int(1.41(PV/RR)).

9.3.8.5. Gauge R&R Charts

The two types of charts that can be plotted are Range Chart and Average Chart, which are basically an R Chart and an X Bar Chart for two factor variables respectively. Range Chart is a graphical representation of repeatability and shows the consistency of the gage variability. Average Chart represents reproducibility (or operator variability) and part variation.

For Range Chart, the difference between maximum and minimum measurements are computed for each trial:

Assuming the Control Range is 6, the control limits are found as:

$$LCL = \overline{\overline{R}} - 3\frac{\overline{\overline{R}}}{d_2}d_3$$

$$T\arg et = \overline{\overline{R}}$$

$$UCL = \overline{\overline{R}} + 3\frac{\overline{\overline{R}}}{d_2}d_3$$

For Average Chart, the overall mean is calculated as:

$$\bar{\bar{X}} = \frac{1}{opt}\sum_{i=1}^{o}\sum_{j=1}^{p}\sum_{k=1}^{t} X_{ij}$$

And the control limits are:

$$LCL = \bar{\bar{X}} - 3\frac{\bar{\bar{R}}}{d_2}\frac{1}{\sqrt{t}}$$

$$\text{Target} = \bar{\bar{X}}$$

$$UCL = \bar{\bar{X}} + 3\frac{\bar{\bar{R}}}{d_2}\frac{1}{\sqrt{t}}$$

If the Unstacked option is selected trials by different operators are represented separately along the X-axis. If the Stacked option is selected, each operator's data is plotted for the same trial value on the X-axis.

9.3.8.6. Gauge R&R Examples

Example 1

Data is given in Figure 12 *Gage Repeatability and Reproducibility Data Collection Sheet* on p. 101 of AIAG (2002). Note that AIAG (2002) reports standard deviations, rather than the variance or study variation.

Open TIMESER and select Statistics 2 → Quality Control → Gauge R&R Analysis. Select *Measurement* (*C22*) as [Data], *Part* (*C23*) as [Part] and *Appraiser* (*S24*) as [Operator]. On Step 2 enter 1 for Method: Average and Range and leave all other parameters unchanged. On the Output Options Dialogue check all options to obtain the following output.

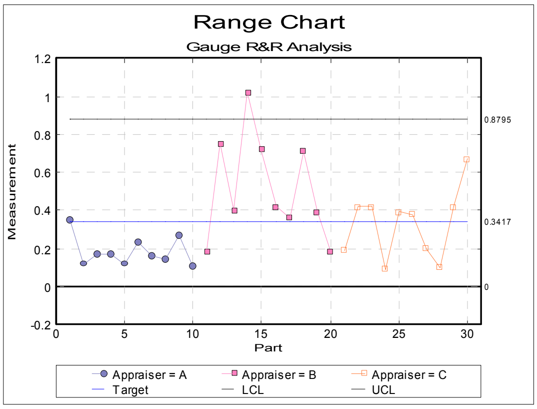
Gauge R&R Analysis

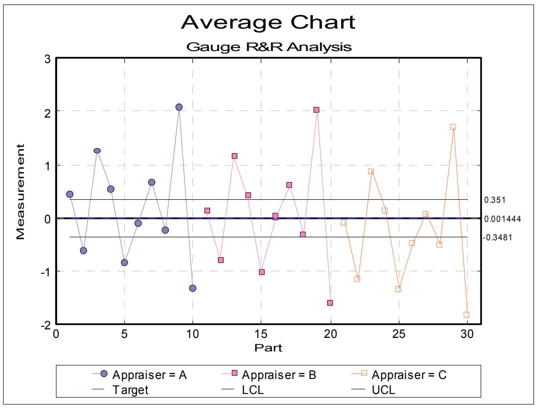
Gauge RR

```
Data: Measurement
Part: Part
Operator: Appraiser
Method: Average and Range
Control Range: 6 x Sigma
```

	Standard Deviation	% Variation	Variation	% Variation
Gauge RR	0.3058	26.68	1.8347	26.68
Repeatability EV	0.2019	17.61	1.2112	17.61
Reproducibility AV	0.2297	20.04	1.3781	20.04
Part PV	1.1045	96.37	6.6267	96.37
Total TV	1.1460	100.00	6.8760	100.00

Number of distinct categories: 5





Example 2

Continuing from the last example, on the Variable Selection Dialogue edit the Confidence Level as 0.9, on the second dialogue enter the following parameters and click [Finish].

- **0** Method: ANOVA
- **1** Interaction Term: Omit
- **0** d2* (not used)
- **1** Extended Output: Yes
- **1** Charts: Stacked
- **0.4** Tolerance
- **6** Control Range

Gauge R&R Analysis

ANOVA

Data: Measurement

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Part	88.362	9	9.818	245.614	0.0000
Operator	3.167	2	1.584	39.617	0.0000
Repeatability	3.118	78	0.040		
Total	94.647	89	1.063		

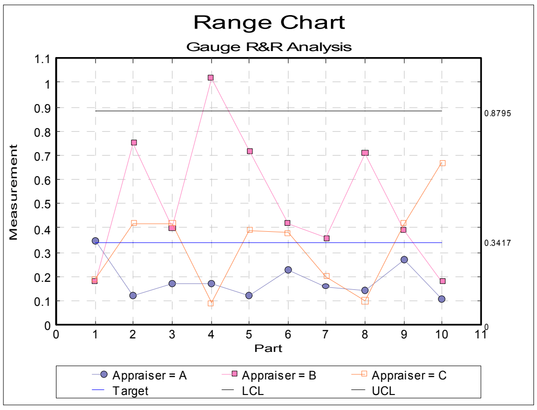
Gauge RR

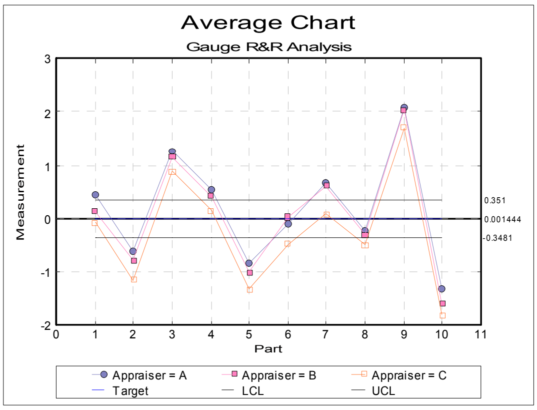
Data: Measurement
Part: Part
Operator: Appraiser
Method: ANOVA
Tolerance: 0.4
Control Range: 6 x Sigma

	Standard Deviation	Lower 90%	Upper 90%	% Variation
Gauge RR	0.3024	0.2351	1.0334	27.86
Repeatability EV	0.1999	0.1769	0.2306	18.42
Reproducibility AV	0.2268	0.1275	1.0138	20.90
Part PV	1.0423	0.7588	1.7170	96.04
Total TV	1.0853	0.8161	1.8111	100.00

	Variation	Lower 90%	Upper 90%	% Variation	% Tolerance
Gauge RR	1.8142	1.4107	6.2002	27.86	453.56
Repeatability EV	1.1996	1.0615	1.3834	18.42	299.90
Reproducibility AV	1.3610	0.7653	6.0827	20.90	340.26
Part PV	6.2540	4.5529	10.3021	96.04	1563.49
Total TV	6.5118	4.8966	10.8667	100.00	

Number of distinct categories: 4


Range Chart
Gauge R&R Analysis
Measurement
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10
11
Part
0.8795
0.3417
0
Appraiser = A
Appraiser = B
Appraiser = C
Target
LCL
UCL


Average Chart
Gauge R&R Analysis
Measurement
3
2
1
0
-1
-2
0
1
2
3
4
5
6
7
8
9
10
11
Part
0.351
0.001444
-0.3481
Appraiser = A
Appraiser = B
Appraiser = C
Target
LCL
UCL

9.4. Survival Analysis

Survival Analysis is useful when the dependent variable represents the time elapsed between an initial event and a termination event. This is most often the case in medical research, where detection of the disease is the initial event and the patient's death is the termination event. This type of data frequently occurs in other disciplines as well, such as engineering (e.g. failure time of components) or social sciences (e.g. survival of marriages). In these types of problems, researchers are often faced with the task of estimating a survival function (the probability that an individual is alive at time t) or a hazard function (the probability of failure at a time period t + Δt, given that the individual has survived until time t).

In principle, one could use standard parametric statistics for describing the average survival times and comparing effects of treatments. However these methods would not consider censored data, that is, data where the termination event has never occurred. Data may be censored because the patient has entered the study too late, patients may have survived the whole study period, or because of patients with whom we have lost contact.

When there are no censored cases in data, it is a relatively straightforward exercise to estimate the survival function. The empirical survival function is defined simply as:

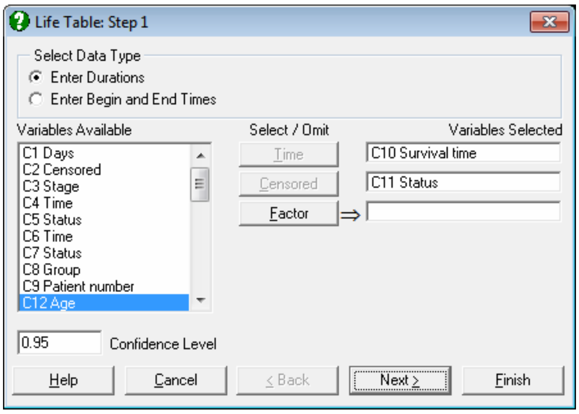
$$S(t) = \frac{\text{Number of cases surviving} \geq t}{\text{Total number of cases}}$$

However, when there are censored cases in data or when we want take into consideration other factors that may affect the survival times (categorical variables such as sex, region, etc. or continuous variables such as temperature, age), then one of the Survival Analysis methods available in this section should be used. UNISTAT will also estimate the empirical survival function when a censor variable is not selected.

9.4.0. Survival Variable Selection

In Survival Analysis, two alternative types of Time Data can be selected. Each row in the data matrix represents a separate case.

Enter Durations: One column containing the duration of each case is selected by clicking on [Time].



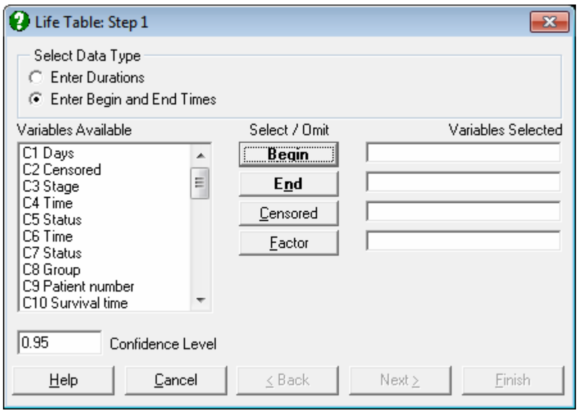
Select Data Type
● Enter Durations
○ Enter Begin and End Times

Variables Available
C1 Days
C2 Censored
C3 Stage
C4 Time
C5 Status
C6 Time
C7 Status
C8 Group
C9 Patient number
C12 Age

Variables Selected
C10 Survival time
C11 Status

0.95 Confidence Level

Enter Begin and End Times: Select one column as the starting time by clicking on [Begin] and another column as the final time by clicking on [End]. The second column is subtracted from the first one to obtain the duration of each case. If subtraction results in some non positive values, these are excluded from the analysis as invalid (missing) cases. It is possible to use data in date format, in which case the number of days between the two dates will be used as the duration data (see 3.0.2.3. Date Data).



Select Data Type
○ Enter Durations
● Enter Begin and End Times

Variables Available
C1 Days
C2 Censored
C3 Stage
C4 Time
C5 Status
C6 Time
C7 Status
C8 Group
C9 Patient number
C10 Survival time

Variables Selected

0.95 Confidence Level

Censored: This variable is selected to show whether the termination time of a case is not known or the termination event has occurred. If the value in the column is zero then it is assumed that the case has been censored, which means that it has been excluded from the study before the termination event has occurred. By default, non-zero values (usually one) indicate that the termination event has occurred for this case. If a censor variable is not selected, then it is assumed that termination event has occurred for all cases.

The Cox Regression procedure provides a facility to change the default value of 0 for censored cases in a dialogue that pops up just after the Variable Selection Dialogue. You should be aware that once a change is made here, it applies to all Survival Analysis procedures throughout the session, which do not have a facility to edit this value.

Factor: One categorical data column may be selected by clicking on [Factor], to produce a table for all or some of the subgroups defined by this variable. Selection of a factor column is compulsory for the Survival Comparison Statistics procedure. In Cox Regression, the [Factor] selection has a slightly different meaning. It is still used to perform analysis on a number of subgroups, but the nature of the analysis changes to what is called a *stratified analysis*. The maximum number of strata is limited to six. For further information see 9.4.4. Cox Regression.

9.4.1. Life Table

The Life Table is useful in analysing data on failure times when some of the cases are censored – that is, when the times of failure are not known. First, the data is grouped into a number of intervals, spanning a time period specified by the user. Let:

- d_j = number of deaths at interval j, and
- c_j = number of censored cases at interval j.
- n_j = number of cases entering interval j.

The average number of cases who are at risk at interval j is then defined as:

$$n_j^{'} = n_j - c_j/2$$

The Life Table estimates the survival function as:

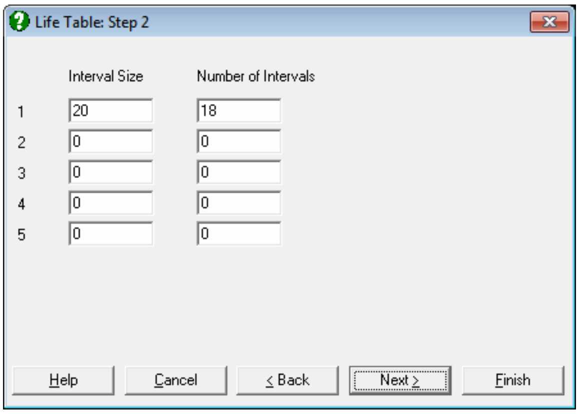
$$S(t) = \prod_{j=1}^{k} \frac{n_j^{'} - d_j}{n_j^{'}}$$

The survival and hazard functions are estimated and they are displayed together with their standard errors and confidence intervals for a user-defined confidence level (the default is 0.95).

9.4.1.1. Life Table Interval Selection

Once the time and optional status and factor variables have been selected (see 9.4.0. Survival Variable Selection), a second dialogue will ask for the intervals into which the data will be grouped.

UNISTAT will suggest a suitable selection that will cover the data in a regular pattern of intervals. It is also possible to enter a number of irregular intervals.



Life Table: Step 2

	Interval Size	Number of Intervals
1	20	18
2	0	0
3	0	0
4	0	0
5	0	0

The interval selection dialogue presents a matrix of text fields with five rows and two columns. For a model with regular intervals (the most common practice) it is sufficient to fill in the first row of the matrix. The first column Interval Size is used to enter the duration of each interval and the second column No of Intervals is for entering the number of intervals of this size. Consider the following input:

Interval Size	No of Intervals
5	4
10	2
0	0
0	0
0	0

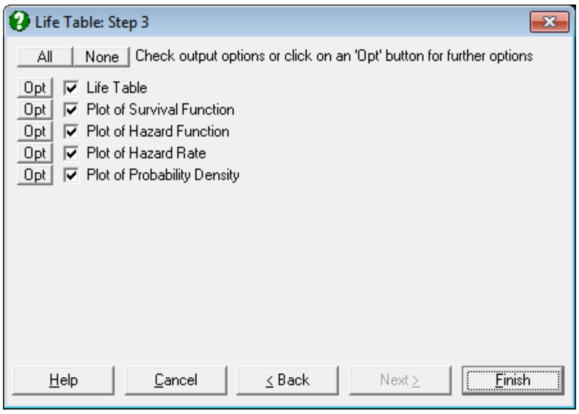
This would result in the following intervals.

Interval Start	
0	4 intervals of size 5 time units
5	
10	
15	
20	2 intervals of size 10 time units.
30	

When the total time span covered by the table does not include the entire data set, only those cases within the table range will be included in the model. If, on the other hand, the specified intervals cover more than the maximum time in data, then the Life Table will not display intervals after the last termination or censoring date.

9.4.1.2. Life Table Output Options

If a factor column has been included in the variable selection stage, then the program displays a further dialogue allowing you to select levels for which output is to be generated.



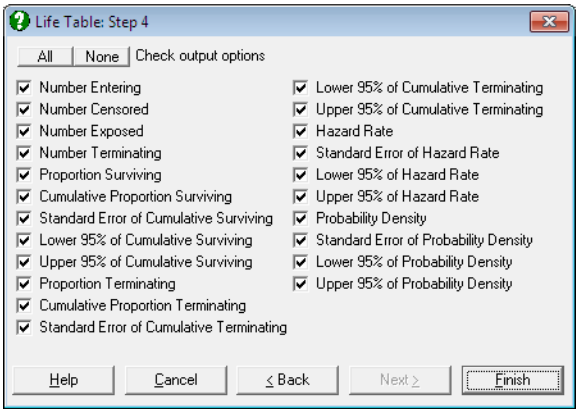
Check output options or click on an 'Opt' button for further options

☑ Life Table
☑ Plot of Survival Function
☑ Plot of Hazard Function
☑ Plot of Hazard Rate
☑ Plot of Probability Density

Next, the model is estimated and an Output Options Dialogue is displayed.

9.4.1.2.1. Life Table

You can select to display one or more of the following columns of the Life Table.



Check output options

☑ Number Entering
☑ Number Censored
☑ Number Exposed
☑ Number Terminating
☑ Proportion Surviving
☑ Cumulative Proportion Surviving
☑ Standard Error of Cumulative Surviving
☑ Lower 95% of Cumulative Surviving
☑ Upper 95% of Cumulative Surviving
☑ Proportion Terminating
☑ Cumulative Proportion Terminating
☑ Standard Error of Cumulative Terminating
☑ Lower 95% of Cumulative Terminating
☑ Upper 95% of Cumulative Terminating
☑ Hazard Rate
☑ Standard Error of Hazard Rate
☑ Lower 95% of Hazard Rate
☑ Upper 95% of Hazard Rate
☑ Probability Density
☑ Standard Error of Probability Density
☑ Lower 95% of Probability Density
☑ Upper 95% of Probability Density

Although the table options could have been grouped under four main categories of (1) Survival Function, (2) Hazard Function, (3) Hazard Rate and (4) Probability

Density (together with their standard errors and confidence intervals), here we prefer to combine them under one table to avoid repetitions. You can always obtain the desired table by unchecking the unwanted table columns.

Interval Start: The time at the start of the interval.

Number Entering:

n_j: The number of cases that enter the interval.

Number Censored:

c_j: The number of cases that are censored in the interval.

Number Exposed:

$$n_j^{'} = n_j - c_j/2$$

This can be considered as the number of cases that are at risk of the terminal event in the interval.

Number Terminating:

d_j: The number terminating is the number of cases that reach the terminal event within the interval.

Proportion Surviving:

$$\frac{n_j^{'} - d_j}{n_j^{'}}$$

The proportion surviving is the proportion of cases that do not reach the terminal event in this interval.

Cumulative Proportion Surviving:

$$S(t) = \prod_{j=1}^{k} \frac{n_j^{'} - d_j}{n_j^{'}}$$

The cumulative proportion surviving is the proportion of cases that have not reached the terminal event by the end of the interval.

Standard Error of Cumulative Surviving:

$$SE_t = S(t)\sqrt{\sum_{j=1}^{k}\frac{d_j}{n_j^{'}(n_j^{'} - d_j)}}$$

The standard error of cumulative proportion surviving is computed from Greenwood's formula (Collett, 1994).

Confidence Intervals of Cumulative Surviving:

$$S(t)^{\mathrm{Exp}(\pm Z_\alpha SE_t^{'})}$$

where the log-transformed standard error is:

$$SE_t^{'} = \frac{SE_t}{S(t)\mathrm{Ln}(S(t))}$$

SE_t (the standard error reported in the table) is not used in computing the confidence intervals, employing the standard Z distribution, because it often leads to values outside the valid range of 0 – 1. The significance level can be set to any value between 0 and 1 from the Variable Selection Dialogue.

Proportion Terminating:

$$\frac{d_j}{n_j^{'}}$$

The proportion of cases that reach the termination event in this interval, which is equal to one minus proportion surviving.

Cumulative Proportion Terminating:

1-S(t): This is the cumulative proportion of cases that have reached the terminal event by the end of the interval and it is equal to one minus cumulative proportion surviving.

Standard Error of Cumulative Terminating:

SE_t: This is identical to the standard error of cumulative proportion surviving.

Confidence Intervals of Cumulative Terminating:

$$1 - S(t)^{\mathrm{Exp}(\pm Z_\alpha SE_t^{'})}$$

This is equal to one minus confidence intervals of surviving.

Hazard Rate:

$$h(t) = \frac{d_j}{(n_j^{'} - d_j / 2)\tau_j}$$

where τ_j is the length of interval j.

The hazard rate is an estimate of the probability per unit time that cases entering the interval will experience the terminal event in the interval.

Standard Error of Hazard Rate:

$$SEh_t = \frac{h(t)\sqrt{1 - [h(t)\tau_j / 2]^2}}{\sqrt{d_j}}$$

The asymptotic standard error of the hazard rate is displayed.

Confidence Intervals of Hazard Rate:

$$h(t) \pm Z_{\alpha/2} SEh_t$$

Confidence intervals are computed from the Z distribution.

Probability Density:

$$p(t) = \frac{S(t-1) - S(t)}{\tau_j}$$

The probability density is an estimate of the probability per unit time of the terminal event occurring in the interval.

Standard Error of Probability Density:

$$SEp_t = \frac{S(t-1)}{\tau_j} \frac{d_j}{n_j^{'}} \sqrt{\sum_{j=1}^{t-1} \frac{d_j}{n_j^{'}(n_j^{'} - d_j)}}$$

The standard error of the probability density is displayed.

Confidence Intervals of Probability Density:

$$p(t) \pm Z_{\alpha/2} SEp_t$$

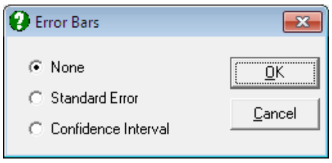
Confidence intervals are computed from the Z distribution.

9.4.1.2.2. Life Table Plots

Four Life Table plots can be displayed. The Edit → Data Series dialogue provides you with necessary controls to edit all aspects of the plot. If a factor column was selected, each subgroup's settings are controlled from a different tab on the same dialogue. There are no limitations on the maximum number of subgroups that can be plotted on one graph, but only the properties of the first nine subgroups can be controlled from the Edit → Data Series dialogue.

For the plot of survival and hazard functions, the line type is set to Step Right by default (see 4.1.1.1.1. Line), following Armitage and Berry (2002) and Altman (1991). But this can be changed to Step Down following Collett (1994), or any other type from the Edit → Data Series → Line dialogue.

It is possible to display standard errors or confidence intervals for each subgroup separately. To do this, first display the graph and then select Edit → Data Series. Clicking on the [Bars…] button, a small dialogue will pop up.



Error Bars

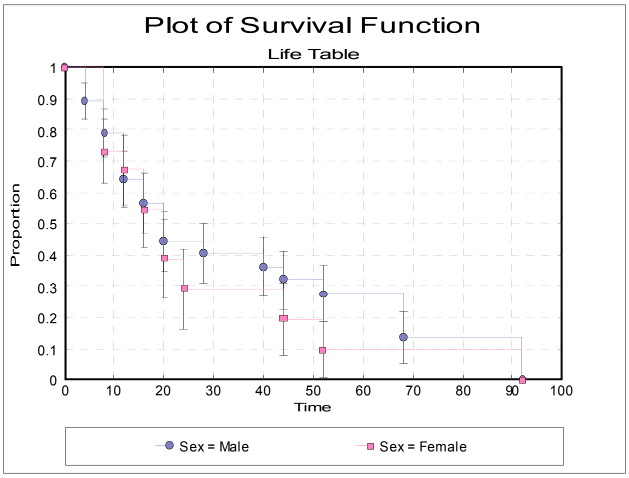
● None

○ Standard Error

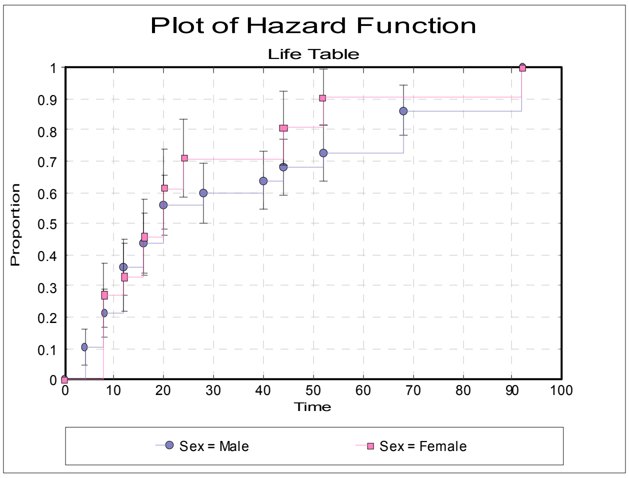
○ Confidence Interval

If the second option Standard Error is selected, then symmetric error bars are drawn for each point of the particular series selected. If the third option Confidence Interval is selected, then asymmetric error bars are drawn for confidence intervals displayed in Life Table.

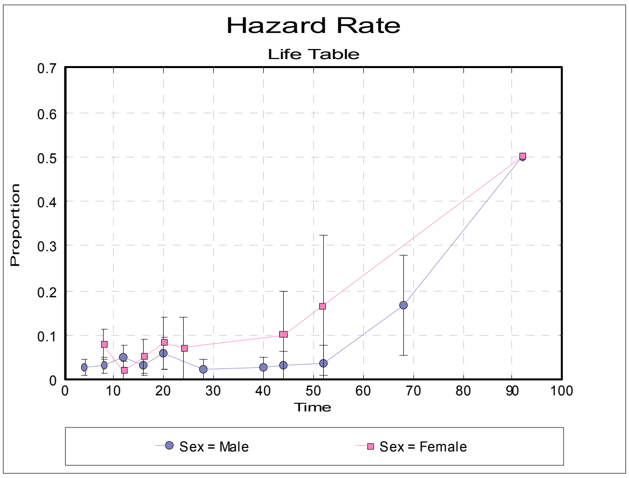
Plot of Survival Function: A plot of the cumulative proportion surviving is displayed.



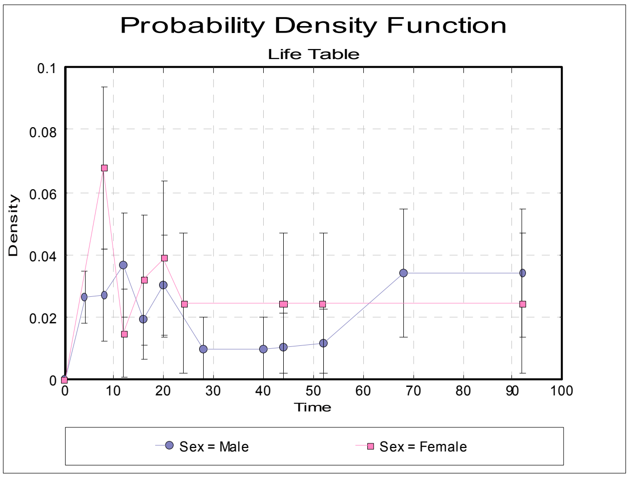
Plot of Hazard Function: A plot of the cumulative proportion terminating is displayed.



Plot of Hazard Rate: A plot of the hazard rate is displayed for each factor level. Select Error Bars in the Edit → Data Series dialogue to display standard errors or confidence intervals of the probability density function as error bars.



Plot of Probability Density: A plot of the probability density is displayed for each factor level. Select Error Bars in the Edit → Data Series dialogue to display standard errors or confidence intervals of the probability density function as error bars.



9.4.1.3. Life Table Example

Data on survival times of patients in a study on multiple myeloma is given in Table 1.3, p. 9, in Collett, D. (1994). Examples 2.2 (p. 17) and 2.5 (p. 28) give the

Life Table estimates and the plot of the survival and hazard functions for this data.

Open SURVIVAL and select Statistics 2 → Survival Analysis → Life Table. From the Variable Selection Dialogue select the data option 1 Enter Durations and *Survival time* (*C10*) as [Time] and *Status* (*C11*) as [Censored]. Enter the following two lines for the interval selection dialogue.

Interval Size	No of Intervals
12	5
36	1
0	0
0	0
0	0

Instead of entering 8 regular intervals of size 12, here we enter 5 intervals of size 12 and one interval of size 36. This is because the results given by Collett in Tables 2.1 and 2.4 are obtained by aggregating the last 3 intervals into one. Selecting Life Table and checking all output options in the next two dialogues, the following results are obtained:

Life Table

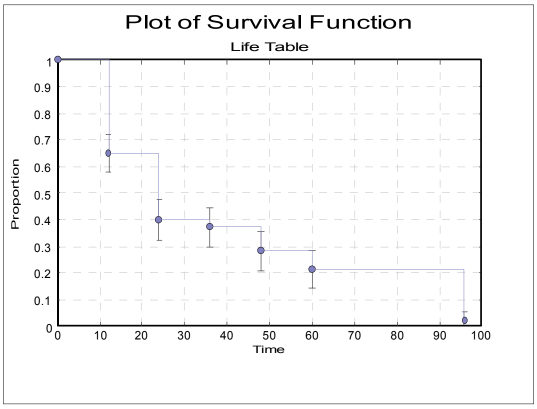
Time Variable: Survival time
Censor Variable: Status
Number of Cases Censored: 12 (25.0%)
Valid Number of Cases: 48, 0 Omitted
Median Survival Time = 19.2800

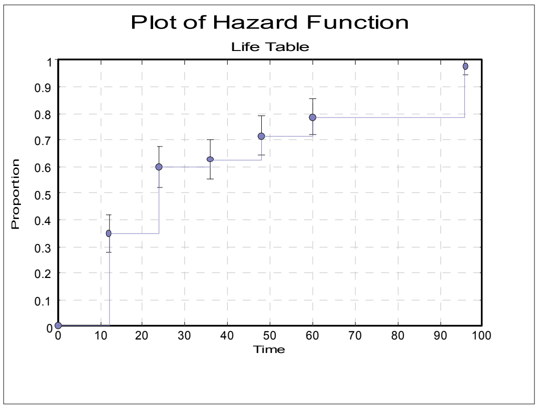
Interval Start	Number Entering	Number Censored	Number Exposed	Number Terminating	Proportion Surviving	Cumulative Proportion Surviving
0	48	4	46.0	16	0.6522	0.6522
12	28	4	26.0	10	0.6154	0.4013
24	14	0	14.0	1	0.9286	0.3727
36	13	1	12.5	3	0.7600	0.2832
48	9	2	8.0	2	0.7500	0.2124
60	5	1	4.5	4	0.1111	0.0236

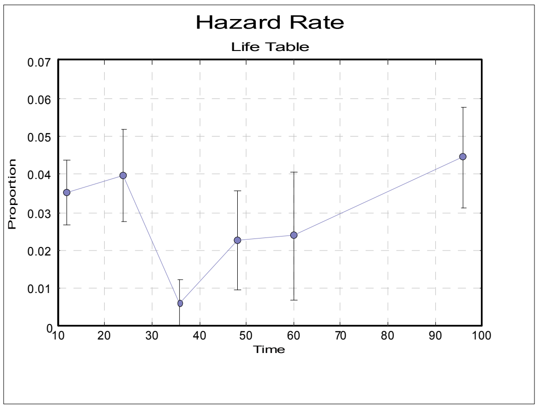
Interval Start	Std Error Cumulative Surviving	Lower 95% Cumulative Surviving	Upper 95% Cumulative Surviving	Proportion Terminating	Cumulative Proportion Terminating	Std Error Cumulative Terminating
0	0.0702	0.4964	0.7704	0.3478	0.3478	0.0702
12	0.0758	0.2543	0.5440	0.3846	0.5987	0.0758
24	0.0756	0.2284	0.5169	0.0714	0.6273	0.0756
36	0.0730	0.1522	0.4294	0.2400	0.7168	0.0730
48	0.0698	0.0955	0.3598	0.2500	0.7876	0.0698
60	0.0324	0.0005	0.1610	0.8889	0.9764	0.0324

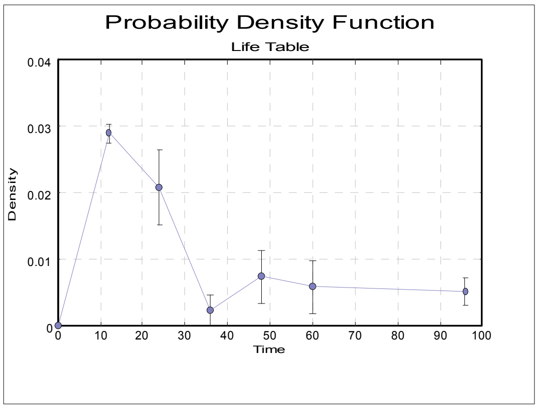
Interval Start	Lower 95% of Cumulative Terminating	Upper 95% of Cumulative Terminating	Hazard Rate	Std Error of Hazard Rate	Lower 95% of Hazard Rate	Upper 95% of Hazard Rate
0	0.2296	0.5036	0.0351	0.0086	0.0183	0.0519
12	0.4560	0.7457	0.0397	0.0122	0.0158	0.0636
24	0.4831	0.7716	0.0062	0.0062	0.0000	0.0183
36	0.5706	0.8478	0.0227	0.0130	0.0000	0.0482
48	0.6402	0.9045	0.0238	0.0167	0.0000	0.0565
60	0.8390	0.9995	0.0444	0.0133	0.0183	0.0706

Interval Start	Probability Density	Sta Error Probability Density	Lower 95% of Probability Density	Upper 95% of Probability Density
0	0.0290	0.0015	0.0261	0.0319
12	0.0209	0.0057	0.0098	0.0320
24	0.0024	0.0023	0.0000	0.0070
36	0.0075	0.0040	0.0000	0.0154
48	0.0059	0.0039	0.0000	0.0136
60	0.0052	0.0019	0.0015	0.0090




Plot of Hazard Function
Life Table
Proportion
Time
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
10
20
30
40
50
60
70
80
90
100


Hazard Rate
Life Table
Proportion
Time
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
10
20
30
40
50
60
70
80
90
100


Probability Density Function
Life Table
Density
Time
0
0.01
0.02
0.03
0.04
0
10
20
30
40
50
60
70
80
90
100

9.4.2. Kaplan-Meier Analysis

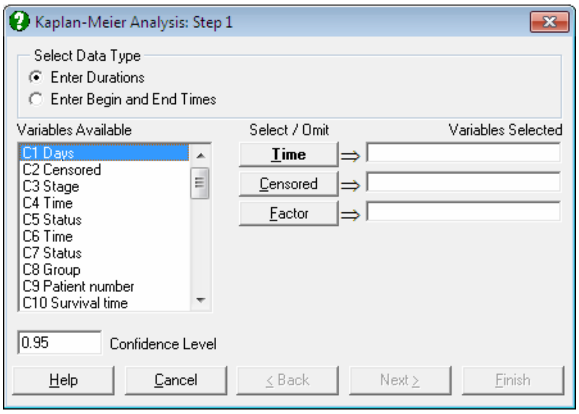
The main difference between Life Table and Kaplan-Meier Analysis is that while cases are aggregated into time intervals in the former, the latter estimates the survival function on individual cases without any aggregation. The Kaplan-Meier estimate of the survival function is given by:

$$S(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$$

where:

- d_j = number of deaths at interval j, and
- n_j = number of cases entering interval j.

This is similar to the Life Table estimate of the survival function except that the number of cases entering interval j here replaces the average at risk in Life Table.



Kaplan-Meier Analysis: Step 1

Select Data Type
● Enter Durations
○ Enter Begin and End Times

Variables Available: C1 Days, C2 Censored, C3 Stage, C4 Time, C5 Status, C6 Time, C7 Status, C8 Group, C9 Patient number, C10 Survival time

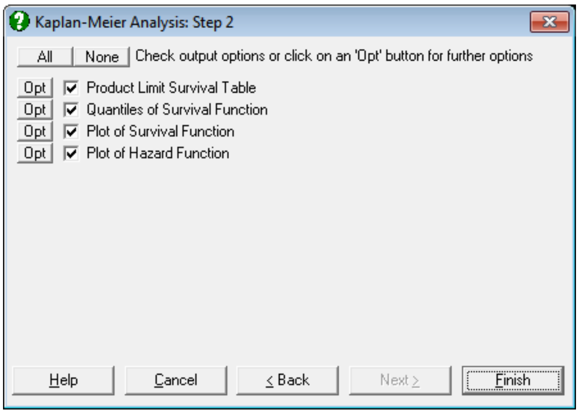
Select / Omit: Time ⇒, Censored ⇒, Factor ⇒

Variables Selected

0.95 Confidence Level

Variables are selected as described at the beginning of this chapter (see 9.4.0. Survival Variable Selection). If a factor variable has been selected, then a further dialogue will allow levels of the factor to be selected for analysis.

The Output Options Dialogue will provide access to the following four options:



Kaplan-Meier Analysis: Step 2

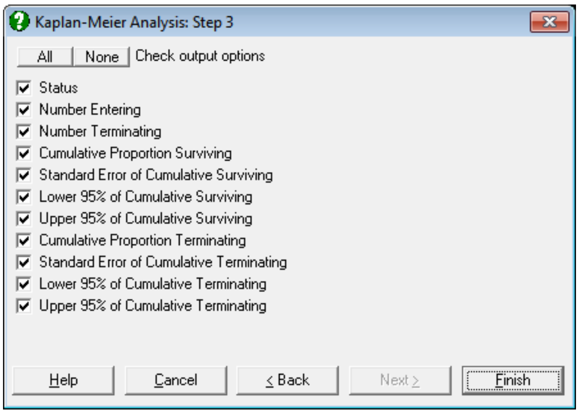
Check output options or click on an 'Opt' button for further options

- ☑ Product Limit Survival Table
- ☑ Quantiles of Survival Function
- ☑ Plot of Survival Function
- ☑ Plot of Hazard Function

9.4.2.1. Product Limit Survival Table

The Kaplan-Meier estimate of the survival function is also called the product limit estimator. The Kaplan-Meier Analysis has the advantage over Life Table analysis in that its results do not depend on grouping of the data into intervals. The product limit method is like a Life Table with a single observation in each interval.

The survival and hazard functions are estimated and they are displayed together with their standard errors and confidence intervals for a user-defined confidence level.



Kaplan-Meier Analysis: Step 3

Check output options

- ☑ Status
- ☑ Number Entering
- ☑ Number Terminating
- ☑ Cumulative Proportion Surviving
- ☑ Standard Error of Cumulative Surviving
- ☑ Lower 95% of Cumulative Surviving
- ☑ Upper 95% of Cumulative Surviving
- ☑ Cumulative Proportion Terminating
- ☑ Standard Error of Cumulative Terminating
- ☑ Lower 95% of Cumulative Terminating
- ☑ Upper 95% of Cumulative Terminating

Status: This indicates whether a case is censored. By default, 0 is censored (the termination time is not known) and non-zero values are uncensored (terminating at this time period).

Number Entering:

n_j: The number of cases that enter the interval.

Number Terminating:

d_j: The number terminating is the number of cases that reach the terminal event within the interval.

Cumulative Proportion Surviving:

$$S(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$$

The cumulative proportion surviving is the proportion of cases that have not reached the terminal event by the end of the interval.

Standard Error of Cumulative Surviving:

$$SE_t = S(t)\sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}}$$

The standard error of cumulative proportion surviving is computed from Greenwood's formula.

Confidence Intervals of Cumulative Surviving:

$$S(t)^{Exp(\pm Z_\alpha SE_t^{'})}$$

where the log-transformed standard error is:

$$SE_t^{'} = \frac{SE_t}{S(t)Log(S(t))}$$

SE_t (the standard error reported in the table) is not used in computing the confidence intervals, employing the standard Z distribution, because it often leads to values outside the valid range of 0 to 1. The significance level can be set to any value between 0 and 1 from the Variable Selection Dialogue.

Cumulative Proportion Terminating:

1-S(t): This is the cumulative proportion of cases that have reached the terminal event by the end of the interval and it is equal to one minus cumulative proportion surviving.

Standard Error of Cumulative Terminating:

SE_t: This is identical to the standard error of cumulative proportion surviving.

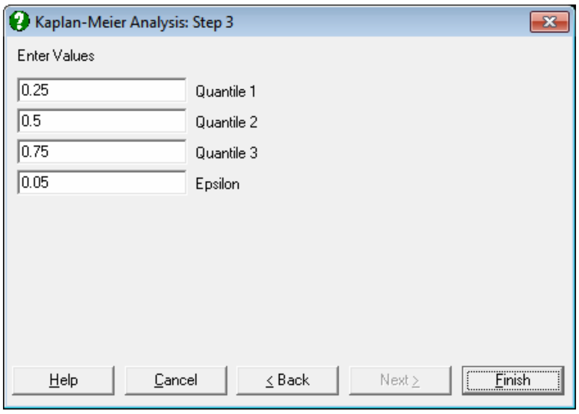
Confidence Intervals of Cumulative Terminating:

$$1 - S(t)^{\mathrm{Exp}(\pm Z_\alpha SE'_t)}$$

This is equal to one minus confidence intervals of surviving.

9.4.2.2. Quantiles of Survival Function

With this procedure it is possible to estimate the mean and up to three quantiles of the survival function. The quantiles are set to quartiles by default, but you can edit these to any values between 0 and 1. The value of epsilon (0.05 by default), which is used in estimating the standard error of quantiles, can also be changed. The mean and quantiles, as well as their standard errors and confidence intervals, are displayed in a table.



0.25 Quantile 1
0.5 Quantile 2
0.75 Quantile 3
0.05 Epsilon

The mean survival time is computed as:

$$\mu = \sum_{j=0}^{k-1} S(t_j)(t_{j+1} - t_j)$$

and its standard error is:

$$SE = \sqrt{\frac{d}{d-1}\sum_{j=1}^{k-1}\frac{a_j^2 d_j}{n_j(n_j - d_j)}}$$

where:

$$a_j = \sum_{i=j}^{k-1} S(t_i)(t_{i+1} - t_{ij})$$

and d is the total number of cases terminating.

The quantile 100p (where p = 0.5 is the median) of the survival function is given as the minimum observed survival time for which the value of the survival function is less than or equal to p. That is:

$$t(p) = \min\{t(j)\,|\,S(t(j)) \le p\}$$

The standard error of a quantile is calculated from:

$$SE_t^{'} = \frac{SE_{t(p)}}{f(t(p))}$$

where

$SE_{t(p)}$ = the standard error of survival function at t(p),

$$f(t(p)) = \frac{S\{u(p)\} - S\{l(p)\}}{l(p) - u(p)}$$

and:

$$u(p) = \max\{t(j)\,|\,S(t(j)) \ge 1 - p + \varepsilon\}$$

$$l(p) = \min\{t(j)\,|\,S(t(j)) \le 1 - p - \varepsilon\}$$

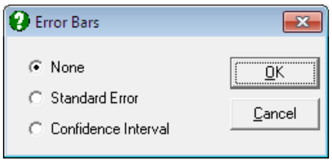
Although the default value for epsilon is 0.05, you can enter any value between 0 and 1.

9.4.2.3. Kaplan-Meier Plots

Survival and hazard functions can be plotted. The Edit → Data Series dialogue provides the necessary controls to edit all aspects of the plot. If a factor column is selected, each subgroup's settings are controlled from a different tab on the same dialogue. There are no limitations on the maximum number of subgroups that can be plotted on one graph, but only the properties of the first nine subgroups can be controlled from the Edit → Data Series dialogue.

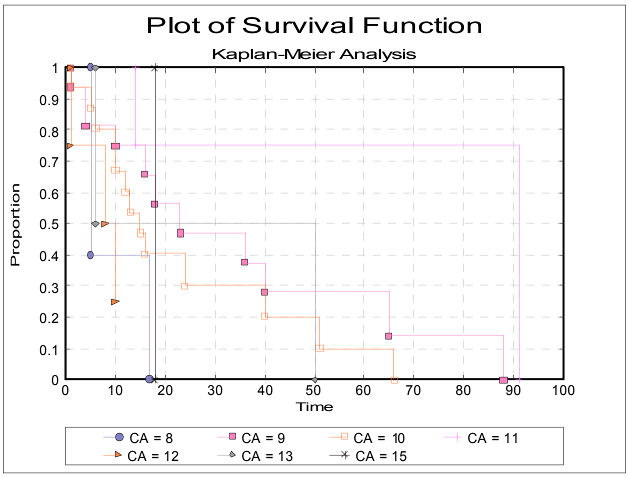
The line type is set to Step Right by default (see 4.1.1.1.1. Line), following Armitage and Berry (2002) and Altman (1991). But this can be changed to Step Down following Collett (1994) from the Edit → Data Series → Line dialogue.

It is possible to display standard errors or confidence intervals for each subgroup separately. To do this, first display the graph and then select Edit → Data Series. Clicking on the [Bars…] button, a small dialogue will pop up.

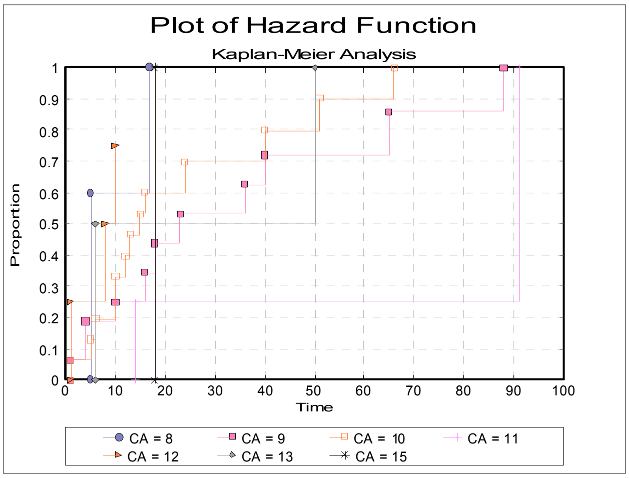

Error Bars

● None
○ Standard Error
○ Confidence Interval

Plot of Survival Function: The cumulative proportion of surviving is plotted against the survival times.



Plot of Hazard Function: The cumulative proportion of terminating is plotted against the survival times.



9.4.2.4. Kaplan-Meier Examples

Example 1

Example 17.1 on p. 578 from Armitage & Berry (2002). Data on survival of patients with diffuse hystiocytic lymphoma by the stage of tumour are given.

Open SURVIVAL and select Statistics 2 → Survival Analysis → Kaplan-Meier Analysis. From the Variable Selection Dialogue click on the data option 1 Enter Durations and select *Days* (*C1*) as [Time], *Censored* (*C2*) as [Censored] and *Stage* (*C3*) as [Factor]. Select Plot of Survival Function as the output option.

Next click on the [Last Procedure Dialogue] button and this time select Product Limit Survival Table. From the next dialogue check only the Stage = 3 box and then select the first 5 boxes from the Output Options Dialogue to obtain the following output:

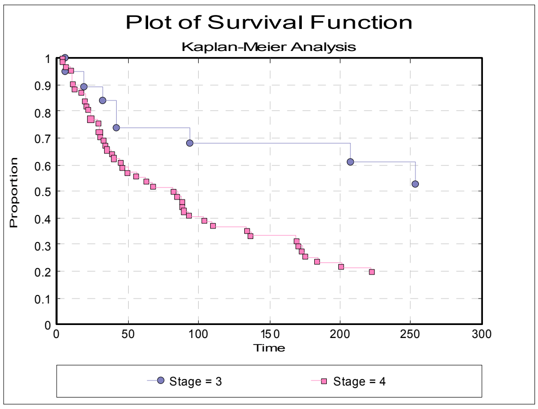
Kaplan-Meier Analysis

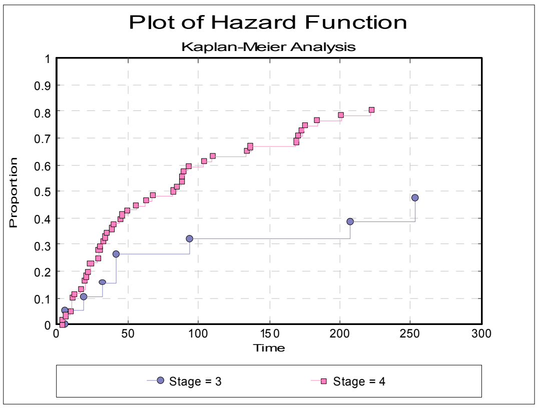
```
Factor variable: Stage = 3
Time Variable: Days
Censor Variable: Censored
Number of Cases Censored: 11 ( 57.9%)
Valid Number of Cases: 19, 61 Omitted
```

Product Limit Survival Table

Time	Status	Number Entering	Number Terminating	Cumulative Proportion Surviving	Standard Error of Cumulative Surviving
6	1	18	1	0.9474	0.0512
19	1	17	2	0.8947	0.0704
32	1	16	3	0.8421	0.0837
42	1	15	4	*	*
42	1	14	5	0.7368	0.1010
43	0	13	5	*	*
94	1	12	6	0.6802	0.1080
126	0	11	6	*	*
169	0	10	6	*	*
207	1	9	7	0.6121	0.1167
211	0	8	7	*	*
227	0	7	7	*	*
253	1	6	8	0.5247	0.1287
255	0	5	8	*	*
270	0	4	8	*	*
310	0	3	8	*	*
316	0	2	8	*	*
335	0	1	8	*	*
346	0	0	8	*	*

The following graphs are obtained by including stages 3 and 4 in the analysis.





Example 2

Time data in weeks to discontinuation of the use of an IUD is given in Table 1.1 (p. 5), in Collett, D. (1994).

1) Example 2.3 Table 2.2 (p.21) gives the cumulative survival function.
2) Example 2.4 Table 2.3 (p.26) gives the cumulative survival function, its standard error and confidence intervals. The 95% confidence intervals reported by Collett are computed by using the standard formula $S(t) \pm Z_{\alpha/2} SE_t$, whereas UNISTAT reports the log-transformed confidence intervals.
3) Example 2.9 (p.34) gives median and its 95% confidence intervals for cumulative survival function.

Open SURVIVAL and select Statistics 2 → Survival Analysis → Life Table. From the Variable Selection Dialogue select the data option 1 Enter Durations and *Survival time* (*C4*) as [Time] and *Status* (*C5*) as [Censored].

Kaplan-Meier Analysis

Time Variable: time
Censor Variable: status
Number of Cases Censored: 9 (50.0%)
Valid Number of Cases: 18, 0 Omitted

Product Limit Survival Table

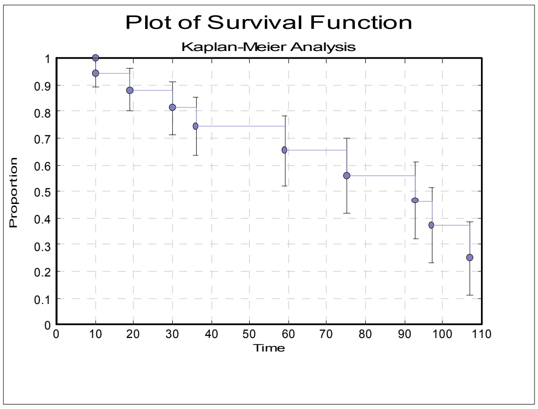
Time	Status	Number Entering	Number Terminating	Cumulative Proportion Surviving	Standard Error of Cumulative Surviving	Lower 95% of Cumulative Surviving
10	1	17	1	0.9444	0.0540	0.6664
13	0	16	1	*	*	*
18	0	15	1	*	*	*
19	1	14	2	0.8815	0.0790	0.6019
23	0	13	2	*	*	*
30	1	12	3	0.8137	0.0978	0.5241
36	1	11	4	0.7459	0.1107	0.4536
38	0	10	4	*	*	*
54	0	9	4	*	*	*
56	0	8	4	*	*	*
59	1	7	5	0.6526	0.1303	0.3438
75	1	6	6	0.5594	0.1412	0.2564
93	1	5	7	0.4662	0.1452	0.1830
97	1	4	8	0.3729	0.1430	0.1209
104	0	3	8	*	*	*
107	1	2	9	0.2486	0.1392	0.0468
107	0	1	9	*	*	*
107	0	0	9	*	*	*

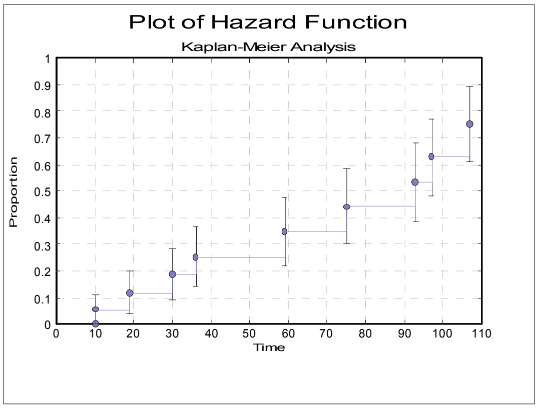
Time	Upper 95% of Cumulative Surviving	Cumulative Proportion Terminating	Standard Error of Cumulative Terminating	Lower 95% of Cumulative Terminating	Upper 95% of Cumulative Terminating
10	0.9920	0.0556	0.0540	0.0080	0.3336
13	*	*	*	*	*
18	*	*	*	*	*
19	0.9691	0.1185	0.0790	0.0309	0.3981
23	*	*	*	*	*
30	0.9363	0.1863	0.0978	0.0637	0.4759
36	0.8970	0.2541	0.1107	0.1030	0.5464
38	*	*	*	*	*
54	*	*	*	*	*
56	*	*	*	*	*
59	0.8432	0.3474	0.1303	0.1568	0.6562
75	0.7804	0.4406	0.1412	0.2196	0.7436
93	0.7097	0.5338	0.1452	0.2903	0.8170
97	0.6310	0.6271	0.1430	0.3690	0.8791
104	*	*	*	*	*
107	0.5313	0.7514	0.1392	0.4687	0.9532
107	*	*	*	*	*
107	*	*	*	*	*

Quantiles of Survival Function

Time Variable: time
Censor Variable: status
Number of Cases Censored: 9 (50.0%)
Valid Number of Cases: 18, 0 Omitted
Epsilon: 0.05

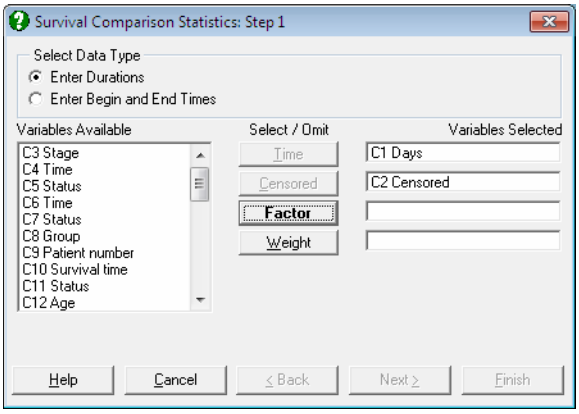
	Value	Standard Error	Lower 95%	Upper 95%
Mean	76.3387	9.4331	57.8502	94.8272
Quantile 1: 25%	107.0000	*	*	*
Quantile 2: 50%	93.0000	17.1311	59.4237	126.5763
Quantile 3: 75%	36.0000	19.9294	*	75.0610





9.4.3. Survival Comparison Statistics

Survival comparison tests are nonparametric rank tests and they are used to test the equality of survival functions for different groups defined by a factor column. Therefore, unlike other survival procedures, the choice of a factor column is compulsory here. It is also possible to select a weights variable for a weighted analysis.



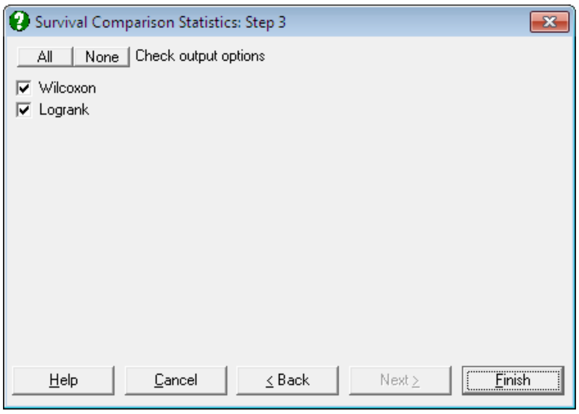
Survival Comparison Statistics: Step 1

Select Data Type
● Enter Durations
○ Enter Begin and End Times

Variables Available
C3 Stage
C4 Time
C5 Status
C6 Time
C7 Status
C8 Group
C9 Patient number
C10 Survival time
C11 Status
C12 Age

Variables Selected
Time: C1 Days
Censored: C2 Censored
Factor:
Weight:

Once the data has been selected (see 9.4.0. Survival Variable Selection) a dialogue will appear facilitating selection of factor levels to be compared. Any number of levels can be selected. UNISTAT will then display an Output Options Dialogue featuring Wilcoxon and Logrank options.



Survival Comparison Statistics: Step 3

Check output options
☑ Wilcoxon
☑ Logrank

9.4.3.1. Wilcoxon Tests: Gehan (Lee Desu), Breslow

This test is appropriate when the hazard functions are not necessarily proportional. The null hypothesis that "the groups do not differ" is tested.

First, all cases from all groups, censored or uncensored, are sorted according to survival times in ascending order. For each unique time period which is not censored, the number of deaths and the number of survivors (including the current time period), censored or uncensored, are computed. For tied time periods containing at least one censored subject, the same entities are also computed. For each time period, the expected value of deaths and its variance are computed and their sum is stored over all time periods, which are then used to compute the test statistic, which is approximately chi-square distributed.

Let us first consider a 2-group case to illustrate the Breslow algorithm. This is more robust than the older Gehan (Lee-Desu) method. Let:

- d_{jA} = deaths in group A at time j, t_j,
- d_j = deaths in all groups at time j,
- n_{jA} = subjects alive in group A just before t_j,
- n_j = subjects alive in all groups just before t_j,
- N = total number of cases

For each time period j we can construct the following table:

	Died	Survived	Total
Group A	d_{jA}	n_{jA} - d_{jA}	n_{jA}
Group B	d_{jB}	n_{jB} - d_{jB}	n_{jB}
Total	d_j	n_j - d_j	n_j

The difference between observed and expected value of deaths, weighted by the total number of individuals at risk, is computed for each time period:

$$U_{jA} = n_j(d_{jA} - n_{jA}d_j/n_j)$$

Expected value of the variance is:

$$\mathrm{Var}(d_{jA}) = \frac{d_j(n_j - d_j)n_{jA}n_{jB}}{(n_j - 1)}$$

Summing over all time periods we obtain scores for each group:

$$U_A = \sum U_{jA}$$

and the overall variance:

$$V_A = \sum \mathrm{Var}(d_{jA})$$

Then the test statistic is computed as:

$$\chi^2 = \frac{U_A{}^2}{V_A}$$

which has r - 1 degrees of freedom, where r is the number of groups (r = 2 and degrees of freedom = 1 in this case). UNISTAT uses an r-group generalisation of this algorithm.

The Gehan (Lee-Desu) statistic also requires computing the individual scores at each time period as follows:

For a censored case at time j:

$$U_j = Unc_j$$

and for an uncensored case:

$$U_j = 2 * Unc_j - UncEq_j + Cen_j - CenEq_j - N$$

where $UncEq_j$ and $CenEq_j$ are the number of uncensored and censored cases at each time period and Unc_j and Cen_j are the number of uncensored and censored cases at all current and previous time periods.

The test statistic is given as:

$$\chi^2 = \frac{(N-1)\sum SS_i^2 / n_i}{\sum U_j^2}$$

where SS_i is the sum of scores for group i.

9.4.3.2. Logrank Test: Mantel-Haenszel (Peto)

This test is appropriate when the hazard functions are proportional. The null hypothesis that “the groups do not differ” is tested. It is analogous to Mantel-Haenszel test for contingency tables.

As in Wilcoxon test (see 9.4.3.1. Wilcoxon Tests: Gehan (Lee Desu), Breslow), a 2 x 2 table is constructed for each time period. Then the expected value of deaths and its variance are computed as:

$$E(d_{jA}) = n_{jA} d_j / n_j$$

$$Var(d_{jA}) = \frac{n_{jA} n_{jB} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

and then they are summed over all periods:

$$O_A = \sum d_{jA}$$

$$E_A = \sum E(d_{jA})$$

$$V_A = \sum Var(d_{jA})$$

The test statistic is computed as:

$$\chi^2 = \frac{(O_A - E_A)^2}{V_A}$$

which has r - 1 degrees of freedom, where r is the number of groups (r = 2 and degrees of freedom = 1 in this case). UNISTAT uses an r-group generalisation of this algorithm.

9.4.3.3. Survival Comparison Tests Examples

Example 1

Example 17.1 on p. 578 from Armitage & Berry (2002). Data on survival of patients with diffuse hystiocytic lymphoma according to stage of tumour are given.

Open SURVIVAL and select Statistics 2 → Survival Analysis → Survival Comparison Statistics. From the Variable Selection Dialogue select the data option 1 Enter Durations and *Days* (*C1*) as [Time], *Censored* (*C2*) as [Censored] and *Stage* (*C3*) as [Factor]. Include both factor levels 3 and 4 and check both output options to obtain the following results:

Survival Comparison Statistics

Wilcoxon

Stage	Total	Died	Censored	% Censored	Score	Mean Score
3	19	8	11	57.89%	-396	-20.8421
4	61	46	15	24.59%	396	6.4918
Total	80	54	26	32.50%	0	

Gehan (Lee-Desu):	
Chi-Square Statistic =	5.5428
Degrees of Freedom =	1
Right-Tail Probability =	0.0186
Breslow:	
Chi-Square Statistic =	5.0998
Degrees of Freedom =	1
Right-Tail Probability =	0.0239

Logrank

Stage	Total	Observed	Expected	(O-E)^2/E	(O-E)^2/V
3	19	8	16.6870	4.5223	6.7097
4	61	46	37.3130	2.0225	6.7097
Total	80	54	54.0000	6.5448	

Mantel-Haenszel (Peto):	
Chi-Square Statistic =	6.7097
Degrees of Freedom =	1
Right-Tail Probability =	0.0096

Example 2

Data on survival times of women with tumours which were negatively or positively stained with PHA is given in Table 1.2 (p. 7), in Collett, D. (1994). Examples 2.11 (p. 40) and 2.12 (p. 44) give the results of Logrank and Wilcoxon tests respectively.

Open SURVIVAL and select Statistics 2 → Survival Analysis → Survival Comparison Statistics. From the Variable Selection Dialogue select the data option 1 Enter Durations and *time* (*C6*) as [Time], *status* (*C7*) as [Censored] and *group* (*C8*) as [Factor]. Include both factor levels 1 and 2 and check both output options to obtain the following results:

Survival Comparison Statistics

Wilcoxon

Group	Total	Died	Censored	% Censored	Score	Mean Score
0	32	21	11	34.38%	159	4.9688
1	13	5	8	61.54%	-159	-12.2308
Total	45	26	19	42.22%	0	

Gehan (Lee-Desu):	
Chi-Square Statistic =	4.5420
Degrees of Freedom =	1
Right-Tail Probability =	0.0331
Breslow:	
Chi-Square Statistic =	4.1800
Degrees of Freedom =	1
Right-Tail Probability =	0.0409

Logrank

Group	Total	Observed	Expected	(O-E)^2/E	(O-E)^2/V
0	32	21	16.4349	1.2681	3.5150
1	13	5	9.5651	2.1788	3.5150
Total	45	26	26.0000	3.4468	

Mantel-Haenszel (Peto):	
Chi-Square Statistic =	3.5150
Degrees of Freedom =	1
Right-Tail Probability =	0.0608

9.4.4. Cox Regression

Apart from time and status variables, data for Survival Analysis often contain measurements on one or more continuous variables, such as temperature, dosage, age or one or more categorical variables such as gender, region, treatment. In such cases it is desirable to construct Life Tables (or survival functions) which reflect the effects of these continuous or categorical variables (which are also called covariates).

A method which combines the elements of nonparametric Life Table analysis and the parametric Regression Analysis was introduced by D R Cox in 1972. It is known as the Cox Regression or Cox's proportional hazards model. The latter reflects a fundamental assumption of this model, namely that the hazard function of an individual in one group is proportional to the hazard function of another in another group at any time period. In graphical terms, this is equivalent to assuming that the hazard curves of different groups do not cross each other.

Let X be a matrix of p independent variables (covariates) for n cases. Then the hazard function for the jth case can be written as:

$$h_j(t) = \mathrm{Exp}(\beta' X_j) h_0(t)$$

where:

- $h_0(t)$ is the baseline hazard function and
- β' is the vector of coefficients.

Rearranging the terms we obtain:

$$\mathrm{Log}\left(\frac{h_j(t)}{h_0(t)}\right) = \beta' X_j$$

Although the right-hand side of this equation is linear, the existence of another unknown on the left-hand side, $h_0(t)$, makes it impossible to solve it using the Linear Regression method. Instead, a maximum likelihood solution for a log-likelihood function that is derived from the above equation is found:

$$\mathrm{Log}\, L(\beta) = \sum_{j=1}^{n} d_j \left\{ \beta' X_j - \mathrm{Log}\left(\sum_{l \in R(t_j)} \mathrm{Exp}(\beta' X_l) \right) \right\}$$

where $R(t_j)$ is the set of cases at risk at the start of time j and d_j is the number of cases that terminate at time j.

The maximum likelihood solution of this proportional hazards model is computed employing a modified Newton-Raphson minimisation algorithm. The nature of this method implies that a solution (convergence) cannot always be achieved. In such cases, you are advised to edit the convergence parameters provided, in order to find the right levels for the particular problem at hand. Three convergence parameters can be edited in a dialogue that pops up just after the Variable Selection Dialogue (see 9.4.4.2. Cox Regression Intermediate Inputs).

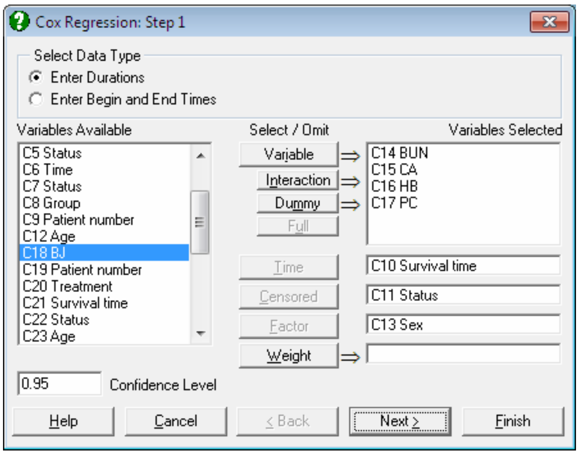
Stratified Analysis: It may be the case that with some data sets the assumption of proportionality of individual hazard functions cannot be maintained. Under such circumstances it may be possible to define some subgroups within which the hazard functions can assumed to be proportional. This type of analysis is called a *stratified analysis* and the subgroups within which the proportionality assumption is maintained are called the *strata*.

The strata are defined by the levels of a factor column, which can be selected from the Variable Selection Dialogue. The program then assumes that different strata have different baseline hazard functions, but that they all have the same coefficient estimates for covariates. A maximum of six strata can be fitted simultaneously.

Baseline Functions at Means of Covariates: In order to maintain numerical stability, UNISTAT always centres the covariates by subtracting their respective means first. Most results, including the coefficient estimates, are invariant to centring. However, the baseline functions are different with and without centring and some sources prefer reporting the centred baseline functions while the others the uncentred ones (baseline functions relative to the origin). UNISTAT makes both sets of results available. See 9.4.4.2. Cox Regression Intermediate Inputs below for how to make this selection. The only other result that is affected by centring is the Linear Fit (XBeta) reported under Case (Diagnostic) Statistics (see **Error! Reference source not found.**).

Predictions: If a case (row) of data does not contain any missing values except for the time variable, then a prediction of survival times is generated for this case. If at least one such case exists in data, then a fourth option Predicted Survival is added to the Baseline Functions dialogue. An additional column of predicted survival times is output for each predicted case.

9.4.4.1. Cox Regression Variable Selection



Select Data Type
● Enter Durations
○ Enter Begin and End Times

Variables Selected: C14 BUN, C15 CA, C16 HB, C17 PC

Time: C10 Survival time

Censored: C11 Status

Factor: C13 Sex

Weight:

0.95 Confidence Level

Variable: Any number of columns containing numeric data can be selected as independent variables (covariates). However, unlike any other regression procedures, a Cox regression can be run without selecting any independent variables. This is called the initial model (or the null model). UNISTAT reports the –2Log Likelihood value for the initial model for any model fitted.

Interaction: This button is used to create independent variables, which are the products of existing numeric variables. If only one variable is highlighted, then the new independent variable created will be the product of the selected column by itself. If two or more variables are highlighted, then the new term will be the product of these variables. Maximum three-way interactions are allowed. Interactions of dummy variables or lags are not allowed. In order to create interaction terms for dummy variables, create interactions first, and then create dummy variables for them. For further information see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables.

Dummy: This button is used to create n or n – 1 new independent (dummy or indicator) variables for a factor column containing n levels. Each dummy variable corresponds to a level of the factor column. A case in a dummy column will have the value of 1 if the factor contains the corresponding level in the same row, and 0 otherwise. If the selected variable is an interaction term, then dummy variables will be created for this interaction term. Up to three-way interactions are allowed and short or long string columns can be selected as factors. It is possible to include all n levels in the analysis or to omit the first or the last level to remove the inherent over-parameterisation of the model. For further information see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables.

Full: This button becomes activated when two or more categorical variables are highlighted. Like the [Dummy] button, it is also used to create dummy variables. The only difference is that this button will create all necessary dummy variables and their interactions to specify a complete model. For instance, if two categorical variables are highlighted, then this button will create two sets of dummy variables representing the main effects and a third set representing the interaction term. Maximum three-way interactions are supported.

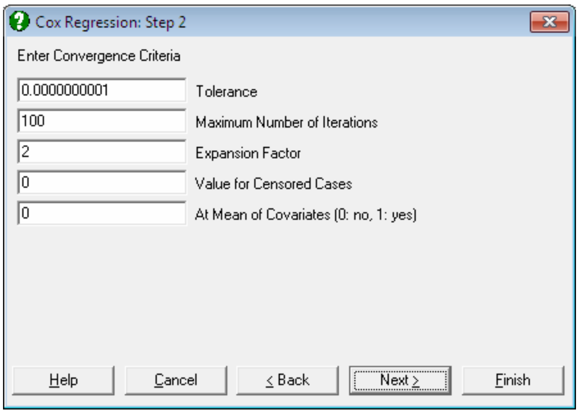
Time: It is compulsory to select at least one column containing numeric data as the time variable. Two types of data can be selected for Cox Regression; (1) Enter Durations where one column is selected containing the duration of each case and (2) Enter Begin and End Times where one column is selected as the starting time and another column as the termination time. In the latter case, the program subtracts the second column from the first one to obtain the duration of each case. If this results in some non positive values, then they are treated as missing cases (see 3.0.2.4. Time Data).

Censored: This variable is selected to show whether the termination time of a case is not known or the termination event has occurred. If the value in the column is zero then it is assumed that the case has been censored, which means that it has been excluded from the study before the termination event has occurred. Non-zero values (usually one) indicate that the termination event has occurred for this case. If a censor variable is not selected, then it is assumed that termination event has occurred for all cases. The default value of 0 indicating the censored cases can be edited to any other value from the next dialogue (see 9.4.4.2. Cox Regression Intermediate Inputs).

Weight: A column containing continuous numeric data can be selected as a weight variable. In this case, the program will normalise this column so that its sum is equal to the valid number of cases and then multiply each covariate by its square root before running the regression.

Factor: A categorical variable containing a limited number of distinct values can be selected as a factor. Unlike other Survival Analysis procedures though, this variable is not used in selecting subgroups to be included in the analysis. All cases are included in the analysis, but a stratified model is fitted. For more information on stratified analysis see the beginning of section 9.4.4. Cox Regression.

9.4.4.2. Cox Regression Intermediate Inputs



Cox Regression: Step 2

Enter Convergence Criteria

0.0000000001 Tolerance

100 Maximum Number of Iterations

2 Expansion Factor

0 Value for Censored Cases

0 At Mean of Covariates (0: no, 1: yes)

Tolerance: This value is used to control the sensitivity of nonlinear minimisation procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Expansion Factor: When the nonlinear minimisation algorithm moves away from a valid range of input parameters, this expansion factor is used to start a new search in a wider area. Under normal circumstances, you do not need to edit this value. If convergence cannot be achieved, then other values of similar (but larger) magnitude can be tried.

Value for Censored Cases: If a column of the spreadsheet has been selected as a censor variable, UNISTAT will consider cases with a zero entry in this column as censored. However, you can specify any value here to indicate which cases are to be treated as censored. Once a change is made here, it will apply to all Survival Analysis procedures throughout the session.

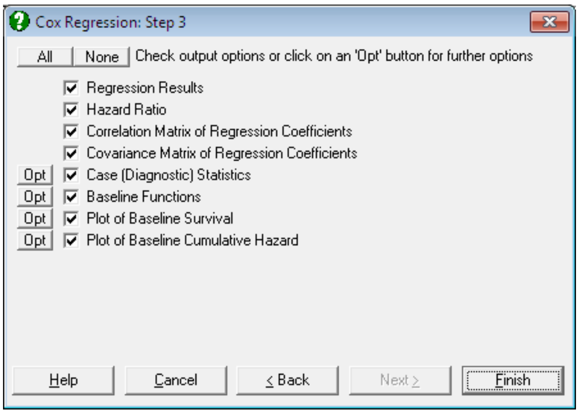
At Mean of Covariates: This box is used to select the type of baseline survival and hazard functions. A zero value results in baseline functions being reported relative to the origin and other values relative to the covariate means.

One other result that is affected by this choice is the linear fit (XBeta) vector reported under the Case (Diagnostic) Statistics option.

Omit Level: This box will appear only when one or more dummy variables have been included in the model from the Variable Selection Dialogue. Three options are available; (0) do not omit any levels, (1) omit the first level and (2) omit the last level. When no levels are omitted, the model will usually be over-parameterised (see 2.1.4. Creating Interaction, Dummy and Lag/Lead Variables).

9.4.4.3. Cox Regression Output Options

When the maximum likelihood model has converged or the maximum number of iterations has been reached without convergence, an Output Options Dialogue will provide access to the following options.



Check output options or click on an 'Opt' button for further options

☑ Regression Results
☑ Hazard Ratio
☑ Correlation Matrix of Regression Coefficients
☑ Covariance Matrix of Regression Coefficients
☑ Case (Diagnostic) Statistics
☑ Baseline Functions
☑ Plot of Baseline Survival
☑ Plot of Baseline Cumulative Hazard

When no covariates have been included in the model, only the -2 Log likelihood value will be reported in the Regression Results output. In this case, a Life Table will be displayed with a survival function that is identical to the Kaplan-Meier estimate of the same model (see 9.4.2. Kaplan-Meier Analysis).

9.4.4.3.1. Cox Regression Coefficient Output

Regression Results: The main regression output displays a table for coefficients of the estimated regression equation, their standard errors, Wald statistics, probability values and confidence intervals for the significance level specified in the Variable Selection Dialogue. If any independent variables have been omitted due to multicollinearity, they are reported at the end of the table. If you do not wish to display these variables enter the following line in the [Options] section of *Documents*\Unistat65\Unistat65.ini file:

```
    DispCollin=0
```

In Stand-Alone Mode, the estimated regression coefficients can be saved to the data matrix by clicking on the UNISTAT icon situated on the Output Medium Toolbar.

Wald Statistic: This is defined as:

$$W_i = \frac{\beta_i^2}{\sigma_i^2}, i = 1, \ldots, k.$$

and has a chi-square distribution with one degree of freedom.

Confidence Intervals: The confidence intervals for regression coefficients are computed from:

$$\beta_i \pm Z_{\alpha/2}\sigma_i, i = 1, \ldots, k.$$

where each coefficient's standard error, σ_i, is the square root of the diagonal element of covariance matrix.

-2 Log-Likelihood Initial Model: This is the value when all independent variables are excluded from the model:

$$\beta_i = 0, i = 1, \ldots, k.$$

-2 Log-Likelihood Final Model: This is –2 times the value of the log likelihood function when convergence is achieved.

Pseudo R-squared: In Cox Regression, an r-squared statistic as in the OLS regression is not available. This is because Cox Regression employs an iterative maximum likelihood estimation method. Equivalent statistics to test the goodness of fit have been proposed using the initial (L_0) and maximum (L_1) likelihood values.

McFadden:

$$R^2_{McF} = 1 - \left(\frac{L_1}{L_0}\right)$$

Adjusted McFadden:

$$R^2_{AdjMcF} = 1 - \left(\frac{L_1 - m}{L_0}\right)$$

Cox & Snell:

$$R_{CS}^{2} = 1 - \left(\frac{L_1}{L_0} \right)^{\left(\frac{2}{n} \right)}$$

Nagelkerke:

$$R_{N}^{2} = \frac{R_{CS}^{2}}{1 - L_0^{\left(\frac{2}{n} \right)}}$$

Likelihood Ratio: This is a test statistic for the null hypothesis that "all regression coefficients for covariates are zero". It is equal to –2 times the difference between the initial and final model likelihood values and has a chi-square distribution with k degrees of freedom (the number of independent variables in the model).

If no covariates have been included in the model, then only the -2 Log likelihood value will be reported.

Hazard Ratio:

Hazard ratio is defined as:

$$\mathrm{Exp}(\beta_i)$$

Values of hazard ratio above 1 indicate an increased hazard as the values of the corresponding covariate increase at any time period, and vice versa for values below 1.

The standard error of the hazard ratio is found as:

$$\sigma_i \mathrm{Exp}(\beta_i)$$

and its confidence intervals as:

$$\mathrm{Exp}(\beta_i \pm Z_{\alpha/2}\sigma_i)$$

which are simply the exponential of the coefficient confidence intervals.

This option is not available when no covariates have been included in the model.

Correlation Matrix of Regression Coefficients: This is a symmetric matrix with unity diagonal elements. The off-diagonal elements give the correlations between the regression coefficients.

This output is not available when no covariates have been included in the model.

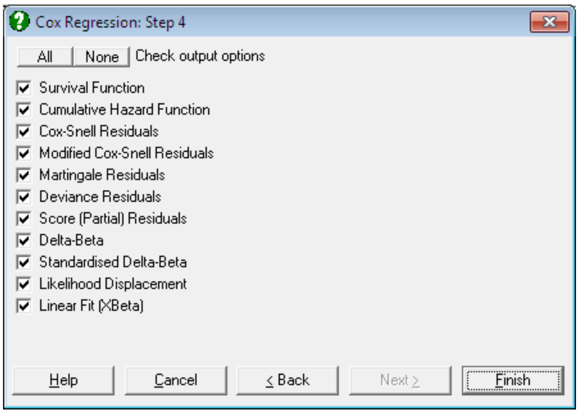
Covariance Matrix of Regression Coefficients: This is a symmetric matrix where diagonal elements are the square of parameter standard errors.

This output is not available when no covariates have been included in the model.

9.4.4.3.2. Cox Regression Case Output

Case (Diagnostic) Statistics: Case statistics are useful in determining the influence of individual observations on the overall fit of the model. For further information see 7.2.1.2.2. Linear Regression Case Output.

The values reported here are not sorted according to strata and the time variable. Therefore, when this output is sent to the Data Processor for further analysis, a case-by-case correspondence with the original data will be maintained.



☑ Survival Function
☑ Cumulative Hazard Function
☑ Cox-Snell Residuals
☑ Modified Cox-Snell Residuals
☑ Martingale Residuals
☑ Deviance Residuals
☑ Score (Partial) Residuals
☑ Delta-Beta
☑ Standardised Delta-Beta
☑ Likelihood Displacement
☑ Linear Fit (XBeta)

Survival Function:

$$S(t_j \mid x) = \left[S_0(t_j)\right]^{\mathrm{Exp}(\beta' X_j)}$$

Cumulative Hazard Function:

$$-\mathrm{Log}(S(t_j \mid x)) = \mathrm{Exp}(\beta' X_j) H_0(t_j)$$

Cox-Snell Residuals:

$$rC_j = -\mathrm{Log}(S(t_j \mid x)) = \mathrm{Exp}(\beta' X_j) H_0(t_j)$$

These are identical to the cumulative hazard function.

Modified Cox-Snell Residuals:

$$r^{'}C_j = 1 - \delta_j + rC_j$$

where $\delta_j = 1$ is if the case terminates at time j and $\delta_j = 0$ if it is censored.

Martingale Residuals:

$$rM_j = \delta_j - rC_j$$

Deviance Residuals:

$$rD_j = \mathrm{Sgn}(rM_j)\sqrt{-2\left(rM_j + \delta_j \mathrm{Log}(\delta_j - rM_j)\right)}$$

Score (Partial) Residuals:

$rS_{ij} = x_{ij} - a_{ij}$ for uncensored cases and missing otherwise for all covariates i = 1,…,p, where:

$$a_{ij} = \frac{\sum_{l \in R(t_j)} x_{il} \mathrm{Exp}(\beta' X_l)}{\sum_{l \in R(t_j)} \mathrm{Exp}(\beta' X_l)}$$

Delta-Beta:

$\Delta' I^{-1}$ where I^{-1} is the p x p covariance matrix of regression coefficients and Δ' is an n x p matrix, elements of which are defined as:

$$d_{ij} = \delta_j(x_{ij} - a_{ij}) + \mathrm{Exp}(\beta' X_j) \sum_{t_k \le t_j} \frac{\delta_k a_{ik}}{\sum_{j \in R(t_j)} \mathrm{Exp}(\beta' X_j)}$$

$$+\operatorname{Exp}(\beta' X_j)x_{ij}\sum_{t_k \le t_j}\frac{\delta_k}{\sum_{j\in R(t_j)}\operatorname{Exp}(\beta' X_j)}$$

for j = 1,…,n, i = 1,…, p where a_{ij} is defined as in score residuals above.

Delta-beta is defined as the change in an estimated coefficient when a case is omitted from the analysis. However, unlike in Linear Regression (see 7.2.1.2.2. Linear Regression Case Output), where an exact estimate can be computed without having to run n regressions, the method employed here is an approximation (see Pettitt & Bin Daud, 1989). The user must be aware that different sources and statistical packages may adopt different approximation methods. The best way to find out which method gives the most accurate results is to run a small example and compute each delta-beta by running n regressions, each time omitting one row from the analysis.

Standardised Delta-Beta:

d_{ij}/σ_i for i = 1,…, p.

This is delta-beta divided by its standard error for each coefficient.

Likelihood Displacement:

$\Delta' I^{-1}\Delta$ where I^{-1} is the m x m covariance matrix of regression coefficients and Δ is an n x m matrix, elements of which are as defined in delta-betas. This statistic reflects the effect of deleting a case on the likelihood function.

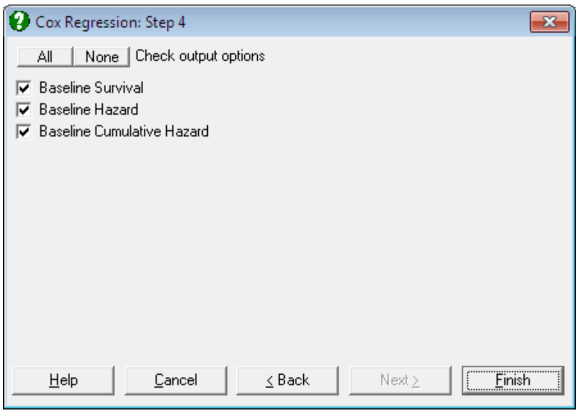
Linear Fit (XBeta):

Defined as:

$$\beta' X_j$$

Alongside baseline functions, this is the only result that is affected by whether the baseline functions are reported centred or relative to the origin (see 9.4.4.2. Cox Regression Intermediate Inputs).

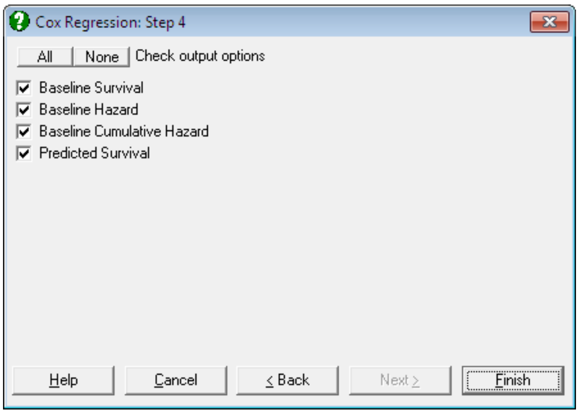
Baseline Functions: Baseline survival, hazard and cumulative hazard functions are reported for each stratum, which are ordered according to the time variable in ascending order. Baseline functions reported here would in general be different for a model where the covariates are centred. If there are no predicted cases, then the Baseline Functions dialogue contains three options.



Check output options

☑ Baseline Survival
☑ Baseline Hazard
☑ Baseline Cumulative Hazard

If at least one case (row) of data does not contain any missing values except for the time variable, then a fourth option Predicted Survival is added to the Baseline Functions dialogue.



Check output options

☑ Baseline Survival
☑ Baseline Hazard
☑ Baseline Cumulative Hazard
☑ Predicted Survival

Baseline Survival:

$$S_0(t_j) = \prod_{k=0}^{j-1} \alpha_k$$

where α_j is obtained by solving the following equation:

$$\sum_{k \in D(t_j)} \frac{\mathrm{Exp}(\beta' X_k)}{1-\alpha_j^{\mathrm{Exp}(\beta' X_k)}} = \sum_{l \in R(t_j)} \mathrm{Exp}(\beta' X_l)$$

where D(t_j) is the set of cases that terminate at time j.

Baseline Hazard:

$$h_0(t_j) = 1 - \alpha_j$$

where α_j is as defined for the baseline survival.

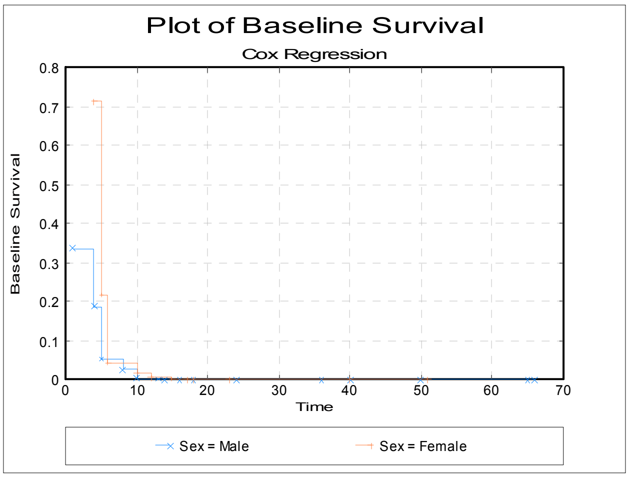
Baseline Cumulative Hazard:

$$-\operatorname{Log}(S_0(t_j))$$

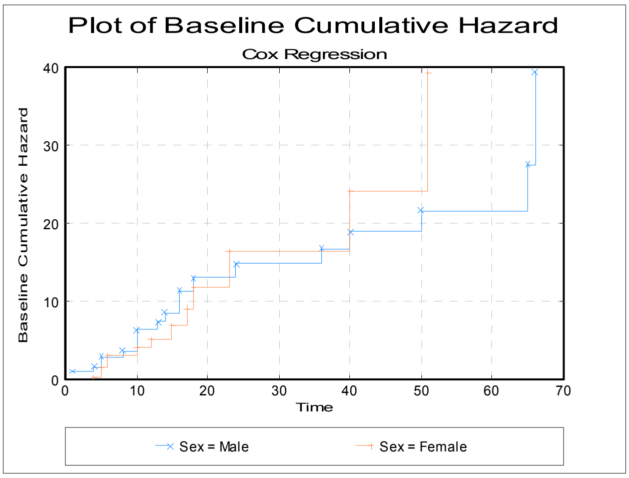
Predicted Survival: This is computed for cases containing no missing values except for the time variable. For each such case, a column of predicted survival times is computed as:

$$S_0(t_j)\hat{}\operatorname{Exp}(\beta' X_k)$$

Plot of Baseline Survival Function: The baseline survival function is plotted against time. The appearance of this graph depends on the selection of estimation at covariate means (see 9.4.4.2. Cox Regression Intermediate Inputs).



Plot of Baseline Cumulative Hazard: The baseline cumulative hazard function is plotted against time. The appearance of this graph depends on the selection of estimation method, at origin or at covariate means (see 9.4.4.2. Cox Regression Intermediate Inputs).



9.4.4.4. Cox Regression Examples

Example 1

Data on survival times of patients in a study on multiple myeloma is given in Table 1.3 (p. 9), in Collett, D. (1994).

Open SURVIVAL and select Statistics 2 → Survival Analysis → Cox Regression. From the Variable Selection Dialogue select the data option 1 Enter Durations and *Survival time* (*C10*) as [Time], *Status* (*C11*) as [Censored] and *Age*, *Sex*, *BUN*, *CA*, *HB*, *PC* and *BJ* (*C12* to *C18*) as [Variable]s.

The Regression Results output reproduces the coefficient estimates and their standard errors as Collett reports them on Table 3.1 (p. 71):

Cox Regression

```
Time Variable: Survival time
Censor Variable: Status
Number of Cases Censored: 12 ( 25.0%)
Valid Number of Cases: 48, 0 Omitted
```

Regression Results

	Coefficient	Standard Error	Wald statistic	Probability	Lower 95%	Upper 95%
Age	-0.0194	0.0279	0.4806	0.4882	-0.0741	0.0354
Sex	-0.2509	0.4023	0.3890	0.5328	-1.0394	0.5376
BUN	0.0208	0.0059	12.3396	0.0004	0.0092	0.0324
CA	0.0131	0.1324	0.0098	0.9211	-0.2465	0.2727
HB	-0.1352	0.0689	3.8537	0.0496	-0.2703	-0.0002
PC	-0.0016	0.0066	0.0587	0.8085	-0.0145	0.0113
BJ	-0.6404	0.4267	2.2529	0.1334	-1.4767	0.1959

-2 Log likelihood:	
Initial Model =	215.9399
Final Model =	199.7009
Pseudo R-squared:	
McFadden =	0.0752
Adjusted McFadden =	0.0104
Cox & Snell =	0.2870
Nagelkerke =	0.2903
Likelihood Ratio Statistic:	
Chi-Square Statistic =	16.2390
Degrees of Freedom =	7
Right-Tail Probability =	0.0230

Example 2

Survival times of patients classified according to age group and whether or not they have had a nephrectomy is given in Table 3.3 (p. 76), in Collett, D. (1994). Example 3.2 (p. 69).

Open SURVIVAL and select Statistics 2 → Survival Analysis → Cox Regression. From the Variable Selection Dialogue select the data option 1 Enter Durations and *Time* (*C27*) as [Time], *Status* (*C28*) as [Censored]. Then highlight *Age group* (*C30*) and click [Dummy] then highlight *Nephrectomy* (*C29*) and click [Dummy] again. This will add two terms to the model: Dummy(*C30 Age Group*) and Dummy(*C29 Nephrectomy*). The next dialogue will ask for a number of model parameters. Leave all entries unchanged as suggested by the program, but change the last one Omit Level? from 0 to 1, to omit the first level of each factor from the model.

The Regression Results output reproduces the coefficient estimates and their standard errors as Collett reports them in Example 3.12 (p. 100):

Cox Regression

Time Variable: Time
Censor Variable: Status
Number of Cases Censored: 4 (11.1%)
Valid Number of Cases: 36, 0 Omitted

Regression Results

	Coefficient	Standard Error	Wald Statistic	Signifi cance	Lower 95%	Upper 95%
Age Group = 2	0.0125	0.4246	0.0009	0.9765	-0.8197	0.8447
3	1.3416	0.5918	5.1396	0.0234	0.1817	2.5014
Nephrectomy = 2	-1.4115	0.5152	7.5044	0.0062	-2.4213	-0.4016

-2 Log likelihood:	
Initial Model =	177.6665
Final Model =	165.5084
Pseudo R-squared:	
McFadden =	0.0684
Adjusted McFadden =	0.0347
Cox & Snell =	0.2866
Nagelkerke =	0.2887
Likelihood Ratio Statistic:	
Chi-Square Statistic =	12.1581
Degrees of Freedom =	3
Right-Tail Probability =	0.0069

Hazard Ratio

	Hazard Ratio	Standard Error	Lower 95%	Upper 95%
Age Group = 2	1.0126	0.4299	0.4406	2.3273
3	3.8250	2.2635	1.1993	12.1996
Nephrectomy = 2	0.2438	0.1256	0.0888	0.6692

Correlation Matrix of Regression Coefficients

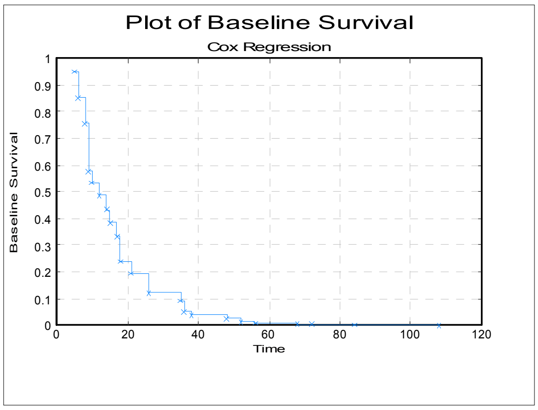
	Age Group = 2	3	Nephrectomy = 2
Age Group = 2	1.0000	0.3119	0.0767
3	0.3119	1.0000	0.0818
Nephrectomy = 2	0.0767	0.0818	1.0000

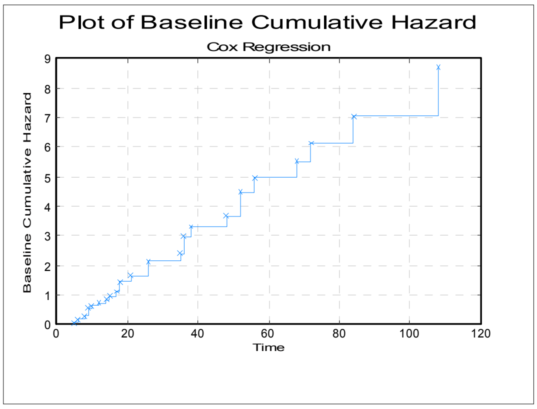
The baseline functions are reported as in Collett Table 3.10 (p. 101):

Cox Regression

Baseline Functions

Time	Baseline Survival	Baseline Hazard	Baseline Cumulative Hazard
5	0.9502	0.0498	0.0510
6	0.8516	0.1038	0.1607
8	0.7552	0.1132	0.2808
9	0.5761	0.2371	0.5514
10	0.5339	0.0733	0.6275
12	0.4861	0.0896	0.7214
14	0.4335	0.1082	0.8359
15	0.3831	0.1163	0.9595
17	0.3326	0.1318	1.1008
18	0.2379	0.2848	1.4360
21	0.1939	0.1851	1.6406
26	0.1198	0.3822	2.1222
35	0.0920	0.2319	2.3860
36	0.0512	0.4430	2.9712
38	0.0369	0.2792	3.2985
48	0.0259	0.2993	3.6543
52	0.0114	0.5598	4.4748
56	0.0070	0.3823	4.9565
68	0.0041	0.4207	5.5024
72	0.0022	0.4673	6.1323
84	0.0009	0.5975	7.0424
108	0.0002	0.8072	8.6886





Example 3

Data on times to removal of a catheter following a kidney infection is given in Table 5.1 (p. 157), in Collett, D. (1994), Example 3.2 (p. 69).

Open SURVIVAL and select Statistics 2 → Survival Analysis → Cox Regression. From the Variable Selection Dialogue select the data option 1 Enter Durations and *Time* (*C32*) as [Time], *Status* (*C33*) as [Censored]. Then select *Age* (*C34*) and *Sex* (*C35*) as [Variable]s. From the Output Options Dialogue select Case (Diagnostic) Statistics and at the next dialogue click All to select all options.

Different types of residuals calculated are reported by Collett in Table 5.2 (p. 157). Collett also reports the approximate estimated delta-betas in Table 5.4 (p. 172) and the likelihood displacement values in Table 5.5 (p.176).

Cox Regression

Case (Diagnostic) Statistics

Row	Survival Function	Cumulative Hazard Function	Cox-Snell Residuals	Modified Cox-Snell Residuals	Martingale Residuals	Deviance Residuals
1	0.7200	0.3286	0.3286	0.3286	0.6714	0.9398
2	0.9245	0.0785	0.0785	0.0785	0.9215	1.8020
3	0.2386	1.4331	1.4331	1.4331	-0.4331	-0.3828
4	0.9104	0.0939	0.0939	0.0939	0.9061	1.7087
5	0.1697	1.7736	1.7736	1.7736	-0.7736	-0.6334
6	0.7322	0.3117	0.3117	1.3117	-0.3117	-0.7895
7	0.7668	0.2655	0.2655	0.2655	0.7345	1.0877
8	0.5836	0.5386	0.5386	0.5386	0.4614	0.5611
9	0.1916	1.6523	1.6523	1.6523	-0.6523	-0.5480
10	0.2409	1.4234	1.4234	1.4234	-0.4234	-0.3751
11	0.2415	1.4207	1.4207	1.4207	-0.4207	-0.3730
12	0.0914	2.3927	2.3927	2.3927	-1.3927	-1.0201
…	…	…	…	…	…	…

Row	Score (Partial) Residuals Age	Score (Partial) Residuals Sex	Delta-Beta Age	Delta-Beta Sex	Standardised Delta-Beta Age	Standardised Delta-Beta Sex
1	-1.0850	-0.2416	0.0020	-0.1977	0.0750	-0.1804
2	14.4930	0.6644	0.0004	0.5433	0.0155	0.4957
3	3.1291	-0.3065	-0.0011	0.0741	-0.0402	0.0677
4	-10.2215	0.4341	-0.0119	0.5943	-0.4528	0.5423
5	-16.5882	-0.5504	0.0049	0.0139	0.1869	0.0127
6	*	*	-0.0005	-0.1192	-0.0206	-0.1088
7	-17.8286	-0.0000	-0.0095	0.1269	-0.3607	0.1158
8	-7.6201	-0.0000	-0.0032	-0.0346	-0.1235	-0.0315
9	17.0910	-0.0000	-0.0073	-0.0733	-0.2771	-0.0669
10	10.2390	-0.0000	0.0032	-0.2023	0.1232	-0.1846
11	2.8575	-0.0000	0.0060	-0.2158	0.2279	-0.1970
12	5.5338	-0.0000	0.0048	-0.1939	0.1829	-0.1770
…	…	…	…	…	…	…

Row	Likelihood Displacement	Linear Fit (XBeta)
1	0.0328	-1.8604
2	0.3387	-4.0852
3	0.0046	-1.7389
…	…	…
9	0.1334	-3.5993
10	0.0353	-4.1156
11	0.0611	-4.5104
12	0.0432	-4.4800
…	…	…

9.5. Fourier Analysis

The Fourier Analysis is used to transform real or complex data, which is assumed to be in the time domain, into the frequency domain. Once a series has been transformed, various further transformations can be carried out. Then the series can be transformed back into the original scale using the Inverse Fourier Transform procedure. A filtering effect can be achieved in this way.

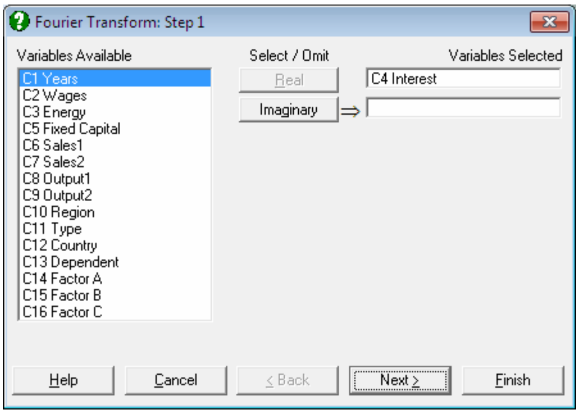
The Fourier Transform is based on an expansion of a complex periodic function of time into a sum of sine and cosine waves.

$$y_p = \sum_{t=0}^{n-1} x_t \mathrm{Cos}\left(2\pi \frac{tp}{n}\right) + i \sum_{t=0}^{n-1} x_t \mathrm{Sin}\left(2\pi \frac{tp}{n}\right)$$

where:

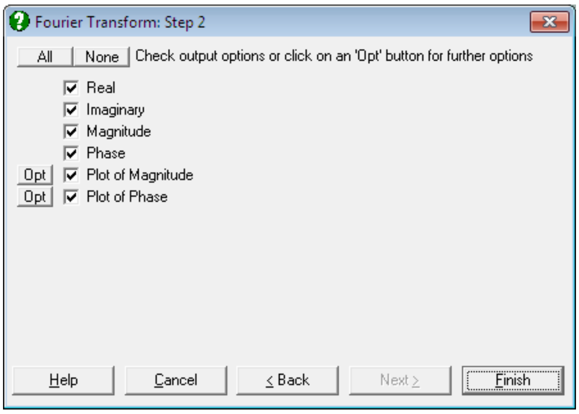
y_p is the pth complex-valued output in the frequency domain, $p = 0,\ldots, n - 1$,
x_t is the tth complex-valued input in the time domain and
n is the number of observations.

This formula requires n^2 computations and applies to any number of observations. For large values of n though, it may take a very long time to compute. Instead, the Fast Fourier Transform (FFT) method is employed here, which requires only $n(\mathrm{Log}_2(n))$ computations (see Elliott, D. F. & K. R. Rao 1982). A restriction brought by the FFT method is that it only works with a number of observations which is a power of 2. If the number of points is not a power of 2, then UNISTAT extends the series by its mean so that it has a number of cases which is equal to the next power of 2.



You may select a column containing the real part by clicking on [Real] and / or a second column containing the imaginary part by clicking on [Imaginary]. One of real or imaginary components or both of them can be selected. In most cases only the real part will need to be selected. When this is the case, the real and imaginary parts of the output will only contain n / 2 distinct values (or (n + 1) / 2 if n is odd), values being symmetric about the midpoint of the series (the Nyquist frequency). For the Inverse Fourier Transform both the real and imaginary parts will usually be needed.

9.5.1. Fourier Transform


Fourier Transform: Step 2

Check output options or click on an 'Opt' button for further options

- ☑ Real
- ☑ Imaginary
- ☑ Magnitude
- ☑ Phase
- Opt ☑ Plot of Magnitude
- Opt ☑ Plot of Phase

The real and imaginary parts of the series in the frequency domain can be displayed, as well as the corresponding magnitude and phase values and their plots.

$$\text{Magnitude}_p = \sqrt{a_p^2 + b_p^2}$$

$$\text{Phase}_p = \text{a}\tan 2(b_p, a_p)$$

where a_p and b_p are the real and complex components in the frequency domain respectively and atan2 function returns the arctangent of b_p / a_p in radians. The following inverse relationships should also hold:

$$a_p = \text{Magnitude}_p \text{Cos}(\text{Phase}_p)$$

$$b_p = \text{Magnitude}_p \text{Sin}(\text{Phase}_p)$$

If you wish to express the phase angle in degrees, rather than radians, then you may use a conversion factor of π / 180. The same conversion can be performed by the program automatically by entering the following line in *Documents\Unistat65\Unistat65.ini* file under the [Options] section:

```
FFTPhaseDegree=1
```

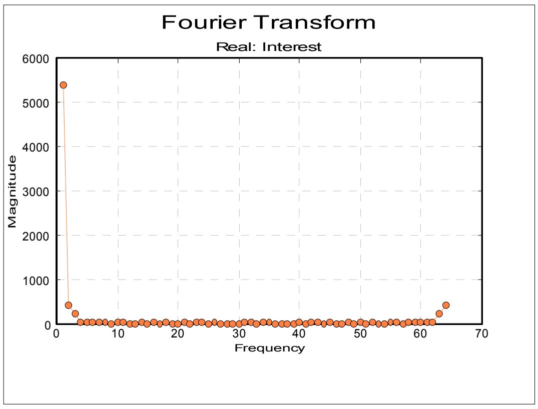
Example

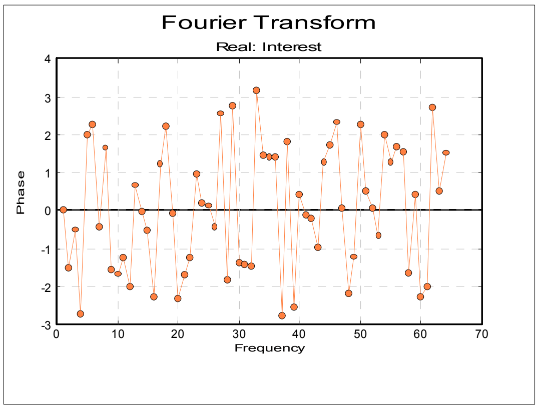
Open DEMODATA and select Statistics 2 → Fourier Analysis → Fourier Transform. From the Variable Selection Dialogue select *Interest* (*C4*) as [Real] to obtain the following results:

Fourier Transform

Real: Interest

	Real	Imaginary	Magnitude	Phase (Radian)
1	5408.4193	0.0000	5408.4193	0.0000
2	21.4881	-416.1937	416.7481	-1.5192
3	223.7763	-122.4555	255.0905	-0.5007
4	-29.7381	-13.8112	32.7888	-2.7068
5	-14.7492	32.7994	35.9630	1.9934
6	-30.7761	37.7633	48.7159	2.2546
7	42.1814	-18.9490	46.2421	-0.4222
8	-3.1800	39.3871	39.5152	1.6514
9	0.4593	-15.9255	15.9321	-1.5420
10	-5.0154	-50.0637	50.3142	-1.6706
11	14.9913	-45.5162	47.9215	-1.2526
12	-10.0253	-21.5218	23.7422	-2.0067
...	…16.0507	…	…	…
48	-24.6684	-34.0238	42.0256	-2.1981
49	4.8234	-13.0566	13.9190	-1.2169
50	-21.2454	25.1352	32.9112	2.2725
51	20.7350	11.7082	23.8122	0.5140
52	40.1326	2.0183	40.1833	0.0502
53	16.0507	-12.5679	20.3857	-0.6643
54	-10.0253	21.5218	23.7422	2.0067
55	14.9913	45.5162	47.9215	1.2526
56	-5.0154	50.0637	50.3142	1.6706
57	0.4593	15.9255	15.9321	1.5420
58	-3.1800	-39.3871	39.5152	-1.6514
59	42.1814	18.9490	46.2421	0.4222
60	-30.7761	-37.7633	48.7159	-2.2546
61	-14.7492	-32.7994	35.9630	-1.9934
62	-29.7381	13.8112	32.7888	2.7068
63	223.7763	122.4555	255.0905	0.5007
64	21.4881	416.1937	416.7481	1.5192

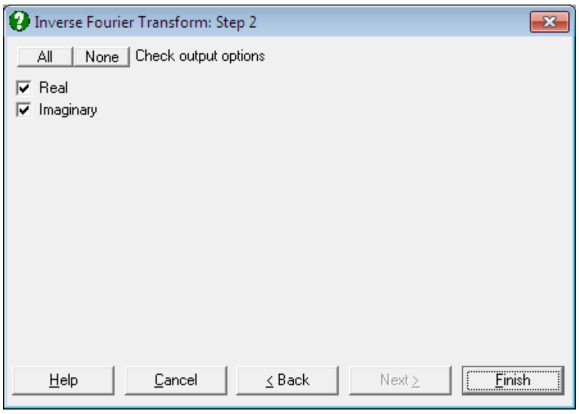




Since there are only 58 rows in the variable *Interest*, it is first padded by the program automatically up to the next power of 2 (64) by the mean value of the variable.

9.5.2. Inverse Fourier Transform

The real and imaginary parts of the series in the time domain is displayed.



Check output options

☑ Real

☑ Imaginary

Example

Carry out the example in Fourier Transform above. If you are using UNISTAT in Stand-Alone Mode, click on the UNISTAT icon on the Output Medium Toolbar to send the output table to UNISTAT spreadsheet. In Excel Add-In Mode select the output matrix as data. Then select Statistics 2 → Fourier Analysis → Inverse Fourier Transform. From the Variable Selection Dialogue select *Real* (*C17*) as [Real] and *Imaginary* (*C18*) as [Imaginary]to obtain the following results:

Inverse Fourier Transform

Real: Real, Imaginary: Imaginary

	Real	Imaginary
1	95.7300	0.0000
2	94.5200	-0.0000
3	93.2900	-0.0000
4	96.9200	-0.0000
5	88.8900	-0.0000
6	98.1600	-0.0000
7	97.0300	-0.0000
8	100.3100	-0.0000
9	102.1200	0.0000
10	102.4800	-0.0000
11	102.3100	-0.0000
12	102.5900	0.0000

	Real	Imaginary
13	83.3500	-0.0000
14	94.6100	0.0000
15	96.5000	0.0000
16	93.6900	0.0000
17	84.7600	0.0000
18	88.0200	0.0000
19	83.7300	0.0000
…	…	…
50	61.7400	-0.0000
51	61.7600	0.0000
52	61.6100	0.0000
53	71.4700	-0.0000
54	66.7900	-0.0000
55	69.0800	-0.0000
56	68.3500	-0.0000
57	72.5300	-0.0000
58	65.7400	-0.0000
59	84.5066	0.0000
60	84.5066	-0.0000
61	84.5066	0.0000
62	84.5066	-0.0000
63	84.5066	-0.0000
64	84.5066	-0.0000

We can see that the real part is the same as *Interest* (*C4*) with the mean value added in rows 59 to 64. The imaginary part is almost zero, except for the accumulated round-off errors.

UNISTAT Statistical Package

Chapter 10
Analysis of Bioassays
(Optional Module)

10.0. Overview

In cases where effectivity of a drug or a substance cannot be assessed by chemical or physical analysis, it can be assessed by testing its effects on biological organisms. Such experiments are called biological assays, or bioassays in short. The most common form of bioassay is where the Potency of one or more *test* preparations is determined by comparison to a *standard* preparation.

This module supports Potency calculations employing Parallel Line Method, Slope Ratio Method and Quantal Response Method, complete with confidence intervals, validity tests and graphical representations. Note that other types of bioassays and effective dose (or ED50) applications can be analysed using UNISTAT's Nonlinear Regression and Logit / Probit / Gompit.

10.1. Parallel Line Method

Balanced, symmetric or unbalanced assays can be analysed. The analysis is based on a regression of the response variable against the natural logarithm of the dose variable. A separate line is fitted on each preparation, subject to a constraint that they should be parallel. An assay is said to be *balanced* when:

1) there is an equal number of cases in each treatment group,
2) there is an equal number of dose groups for each preparation and
3) successive dose levels are the same for all preparations.

An assay fulfilling the first two conditions but having different dose levels for different preparations (yet having the same ratio of successive dose levels) will be called *symmetric*. Assays not fulfilling one or more of these conditions will be called *asymmetric* or *unbalanced*.

For validity tests, the following Analysis of Variance (ANOVA) options are available:

1) Completely randomised design
2) Randomised block design
3) Latin squares design
4) Twin and triple crossover designs

The *unbalanced* assays can only be analysed using the Completely Randomised Design option. All other options require *symmetric* or *balanced* assays. In most cases, the program will detect whether an assay is *unbalanced*, *symmetric* or *balanced* and apply the relevant algorithm automatically.

The data sets to be analysed according to *European Pharmacopoeia* (1997-2008) Parallel Line Method should be balanced.

10.1.1. Data Preparation

Data is usually given in the form of a table where measurements corresponding to different preparations and dose levels are in separate columns (i.e. treatment groups). Let:

h be the number of preparations (including the standard preparation),
k be the number of treatments and
n be the number of cases in each treatment group. Then, it follows that
d = k / h is the number of dose levels.

The standard is always the first preparation in a column of data.

Consider the following hypothetical 3-dose / 2-preparation example where h = 2, k = 6 and n = 4 and suppose the dose levels are given as 0.125, 0.25 and 0.5.

	Preparations					
	Standard			**Unknown**		
Cases	**Dose 1**	**Dose 2**	**Dose 3**	**Dose 1**	**Dose 2**	**Dose 3**
1	1.3	2.1	4.1	1.5	2.0	3.9
2	1.7	2.3	4.2	1.1	1.9	4.6
3	1.1	2.7	3.9	0.9	2.1	4.0
4	1.5	2.2	4.3	1.0	2.2	3.7

Although this is a well-defined data set for a bioassay, it should be first transformed into a more convenient format for analysis using a statistical package. This is done by stacking all response measurements in a single column. It is also necessary to create a number of categorical data columns (or *factors*) to keep track of which measurement belongs to which preparation, to which dose group and to which treatment case.

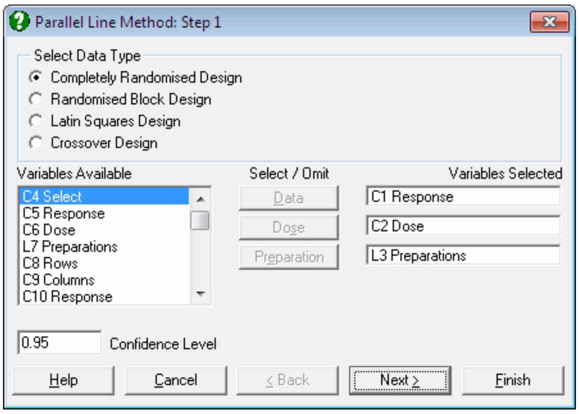
For analysis with UNISTAT, the data for the above example should be entered as follows:

Data	Dose	Preparation	Rows	Columns
1.3	.125	Standard	1	1
1.7	.125	Standard	2	1
1.1	.125	Standard	3	1
1.5	.125	Standard	4	1
2.1	.25	Standard	1	2
2.3	.25	Standard	2	2
2.7	.25	Standard	3	2
2.2	.25	Standard	4	2
4.1	.5	Standard	1	3
4.2	.5	Standard	2	3
3.9	.5	Standard	3	3
4.3	.5	Standard	4	3
1.5	.125	Unknown	1	4
1.1	.125	Unknown	2	4
0.9	.125	Unknown	3	4
1.0	.125	Unknown	4	4
2.0	.25	Unknown	1	5
1.9	.25	Unknown	2	5
2.1	.25	Unknown	3	5
2.2	.25	Unknown	4	5
3.9	.5	Unknown	1	6
4.6	.5	Unknown	2	6
4.0	.5	Unknown	3	6
3.7	.5	Unknown	4	6

The *Dose* column contains the actual dose units for all preparations, instead of dose group numbers. This information is needed in Potency and Plot of Treatment Means options. Also, the column *Rows* is needed for Randomised Block, Latin Squares Design and Crossover Design and *Columns* is needed for Latin Squares Design and Crossover Design. In Stand-Alone Mode, *Rows* and *Columns* variables can be generated automatically by using UNISTAT spreadsheet functions **Level(4)** and **Level(4);B** respectively. (see 3.4.2.5. Statistical Functions).

10.1.2. Parallel Line Variable Selection

Once the data is arranged as described above, select Bioassay → Parallel Line Method from UNISTAT menus. A Variable Selection Dialogue will pop up.



Parallel Line Method: Step 1

Select Data Type
● Completely Randomised Design
○ Randomised Block Design
○ Latin Squares Design
○ Crossover Design

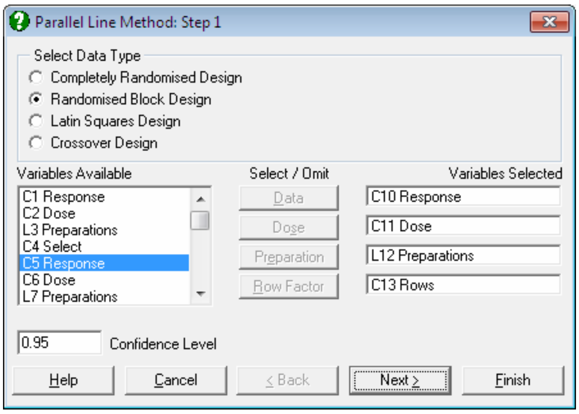
Variables Available: C4 Select, C5 Response, C6 Dose, L7 Preparations, C8 Rows, C9 Columns, C10 Response

Variables Selected:
Data: C1 Response
Dose: C2 Dose
Preparation: L3 Preparations

0.95 Confidence Level

Data columns available for selection are listed on the left. Variables are referred to by their column numbers, which are prefixed by a single letter representing the type of data. For instance, in the above example *C1*, *C2* and *C4* are numeric columns, whereas *L3* means that column three contains Long Strings. Columns containing Short Strings (up to 8 characters) are prefixed by (S). Other data types that will probably not be used in bioassays are date (D) and time (T). If Column Labels have been entered, they will also appear in the list next to the column numbers.

The frame Select Data Type (at the top) displays options for the type of Analysis of Variance to be performed. The number of variables to be selected is different for these types of analyses. When the second option Randomised Block Design is selected, four variables will need to be selected.



Parallel Line Method: Step 1

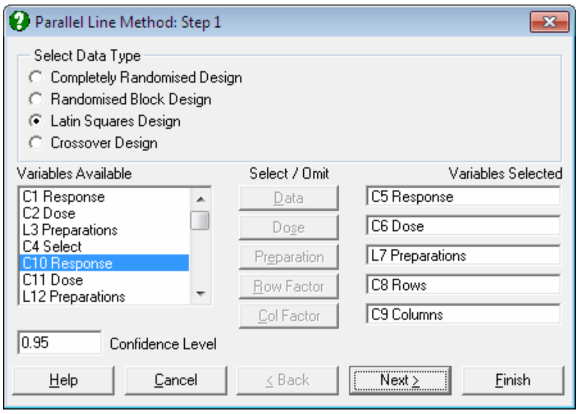
Select Data Type
○ Completely Randomised Design
● Randomised Block Design
○ Latin Squares Design
○ Crossover Design

Variables Available: C1 Response, C2 Dose, L3 Preparations, C4 Select, C5 Response, C6 Dose, L7 Preparations

Variables Selected:
Data: C10 Response
Dose: C11 Dose
Preparation: L12 Preparations
Row Factor: C13 Rows

0.95 Confidence Level

The third and fourth options Latin Squares Design and Crossover Design require selection of five variables.



Parallel Line Method: Step 1

Select Data Type

- ○ Completely Randomised Design
- ○ Randomised Block Design
- ● Latin Squares Design
- ○ Crossover Design

Variables Available: C1 Response, C2 Dose, L3 Preparations, C4 Select, C10 Response, C11 Dose, L12 Preparations

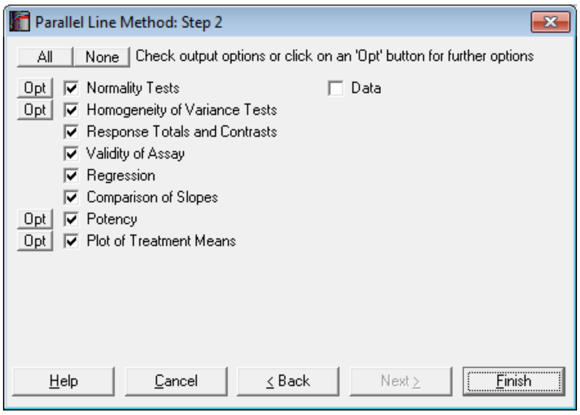
Select / Omit	Variables Selected
Data	C5 Response
Dose	C6 Dose
Preparation	L7 Preparations
Row Factor	C8 Rows
Col Factor	C9 Columns

0.95 Confidence Level

After selecting the analysis type, you will need to assign tasks to variables by sending them to the boxes on the right. To do this, highlight the variable on the left list and click on the desired task button (i.e. one of the command buttons in the middle of the dialogue). Likewise, you can deselect an already selected variable by highlighting it on the right list first and then clicking its task button.

When all variables are selected, click the [Next] button to proceed to Output Options Dialogue.

10.1.3. Parallel Line Output Options



Parallel Line Method: Step 2

All | None | Check output options or click on an 'Opt' button for further options

- Opt ☑ Normality Tests
- Opt ☑ Homogeneity of Variance Tests
- ☑ Response Totals and Contrasts
- ☑ Validity of Assay
- ☑ Regression
- ☑ Comparison of Slopes
- Opt ☑ Potency
- Opt ☑ Plot of Treatment Means
- □ Data

Output options that have further options under them (i.e. they have further dialogues and windows to display) then an [Opt] button is placed to the left of their check boxes. When you click [Finish] without clicking on an [Opt] button first, the program will generate output with the default values. If you want to change the default values, you can click on the [Opt] button to display the further dialogues for this particular output option. Then you can either obtain this particular output option on its own by clicking [Finish], or click [Back] to display the Output Options Dialogue again and output all selected options together.

[Opt] buttons on this dialogue will allow you to choose from four different types of Normality Tests, five Homogeneity of Variance Tests, enter assigned potencies for test preparations and edit the plot of treatment means in UNISTAT's Graphics Editor.

Data

This new output option (which is available for all three bioassay analysis methods supported here) will enable the user to include a printout of the data used in the analysis as part of the output. This may be useful in fulfilling reporting and data integrity requirements.

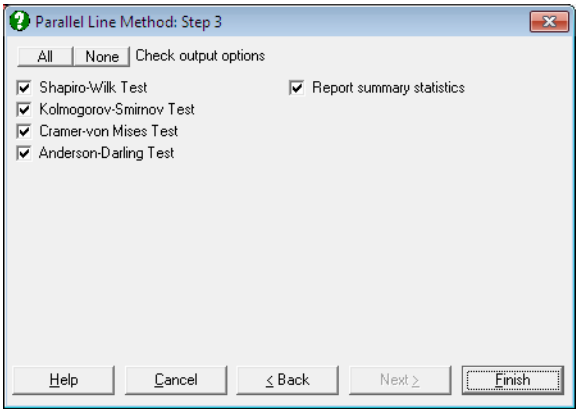
Data

Preparations	Response	Dose
Standard S	300.0000	0.2500
Standard S	310.0000	0.2500
Standard S	330.0000	0.2500
Standard S	289.0000	1.0000
Standard S	221.0000	1.0000
Standard S	267.0000	1.0000
Preparation T	310.0000	0.2500
Preparation T	290.0000	0.2500
Preparation T	360.0000	0.2500
Preparation T	230.0000	1.0000
Preparation T	210.0000	1.0000
Preparation T	280.0000	1.0000
Preparation U	250.0000	0.2500
Preparation U	268.0000	0.2500
Preparation U	273.0000	0.2500
Preparation U	236.0000	1.0000
Preparation U	213.0000	1.0000
Preparation U	283.0000	1.0000

10.1.3.1. Normality Tests for Bioassays

One of the basic assumptions of Parallel Line Method is that for each treatment group (i.e. a unique dose-preparation combination), observations are normally distributed.

Earlier versions of UNISTAT featured a classic Shapiro-Wilk normality test (1965, 1968) as recommended by *European Pharmacopoeia* (1997-2008). However, this test was shown to be inaccurate and substantially revised by Royston (1995). Also, there are other normality tests which are more powerful than Shapiro-Wilk, such as Cramer-von Mises and Anderson-Darling. Accordingly, we provide the four most commonly used normality tests as part of the Parallel Line Method (see 6.3.3. Normality Tests). A futher dialogue enables you to display all or any of the four normality tests supported.



☑ Shapiro-Wilk Test
☑ Kolmogorov-Smirnov Test
☑ Cramer-von Mises Test
☑ Anderson-Darling Test
☑ Report summary statistics

If you still wish to use the classic Shapiro-Wilk (1965) and its accompanying overall normality tests as in earlier version of UNISTAT, then you can do so by entering the following line in *Documents*\Unistat65\Unistat65.ini file under the [Options] section:

```
OverallNormality=1
```

In classic Shapiro-Wilk test, observations are arranged in ascending order for each treatment group and then the following sum is found:

$$b = \sum_{i=1}^{k} a_{n-i+1}(y_{n-i+1} - y_i)$$

The test statistic for each sample is:

$$W = \frac{b^2}{S^2}$$

where S^2 is the sum of squared differences from the mean, and a_i i = 1, …, k are the coefficients given by the authors.

If all sample sizes are between 7 and 20 (inclusive), an overall test of normality, which is based on the normal distribution, is also performed according to Shapiro & Wilk (1968).

First, the following ratio is calculated for each sample:

$$V_i = \log\left(\frac{W_i\text{-}a}{1\text{-}W_i}\right)$$

and:

$$H_i = q + mV_i$$

where a, q and m are the coefficients for k degrees of freedom given by the authors in Shapiro & Wilk (1968).

The test statistic is defined as:

$$Z = \frac{\sum H_i}{\sqrt{k}}$$

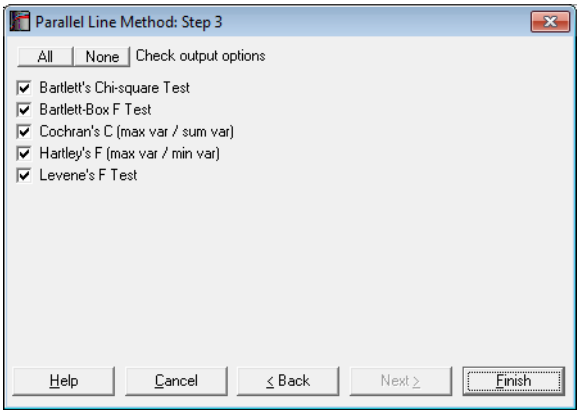
with a 1-tail probability from the normal distribution.

10.1.3.2. Homogeneity of Variance Tests

Another basic assumption of Parallel Line Method is that variances for different treatment groups are not significantly different from each other.

Earlier versions of UNISTAT featured Bartlett's chi-square test as recommended by *European Pharmacopoeia* (1997-2008), and Hartley's F test. Here we provide three more homogeneity of variance tests. The computationally demanding Levene's test is considered to be more powerful than other homogeneity of variance tests. For a detailed description of these tests see 7.4.2.1. Homogeneity of Variance Test Results.

A futher dialogue enables you to display all or any of the five homogeneity of variance tests supported.



Check output options

☑ Bartlett's Chi-square Test
☑ Bartlett-Box F Test
☑ Cochran's C (max var / sum var)
☑ Hartley's F (max var / min var)
☑ Levene's F Test

10.1.3.3. Response Totals and Contrasts

These are the intermediate values calculated directly from raw data and they are used in computing all output statistics. Here we report these values in order to help the user with validating the final results.

First d (number of doses) rows of the table report the sums of all cases in each treatment group. Let:

$$S_{ij} = \sum_{k=1}^{n} X_{ijk}, j = 1, \ldots, d, i = 1, \ldots, h$$

represent the sum of cases for the jth dose and the ith preparation in the table.

The next row *Total* is the sum of these values over dose for each preparation:

$$T_i = \sum_{j=1}^{d} S_{ij}, i = 1, \ldots, h.$$

The contrasts are then calculated for each preparation (i = 1, … , h) as follows:

No of doses	**Linear Contrast (L_i)**	**Quadratic Contrast (Q_i)**	**Cubic Contrast (J_i)**
2	$S_{2i} – S_{1I}$		
3	$S_{3i} – S_{1i}$	$S_{1i} – 2S_{2i} + S_{3i}$	
4	$3S_{4i} + S_{3i} – S_{2i} – 3S_{1i}$	$S_{1i} – S_{2i} – S_{3i} + S_{3i}$	$3S_{2i} – S_{1i} – S_{4i} – 3S_{3I}$

10.1.3.4. Validity of Assay

This output option displays an Analysis of Variance (ANOVA) table, which is used to test the Validity of Assay. Also, the residual sum of squares and its degrees of freedom are used in estimating the confidence limits for the Potency (see 10.1.3.7. Potency). The three basic tests performed are (i) significance of regression, (ii) parallelism and (iii) linearity. The table may have different entries in its rows depending on the number of doses and / or the ANOVA model employed.

The notation below is for balanced designs as given by *European Pharmacopoeia* (1997-2008). For unbalanced designs, the only difference is that the sums are taken up to the maximum number of observation in each treatment group. See 10.2. Slope Ratio Method, section Validity of Assay for a general unbalanced formulation.

Let us first define the three key entries of all ANOVA tables, namely, the Constant term:

$$K = \frac{1}{nk}\sum_{i=1}^{h} T_i$$

which is the sum total of all cases divided by the total number of treatment groups, the Treatments term:

$$M = \frac{1}{n}\sum_{i=1}^{h} T_i^2 - K$$

which is the sum of all squared treatment totals minus the constant term, and the Total term:

$$T = \sum_{k=1}^{n}\sum_{j=1}^{d}\sum_{i=1}^{h} X_{ijk}^2 \text{-} K$$

which is the sum of all squared cases minus the constant term. The rest of table entries are defined as follows:

	Degrees of Freedom	**2-dose**	**3-dose**	**4-dose**
Preparations	h - 1	$\frac{\sum T_i^2}{dn} - K$	$\frac{\sum T_i^2}{dn} - K$	$\frac{\sum T_i^2}{dn} - K$
Linear Regression	1	$E = \frac{(\sum L_i)^2}{2nh}$	$E = \frac{(\sum L_i)^2}{2nh}$	$E = \frac{(\sum L_i)^2}{20nh}$
Non-parallelism	h - 1	$\frac{\sum L_i^2}{2n} - E$	$\frac{\sum L_i^2}{2n} - E$	$\frac{\sum L_i^2}{20n} - E$
Non-linearity	h for 3-dose 2h for 4-dose		$\frac{\sum Q_i^2}{6n}$	$\frac{\sum Q_i^2}{4n} + \frac{\sum J_i^2}{20n}$
Quadratic Regression	1		$Q = \frac{(\sum Q_i)^2}{6nh}$	$Q = \frac{(\sum Q_i)^2}{4nh}$
Difference of Quadratics	h - 1		$\frac{\sum Q_i^2}{6n} - Q$	$\frac{\sum Q_i^2}{4n} - Q$
Residual	h			$\frac{\sum J_i^2}{20n}$
Treatments	k - 1	M	M	M
Residual		R	R	R
Total	nk - 1	T	T	T

The following relationships should always be true:

1) Degrees of freedom and sum of squares for the first four rows (i.e. Preparations, Linear Regression, Non-parallelism and Non-linearity) should always add up to Treatments,
2) Quadratic Regression, Difference of Quadratics and their Residuals should always add up to Non-linearity,

3) Treatments and their Residuals should always add up to Total.

All four ANOVA methods supported here differ only in their residual terms, where the residual for the latter two designs contain a Row Block term, which is defined as:

$$R_{Row} = \frac{\sum_{k=1}^{n} R_k^2}{k} - K$$

where:

$$R_k = \sum_{j=1}^{d} \sum_{i=1}^{h} X_{ijk}, k = 1, \ldots, n.$$

The Latin Squares Design also contains a Column Block term, which is defined as:

$$R_{Column} = \frac{\sum_{j=1}^{d} \sum_{i=1}^{h} C_{ij}^2}{k} - K$$

where:

$$C_{ij} = \sum_{k=1}^{n} X_{ijk}, i = 1, \ldots, h, j = 1, \ldots, d.$$

Although the sum of squares and degrees of freedom here are not different from that of Completely Randomised Design, F-statistics and their probability values will be different, as these values are based on the residual sum of squares and degrees of freedom found after removing the effects of Row and Column blocks.

The Crossover Design is different from others in that the column factor usually represents time periods for different treatments and the model includes interaction terms between this variable and others. If the *Dose* variable has only two values for each preparation then the model is called Twin Crossover Design and it contains interactions between the column factor and Preparations, Linear Regression and Non-Parallelism terms. If there are three dose levels for each preparation, the model is called Triple Crossover Design and it contains additional interactions between the column factor and Quadratic Regression and Quadratic Difference terms.

The Validity of Assay output has different entries for different ANOVA Designs and for different number of dose units used. See 10.2.3. Slope Ratio Examples for different types.

10.1.3.5. Regression

A separate regression line is fitted on each preparation against the natural logarithm of dose. The Common Regression is obtained by pooling the difference sum of squares for all preparations and the Total Regression by regressing the response variable on the log of dose variable, without distinguishing between preparations. The slopes of Common Regression and Total Regression are not the same when the design is not balanced (i.e. when it is symmetric or asymmetric). In that case, the slope of Common Regression is used as the common slope in Potency calculations.

The output from this procedure is similar to the first part of output from Heterogeneity of Regression procedure (see 7.4.5. Heterogeneity of Regression).

10.1.3.6. Comparison of Slopes

If an assay with two or more test preparations is found to depart from parallelism significantly, then we ask the question which test preparation's slope differs from the slope of the standard preparation. A Dunnett's multiple comparison test is performed to answer this question.

This output option is equivalent to part of the analysis from Heterogeneity of Regression procedure. Running the same bioassay data set with this procedure, however, may produce slightly different results, since while Heterogeneity of Regression procedure is always based on a 1-way ANOVA model, the current procedure is based on the residual sum of squares and its degrees of freedom as computed for the Validity of Assay output.

European Pharmacopoeia (1997-2008) employs a slightly different algorithm, which is based on linear contrasts as a proxy for the slopes. The two approaches are identical and produce the same probability values. Although we report here the slopes test by default, the linear contrast test output can be displayed instead by entering the following line in *Documents\Unistat65\Unistat65.ini* file under the [Bioassay] section:

```
ParalEuroPharma=1
```

This *Unistat65.ini* line also affects the Potency output below.

For *European Pharmacopoeia* (1997-2008) linear contrast test we first calculate a test statistic q' for each test preparation:

$$q' = \frac{L_1\text{-}L_i}{2\sqrt{ns^2}} \quad i = 2, \ldots, h \text{ for a 2 or 3 - dose model}$$

$$q' = \frac{L_1\text{-}L_i}{2\sqrt{10ns^2}} \quad i = 2, \ldots, h \text{ for a 4 - dose model}$$

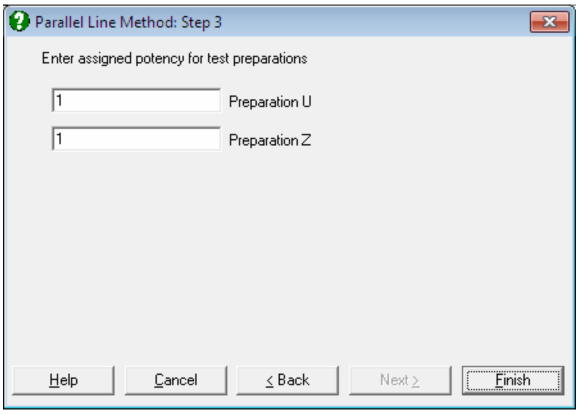
where:

L_1 is the linear contrast for the standard preparation,
L_i is the linear contrast for the ith test preparation and
s^2 is the residual mean square value from the ANOVA table (i.e. sum of squares for the overall residual term divided by its degrees of freedom)

The two-tailed probability for the test statistic is generated using an algorithm developed by Charles Dunnett for α significance level, (h - 1) number of groups. The degrees of freedom is equal to that of the overall residual term of the ANOVA table.

10.1.3.7. Potency

By default, each test preparation is assigned a Potency of unity, in which case the reported potency is the relative potency ratio. If you want to change this click the [Opt] button situated to the left of the Potency option. Then a further dialogue pops up asking for entry of assigned potency for each test preparation.



Parallel Line Method: Step 3

Enter assigned potency for test preparations

1 Preparation U

1 Preparation Z

By default, the program calculates the potency ratio and its confidence limits employing the generalised algorithm given in Finney (1978), which can work with unbalanced, symmetric and balanced designs. Alternatively, the more restrictive algorithm for balanced assays (see *European Pharmacopoeia* 1997-2008) can be employed by entering the following line in *Documents\Unistat65\Unistat65.ini* file under the [Bioassay] section:

```
ParalEuroPharma=1
```

This line also affects the Comparison of Slopes output above. For balanced designs both methods produce exactly the same estimates.

The logarithm of potency ratio is estimated for each test preparation, using the Common Regression slope b.

$$M_i^{'} = \overline{x}_s - \overline{x}_i - \frac{\overline{y}_s - \overline{y}_i}{b} \quad i = 2, \ldots, h$$

and:

$$M = \mathrm{Log}(A_i) + M_i^{'} \quad i = 2, \ldots, h$$

where:

$$\overline{y}_i = \frac{S_i}{N_i} \quad i = 1, \ldots, h$$

are the preparation means and A_i is the assigned potency of each test preparation. The estimated potency is the antilog of M, Exp(M).

The method of estimating the confidence interval for potency is based on *Fieller's Theorem* (see Finney 1978, p. 80). Let us first define the correction factor g as:

$$g = \frac{s^2 t_{\alpha,df}^2}{E}$$

where E is the sum of squares for the Linear Regression term and s^2 is the residual mean squares from the ANOVA table. $t_{\alpha,df}$ is the critical value from the t-distribution with the same degrees of freedom. The confidence limits computed below are reliable for $g < 1$. If this condition is not fulfilled, the program will issue a warning message, but still display the confidence limits computed.

The log of confidence limits for the potency ratio of each test preparation is defined as:

$$M_{iL}, M_{iU} = \left(M_i^{'} \pm \frac{st_{\alpha,df}}{b} \sqrt{(1-g)\left(\frac{1}{N_S} + \frac{1}{N_{Ti}}\right) + \frac{\left(bM_i^{'}\right)^2}{E}} \right) / (1-g)$$

where the variance of M_i is:

$$V_i = \frac{s^2}{b^2}\left(\frac{1}{N_S} + \frac{1}{N_{Ti}} + \frac{\left(bM_i^{'}\right)^2}{E} \right)$$

Weights are computed after the estimated potency and its confidence interval are found:

$$W_i = \frac{4t_{\alpha,df}^2}{\left(M_{iU} - M_{iL}\right)^2}$$

and % Precision is:

$$P_i = 100\frac{M_{iL}}{M_i}$$

According to *European Pharmacopoeia* (1997) the common slope b is calculated as:

$$b = \frac{\sum L_i}{(d-1)Znh} \text{ for a 2 or 3 - dose model and}$$

$$b = \frac{\sum L_i}{10Znh} \text{ for a 4 - dose model.}$$

Where:

$Z = \text{Log}(\text{dose}_{i+1}) - \text{Log}(\text{dose}_i)$, $i = 2, \ldots, d$

is the log of successive dose ratios. Also define a correction factor:

$$C = \frac{E}{E - s^2 t_{\alpha,df}^2}$$

where E is the sum of squares for the Linear Regression term and s^2 is the residual mean squares and $t_{\alpha,df}$ is the critical value from the t-distribution with degrees of freedom of the overall residual term from the ANOVA table.

The log of the corrected potency estimate and its confidence intervals are computed as:

$$\mathrm{Log}(A_i) + CM_i^{'} \pm \sqrt{(C-1)\left(CM_i^{'2} + 2H\right)} \quad i = 2, \ldots, h$$

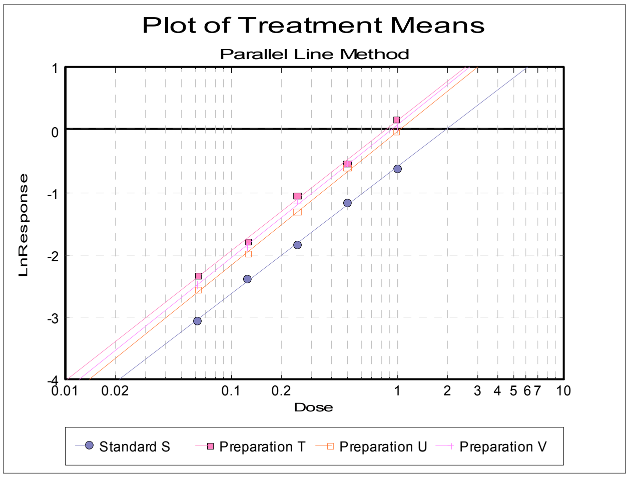
where:

$$H = \frac{E}{b^2 dn}$$

The only difference from the default output here is reporting of C and H constants for validation purposes, where C = 1 / (1 - g).

10.1.3.8. Plot of Treatment Means

This option generates a Plot of Treatment Means against the log of dose. It provides a visual means of inspecting the data, enabling the user to notice immediately whether there is something substantially wrong with the data. In the following example, for instance, the slope of *Preparation T* is quite different from that of *Standard* and *Preparation U*.



Clicking the [Opt] button situated to the left of the Plot of Treatment Means option will place the graph in UNISTAT's Graphics Editor. Each preparation will be plotted as one data series, with as many points as the number of doses applied. A line of best fit will be drawn for each series, including the standard and all test preparations.

The plot can be further customised and annotated using the tools available under UNISTAT Graphics Window's Edit menu.

10.1.4. Parallel Line Examples

The following Parallel Line Method examples are based on different Analysis of Variance methods. The data sets were entered into UNISTAT's spreadsheet and the necessary data manipulations made by using UNISTAT spreadsheet functions (see 10.1.1. Data Preparation). The final data sets were saved in two files; BIOPHARMA6 which contains examples from *European Pharmacopoeia* (2008, the 6th edition) and BIOFINNEY containing examples from Finney (1978).

10.1.4.1. Completely Randomised Design with 2 Doses and 3 Preparations

Data is given in Table 5.1.1-I. on p. 582 of *European Pharmacopoeia* (2008).

Open BIOPHARMA6 and select Bioassay → Parallel Line Method. From the Variable Selection Dialogue select the first option Completely Randomised Design and then select columns *C1*, *C2* and *L3* respectively as [Data], [Dose] and [Preparation]. Click [Next] to proceed to Output Options Dialogue. Click [All] to perform all tests in one go and click [Finish]. The following output is obtained:

Parallel Line Method

Completely Randomised Design

Normality Tests

Smaller probabilities indicate non-normality.
* Lilliefors probability = 0.2 means 0.2 or greater.

Dose×Preparations	Valid Cases	Mean	Standard Deviation
0.25 × Standard S	10	332.0000	32.0416
0.25 × Preparation T	10	323.9000	26.9256
0.25 × Preparation U	10	282.2000	29.2339
1 × Standard S	10	248.4000	21.9960
1 × Preparation T	10	244.0000	26.8080
1 × Preparation U	10	250.0000	28.0119

Dose×Preparations	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	* Probability
0.25 × Standard S	0.9565	0.7451	0.1538	0.2000
0.25 × Preparation T	0.9471	0.6348	0.1429	0.2000
0.25 × Preparation U	0.8940	0.1878	0.2235	0.1639
1 × Standard S	0.9302	0.4494	0.2135	0.2000
1 × Preparation T	0.9475	0.6390	0.1446	0.2000
1 × Preparation U	0.9515	0.6864	0.1324	0.2000

Dose×Preparations	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
0.25 × Standard S	0.0331	0.7759	0.2218	0.7658
0.25 × Preparation T	0.0333	0.7721	0.2311	0.7326
0.25 × Preparation U	0.0895	0.1360	0.5030	0.1549
1 × Standard S	0.0579	0.3692	0.3494	0.3962
1 × Preparation T	0.0341	0.7582	0.2337	0.7232
1 × Preparation U	0.0278	0.8580	0.2055	0.8201

Homogeneity of Variance Tests

	Test Statistic	Probability	
Bartlett's Chi-square Test	1.2810	0.9369	
Bartlett-Box F Test	0.2575	0.9362	
Cochran's C (max var / sum var)	0.2235	1.0000	
Hartley's F (max var / min var)	2.1220	0.0500	`p > 0.05`
Levene's F Test	0.3738	0.8644	

Response Totals and Contrasts

Dose	Standard S	Preparation T	Preparation U	Total
0.25	3320.0000	3239.0000	2822.0000	
1	2484.0000	2440.0000	2500.0000	
Total	5804.0000	5679.0000	5322.0000	16805.0000
Linear Contrast	-836.0000	-799.0000	-322.0000	-1957.0000

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	4706800.417	1	4706800.417		
Preparations	6256.633	2	3128.317	4.086	0.0223
Linear Regression	63830.817	1	63830.817	83.377	0.0000
Non-parallelism	8218.233	2	4109.117	5.367	0.0075
Treatments	78305.683	5	15661.137		
Residual	41340.900	54	765.572		
Total	119646.583	59	2027.908		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	248.4000	-60.3047	13594.4000	0.7199
Preparation T	244.0000	-57.6357	12992.9000	0.7107
Preparation U	250.0000	-23.2274	14753.6000	0.2600

Common Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	257.5833	-47.0559	49559.1333	0.5629
Preparation T	251.3333			
Preparation U	233.4833			

Comparison of Slopes

Comparison	Difference	Standard Error	q Stat	Table q
Preparation U – Standard S	37.0773	12.6231	2.9372	2.2713
Preparation T – Standard S	2.6690	12.6231	0.2114	2.2713

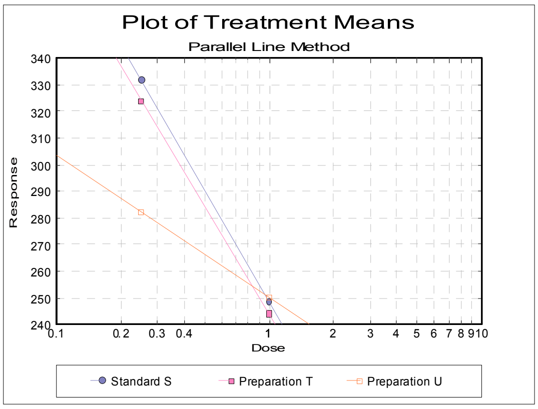
Comparison	Probability	Lower 95%	Upper 95%	Result
Preparation U – Standard S	0.0093	8.4061	65.7484	**
Preparation T – Standard S	0.9678	-26.0022	31.3401	

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	1.0000	1.1420	0.7836	1.6869
Preparation U	1.0000	1.6689	1.1481	2.5550

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0348	19.7070	68.6179
Preparation U	0.0377	8.1229	68.7960

G =	0.0482
C =	1.0507



Looking at the plot of treatment means we can see that *Preparation U* line is not parallel to *Standard S* and *Preparation T* lines. This can also be picked up from the non-parallelism test in Validity of Assay (0.0075), which is significant at 5% level. The Comparison of Slopes test also reports a significant difference between *Preparation U* and *Standard S* slopes.

This assay can still be useful by omitting *Preparation U* and performing the analysis for *Standard S* and *Preparation U*. In Excel Add-In Mode, you can simply select the block A1:C41 and repeat the analysis. In Stand-Alone Mode, you can define a Select Row column to omit these rows from the analysis, without actually deleting them from the spreadsheet. To do this, click somewhere on column 4, and select Data → Select Row option from UNISTAT's spreadsheet menus. The colour of *C4* will change. This indicates that all rows with a 0 entry in this column will be omitted from the subsequent analyses.

When the analysis is repeated without *Preparation U*, the following results are obtained:

Parallel Line Method

Rows 41-60 Omitted
Selected by C4 Select
Completely Randomised Design

Normality Tests

Smaller probabilities indicate non-normality.
* Lilliefors probability = 0.2 means 0.2 or greater.

Dose×Preparations	Valid Cases	Mean	Standard Deviation
0.25 × Standard S	10	332.0000	32.0416
0.25 × Preparation T	10	323.9000	26.9256
1 × Standard S	10	248.4000	21.9960
1 × Preparation T	10	244.0000	26.8080

Dose×Preparations	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	* Probability
0.25 × Standard S	0.9565	0.7451	0.1538	0.2000
0.25 × Preparation T	0.9471	0.6348	0.1429	0.2000
1 × Standard S	0.9302	0.4494	0.2135	0.2000
1 × Preparation T	0.9475	0.6390	0.1446	0.2000

Dose×Preparations	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
0.25 × Standard S	0.0331	0.7759	0.2218	0.7658
0.25 × Preparation T	0.0333	0.7721	0.2311	0.7326
1 × Standard S	0.0579	0.3692	0.3494	0.3962
1 × Preparation T	0.0341	0.7582	0.2337	0.7232

Homogeneity of Variance Tests

	Test Statistic	Probability	
Bartlett's Chi-square Test	1.1985	0.7534	
Bartlett-Box F Test	0.4029	0.7509	
Cochran's C (max var / sum var)	0.3475	0.6641	
Hartley's F (max var / min var)	2.1220	0.0500	p > 0.05
Levene's F Test	0.4381	0.7271	

Response Totals and Contrasts

Dose	Standard S	Preparation T	Total
0.25	3320.0000	3239.0000	
1	2484.0000	2440.0000	
Total	5804.0000	5679.0000	11483.0000
Linear Contrast	-836.0000	-799.0000	-1635.0000

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	3296482.225	1	3296482.225		
Preparations	390.625	1	390.625	0.529	0.4718
Linear Regression	66830.625	1	66830.625	90.491	0.0000
Non-parallelism	34.225	1	34.225	0.046	0.8308
Treatments	67255.475	3	22418.492		
Residual	26587.300	36	738.536		
Total	93842.775	39	2406.225		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	248.4000	-60.3047	13594.4000	0.7199
Preparation T	244.0000	-57.6357	12992.9000	0.7107

Common Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	249.3250	-58.9702	26621.5250	0.7151
Preparation T	243.0750			

Comparison of Slopes

Comparison	Difference	Standard Error	q Stat	Table q
Preparation T – Standard S	2.6690	12.3983	0.2153	2.0281

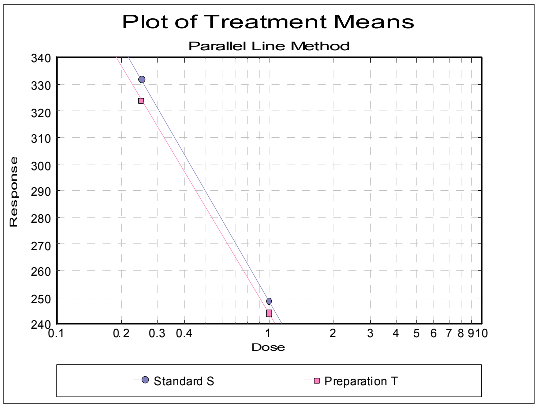
Comparison	Probability	Lower 95%	Upper 95%	Result
Preparation T – Standard S	0.8308	-22.4759	27.8138	

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	1.0000	1.1118	0.8250	1.5136

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0214	34.6983	74.2012

G =	0.0455
C =	1.0476



The estimated potency was calculated with the default assigned potency value of 1 for *Preparation U*.

In Stand-Alone Mode, do not forget to reset column 4, otherwise the Select Row function will be effective in subsequent procedures you run. To do this, click somewhere on column 4, and select Data → Select Row option again, or select Formula → Quick Formula from the menu and enter *data*. The colour of *C4* will change back to its original value.

10.1.4.2. Completely Randomised Design with 5 Doses and 4 Preparations

Data is given in Table 5.1.4-I. on p. 585 of *European Pharmacopoeia* (2008).

Open BIOPHARMA6 and select Bioassay → Parallel Line Method. From the Variable Selection Dialogue select the first option Completely Randomised Design and then select columns *C15*, *C16* and *L17* respectively as [Data], [Dose] and [Preparation]. Click [Next] to proceed to Output Options Dialogue. If you do not want to display all normality tests click on the [Opt] button situated to the left of Normality Tests option. Click [None] and then check the Anderson-Darling Test and Report summary statistics boxes. Then click [Back] and [Finish] to display the following output:

Parallel Line Method

Completely Randomised Design

Normality Tests

Smaller probabilities indicate non-normality.

Dose×Preparations	Valid Cases	Mean	Standard Deviation	Anderson-Darling Test	Probability
0.0625 × Standard S	3	-3.0745	0.0884	0.2663	0.3634
0.125 × Standard S	3	-2.3963	0.0960	0.2303	0.4841
0.25 × Standard S	3	-1.8351	0.0377	0.1976	0.5929
0.5 × Standard S	3	-1.1664	0.1318	0.3365	0.2031
1 × Standard S	3	-0.6352	0.0293	0.1941	0.6090
0.0625 × Preparation T	3	-2.3435	0.0181	0.4878	0.0565
0.125 × Preparation T	3	-1.7891	0.0628	0.1896	0.6303
0.25 × Preparation T	3	-1.0725	0.0417	0.2231	0.5077
0.5 × Preparation T	3	-0.5503	0.1416	0.1896	0.6304
1 × Preparation T	3	0.1691	0.1422	0.2501	0.4141
0.0625 × Preparation U	3	-2.5719	0.1036	0.3560	0.1719
0.125 × Preparation U	3	-2.0017	0.0710	0.2307	0.4827
0.25 × Preparation U	3	-1.3045	0.0181	0.3835	0.1353
0.5 × Preparation U	3	-0.6183	0.0912	0.2070	0.5544

Dose×Preparations	Valid Cases	Mean	Standard Deviation	Anderson-Darling Test	Probability
1 × Preparation U	3	-0.0480	0.0940	0.1910	0.6236
0.0625 × Preparation V	3	-2.4852	0.0275	0.4878	0.0565
0.125 × Preparation V	3	-1.8745	0.1040	0.4518	0.0768
0.25 × Preparation V	3	-1.1606	0.0206	0.3403	0.1967
0.5 × Preparation V	3	-0.5539	0.0713	0.4738	0.0637
1 × Preparation V	3	0.0468	0.0165	0.4238	0.0975

Homogeneity of Variance Tests

	Test Statistic	Probability
Bartlett's Chi-square Test	25.6778	0.1394
Bartlett-Box F Test	1.3733	0.1319
Cochran's C (max var / sum var)	0.1514	0.8834
Hartley's F (max var / min var)	74.4176	
Levene's F Test	2.1145	0.0230

Response Totals and Contrasts

Dose	Standard S	Preparation T	Preparation U	Preparation V	Total
0.0625	-9.2236				
0.125	-7.1888				
0.25	-5.5054				
0.5	-3.4992				
1	-1.9055				
0.0625		-7.0305			
0.125		-5.3672			
0.25		-3.2176			
0.5		-1.6508			
1		0.5072			
0.0625			-7.7158		
0.125			-6.0051		
0.25			-3.9135		
0.5			-1.8550		
1			-0.1439		
0.0625				-7.4555	
0.125				-5.6234	
0.25				-3.4819	
0.5				-1.6618	
1				0.1404	
Total	-27.3225	-16.7590	-19.6333	-18.0822	-81.7970

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	111.513	1	111.513		
Preparations	4.475	3	1.492	223.395	0.0000
Linear Regression	47.584	1	47.584	7125.912	0.0000
Non-parallelism	0.019	3	0.006	0.933	0.4339
Non-linearity	0.074	12	0.006	0.926	0.5307
Treatments	52.152	19	2.745		
Residual	0.267	40	0.007		
Total	52.419	59	0.888		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	-0.5998	0.8813	0.0904	0.9920
Preparation T	0.1355	0.9037	0.1208	0.9898
Preparation U	-0.0226	0.9278	0.0843	0.9933
Preparation V	0.0714	0.9211	0.0459	0.9963

Common Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	-0.5621	0.9085	0.3600	0.9925
Preparation T	0.1422			
Preparation U	-0.0495			
Preparation V	0.0539			

Comparison of Slopes

Comparison	Difference	Standard Error	q Stat	Table q
Preparation U - Standard S	0.0466	0.0304	1.5296	2.4415
Preparation V - Standard S	0.0398	0.0304	1.3074	2.4415
Preparation T - Standard S	0.0224	0.0304	0.7365	2.4415

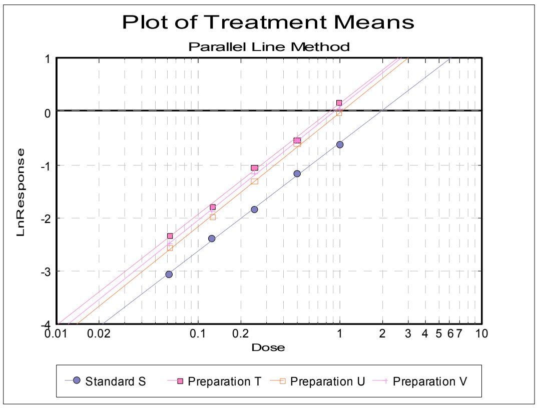
Comparison	Probability	Lower 95%	Upper 95%	Result
Preparation U - Standard S	0.3036	-0.0278	0.1209	
Preparation V - Standard S	0.4262	-0.0345	0.1141	
Preparation T - Standard S	0.8033	-0.0519	0.0967	

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	1.0000	2.1710	2.0272	2.3270
Preparation U	1.0000	1.7581	1.6435	1.8820
Preparation V	1.0000	1.9701	1.8406	2.1103

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0012	181.8564	93.3790
Preparation U	0.0011	287.1655	93.4785
Preparation V	0.0011	224.6942	93.4289

G =	0.0006
C =	1.0006



All samples have an assigned potency of 20 µg protein/ml. Next click on the [Last Procedure Dialogue] button on the Output Medium Toolbar. This will display the Output Options Dialogue again. Click the [Opt] button situated to the left of the Potency option, enter 20 for each test preparation and click [Finish].

Parallel Line Method

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	20.0000	43.4197	40.5448	46.5397
Preparation U	20.0000	35.1628	32.8697	37.6403
Preparation V	20.0000	39.4018	36.8126	42.2058

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0012	0.4546	93.3790
Preparation U	0.0011	0.7179	93.4785
Preparation V	0.0011	0.5617	93.4289

G =	0.0006
C =	1.0006

10.1.4.3. Randomised Block Design with 4 Doses and 2 Preparations

Data is given in Table 5.1.3.-I on p. 585 of *European Pharmacopoeia* (2008).

Open BIOPHARMA6 and select Bioassay → Parallel Line Method. From the Variable Selection Dialogue select the second option Randomised Block Design and the select columns *C10*, *C11*, *L12* and *C13* respectively as [Data], [Dose], [Preparation] and [Row Factor]. Click [Next] to proceed to the Output Options Dialogue.

Parallel Line Method

Randomised Block Design

Normality Tests

Smaller probabilities indicate non-normality.
* Lilliefors probability = 0.2 means 0.2 or greater.

Dose×Preparations	Valid Cases	Mean	Standard Deviation
1 × Standard S	5	246.6000	6.7305
1 × Preparation T	5	237.4000	6.4653
1.5 × Standard S	5	203.0000	6.1644
1.5 × Preparation T	5	195.4000	7.5033
2.25 × Standard S	5	162.4000	17.2714
2.25 × Preparation T	5	150.4000	5.5946
3.375 × Standard S	5	107.4000	5.7706
3.375 × Preparation T	5	104.4000	7.2319

Dose×Preparations	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	* Probability
1 × Standard S	0.7977	0.0777	0.3237	0.0942
1 × Preparation T	0.9171	0.5116	0.1982	0.2000
1.5 × Standard S	0.7607	0.0373	0.3418	0.0568
1.5 × Preparation T	0.9649	0.8416	0.2156	0.2000
2.25 × Standard S	0.9904	0.9809	0.1729	0.2000
2.25 × Preparation T	0.8523	0.2018	0.2660	0.2000
3.375 × Standard S	0.8977	0.3974	0.2724	0.2000
3.375 × Preparation T	0.9718	0.8866	0.2125	0.2000

Dose×Preparations	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
1 × Standard S	0.1018	0.0770	0.5636	0.0681
1 × Preparation T	0.0413	0.5834	0.2658	0.5150
1.5 × Standard S	0.1095	0.0592	0.6235	0.0447
1.5 × Preparation T	0.0381	0.6470	0.2243	0.6507
2.25 × Standard S	0.0238	0.8949	0.1660	0.8710
2.25 × Preparation T	0.0666	0.2535	0.4027	0.2095
3.375 × Standard S	0.0559	0.3616	0.3395	0.3235
3.375 × Preparation T	0.0355	0.6990	0.2191	0.6722

Homogeneity of Variance Tests

	Test Statistic	Probability	
Bartlett's Chi-square Test	9.7854	0.2011	
Bartlett-Box F Test	1.4146	0.1953	
Cochran's C (max var / sum var)	0.5000	0.0039	
Hartley's F (max var / min var)	9.5304	0.0500	p > 0.05
Levene's F Test	1.3017	0.2813	

Response Totals and Contrasts

Dose	Standard S	Preparation T	Total
1	1233.0000	1187.0000	
1.5	1015.0000	977.0000	
2.25	812.0000	752.0000	
3.375	537.0000	522.0000	
Total	3597.0000	3438.0000	7035.0000
Linear Contrast	-2291.0000	-2220.0000	-4511.0000
Quadratic Contrast	-57.0000	-20.0000	-77.0000
Cubic Contrast	-87.0000	10.0000	-77.0000

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	1237280.625	1	1237280.625		
Preparations	632.025	1	632.025	11.722	0.0019
Linear Regression	101745.605	1	101745.605	1887.111	0.0000
Non-parallelism	25.205	1	25.205	0.467	0.4998
Non-linearity	259.140	4	64.785	1.202	0.3321
Quadratic Regression	148.225	1	148.225	2.749	0.1085
Quadratic Difference	34.225	1	34.225	0.635	0.4323
Residual	76.690	2	38.345		
Treatments	102661.975	7	14665.996		
Blocks(Rows)	876.750	4	219.188	4.065	0.0101
Residual	1509.650	28	53.916		
Total	105048.375	39	2693.548		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	248.5800	-113.0060	1897.7400	0.9651
Preparation T	238.5000	-109.5039	747.8000	0.9851

Common Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	247.5150	-111.2549	2670.7450	0.9744
Preparation T	239.5650			

Comparison of Slopes

Comparison	Difference	Standard Error	q Stat	Table q
Preparation T - Standard S	3.5022	5.1221	0.6837	2.0484

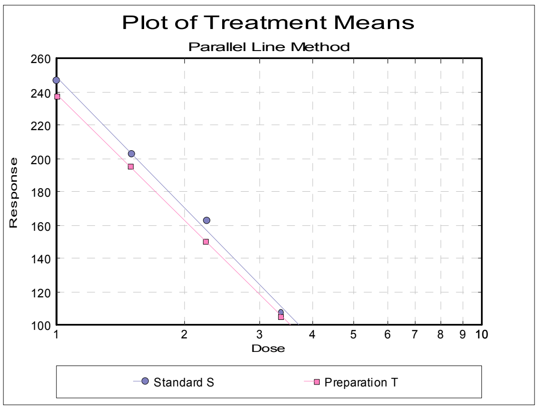
Comparison	Probability	Lower 95%	Upper 95%	Result
Preparation T - Standard S	0.4998	-6.9901	13.9944	

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	1.0000	1.0741	1.0291	1.1214

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0004	1971.4511	95.8129

G =	0.0022
C =	1.0022



The assigned potency for the test preparation is 20,000 IU/vial and we also need to apply a correction factor of 0.89512 because dilutions were not exactly equipotent on the basis of the assigned potency. Next click on the [Last Procedure Dialogue] button on the Output Medium Toolbar. This will display the Output Options Dialogue again. Click the [Opt] button situated to the left of the Potency option, enter 20000 * 0.89512 and click [Finish]. You can also enter the complete expression as:

20000 * (670*16.7/25)/(20000*1/40)

as UNISTAT numeric input boxes now accept formulas.

Parallel Line Method

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	17902.4000	19228.4755	18423.3508	20075.1776

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0004	0.0000	95.8129

G =	0.0022
C =	1.0022

10.1.4.4. Latin Squares Design with 3 Doses and 2 Preparations

Data is given in Table 5.1.2.-II on p. 584 of *European Pharmacopoeia* (2008).

The entry and transformation of this data set is more complicated than the two previous examples. In order to assign the correct dose levels and preparation groups, information given in Table 5.1.2.-I is essential. Ensure that the way the factor columns are created is understood well.

Open BIOPHARMA6 and select Bioassay → Parallel Line Method. From the Variable Selection Dialogue select the third option Latin Squares Design and then select columns *C5*, *C6*, *L7*, *C8* and *C9* respectively as [Data], [Dose], [Preparation], [Row Factor] and [Column Factor]. Click [Next] to proceed to Output Options Dialogue.

Parallel Line Method

Latin Squares Design

Normality Tests

Smaller probabilities indicate non-normality.
* Lilliefors probability = 0.2 means 0.2 or greater.

Dose×Preparations	Valid Cases	Mean	Standard Deviation
1 × Standard S	6	158.6667	6.5929
1 × Preparation T	6	156.1667	4.7081
1.5 × Standard S	6	176.5000	5.3572
1.5 × Preparation T	6	174.6667	8.6641
2.25 × Standard S	6	194.5000	4.8477
2.25 × Preparation T	6	195.5000	4.0373

Dose×Preparations	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	* Probability
1 × Standard S	0.8581	0.1827	0.3050	0.0851
1 × Preparation T	0.8118	0.0749	0.2922	0.1177
1.5 × Standard S	0.8965	0.3536	0.2038	0.2000
1.5 × Preparation T	0.9757	0.9284	0.1639	0.2000
2.25 × Standard S	0.9879	0.9835	0.1364	0.2000
2.25 × Preparation T	0.8255	0.0984	0.3115	0.0703

Dose×Preparations	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
1 × Standard S	0.0849	0.1450	0.4756	0.1438
1 × Preparation T	0.0886	0.1276	0.5470	0.0901
1.5 × Standard S	0.0544	0.3902	0.3341	0.3689

Dose×Preparations	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
1.5 × Preparation T	0.0274	0.8514	0.1760	0.8637
2.25 × Standard S	0.0210	0.9388	0.1514	0.9165
2.25 × Preparation T	0.1046	0.0740	0.5613	0.0818

Homogeneity of Variance Tests

	Test Statistic	Probability	
Bartlett's Chi-square Test	3.7817	0.5813	
Bartlett-Box F Test	0.7637	0.5760	
Cochran's C (max var / sum var)	0.3588	0.2313	
Hartley's F (max var / min var)	4.6053	0.0500	$p > 0.05$
Levene's F Test	1.5818	0.1954	

Response Totals and Contrasts

Dose	Standard S	Preparation T	Total
1	952.0000	937.0000	
1.5	1059.0000	1048.0000	
2.25	1167.0000	1173.0000	
Total	3178.0000	3158.0000	6336.0000
Linear Contrast	215.0000	236.0000	451.0000
Quadratic Contrast	1.0000	14.0000	15.0000

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	1115136.000	1	1115136.000		
Preparations	11.111	1	11.111	0.535	0.4730
Linear Regression	8475.042	1	8475.042	408.108	0.0000
Non-parallelism	18.375	1	18.375	0.885	0.3581
Non-linearity	5.472	2	2.736	0.132	0.8773
Quadratic Regression	3.125	1	3.125	0.150	0.7022
Quadratic Difference	2.347	1	2.347	0.113	0.7402
Treatments	8510.000	5	1702.000		
Blocks(Rows)	412.000	5	82.400	3.968	0.0116
Blocks(Columns)	218.667	5	43.733	2.106	0.1069
Residual	415.333	20	20.767		
Total	9556.000	35	273.029		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	158.6389	44.1879	478.3611	0.8895
Preparation T	155.7778	48.5040	573.1111	0.8901

Common Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	157.7639	46.3460	1069.8472	0.8879
Preparation T	156.6528			

Comparison of Slopes

Comparison	Difference	Standard Error	q Stat	Table q
Preparation T - Standard S	4.3160	4.5883	0.9407	2.0860

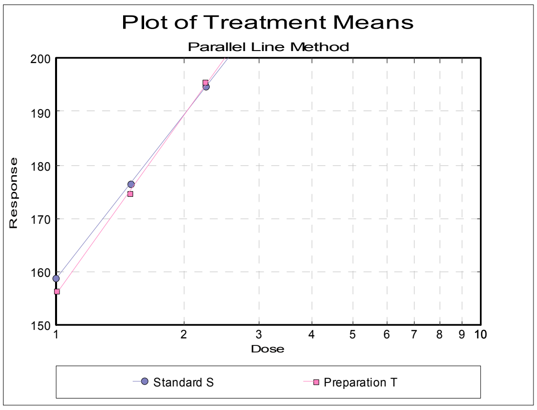
Comparison	Probability	Lower 95%	Upper 95%	Result
Preparation T - Standard S	0.3581	-5.2551	13.8871	

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	1.0000	0.9763	0.9112	1.0456

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0011	963.9008	93.3289

G =	0.0107
C =	1.0108



The assigned potency for the test preparation is 5600 IU/mg and we also need to apply a correction factor of 0.99799 because dilutions were not exactly equipotent on the basis of the assigned potency. Next click on the [Last Procedure Dialogue] button on the Output Medium Toolbar. This will display the Output Options Dialogue again. Click the [Opt] button situated to the left of the Potency option, enter 5600 * 0.99799 and click [Finish]. You can also enter the complete expression as:

5600 * (4855*25.2/24.5)/(5600*21.4/23.95)

as UNISTAT numeric input boxes now accept formulas.

Parallel Line Method

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	5588.7440	5456.3512	5092.3546	5843.3456

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0011	0.0000	93.3289

G =	0.0107
C =	1.0108

10.1.4.5. Twin Crossover Design

Data is given in Table 5.1.5-II. on p. 586 of *European Pharmacopoeia* (2008).

Open BIOPHARMA6 and select Bioassay → Parallel Line Method. From the Variable Selection Dialogue select the fourth option Crossover Design and select columns *C18*, *C19*, *L20*, *C21* and *C22* respectively as [Data], [Dose], [Preparation], [Row Factor] and [Column Factor]. Click [Next] to proceed to Output Options Dialogue. Click the [Opt] button situated to the left of the Potency option. Enter 40 as the assigned potency for the unknown. Click [Back] to get back to output options, click [All] to perform all tests in one go and then click [Finish]. The following output is obtained:

Parallel Line Method

Crossover Design

Normality Tests

Smaller probabilities indicate non-normality.
* Lilliefors probability = 0.2 means 0.2 or greater.

Dose×Preparations	Valid Cases	Mean	Standard Deviation
1 × Standard	16	101.1875	30.5160
1 × Test	16	91.5625	26.4448
2 × Standard	16	68.1250	18.8428
2 × Test	16	77.5625	29.8060

Dose×Preparations	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	* Probability
1 × Standard	0.9507	0.5009	0.1380	0.2000
1 × Test	0.9229	0.1876	0.1453	0.2000
2 × Standard	0.9112	0.1216	0.1901	0.1220
2 × Test	0.9278	0.2253	0.1260	0.2000

Dose×Preparations	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
1 × Standard	0.0470	0.5319	0.2966	0.5482
1 × Test	0.0542	0.4284	0.4079	0.3070
2 × Standard	0.0806	0.1890	0.4974	0.1809
2 × Test	0.0636	0.3197	0.4091	0.3049

Homogeneity of Variance Tests

	Test Statistic	Probability
Bartlett's Chi-square Test	3.7999	0.2839
Bartlett-Box F Test	1.2736	0.2815
Cochran's C (max var / sum var)	0.3240	0.6856
Hartley's F (max var / min var)	2.6228	
Levene's F Test	2.1683	0.1011

Response Totals and Contrasts

Dose	Standard	Test	Total
Days: 1			
1	765.0000	719.0000	
2	557.0000	579.0000	
Total	1322.0000	1298.0000	2620.0000
Days: 2			
1	854.0000	746.0000	
2	533.0000	662.0000	
Total	1387.0000	1408.0000	2795.0000
Preparations			
Total	2709.0000	2706.0000	5415.0000
Linear Contrast			
Days: 1	-208.0000	-140.0000	-348.0000
Days: 2	-321.0000	-84.0000	-405.0000
Total	-529.0000	-224.0000	-753.0000

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	458159.7656	1	458159.7656		
Non-parallelism	1453.5156	1	1453.5156	1.0638	0.3112
Days×Preparations	31.6406	1	31.6406	0.0232	0.8801
Days×Linear Regression	50.7656	1	50.7656	0.0372	0.8485
Error Between	38258.8125	28	1366.3862		
Blocks(Rows)	39794.7344	31	1283.7011		
Preparations	0.1406	1	0.1406	0.0010	0.9747
Linear Regression	8859.5156	1	8859.5156	64.5324	0.0000
Days	478.5156	1	478.5156	3.4855	0.0724
Days×Non-parallelism	446.2656	1	446.2656	3.2506	0.0822
Error Within	3844.0625	28	137.2879		
Total	53423.2344	63	847.9878		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard	101.1875	-47.6991	19294.1875	0.3119
Test	91.5625	-20.1977	23815.8750	0.0618

Common Regression

	Intercept	Slope	Residual SS	R-squared
Standard	96.4219	-33.9484	44563.5781	0.1658
Test	96.3281			

Comparison of Slopes

Comparison	Difference	Standard Error	q Stat	Table q
Test - Standard	27.5014	8.4520	3.2538	2.0484

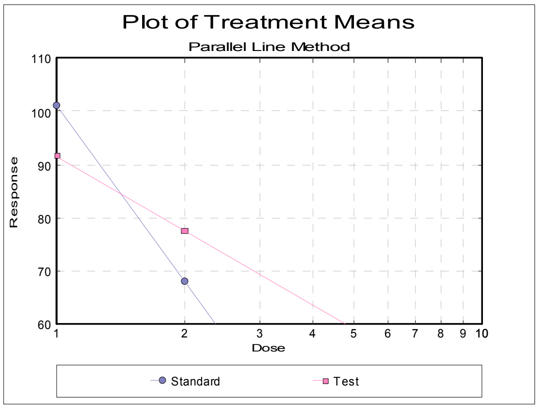
Comparison	Probability	Lower 95%	Upper 95%	Result
Test - Standard	0.0030	10.1882	44.8146	**

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Test	40.0000	40.1106	33.4162	48.1646

Test Preparation	Variance	Weight	% Precision
Test	0.0074	0.0772	83.3102

G =	0.0650
C =	1.0695



Although the plot of treatment means and the Comparison of Slopes test seem to indicate deviation from parallelism, the non-parallelism test in Validity of Assay (0.3112) is not significant at 5% level.

10.1.4.6. Triple Crossover Design

Table 10.3.1. on p. 205 from Finney, D. J. (1978) is an example with three dose levels and two preparations.

Open BIOFINNEY and select Bioassay → Parallel Line Method. From the Variable Selection Dialogue select the fourth option Crossover Design and select columns *C15*, *C16*, *S17*, *C18* and *C19* respectively as [Data], [Dose], [Preparation], [Row Factor] and [Column Factor]. Click [Next] to proceed to Output Options Dialogue. Click on the [Opt] button situated to the left of Normality Tests option and check the Shapiro-Wilk Test and Report summary statistics boxes. Then click [Back] and [Finish].

Parallel Line Method

Crossover Design

Normality Tests

Smaller probabilities indicate non-normality.

Dose×Preparations	Valid Cases	Mean	Standard Deviation	Shapiro-Wilk Test	Probability
0.5 × Standard	10	3.7120	0.1946	0.9485	0.6508
1.25 × Test	10	4.0700	0.3391	0.9285	0.4330
1 × Standard	10	3.2550	0.7950	0.9026	0.2341
2.5 × Test	10	3.4630	0.5483	0.9046	0.2459
2 × Standard	10	3.3970	0.4072	0.9038	0.2408
5 × Test	10	3.2780	0.3720	0.9058	0.2532

Homogeneity of Variance Tests

	Test Statistic	Probability	
Bartlett's Chi-square Test	18.0814	0.0028	
Bartlett-Box F Test	3.6473	0.0027	
Cochran's C (max var / sum var)	0.4548	0.0035	
Hartley's F (max var / min var)	16.6848	0.0100	p < 0.01
Levene's F Test	6.6315	0.0001	

Response Totals and Contrasts

Dose	Standard	Test	Total
Days: 1			
1	18.7200	19.8100	
2	14.6900	19.0100	
3	17.0300	16.1400	
Total	50.4400	54.9600	105.4000
Days: 2			
1	18.4000	20.8900	
2	17.8600	15.6200	
3	16.9400	16.6400	
Total	53.2000	53.1500	106.3500
Preparations			
Total	103.6400	108.1100	211.7500
Linear Contrast			
Days: 1	-1.6900	-3.6700	-5.3600
Days: 2	-1.4600	-4.2500	-5.7100
Total	-3.1500	-7.9200	-11.0700
Quadratic Contrast			
Days: 1	6.3700	-2.0700	4.3000
Days: 2	-0.3800	6.2900	5.9100
Total	5.9900	4.2200	10.2100

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	747.3010	1	747.3010		
Non-parallelism	0.5688	1	0.5688	1.6032	0.2176
Quadratic Regression	0.8687	1	0.8687	2.4484	0.1307
Days×Preparations	0.3481	1	0.3481	0.9810	0.3318
Days×Linear Regression	0.0031	1	0.0031	0.0086	0.9267
Days×Quadratic Difference	1.9026	1	1.9026	5.3624	0.0294
Error Between	8.5153	24	0.3548		
Blocks(Rows)	12.2066	29	0.4209		
Preparations	0.3330	1	0.3330	4.7403	0.0395
Linear Regression	3.0636	1	3.0636	43.6087	0.0000
Days	0.0150	1	0.0150	0.2141	0.6477
Quadratic Difference	0.0261	1	0.0261	0.3716	0.5478
Days×Non-parallelism	0.0164	1	0.0164	0.2335	0.6333
Days×Quadratic Regression	0.0216	1	0.0216	0.3075	0.5844
Error Within	1.6861	24	0.0703		
Total	17.3685	59	0.2944		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard	3.4547	-0.2272	8.1200	0.0576
Test	4.1272	-0.5713	5.2830	0.3725

Common Regression

	Intercept	Slope	Residual SS	R-squared
Standard	3.4547	-0.3993	13.9718	0.1798
Test	3.9695			

Comparison of Slopes

Comparison	Difference	Standard Error	q Stat	Table q
Test – Standard	-0.3441	0.1209	2.8455	2.0639

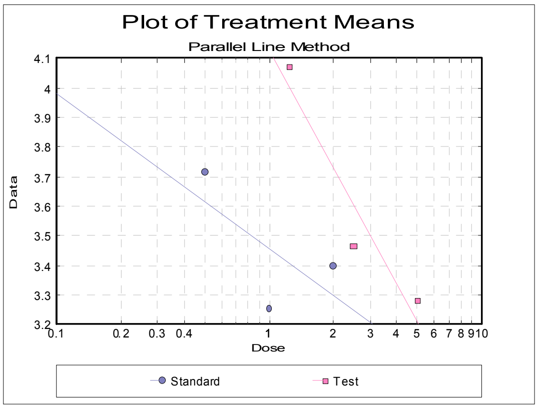
Comparison	Probability	Lower 95%	Upper 95%	Result
Test – Standard	0.0089	-0.5937	-0.0945	**

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Test	1.0000	0.2754	0.1783	0.3923

Test Preparation	Variance	Weight	% Precision
Test	0.0326	372.0902	64.7516

G =	0.0977
C =	1.1083



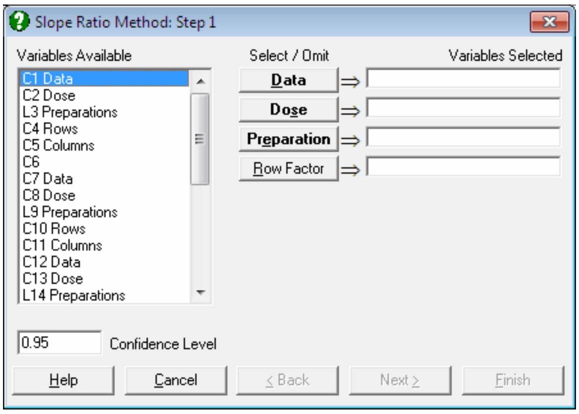
10.2. Slope Ratio Method

This is a general purpose procedure that can be used to analyse balanced or unbalanced assays with blanks (0-dose treatments), plate (row) effects and unlimited numbers of dose levels and test preparations. The algorithm is based on Finney (1978). The Slope Ratio Method specification given in *European Pharmacopoeia* (1997-2008), is a restricted special case of this procedure.

10.2.1. Slope Ratio Variable Selection

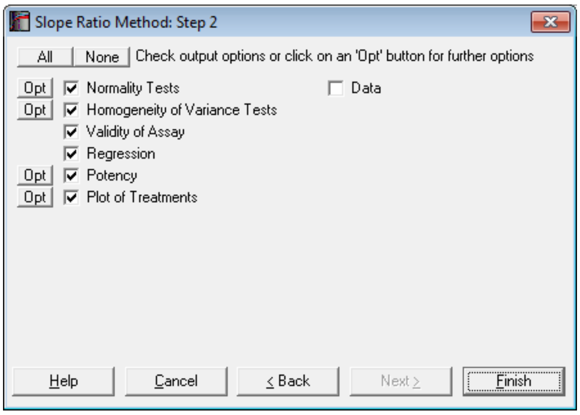
The data format is as in Parallel Line Method (see 10.1.1. Data Preparation). Measurement data is stacked in a single column, a second column contains the dose level for each measurement and another categorical column indicates which preparation a particular measurement belongs to. An optional row factor can be entered to keep track of the replicates.

Designs can be unbalanced, i.e. the number of replicates for each dose-preparation combination may be different, dose levels for standard and test preparations may be different, there can be more than one test preparation, but the first preparation should always be the standard. It is compulsory to select at least three columns [Data], [Dose] and [Preparation]. The optional [Row Factor] column is usually used to isolate a plate effect (the replicates) and when one is selected, the program assumes that all dose/treatment groups (or cells) have an equal number of replicates.



0.95 Confidence Level

10.2.2. Slope Ratio Output Options



Check output options or click on an 'Opt' button for further options

☑ Normality Tests
☑ Homogeneity of Variance Tests
☑ Validity of Assay
☑ Regression
☑ Potency
☑ Plot of Treatments
□ Data

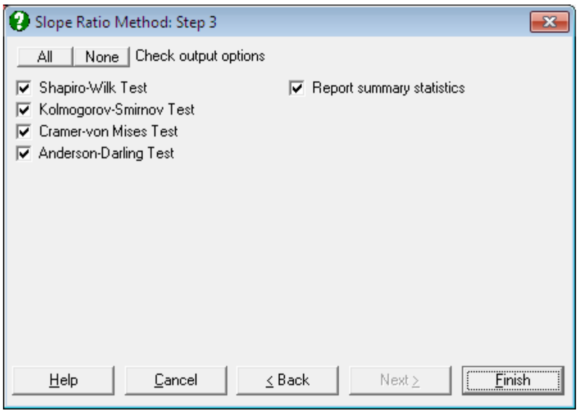
Let X_{ijk} and Y_{ijk} be the dose and response values for the ith preparation (i = Blank, Standard, Test 1, …, Test n - 1) and the jth dose level of the kth replicate. First, all dose readings are transformed as:

$$X_{ijk} = \mathrm{Ln}(X_{ijk})$$

where the logarithm is natural (e-based).

10.2.2.1. Normality Tests

As in Parallel Line Method, you can select to display all or any of the four most commonly used normality tests; Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling (see 10.1.3.1. Normality Tests for Bioassays).

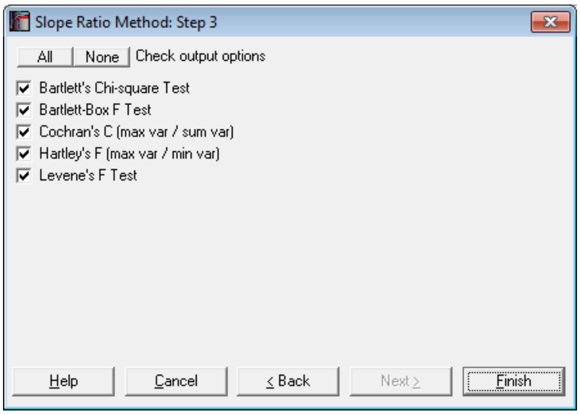


Check output options

☑ Shapiro-Wilk Test
☑ Kolmogorov-Smirnov Test
☑ Cramer-von Mises Test
☑ Anderson-Darling Test
☑ Report summary statistics

10.2.2.2. Homogeneity of Variance Tests

Five alternative homogeneity of variance tests are performed for unique dose-preparation (treatment) groups as described in section 10.1.3.2. Homogeneity of Variance Tests.



☑ Bartlett's Chi-square Test
☑ Bartlett-Box F Test
☑ Cochran's C (max var / sum var)
☑ Hartley's F (max var / min var)
☑ Levene's F Test

10.2.2.3. Validity of Assay

This output option displays an Analysis of Variance (ANOVA) table, which is used in testing the Validity of Assay. The standard significance tests performed are (i) regression, (ii) intercept and (iii) non-linearity. The overall non-linearity test is also broken down to individual tests for each preparation. If blanks (entries with a 0 dose level) exist, there will be an additional term for them. If a [Row Factor] was selected it will appear in the table as a main effect.

Define a *cell* as a unique combination of dose levels and preparations. For each cell calculate:

$$\overline{Y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\overline{X}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$$

$$\overline{X}_{ij}^2 = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}^2$$

Define the overall mean as:

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

where N is the total number of observations. Also define $\hat{a}_i$ and $\hat{b}_i$ as the intercept and slope for each preparation from Separate Regression and $\hat{\hat{b}}_i$ as the slope for each preparation from Common Regression (see 10.2.2.4. Regression).

The following definitions are used in calculating the blanks effect:

$$c_i = \frac{1}{n_i} + \frac{\overline{X}_i^2}{Sxx_i}$$

where Sxx_i is as defined in Separate Regression and:

$$c = \frac{1}{\sum_{i=1}^{n} \frac{1}{c_i}}$$

$$q = \sum_{i=1}^{n} \frac{\hat{a}_i}{c_i}$$

Also define the number of unique dose-preparation combinations excluding blanks as:

$$D = \sum_{i=1}^{n} \sum_{j=1}^{n_i} n_{ij}$$

The ANOVA table is then constructed as follows.

Due to	**Degrees of Freedom**		**Sum of Squares**
Plate	K – 1	SSP	$\sum_{k=1}^{K} n_k(\overline{Y}_k - \overline{Y})^2$
Between Doses	D – 1 + B	SSD	$\sum_{i=1}^{n}\sum_{j=1}^{n_i} n_{ij}(\overline{Y}_{ij} - \overline{Y})^2$
Blanks	B = 0 or 1	SSB	$\frac{(\overline{Y}_B - cq)^2}{1/n_B + c}$
Regression	n	SSR	$\sum_{i=1}^{n} \hat{b}_i \sum_{j=1}^{n_i} n_{ij}\overline{X}_{ij}(\overline{Y}_{ij} - \overline{Y})$
Intercept	n - 1	SSD - SSB - SSR - SSL	
Non-linearity	D – 2n	SSL	$\sum_{i=1}^{n} SSL_i$
Non-linearity for Preparation$_i$	D / n – 2	SSL$_i$	$\sum_{j=1}^{n_i} n_{ij}(\overline{Y}_{ij} - \hat{a}_i - \hat{b}_i\overline{X}_{ij})^2$
Residual	N – D – B – (K – 1)	SSE – SSP	$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij})^2$ - SSP
Total	N - 1	SST	$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y})^2$

10.2.2.4. Regression

Calculate for i = S, T_1, …, Tn - 1:

$$\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} X_{ijk}$$

$$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{jk}} Y_{ijk}$$

$$Sxx_i = \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \left(X_{ijk} - \overline{X}_i\right)^2$$

$$Syy_i = \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \overline{Y}_i\right)^2$$

$$Sxy_i = \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \left(X_{ijk} - \overline{X}_i\right)\left(Y_{ijk} - \overline{Y}_i\right)$$

The estimated parameters of the line of best fit for each preparation (i = S, T_1, …, Tn - 1) are:

Slope: $\hat{b}_i = Sxy_i / Sxx$

R-squared: $R_i^2 = \left(Sxy_i\right)^2 / \left(Sxx \times Syy_i\right)$

Residual sum of squares: $RSS_i = Syy_i\left(1 - R_i^2\right)$

Standard error of slope: RSS_i / Sxx

This information is displayed in the Separate Regression table and used in drawing the best fit lines in Plot of Treatments.

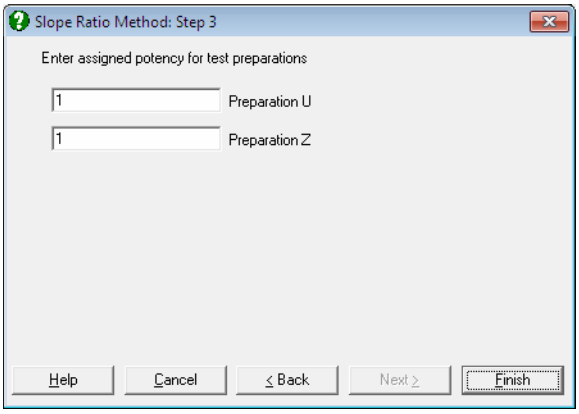
The Common Regression is obtained from a multivariate regression run, after transforming the data into the following form first.

	Dependent Variable	Independent Variables		
		Standard	Test 1	Test n
Blank	Y_{0jk}	0	0	0
Replicates	…	…	…	…
Standard	Y_{Sjk}	X_{Sjk}	0	0
Replicates	…	…	…	…
Test 1	Y_{1jk}	0	X_{1jk}	0
Replicates	…	…	…	…
Test n	Y_{njk}	0	0	X_{njk}
Replicates	…	…	…	…

The estimated parameters are displayed in Common Regression table.

10.2.2.5. Potency

By default, each test preparation is assigned a potency of unity. If you want to change this click the [Opt] button situated to the left of the Potency option. In this case, a further dialogue pops up asking for entry of assigned potency for each test preparation.



Slope Ratio Method: Step 3

Enter assigned potency for test preparations

1 Preparation U

1 Preparation Z

For each test preparation, the potency ratio is calculated as follows:

$$M_i = \frac{\hat{b}_i}{\hat{b}_S}, i = T_1, \ldots, T_{n-1}$$

For confidence intervals of M first define Vss, Vii, Vsi, i = T_1, …, T_{n-1} as the values corresponding to elements of $(X'X)^{-1}$ matrix from the Common Regression run. First define:

$$g = \frac{s^2 t^2_{\alpha,df} Vss}{\hat{b}^2_S}$$

where s^2 is the residual mean squares and $t_{\alpha,df}$ is the critical value from the t-distribution with degrees of freedom of the overall residual term from the ANOVA table. Note that if divided by s^2, Vss, Vii, Vsi give the variance / covariance matrix of the Common Regression coefficients.

Then the confidence interval for potency ratio of each test preparation is calculated using *Fieller's Theorem* (see Finney 1978, p. 156):

$$M_{iL}, M_{iU} = \left(M_i - \frac{gVsi}{Vss} \pm \frac{t_{\alpha,df} s}{\hat{b}_S} \sqrt{G_i} \right) / (1 - g)$$

where the variance of M$_i$ is:

$$G_i = Vii - 2M_i Vsi + M_i^2 Vss - g(Vii - Vsi^2 / Vss)$$

and the approximate variance of M$_i$ is (when g is negligible):

$$V_i = \frac{s^2}{\hat{b}^2_S} \left[Vii - 2M_i Vsi + M_i^2 Vss \right]$$

M$_i$ is the relative potency and M$_{iL}$ and M$_{iU}$ are the confidence limits for the relative potency. The estimated potency and its confidence interval are obtained by multiplying these relative values by the assigned potency supplied by the user for each test preparation separately.

Weights are computed after the estimated potency and its confidence interval are found:

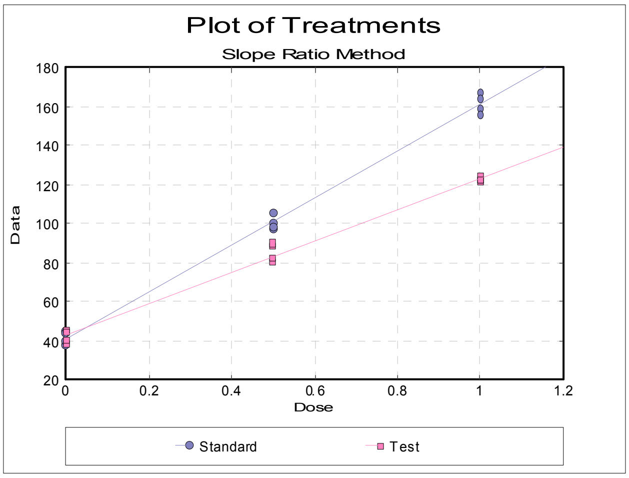
$$W_i = \frac{t^2_{\alpha,df}}{(M_{iU} - M_{iL})^2}$$

and % Precision is:

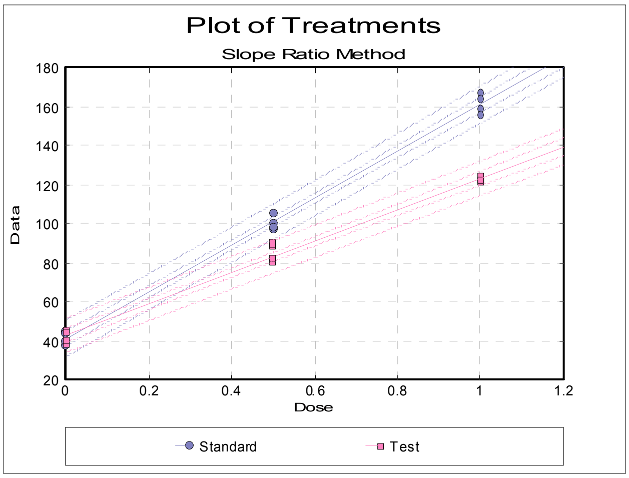
$$P_i = 100 \frac{M_{iL}}{M_i}$$

10.2.2.6. Plot of Treatments

This option generates a Plot of Treatments against dose levels. Standard and each test preparation are plotted in separate series and a line of best fit is drawn for each one of them. The coefficients of lines are as in Separate Regression output.



Clicking the [Opt] button situated to the left of the Plot of Treatments option will place the graph in UNISTAT's Graphics Editor. The plot can be further customised and annotated using the tools available under the UNISTAT Graphics Window's Edit menu.



The same plot is drawn here using the X-Y Plots procedure, this time with confidence intervals for regression lines included.

10.2.3. Slope Ratio Examples

Example 1

Data is given in Table 5.2.1-I on p. 588 of *European Pharmacopoeia* (2008). The data is rearranged as described in section 10.1.1. Data Preparation and saved in columns 24-26 of BIOPHARMA6.

Although the data set contains blanks (0 dose treatments), they need to be removed from the analysis. In Excel Add-In Mode, you can simply select the block X10:Z57. In Stand-Alone Mode, you can define *C26* as a Select Row column to omit these rows from the analysis, without actually deleting them from the spreadsheet. To do this, click somewhere on column 26, and select Data → Select Row option from UNISTAT's spreadsheet menus. The colour of *C26* will change. This indicates that all rows with a 0 entry in this column will be omitted from subsequent analyses.

Select Bioassay → Slope Ratio Method. In Stand-Alone Mode select columns *C23*, *C24* and *L25* respectively as [Data], [Dose] and [Preparation] from the Variable Selection Dialogue. In Excel Add-In Mode, you will need to select the three highlighted columns in the same order. Click [Next] to proceed to Output Options Dialogue. If you do not want to display all normality tests click on the [Opt] button situated to the left of Normality Tests option. Click [None] and then check the Shapiro-Wilk Test and Report summary statistics boxes. Then click [Back] and [Finish].

In Stand-Alone Mode, do not forget to reset column 4 after you finish this example, otherwise the Select Row function will be effective in subsequent procedures you run. To do this, click somewhere on column 4, and select Data → Select Row option again, or select Formula → Quick Formula from the menu and enter *data*. The colour of *C26* will change back to its original value.

Slope Ratio Method

Rows 1-8 Omitted
Selected by C26 Select

Normality Tests

Smaller probabilities indicate non-normality.
* Lilliefors probability = 0.2 means 0.2 or greater.

Dose×Preparations	Valid Cases	Mean	Standard Deviation	Shapiro-Wilk Test	Probability
1 × Standard S	8	0.1351	0.0025	0.8969	0.2707
2 × Standard S	8	0.2176	0.0021	0.8816	0.1952
3 × Standard S	8	0.2996	0.0027	0.8269	0.0551
1 × Preparation T	8	0.1200	0.0011	0.8599	0.1199
2 × Preparation T	8	0.1898	0.0012	0.8042	0.0318
3 × Preparation T	8	0.2554	0.0018	0.9255	0.4763

Homogeneity of Variance Tests

	Test Statistic	Probability	
Bartlett's Chi-square Test	8.5820	0.1269	p > 0.05
Bartlett-Box F Test	1.7315	0.1239	
Cochran's C (max var / sum var)	0.3079	0.3345	
Hartley's F (max var / min var)	6.2344	0.0500	
Levene's F Test	2.2830	0.0635	

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	1.976	1	1.976		
Regression	0.192	2	0.096	24849.565	0.0000
Intercept	0.000	1	0.000	0.001	0.9780
Non-linearity	0.000	2	0.000	2.984	0.0614
Standard S Non-linearity	0.000	1	0.000	0.086	0.7702
Preparation T Non-linearity	0.000	1	0.000	5.882	0.0197
Treatments	0.192	5	0.038		
Residual	0.000	42	0.000		
Total	0.192	47	0.004		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	0.0530	0.0822	0.0001	0.9989
Preparation T	0.0530	0.0677	0.0001	0.9992

Common Regression

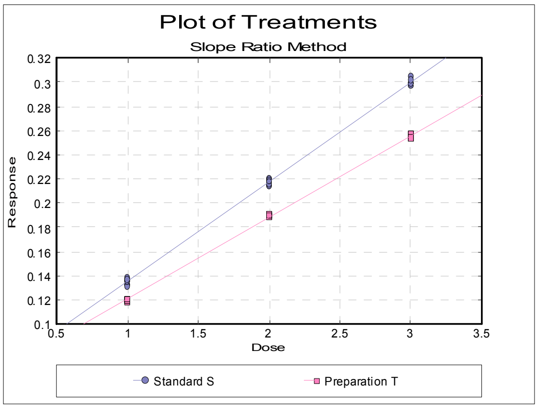
	Intercept	Slope	Residual SS	R-squared
Standard S	0.0530	0.0822	0.0002	0.9990
Preparation T		0.0677		

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	1.0000	0.8231	0.8171	0.8292

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0000	110860.3771	99.2644

G =	0.0001
C =	1.0001



Example 2

Data is given in Table 5.2.2-I on p. 589 of European Pharmacopoeia (2008).

Open BIOPHARMA6 and select Bioassay → Slope Ratio Method. The blank preparation is already omitted from this data set. From the Variable Selection Dialogue select columns *C27*, *C28* and *L29 Preparations* respectively as [Data],

[Dose] and [Preparation]. Click [Next] to proceed to Output Options Dialogue. Click on the [Opt] button situated to the left of Normality Tests option, click [None] and then check the Cramer-von Mises Test and Report summary statistics boxes and click [Back]. Click the [Opt] button situated to the left of the Potency option. Enter the assigned potency value 15 for both preparations, click [Back] and [Finish].

Slope Ratio Method

Normality Tests

Smaller probabilities indicate non-normality.

Dose×Preparations	Valid Cases	Mean	Standard Deviation	Cramer-von Mises Test	Probability
1 × Standard S	2	18.0000	0.0000	*	*
2 × Standard S	2	23.6500	1.2021	0.0419	0.4774
3 × Standard S	2	30.4000	0.0000	*	*
4 × Standard S	2	36.1500	0.6364	0.0419	0.4774
1 × Preparation T	2	15.9500	1.2021	0.0419	0.4774
2 × Preparation T	2	23.6500	0.7778	0.0419	0.4774
3 × Preparation T	2	28.1500	1.0607	0.0419	0.4774
4 × Preparation T	2	36.1000	2.4042	0.0419	0.4774
1 × Preparation U	2	15.5500	0.2121	0.0419	0.4774
2 × Preparation U	2	19.4000	1.1314	0.0419	0.4774
3 × Preparation U	2	23.6500	0.7778	0.0419	0.4774
4 × Preparation U	2	27.2000	0.2828	0.0419	0.4774

Homogeneity of Variance Tests

	Test Statistic	Probability
Bartlett's Chi-square Test	5.2396	0.8129
Bartlett-Box F Test	0.5751	0.8146
Cochran's C (max var / sum var)	0.4510	0.2364
Hartley's F (max var / min var)	128.4444	
Levene's F Test		

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	14785.770	1	14785.770		
Regression	1087.665	3	362.555	339.498	0.0000
Intercept	3.474	2	1.737	1.626	0.2371
Non-linearity	5.065	6	0.844	0.791	0.5943
Standard S Non-linearity	0.446	2	0.223	0.209	0.8144
Preparation T Non-linearity	4.453	2	2.227	2.085	0.1670
Preparation U Non-linearity	0.166	2	0.083	0.078	0.9257
Treatments	1096.205	11	99.655		
Residual	12.815	12	1.068		
Total	1109.020	23	48.218		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard S	11.7500	6.1200	2.2960	0.9939
Preparation T	9.7250	6.4950	13.4085	0.9692
Preparation U	11.6500	3.9200	2.1760	0.9860

Common Regression

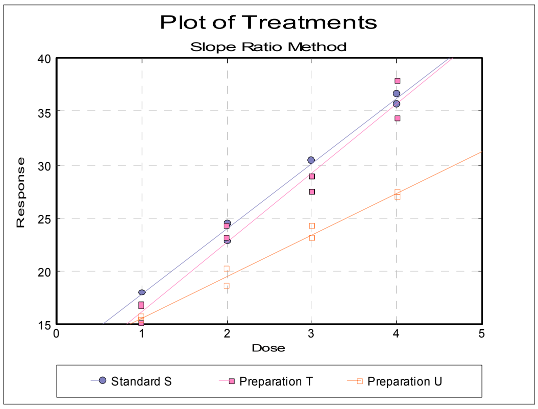
	Intercept	Slope	Residual SS	R-squared
Standard S	11.0417	6.3561	21.3544	0.9807
Preparation T		6.0561		
Preparation U		4.1228		

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	15.0000	14.2920	13.3681	15.2711
Preparation U	15.0000	9.7295	8.8542	10.6088

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0008	5.2437	93.5355
Preparation U	0.0007	6.1678	91.0034

G =	0.0056
C =	1.0056



Example 3

Table 7.10.2. on p. 161 from Finney, D. J. (1978) is an example with blanks, four replicates and two preparations.

Open BIOFINNEY and select Bioassay → Slope Ratio Method. From the Variable Selection Dialogue select columns *C12 Data*, *C13 Dose* and *S14 Preparations* respectively as [Data], [Dose] and [Preparation]. Click [Next] to proceed to Output Options Dialogue. Click [All] to select all output options and then click [Finish]. The potency ratio and its confidence limits are calculated with the default assigned potency of 1. The following output is obtained.

Slope Ratio Method

Normality Tests

Smaller probabilities indicate non-normality.
* Lilliefors probability = 0.2 means 0.2 or greater.

Dose×Preparations	Valid Cases	Mean	Standard Deviation
0 × Blank	4	41.7500	3.3040
0.5 × Standard	4	100.0000	3.5590
1 × Standard	4	161.5000	4.9329
0.5 × Test	4	85.0000	4.7610
1 × Test	4	122.2500	1.2583

Dose×Preparations	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	* Probability
0 × Blank	0.9157	0.5130	0.2521	0.2000
0.5 × Standard	0.8947	0.4051	0.2500	0.2000
1 × Standard	0.9646	0.8081	0.1939	0.2000

Dose×Preparations	Shapiro-Wilk Test	Probability	Kolmogorov-Smirnov Test	* Probability
0.5 × Test	0.9110	0.4877	0.2357	0.2000
1 × Test	0.8949	0.4064	0.3287	0.1554

Dose×Preparations	Cramer-von Mises Test	Probability	Anderson-Darling Test	Probability
0 × Blank	0.0443	0.5124	0.2706	0.4502
0.5 × Standard	0.0518	0.3989	0.3151	0.3280
1 × Standard	0.0304	0.7840	0.1973	0.7044
0.5 × Test	0.0463	0.4814	0.2783	0.4263
1 × Test	0.0676	0.2335	0.3610	0.2343

Homogeneity of Variance Tests

	Test Statistic	Probability
Bartlett's Chi-square Test	4.3575	0.3598
Bartlett-Box F Test	1.1126	0.3498
Cochran's C (max var / sum var)	0.3372	0.8137
Hartley's F (max var / min var)	15.3684	
Levene's F Test	3.3168	0.0390

Validity of Assay

Due To	Sum of Squares	DoF	Mean Square	F-Stat	Prob
Constant	208488.200	1	208488.200		
Regression	31456.914	2	15728.457	1089.731	0.0000
Blanks	2.161	1	2.161	0.150	0.7043
Intercept	34.225	1	34.225	2.371	0.1444
Non-linearity	0.000	0			
Standard Non-linearity	0.000	0			
Test Non-linearity	0.000	0			
Treatments	31493.300	4	7873.325		
Residual	216.500	15	14.433		
Total	31709.800	19	1668.937		

Separate Regression

	Intercept	Slope	Residual SS	R-squared
Standard	38.5000	123.0000	111.0000	0.9855
Test	47.7500	74.5000	72.7500	0.9745

Common Regression

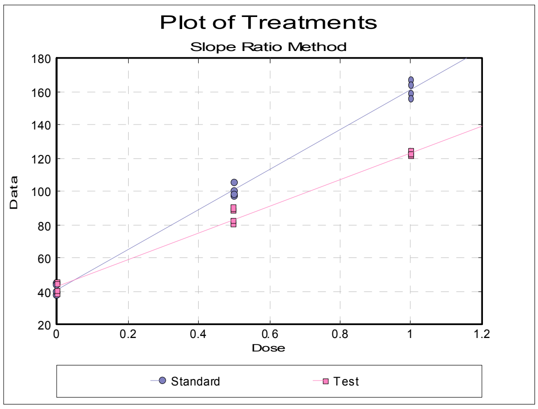
	Intercept	Slope	Residual SS	R-squared
Standard	42.1429	118.6286	252.8857	0.9920
Test		81.2286		

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Test	1.0000	0.6847	0.6464	0.7236

Test Preparation	Variance	Weight	% Precision
Test	0.0003	3046.6783	94.3986

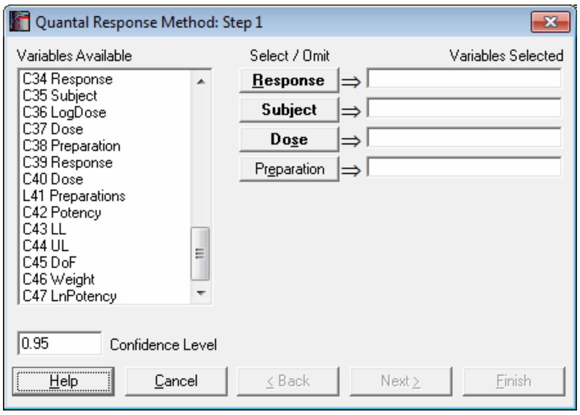
G =	0.0021
C =	1.0021



10.3. Quantal Response Method

Parallel line models can be fitted using one of logit, probit, gompit (clolglog) or loglog link functions. Asymmetric dose structures and multiple test preparations are supported. The parallelism and linearity tests are performed. Output includes estimates of effective dose (or lethal dose) for any user-defined percentile (including ED50, or LD50), the potency ratio and their confidence limits.

10.3.1. Quantal Response Variable Selection


Quantal Response Method: Step 1

Variables Available

C34 Response
C35 Subject
C36 LogDose
C37 Dose
C38 Preparation
C39 Response
C40 Dose
L41 Preparations
C42 Potency
C43 LL
C44 UL
C45 DoF
C46 Weight
C47 LnPotency

0.95 Confidence Level

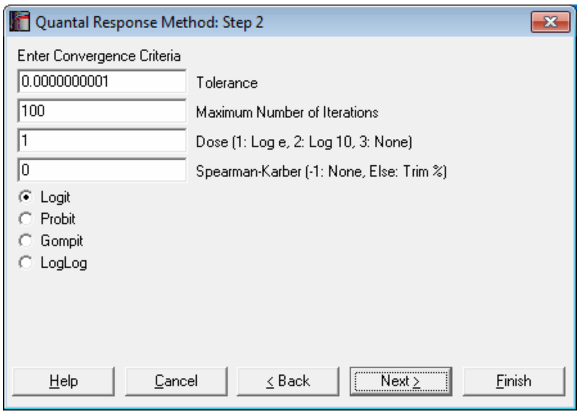
The first variable [Response] represents the number subjects responding positively (or negatively) to the test and the second [Subject] contains the total number of subjects in that group. Therefore, the following relation should hold for each case:

$$0 \leq \text{Response} \leq \text{Subject}$$

If some cases do not conform to this, then the analysis will be aborted.

As in other bioassay procedures, a [Dose] variable should also be selected. However, the choice of a [Preparation] variable is optional. If a [Preparation] variable is not selected, then an ED50 estimate can still be computed, fitting a single line (instead of parallel lines) on all data points.

The next dialogue asks for the following convergence and model parameters.



Quantal Response Method: Step 2

Enter Convergence Criteria

0.0000000001 Tolerance

100 Maximum Number of Iterations

1 Dose (1: Log e, 2: Log 10, 3: None)

0 Spearman-Karber (-1: None, Else: Trim %)

● Logit

○ Probit

○ Gompit

○ LogLog

Tolerance: This value is used to control the sensitivity of the maximum likelihood procedure employed. Under normal circumstances, you do not need to edit this value. If a convergence cannot be achieved, then larger values of this parameter can be tried by removing one or more zeros.

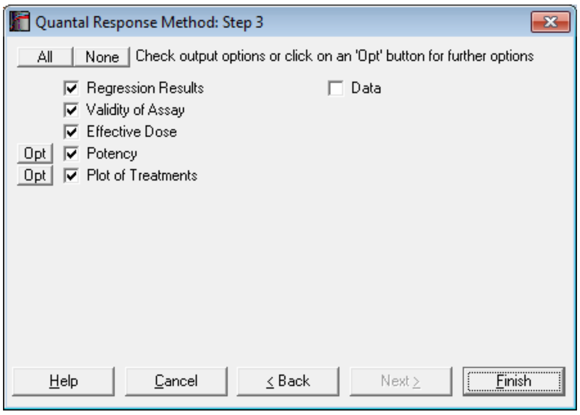
Maximum Number of Iterations: When convergence cannot be achieved with the default value of 100 function evaluations, a higher value can be tried.

Dose Transformation: It is possible to transform the dose variable by natural (default) or 10-based logarithm or leave it untransformed.

Spearman-Karber: When a non negative percentage is entered, ED50 and its confidence limits are also computed using the Spearman-Karber method. When this box contains a negative value, Spearman-Karber results are not reported.

Link Function: Select the model to be estimated; logit, probit, gompit (cloglog) or loglog.

10.3.2. Quantal Response Output Options



Check output options or click on an 'Opt' button for further options

☑ Regression Results
☑ Validity of Assay
☑ Effective Dose
☑ Potency
☑ Plot of Treatments
□ Data

10.3.2.1. Regression Results

The maximum likelihood model is constructed as a regression without a constant term (i.e. through the origin), with independent variables consisting of the transformed dose variable and a set of m dummy variables created from the preparations variable. When the convergence is achieved, the coefficient for the dose variable represents the estimated common slope and coefficients for the dummy variables represent the estimated intercept for each preparation.

The dependent variable is obtained from the response and subject variables. Let $\hat{Y}_j$ be the expected value for case j. Then:

Logit:

$$F_j = \frac{\mathrm{Exp}(\hat{Y}_j)}{1 + \mathrm{Exp}(\hat{Y}_j)}$$

Probit:

Fj is the cumulative normal probability at $\hat{Y}_j$

Gompit (cloglog):

$$Fj = 1 - \mathrm{Exp}(-\mathrm{Exp}(\hat{Y}_j))$$

Loglog:

$$Fj = \mathrm{Exp}(-\mathrm{Exp}(-\hat{Y}_j))$$

For further details see 7.2.5.1. Logit / Probit / Gompit Model Description.

A Newton-Raphson type maximum likelihood algorithm is employed to minimise the negative of the log likelihood function. The nature of this method implies that a solution (convergence) cannot always be achieved. In such cases, you are advised to edit the convergence parameters provided, in order to find the right levels for the particular problem at hand.

10.3.2.2. Validity of Assay

Three chi-square tests are performed:

1) Pearson's overall goodness of fit test:

$$\chi^2 = \sum_{j=1}^{n} \frac{(r_j - \hat{E}_j)^2}{\hat{E}_j(1 - \hat{Y}_j)}$$

where:

$$\hat{E}_j = r_j \hat{Y}_j$$

is the expected frequency for case j. The test statistic has (n – m -1) degrees of freedom.

2) Non-linearity test:

$$\chi^2 = \sum_{i=1}^{m} S_{yy} - \sum_{i=1}^{m} \frac{S_{xy}^2}{S_{xx}}$$

where Sxx, Syy and Sxy are as defined in Finney, D. J. (1978) p. 372. The test statistic has (n – 4) degrees of freedom.

3) Non-parallelism test:

$$\chi^2 = \sum_{i=1}^{m} \frac{S_{xy}^2}{S_{xx}} - \frac{\sum S_{xy}^2}{\sum S_{xx}}$$

The test statistic has (m – 1) degrees of freedom.

10.3.2.3. Effective Dose (or Lethal Dose)

By default, ED50 (or LD50) values and their fiducial confidence limits are computed for all preparations. If a non negative percentage is entered in the Spearman-Karber box, ED50 values, their confidence limits and actual percentage trim used are also displayed for each preparation using this method. If the [Preparation] variable is not selected or contains only one value, then an ED50 estimate will still be calculated, fitting a single line (instead of parallel lines) on all data points. Let d be the user-supplied effective dose (or lethal dose) quantile. Then for the logit model compute:

$$Y = \mathrm{Log}\left(\frac{1-d}{d}\right)$$

and for the probit model:

Y = Critical value of (1 - d) from inverse standard normal distribution.

The effective dose for preparation i is then found as:

$$M_i = \frac{\hat{a}_i - Y}{\hat{b}}$$

where $\hat{a}_i$ is the intercept for preparation i and $\hat{b}$ is the common slope.

To calculate the confidence limits of M_i first define:

$$g = \frac{Z_\alpha^2 V_b}{\hat{b}^2}$$

where V_b is the variance of common slope and Z_α is the critical value from normal distribution.

The confidence interval for potency ratio of each test preparation is defined as:

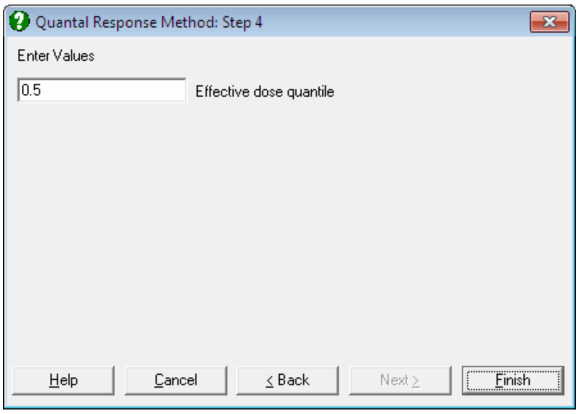
$$M_{iL}, M_{iU} = \left(M_i - \frac{gVsi}{Vss} \pm \frac{Z_\alpha}{b}\sqrt{G_i}\right)/(1-g)$$

where:

$$G_i = Vii - 2M_i Vsi + M_i^2 Vss - g(Vii - Vsi^2 / Vss)$$

and Vss, Vii and Vsi are the elements of covariance matrix of regression coefficients for standard and preparation i.

The trimmed Spearman-Karber (or Kaerber) and its confidence interval are computed as described in Hamilton at al (1977). If the consecutive response values are not monotonically increasing (or decreasing) their average is used. If the trim entered by the user has no solution, then the minimum trim estimated by the program is used. For each preparation, the percentage trim entered and the trim used by the program are displayed.

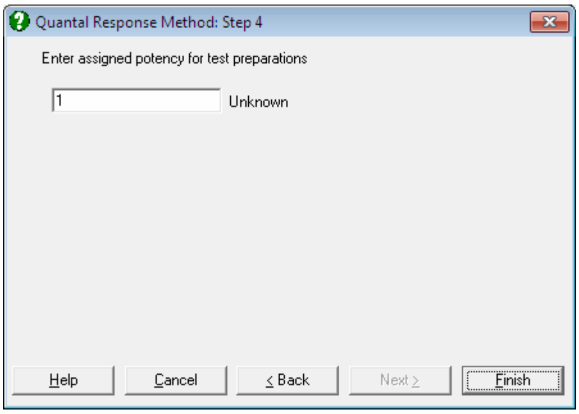


Effective dose quantile: 0.5

If you wish to compute other effective dose values then, on the Output Options Dialogue, click the [Opt] button situated to the left of the Effective Dose option. A further dialogue pops up asking for entry of a value between 0 and 1. The program will then output the effective dose and its confidence limits for this value, as well as its complementary value, for all preparations. For instance, if 0.9 is entered, ED10 and ED90 values will be computed and the output will look like as follows:

	Effective Dose	Lower 95%	Upper 95%
Standard ED10	4.4731	3.5712	5.2983
ED90	28.2233	23.6956	35.6538
Unknown ED10	6.6911	5.2925	8.0338
ED90	42.2176	34.4987	55.0306

10.3.2.4. Potency



Quantal Response Method: Step 4

Enter assigned potency for test preparations

1 Unknown

The relative potency is for test preparation i is found as:

$$M_i = \frac{\hat{a}_i - \hat{a}_s}{\hat{b}}$$

where $\hat{a}_i$ and $\hat{a}_s$ are the intercepts for test i and standard preparations and $\hat{b}$ is the common slope.

To calculate the confidence limits of M_i first define:

$$g = \frac{Z_\alpha^2 V_b}{\hat{b}^2}$$

where V_b is the variance of common slope and Z_α is the critical value from normal distribution.

First define:

$$V_{11} = V_{ss} + V_{ii}\text{-}2V_{si}$$

$$V_{12} = V_{bi} + V_{bs}$$

The fiducial confidence interval for potency ratio of each test preparation is defined as:

$$M_{iL}, M_{iU} = \left(M_i - \frac{gV_{12}}{Vss} \pm \frac{Z_\alpha}{b}\sqrt{G_i} \right) / (1-g)$$

where:

$$G_i = V_{11} - 2M_i V_{12} + M_i^2 Vss - g(V_{11} - V_{12}^2 / Vss)$$

M_i is the relative potency and M_{iL} and M_{iU} are the confidence limits for the relative potency. The estimated potency and its confidence interval are obtained by multiplying these relative values by the assigned potency supplied by the user for each test preparation separately.

The approximate variance of M_i is:

$$V_i = \frac{s^2}{b^2}\left(V_{11} - 2M_i V_{12} + M_i^2 Vss\right)$$

Weights are computed after the estimated potency and its confidence interval are found:

$$W_i = \frac{4Z_\alpha^2}{\left(M_{iU} - M_{iL}\right)^2}$$

and % Precision is:

$$P_i = 100\frac{M_{iL}}{M_i}$$

10.3.2.5. Plot of Treatments

Response ratios are plotted on a logit, probit or gompit Y-axis (see Scale Type), versus log of dose, according to the model selected. A line of best fit is also drawn for each preparation. If you want to edit the properties of the graph, you can send it to Graphics Editor by clicking on the [Opt] button situated to the left of the plot option. The Edit → Data Series dialogue on the graphics window menu provides you with necessary controls to edit all aspects of the plot.

10.3.3. Quantal Response Examples

Example 1

Data is given in *European Pharmacopoeia* (2008), Table 5.3.1.-I on p. 589.

Open BIOPHARMA6 and select Bioassay → Quantal Response Method. From the Variable Selection Dialogue select columns *C30* to *C33* respectively as [Response], [Subject], [Dose] and [Preparation]. Click [Next], select Probit model and leave other entries unchanged. On the Output Options Dialogue. Click the [Opt] button situated to the left of the Potency option. Enter 140 as the assigned potency for the unknown. Click [Back] to get back to output options, click [All] to perform all tests in one go and then click [Finish].

Quantal Response Method

Model selected: Probit

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Common Slope	2.4011	0.4170	5.7582	0.0000	1.5838	3.2183
Intercept 1	-2.0504	0.4086	-5.0176	0.0000	-2.8513	-1.2495
Intercept 2	-1.7208	0.3829	-4.4945	0.0000	-2.4713	-0.9704

Validity of Assay

	Chi-Square	DoF	Probability
Goodness of Fit	1.9225	5	0.8598
Non-linearity	1.9215	4	0.7502
Non-parallelism	0.0010	1	0.9743

Effective Dose

	Effective Dose	Lower 95%	Upper 95%	Trim Entered	Trim Used
Standard S ED50	2.3489	1.9291	2.8956		
Spearman-Karber	2.3391	1.8891	2.8961	0.00%	9.09%
Preparation T ED50	2.0477	1.6716	2.5166		
Spearman-Karber	1.9703	1.5878	2.4450	0.00%	9.09%

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	140.0000	160.5974	120.9660	215.1559

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0170	0.0017	75.3225

G =	0.1131
C =	1.1275

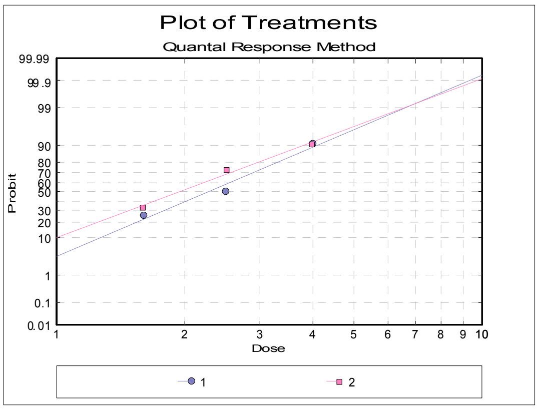


Table 5.3.2.-I. also gives the results for logit and gompit methods. Click on the [Last Procedure Dialogue] button on the Output Medium Toolbar. This will display the Output Options Dialogue again. Click [Back], select the Logit model, click [Next] and select only the Potency output option.

Quantal Response Method

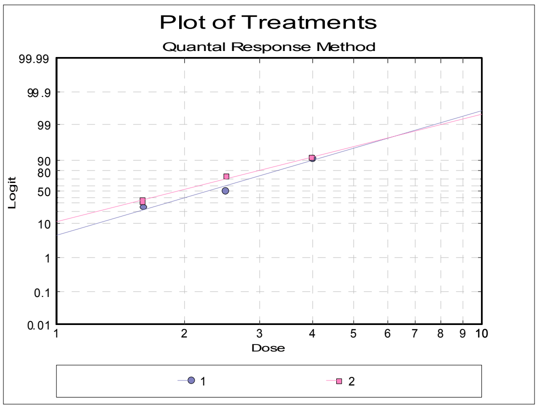
Model selected: Logit

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	140.0000	162.8590	121.1311	221.1056

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0172	0.0015	74.3779

G =	0.1455
C =	1.1703



Select the Gompit model and repeat the analysis.

Quantal Response Method

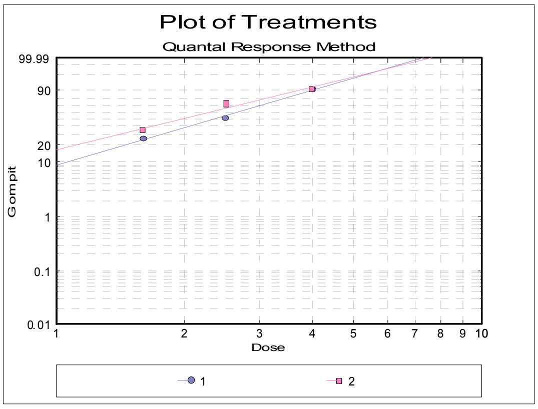
Model selected: Gompit

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Preparation T	140.0000	158.3126	118.7082	213.2961

Test Preparation	Variance	Weight	% Precision
Preparation T	0.0174	0.0017	74.9834

G =	0.1179
C =	1.1336



Example 2

Data is given in *European Pharmacopoeia* (2008), Table 5.3.3.-I on p. 591.

Open BIOPHARMA6 and select Bioassay → Quantal Response Method. From the Variable Selection Dialogue select columns *C34* to *C36*, *Response*, *Subject*, *LogDose* respectively as [Response], [Subject], [Dose]. You do not need to select a [Preparation] variable, though you can also select *C38 Preparation* as [Preparation] as it contains only one value. Click [Next], select Probit model and select the dose transformation None, since the data is already logged base 10. On the Output Options Dialogue click [Finish]. The following output is obtained:

Quantal Response Method

Model selected: Probit

Regression Results

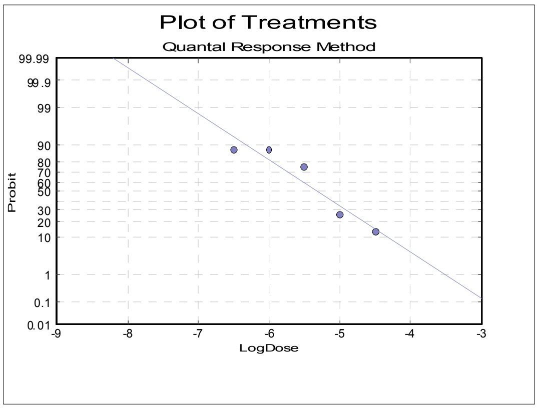
	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Common Slope	-1.4880	0.3063	-4.8580	0.0000	-2.0883	-0.8877
Intercept 1	-7.9314	1.6586	-4.7820	0.0000	-11.1822	-4.6806

Validity of Assay

	Chi-Square	DoF	Probability
Goodness of Fit	2.7112	8	0.9512
Non-linearity	2.7112	8	0.9512

Effective Dose

	Effective Dose	Lower 95%	Upper 95%
1 ED50	-5.3302	-5.6568	-5.0022



Remember that the dose data were already logged base 10. Applying the back-transformation (again base 10):

$$-M_T + \text{Log}(1000/50)$$

and reversing the limits we obtain:

	Effective Dose	Lower 95%	Upper 95%
1 ED50	6.6313	6.3033	6.9578

Example 3

Table 18.2.1. on p. 376 from Finney, D. J. (1978) gives data for an unbalanced assay with one test preparation. Finney gives the results in Table 18.3.1. on p. 380.

Open BIOFINNEY and select Bioassay → Quantal Response Method. From the Variable Selection Dialogue select columns *C20* to *C23* respectively as [Response], [Subject], [Dose] and [Preparation]. Click [Next], select Probit model and leave other entries unchanged. On the Output Options Dialogue click the [Opt] button situated to the left of the Potency option, enter 20, click [Back] and then click [Finish] to obtain the following output.

Quantal Response Method

Model selected: Probit

Regression Results

	Coefficient	Standard Error	Z-Statistic	2-Tail Probability	Lower 95%	Upper 95%
Common Slope	1.3914	0.1234	11.2782	0.0000	1.1496	1.6332
Standard	-3.3660	0.3061	-10.9960	0.0000	-3.9660	-2.7661
Unknown	-3.9263	0.3520	-11.1554	0.0000	-4.6162	-3.2365

Validity of Assay

	Chi-Square	DoF	Probability
Goodness of Fit	5.7033	11	0.8924
Non-linearity	5.4208	10	0.8614
Non-parallelism	0.2824	1	0.5951

Effective Dose

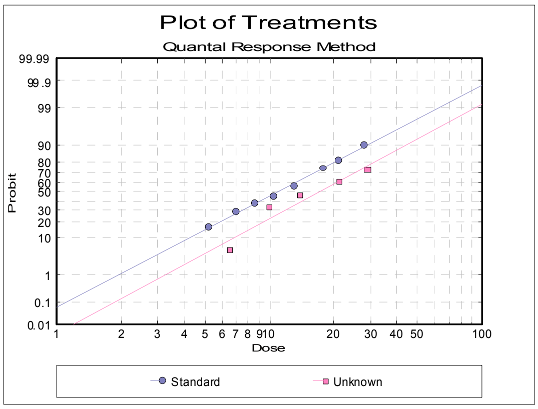
	Effective Dose	Lower 95%	Upper 95%
Standard ED50	11.2359	10.0319	12.6068
Unknown ED50	16.8071	14.5334	19.5569

Potency

Test Preparation	Assigned Potency	Estimated Potency	Lower 95%	Upper 95%
Unknown	20.0000	13.3704	11.0678	16.0812

Test Preparation	Variance	Weight	% Precision
Unknown	0.0086	0.6113	82.7786

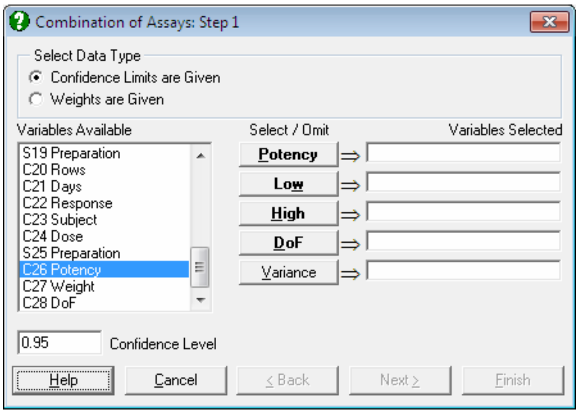
G =	0.0295
C =	1.0304



10.4. Combination of Assays

Weighted mean potency and its confidence limits are calculated from two or more assay results. Two different types of assay results can be combined according to *European Pharmacopoeia* (1997-2008) and *United States Pharmacopoeia* (2009).

10.4.1. Variable Selection

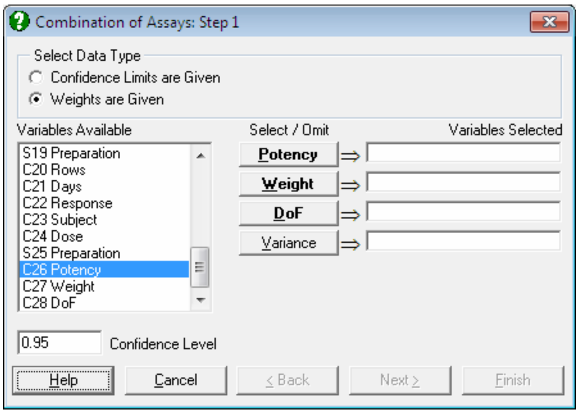


Combination of Assays: Step 1

Select Data Type
● Confidence Limits are Given
○ Weights are Given

0.95 Confidence Level

1) **Confidence Limits are Given:** Use this option when the confidence interval of individual assay potencies are known. Potencies, their lower and upper limits and the degrees of freedom should be entered in separate columns and they are selected for analysis by clicking on [Potency], [Low], [High] and [DoF] respectively. These are the compulsory variables. You can optionally select a column containing variances by clicking on [Variance], in which case a Bartlett test of homogeneity of variances will also be performed.



Combination of Assays: Step 1

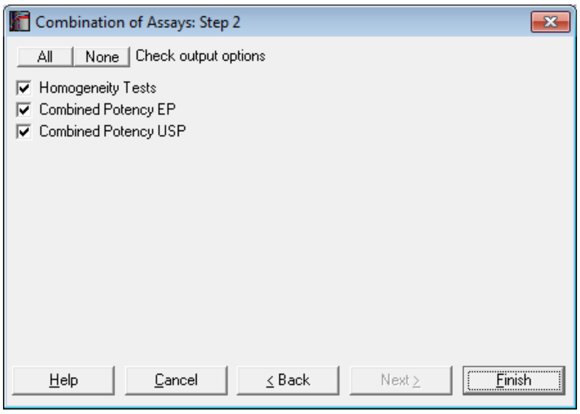
Select Data Type
○ Confidence Limits are Given
● Weights are Given

Variables Available
S19 Preparation
C20 Rows
C21 Days
C22 Response
C23 Subject
C24 Dose
S25 Preparation
C26 Potency
C27 Weight
C28 DoF

0.95 Confidence Level

2) **Weights are Given:** If the confidence limits of potencies are not available, but their weights are, then select this data option. In this case, potencies, their weights and their degrees of freedom should be entered in separate columns and selected by clicking on [Potency], [Weight] and [DoF] respectively. Again, the choice of a column containing variances is optional.

10.4.2. Output Options



Combination of Assays: Step 2

Check output options
☑ Homogeneity Tests
☑ Combined Potency EP
☑ Combined Potency USP

If the first data option Confidence Limits are Given is selected, the weights are computed using the confidence limits of each estimated potency:

$$W_i = \frac{t^2_{\alpha,df_i}}{\left(M_{iU} - M_{iL}\right)^2}$$

where the t-statistic has the same degrees of freedom as it was used in calculating the potency.

For both data options the weighted mean potency is defined as:

$$\overline{M} = \frac{\sum_{i=1}^{n} W_i M_i}{\sum_{i=1}^{n} M_i}$$

10.4.2.1. Homogeneity Tests

Homogeneity of Means

The following chi-square statistic is tested:

$$\chi^2 = \sum_{i=1}^{n} W_i (M_i - \overline{M})^2$$

with n - 1 degrees of freedom. If the probability value displayed is greater than the given significance level (usually 0.05) then the test is said to be satisfactory.

Homogeneity of Variances: Bartlett's Chi-Square Test

This test statistic is calculated only when the optional [Variance] variable is selected.

$$C = N\text{Log}(P) - Q$$

where n is the number of assays, N is the total degrees of freedom and:

$$P = \sum_{i=1}^{n} \frac{n_i V_i}{N}$$

$$Q = \sum_{i=1}^{n} n_i \text{Log}(V_i)$$

where n_i is the degrees of freedom for each assay. Next, the following term is computed:

$$C = 1 + \frac{1}{3(n-1)}\left(\sum \frac{1}{n_i} - \frac{1}{N}\right)$$

and the test statistic is obtained as:

$$B_C = \frac{B}{C}$$

which is approximately chi-square distributed with n - 1 degrees of freedom. If the probability value displayed is greater than the given significance level (usually 0.05) then the test is said to be satisfactory.

10.4.2.2. Combined Potency EP

European Pharmacopoeia (1997-2008) gives calculations for weighted and unweighted means as two alternatives. The Weighted Mean Potency method (which is more robust) should be used if:

1) Homogeneity of Means test is satisfactory,
2) Bartlett's Homogeneity of Variance test is satisfactory and
3) g < 1 for each assay.

If one or more of these criteria fail, the Unweighted Mean Potency results can be used.

Weighted Mean Potency

The standard error of weighted mean potency is defined as:

$$S_{\overline{M}} = \frac{1}{\sqrt{\sum_{i=1}^{n} W_i}}$$

The confidence limits of the weighted mean potency are then:

$$\overline{M} \pm t_{\alpha,df} S_{\overline{M}}$$

where the degrees of freedom for the t-statistic is the sum of the number of degrees of freedom for the error mean squares in the individual assays (n_i).

Unweighted Mean Potency

This is defined as the arithmetic mean of individual assay potencies:

$$\overline{M} = \frac{1}{n}\sum_{i=1}^{n} M_i$$

and its variance is:

$$s_{\overline{M}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i - \overline{M})^2$$

The confidence limits of the unweighted mean potency are then:

$$\overline{M} \pm t_{\alpha,df}\sqrt{\frac{S_{\overline{M}}}{n}}$$

where the degrees of freedom for the t-statistic is n - 1.

10.4.2.3. Combined Potency USP

USP (*US Pharmacopoeia*, 2009) computes the weighted mean potency as shown above, however it differs slightly in the definition of its confidence interval. USP also introduces a Semi Weighted Mean Potency method.

Weighted Mean Potency

USP adds a correction term to the standard error of weighted mean potency as follows:

$$S'_{\overline{M}} = S_{\overline{M}}\sqrt{1 + \frac{4}{\left(\sum W_i\right)^2}\sum\frac{W_i\left(\sum W_i - W_i\right)}{n'_i}}$$

where:

$$n'_i = n_i - 4\frac{n-2}{n-1}$$

The confidence limits of the weighted mean potency are:

$$\overline{M} \pm t_{\alpha,df} S'_{\overline{M}}$$

where the degrees of freedom for the t-statistic is:

$$df = \frac{\left(\sum W_i\right)^2}{\sum W_i^2 / n_i}$$

Semi Weighted Mean Potency

When the chi-square test for combined mean potency has a value p ≥ α, USP advises use of semi weighted mean potency. The variance of the heterogeneity between assays is calculates as:

$$\nu = \frac{\sum M_i^2 - \frac{1}{n}\left(\sum M_i\right)^2}{n-1} - \frac{\sum V_i}{n}$$

where:

$$V_i = \frac{1}{W_i}$$

If $\nu < 0$, then it is re-calculated by omitting the term following the minus sign. A semi weight is then defined as:

$$W_i^{'} = \frac{1}{(V_i + \nu)}$$

The semi weighted mean potency and its confidence interval is then calculated as in the Weighted Mean Potency section above.

10.4.3. Example

Data is given in *European Pharmacopoeia* (2008), Table 6.4.1.-I on p. 594.

Open BIOPHARMA6 and select Bioassay → Combine Assays. From the Variable Selection Dialogue select the second data option Weights are Given. Then select *C47 LnPotency* as [Potency], column *C46* as [Weight] and *C45* as [DoF]. Click [Next] to proceed to Output Options Dialogue. Check both output options and click [Finish] to obtain the following output.

Combination of Assays

Homogeneity Tests

	Chi-Square	DoF	Probability
Mean	4.4274	5	0.4897
Variance (Bartlett)			

Combined Potency

	Potency	Lower 95%	Upper 95%
Weighted Mean	9.8085	9.7951	9.8218
Unweighted Mean	9.8088	9.8087	9.8089

UNISTAT Statistical Package

Appendix

A.1. Continuous Distributions

A.1.1. Normal

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}\text{Exp}\left(-\frac{(X-\mu)^2}{2\sigma^2}\right),\ -\infty<X<\infty$$

Mean = μ
Variance = σ^2

Parameters:

μ: mean, $-\infty<\mu<\infty$
σ: standard deviation, $\sigma>0$

A.1.2. Student's t

$$f(X)=\frac{1}{\sqrt{n\pi}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{X^2}{n}\right)^{-(n+1)/2},\ -\infty<X<\infty$$

Mean = 0
Variance = n/(n-2), n > 2
where X has a standard normal distribution. Mean and standard deviation parameters are needed if X has a non standard normal distribution.

Parameters:

μ: mean, $-\infty<\mu<\infty$
σ: standard deviation, $\sigma>0$
n: degrees of freedom, $n\in N$

A.1.3. Chi-Square

$$f(X)=\frac{X^{(n/2)-1}\text{Exp}(-X/2)}{2^{(n/2)}\Gamma(n/2)},\ X\geq 0$$

Mean = n
Variance = 2n

Parameter:

n: degrees of freedom, $n \in N$

A.1.4. F

$$f(X) = \frac{\Gamma[(p+q)/2]p^{p/2}q^{q/2}}{\Gamma(p/2)\Gamma(q/2)} X^{(p/q)^{-1}} (pX+q)^{-(p+q)/2}, \ X > 0$$

Mean = q/(q-2), q > 2
Variance = $2q^2(p+q-2)/p(q-2)^2(q-4)$, q > 4

Parameters:

p: numerator degrees of freedom, $p \in N$
q: denominator degrees of freedom, $q \in N$

A.1.5. Beta

$$f(X) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} X^{\alpha-1}(1-X)^{\beta-1}, \ 0 < X < 1$$

Mean = α/(α+ß)
Variance = $\alpha ß/((\alpha+ß)^2(\alpha+ß+1))$

Parameters:

α: alpha, $\alpha > 0$
ß: beta, $\beta > 0$

A.1.6. Gamma

$$f(X) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} X^{\alpha-1} \mathrm{Exp}\left(-\frac{X}{\beta}\right), \ X \geq 0$$

Mean = αß
Variance = $\alpha ß^2$

Parameters:

α: alpha shape, $\alpha > 0$
ß: beta scale, $\beta > 0$

A.1.7. Uniform

$$f(X) = \frac{1}{M - N},\ M \le X \le N$$

Mean = (M+N)/2
Variance = $(N\text{-}M)^2/12$

Parameters:

M: lower bound
N: upper bound, $-\infty < M < N < \infty$

A.1.8. Triangular

$$f(X) = \frac{2(X - M)}{(N - M)(O - M)},\ M \le X \le O$$

Mean = (M+O+N)/3
Variance = $(M^2+O^2+N^2\text{-}MO\text{-}MN\text{-}ON)/18$

Parameters:

M: lower bound
O: centre
N: upper bound, $-\infty < M < O < N < \infty$

A.1.9. Lognormal

$$f(X) = \frac{1}{X\sigma\sqrt{2\pi}} \mathrm{Exp}\left(-\frac{1}{2}\left(\frac{\mathrm{Log}(X) - \mu}{\sigma}\right)^2\right),\ X > 0$$

Mean = $\mathrm{Exp}(\mu+\sigma^2/2)$
Variance = $\mathrm{Exp}(2\mu+\sigma^2)(\mathrm{Exp}(\sigma^2)\text{-}1)$

Parameters:

μ: mean, $-\infty < \mu < \infty$
σ: standard deviation, $\sigma > 0$

A.1.10. Exponential

$$f(X) = \beta \mathrm{Exp}(-\beta X),\ X \ge 0$$

Mean = 1/ß
Variance = $1/ß^2$

Parameter:

ß: 1/mean, $\sigma > 0$

A.1.11. Erlang

$$f(X) = \beta^{\alpha} X^{\alpha-1} \frac{\text{Exp}(-\beta X)}{(\alpha - 1)!},\ X \geq 0$$

Mean = 1/ß
Variance = $\alpha/ß^2$

Parameters:

α: alpha shape, $\alpha \in N$
ß: beta scale, $\beta > 0$

A.1.12. Weibull

$$f(X) = \alpha\beta X^{\alpha-1}\text{Exp}(-\beta X^{\alpha}),\ X > 0$$

Mean = $(1/ß)^{1/\alpha}$ G(1+1/α)
Variance = $ß^{-2/\alpha}$ [G(1+2/α)-G^2(1+1/α)]

Parameters:

α: alpha shape, $\alpha > 0$
ß: beta, $\beta > 0$

A.2. Discrete Distributions

A.2.1. Poisson

$$p(X) = \frac{\text{Exp}(-\mu)\mu^{X}}{X!},\ X = 0, 1, 2, \ldots$$

Mean = μ
Variance = μ

Parameter:

μ: mean, $\mu > 0$

A.2.2. Bernoulli

$$p(X) = p^{X}(1-p)^{1-X},\ X = 0, 1$$

Mean = p
Variance = p(1-p)

Parameter:

p: probability of success, $0 \le p \le 1$

A.2.3. Binomial

$$p(X) = \binom{N}{X} p^{X}(1-P)^{N-X},\ X = 0, 1, 2, \ldots, N$$

Mean = Np
Variance = Np(1-p)

Parameters:

N: number of trials
p: probability of success, $0 \le p \le 1$

A.2.4. Negative Binomial

$$p(X) = \binom{N-1}{X-1} p^{M}(1-P)^{X-M},\ X = M,\ M+1,\ ...,\ N$$

Mean = M/p
Variance = M(1-p)/p^2

Parameters:

M: number of successes
p: probability of success, $0 \le p \le 1$

Alternative Notation

$$p(X) = \left(1+\frac{\mu}{k}\right)^{-k} \binom{k+X-1}{X} \left(\frac{\mu}{\mu+k}\right)^{X}$$

Mean = μ
Variance = μ+μ-Squared/k

Parameters:

μ: arithmetic mean
k: exponent

A.2.5. Geometric

$$p(X) = p(1-p)^{X-1},\ X = 1,\ 2,\ ...$$

Mean = 1/p
Variance = (1-p)/p^2

Parameter:

p: probability of success, $0 \le p \le 1$

A.2.6. Hypergeometric

$$p(X) = \frac{\binom{M}{X}\binom{O-M}{N-X}}{\binom{O}{N}}$$

$X = 0,1, ..., N$ if $N \le M$
$X = 0,1, ..., M$ if $N > M$

Mean = NM/O
Variance = N M/O (O-M)/O (O-N)/(O-1)

Parameters:

M: number of defects in sample
O: population size
N: sample size

A.2.7. Discrete Uniform

$$p(X) = \frac{1}{N-M+1},\ X = M,\ M+1,\ ...,\ N$$

Mean = (M+N)/2
Variance = $((N\text{-}M+1)^2\text{-}1)/12$

Parameters:

M: lower bound
N: upper bound

UNISTAT Statistical Package

References

Altman, Douglas (1991). *Practical Statistics for Medical Research.* Chapman and Hall.

Armitage, P. & G. Berry (2002). *Statistical Methods in Medical Research* (Fourth Edition). Blackwell Scientific Publications, Oxford.

Automotive Industry Action Group (AIAG) (2002). *Measurement Systems Analysis Reference Manual*, 3rd edition. Chrysler, Ford, General Motors Supplier Quality Requirements Task Force, 125.

Bangdiwala S.I., Haedo A.S., Natal M.L. (2008). "The agreement chart as an alternative to the receiver operating characteristic curve for diagnostic tests". *Journal of Clinical Epidemiology*, 61, pp. 866 – 874.

Banks, Jerry (1989). *Principles of Quality Control.* John Wiley & Sons.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4), 1088-1101.

Bissell, A. F. (1990). "How Reliable is Your Capability Index?". *Applied Statistics*, 30, 331-340.

Bland J. M., Altman D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8, 135-160.

Bland J. M., Altman D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics*, 17, 571 – 582.

Bliss, C. I. (1967). *Statistics in Biology*, Volume 1. New York: McGraw-Hill.

Blom, G. (1958). *Statistical Estimates and Transformed Beta-variables*. John Wiley & Sons.

Bolton, S. (1990). *Pharmaceutical Statistics.* Marcel Dekker NY.

Bowerman, Bruce L. & Richard T. O'Connell (1987). *Time Series Forecasting* (Second Edition). Duxbury Press.

Box, G. E. P. and Cox, D. R. (1964), "An Analysis of Transformations". *Journal of the Royal Statistical Society*, 211-243, discussion 244-252.

Boyles, R. A. (1991). "The Taguchi Capability Index". *Journal of Quality Technology*, 23, pp. 107-126.

Chan, L.J., Cheng S.K. and Spiring, F.A. (1989). “A New Measure of Process Capability: Cpm”. *Journal of Quality Technology*, Vol.. 20, No. 3, July, p. 16.

Chou, Y., Polansky, A.M. and Mason R.L. (1998). "Transforming Nonnormal Data to Normality in Statistical Process Control". *Journal of Quality Technology*, 30, April, 133-141.

Cohen, L. & M. Holliday (1983). *Statistics for Social Scientists*. Harper & Row.

Cohen, J. A. (1968). “Weighted Kappa”, *Psychological Bulletin*, pp. 213-220.

Cohen, J. A. (1960). “A Coefficient of Agreement for Nominal Scales”. *Educational and Psychological Measurement*. pp. 37-46.

Colditz, G. A. Brewer, C. S. Berkey, M. E. Wilson, E. Burdick, H. Fineberg, V. and Mosteller, F. (1994), "Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature". *Journal of the American Medical Association*, 271(2), 698–702.

Collett, D. (1994). Modelling Survival Data in Medical Research. Chapman & Hall, London.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (Third Edition). John Wiley & Sons.

Dallal, G. E. and Wilkinson, L. (1986). "An analytic approximation to the distribution of Lilliefors’ test statistic for normality”. *The American Statistician*, 40(4), pp. 294-296 (Correction: 41: p. 248).

Delong E.R., Delong D.M., Clarke-Pearson D.L. (1988). "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach". *Biometrics*, 44(3), pp. 837-845.

DerSimonian, R. Laird, N. (1986). "Meta-analysis in clinical trials". *Controlled Clinical Trials*; 7: 177-188.

Duncan, A. J. (1974), *Quality Control and Industrial Statistics* (Fourth Edition), Homewood, Illinois.

Egger, M., Smith, G. D., Schnedier, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal*, 315, 629-634.

Elliott, D. F. & K. R. Rao (1982). *Fast Transforms: Algorithms, Analyses, Applications*. Academic Press, New York.

Elliot, M. A., J. S. Reisch, N. P. Campbell (1989). "Benchmark Data Sets for Evaluating Microcomputer Statistical Programs", *Collegiate Microcomputer*, Volume VII, No 4, November.

European Pharmacopoeia (1997). 3rd Edition, "Statistical Analysis of Results of Biological Assays and Tests".

European Pharmacopoeia (2008). 6th Edition, "Statistical Analysis of Results of Biological Assays and Tests".

Finney, D. J. (1978), *Statistical Method in Biological Assay*, Griffin, London.

Fleiss, J. L. (1971). "Measuring Nominal Scale Agreement Among Many Raters", *Psychological Bulletin*, 1971, pp. 378-382.

Fleiss, J. L., Cohen, J., Everitt, B. S., (1969). "Large sample Standard Errors of Kappa and Weighted Kappa", *Psychological Bulletin*. Vol. 72, No 5, pp. 323-327.

Fleiss, J. L., Nee, J. C. M. and Landis, J. R. (1979). "Large Sample Variance of Kappa in the Case of Different Sets of Raters". *Psychological Bulletin*, pp. 974-977.

Gardner Martin J., Altman, Douglas G. et al. (2000). *Statistics with Confidence*. 2nd edition. British Medical Journal Publications, London.

Greene, W. H. (2012). *Econometric Analysis*. Seventh Ed. Pearson International, New York.

Hamilton, M.A., R.C. Russo, and R.V. Thurston "Trimmed Spearman-Karber Method for Estimating Median Lethal Concentrations in Toxicity Bioassays," *Environ. Sci. Technol.*, 1977, 11(7), pp. 714-719; Correction: 1978, 12(4), pp. 417.

Hand, D. J. (1981). *Discrimination and Classification*, John Wiley & Sons, New York.

Higgins, J. P. T., and Thompson, S. G. (2002). "Quantifying heterogeneity in a meta analysis". *Statistics in Medicine*, 21, 1539-1558.

ISIS-2 (1988). (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*; 2:349–60.

Kotz, S. and Johnson, N.L. (1993). *Process Capability Indices*, Chapman and Hall, London.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*, Sage.

Larson, H. J. (1982). *Introduction to Probability Theory and Statistical Inference* (Third Edition). John Wiley & Sons.

Liddell, F. D. K. (1983). "Simplified exact analysis of case-referent studies; matched pairs; dichotomous exposure.", *Journal of Epidemiology and Community Health*, Vol. 37, pp. 82-84.

Lilliefors, H. W. (1967). "On the Kolmogorov-Smirnov tests for normality with mean and variance unknown". *Journal of the American Statistical Association*, 62, pp. 399-402.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press.

McGill, R., Tukey, J. W. and Larsen, W. A. (1978). "Variations of Box Plots". *American Statistician*. Vol 32, No 1.

McGraw, K. O. & Wong, S. P. (1996). "Forming some inferences about some intraclass correlation coefficients". *Psychological Methods*, 1, 30-46.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley & Sons.

Mehta, C. R. and Patel, N. R. (1983). "A Network Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables," *Journal of the American Statistical Association*, 78, 427–434.

Montgomery, D. C. (1991). *Design and Analysis of Experiments* (Third Edition). John Wiley & Sons.

Montgomery, D. C. (2009). *Statistical Quality Control: A Modern Introduction*, (Sixth Edition). John Wiley & Sons.

Morrison, D. F. (1990), *Multivariate Statistical Methods* (Third Edition). McGraw-Hill.

Newcombe, R. G. (1998), "Two-sided confidence intervals for the single proportion: Comparison of seven methods". *Statistics in Medicine*, 17, 857-872.

Pearson, E. S. and Hartley H. O. (1951). "Charts for the power function for analysis of variance tests, derived from the noncentral F distribution". *Biometrika*, Vol. 38, pp. 112-130.

Pearson, E. S. and Hartley H. O. (1954). "Tables for Statisticians", *Biometrika*, Vol. 1, Cambridge.

Pettitt, A. N., and Bin Daud, I. (1989). "Case-weighted measures of influence for proportional hazards regression". *Applied Statistics,* Vol. 38, pp. 51-67

Polansky, A. M., Chou, Y.-M., and Mason, R. L. (1999). "An Algorithm for Fitting Johnson Transformations to Non-normal Data". *Journal of Quality Technology*, Vol 31, No 3 pp. 345-350.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1992). Numerical Recipes, The Art of Scientific Computing. Second Edition, Cambridge.

Ratkowski, D. A. (1983). Nonlinear Regression Modeling: A *Practical Unified Approach*. Marcell Dekker.

Royston, P. (1995). "A Remark on Algorithm AS 181: The W Test for Normality". *Applied Statistics*, Vol. 44, pp. 547-551.

Shrout, P. E., and Fleiss, J. L. (1979). "Intraclass correlations: Uses in assessing rater reliability". *Psychological Bulletin* 86 (2): 420–428. doi:10.1037//0033-2909.86.2.420.

Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons", *Journal of the American Statistical Association*, Vol. 69, pp. 730-737.

Stephens, M. A. (1986). "Tests based on EDF statistics". In: D'Agostino, R. B. and Stephens, M.A., eds.: *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

Sachs, L. (1984). *Applied Statistics: A Handbook of Techniques*. Springer Verlag, New York.

Scott I.A., Greenburg P.B., Poole P.J. (2008). "Cautionary tales in the clinical interpretation of studies of diagnostic tests". *Internal Medicine Journal*, 38, pp. 120 - 129.

Shapiro, S. S. and M. N. Wilk (1965). “An Analysis of Variance Test for Normality”. *Biometrika*, Vol. 52, p. 606.

Shapiro, S. S. and M. N. Wilk (1968). “The joint assessment of normality of several independent samples”. *Technometrics*, Vol. 10, pp. 825-839.

Simel D., Samsa G., Matchar D. (1991). "Likelihood ratios with confidence: Sample size estimation for diagnostic test studies". *Journal of Clinical Epidemiology*, 44, pp. 763 - 770.

Smith, B. T. et al., (1976). *Matrix Eigensystem Routines - EISPACK*. Springer Verlag, New York.

Sprent, P. (1993). *Applied Nonparametric Statistical Methods* (Second Edition). Chapman & Hall, London.

Stevens, J. (1986), *Applied Multivariate Statistics for the Social Sciences*. Lawrence Erlbaum Assoc, New Jersey

Tabachnick, B. G. & L. S. Fidell (1989). *Using Multivariate Statistics* (Second Edition). Harper Collins, New York.

Theil, H. (1971). *Principles of Econometrics*. New York, Wiley.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. John Wiley & Sons, New York.

United States Pharmacopeia (2009), USP C111, <111> Design and Analysis of Biological Assays.

Wadsworth, H. M. (1990). *Handbook of Statistical Methods for Engineers and Scientists*, McGraw-Hill.

Wheeler D. J. and Chambers D. S. (1992). *Understanding Statistical Process Control*, Second Edition, SPC Press, Inc.

Winer, B. J. (1970). *Statistical Principles in Experimental Design*. McGraw-Hill.

Wright P. A. (1995). “A process capability index sensitive to skewness”. *Journal of statistical computation and simulation*,1995, vol. 52, no3, pp. 195-203.

Zar, Jerrold. H. (2010). *Biostatistical Analysis* (Fifth Edition). Pearson Educational.

Zhang, N. F., Stenback, G. A., Wardrop, D. M. (1990). "Interval Estimation of Process Capability Index Cpk". *Comm. in Stat. Theory and Methods*, 19, pp. 4455-4470.

UNISTAT Statistical Package

Index

2 x 2 table, 357, 358, 360, 419, 432, 443, 445, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 459, 497, 501, 502, 503, 508, 509, 510, 695, 697, 1056
2D Plots, 200
2x2 table, 558, 559, 561, 570, 573, 583, 588
3D Bar chart, 33, 248
3D Histogram, 15, 88, 267, 283, 314, 335, 336, 345
3D Plots, 59, 160, 224, 225, 253

A

A posteriori tests. See Multiple comparisons
ACCDB files, 140
Accuracy, 458, 697
Actual and fitted values, 617, 619, 647, 693, 723, 724, 773, 780
Adjusted fitted values, 615
Adjusted McFadden, 691, 1066
Adjusted means. See Least squares means
Adjusted residuals, 500
Adjusted R-squared, 620, 624, 625, 634, 635, 748, 776, 781
Aggregate, 162, 167, 663, 666, 667, 668, 1043
Agresti-Coull interval, 434
Alpha error, 952
Altman, D. G., 281, 282, 367, 371, 383, 412, 426, 445, 446, 447, 450, 451, 455, 1181
Analysis of covariance, 815, 817
Anderson-Darling test, 324, 397, 405, 406, 409, 734, 988, 1000, 1007, 1008, 1010, 1013, 1014, 1093, 1129
ANOVA, 739, 759
  Classic experimental approach, 744, 745, 747, 752, 753, 754, 755, 756, 757, 758, 764, 765
  Crossover designs, 745, 757, 759, 1087, 1089, 1091, 1098, 1121, 1124, 1137
  Friedman two-way, 474, 475, 476, 477, 478, 480, 485, 797
  Graeco-Latin squares designs, 742, 753, 759
  Hierarchical approach, 764
  Kruskal-Wallis one-way, 34, 461, 462, 464, 465, 466, 467, 468, 796
  Latin squares designs, 742, 751, 753, 759, 1087, 1089, 1091, 1098, 1118
  Nested designs, 595, 756, 759, 762, 763
  Nested factors, 756, 761, 762, 763, 767
  of regression, 605, 612, 620, 624, 625, 646, 650, 780, 781, 812, 813, 814
  Quade two-way, 480, 481, 797
  Randomised block designs, 741, 744, 745, 759
  Regression approach, 749, 764, 766, 780
  Repeated measures, 595, 745, 746, 747, 748, 749, 750, 755, 756, 757, 758, 759, 761, 762, 763, 765, 766, 768, 772, 888
  Sample size and power estimation, 550, 552
  Split-plot designs, 754, 757, 759, 762, 763
  Variable selection, 760, 767
Area chart, 33, 201, 216, 241, 245, 247
Area enclosed, 210, 211, 257, 306, 663, 686, 688, 699, 701
ARIMA, 876, 878, 880, 903, 904, 905, 906, 908, 910, 912, 915
  Autocorrelations, 905, 906, 907, 908, 909, 911
  Autoregressive parameters, 909
  Forecasting, 913
  Lambda, 904
  Moving average parameters, 903, 909
  Nonseasonal differencing, 903, 906
  Partial autocorrelations, 906, 907
  Residuals, 911
  Seasonal differencing, 903, 904, 906
Armitage, P, 364, 1179
ASCII, 151
Aspect ratio, 59, 64, 71, 90
Attribute control charts, 932, 957, 958, 960, 963, 965
  C chart, 957, 958
  Np chart, 960, 962
  P chart, 960, 962, 963, 964, 965
  U chart, 960
AUC, 663, 699, 700, 701
Autocorrelations, 905, 906, 907, 908, 909, 911
Automation, 42
Average effects, 15, 672
Average linkage, 828, 831
Avg(), 162, 167, 168, 183, 196
Axis scale
  Gompit, 79, 978, 1152
  Linear, 75, 78

- Log base 10, 76, 200, 224, 314
- Log base 2, 77
- Log base e, 77, 200, 224, 314
- Logit, 78, 1152
- Loglog, 78, 79, 80
- Probit, 79, 325, 619, 774, 1145, 1152
- Reciprocal, 78, 200, 224
- User-defined, 77

B

- Back transformation, 773, 794
- Background mode, 22
- Bar chart, 33, 68, 201, 216, 240, 241, 242, 243, 244, 245, 248, 895
- Bar plot, 233
- Bartlett’s chi-square test, 785, 786, 790, 791, 1159, 1161, 1162
- Bartlett-Box F-test, 786
- Base category, 709, 710, 712, 713, 721
- Baseline
  - Cumulative hazard, 1072, 1076
  - Hazard, 1060, 1061, 1072, 1076
  - Survival, 1064, 1070, 1071, 1072, 1076
- Baseline functions, 1061, 1064, 1070, 1075, 1076
- Begg-Mazumdar rank correlation, 578
- Bernoulli distribution, 299, 1173
- Beta distribution, 299, 1170
- Beta error, 935, 953
- Binary
  - Marginal effects, 672, 682
- Binomial, 281, 282, 300, 303, 308, 309, 330, 427, 428, 435, 439, 450, 662, 684, 962, 964, 1173, 1174
  - Distribution, 281, 282, 303, 330, 427, 450, 662, 962, 964, 1173
  - Exact, 427, 428
  - Test, 426, 432, 436, 437, 438, 439, 450, 455, 508
- Binomial distribution, 299
- Bioassay, 1087, 1128, 1145, 1159
  - Combination of, 1159
  - Parallel line method, 60, 1087, 1088, 1089, 1093, 1094, 1104, 1110, 1114, 1118, 1121, 1124, 1128, 1129
  - Quantal response method, 60, 1087, 1145, 1153, 1156, 1157
  - Radioimmunoassay, 649
  - Slope ratio method, 60, 1096, 1128, 1137, 1139, 1142
- Bitmap, 53, 64, 89, 199
- Bivariate regression, 214, 812, 815
- Bland-Altman plot, 15, 88, 267, 314, 338, 343, 345
- Bliss, C.I., 271
- Block distance, 826
- Blom scores, 325
- Blom, G, 325, 1179
- Bolton, S., 757, 1179
- Bonferroni, 795, 806, 807
- Box and whisker plot, 33, 263, 264, 283, 314, 315, 316
- Box-Cox regression, 730, 731, 733, 735, 1005, 1011, 1012
- Box-Cox transformation, 730, 733, 734, 985, 987, 988, 1005, 1010, 1011, 1012, 1013, 1014, 1024
- Breslow test, 1055, 1058, 1059
- Breslow-Day test, 511
- Brown's exponential forecasting, 918, 919, 920
- Brushing, 60, 201, 618, 724

C

- C chart, 957, 958
- Calculate procedure, 121, 196
- Canonical correlations, 825, 826, 851, 883, 884, 885
  - Canonical variables, 884, 886
  - Transformation coefficients, 884
- Canonical discriminant, 844, 850, 851, 853, 855, 856, 860
- Categorical plot, 200
- Centroid linkage, 828, 832
- Character plots, 46, 50, 97, 268, 297, 298, 905, 906, 914
- Chart
  - Weibull, 979
- Chart gallery, 87, 88, 199, 314, 345
- Chebychev distance, 826
- Chernoff faces icon plot, 899
- Chi-square
  - Distribution, 419, 462, 471, 474, 475, 483, 486, 535, 690, 713, 721, 786, 851, 1066, 1067
  - One sample test, 398
  - Two sample test, 312, 330, 399
- Chi-square distribution, 299
- Class interval, 268, 293, 294, 312, 327, 328, 329, 332, 335

Classic experimental approach ANOVA, 744, 745, 747, 752, 753, 754, 755, 756, 757, 758, 764, 765
Classical multidimensional scaling, 862
Classification by case, 852, 856, 860
Classification by group, 689, 714, 852, 856, 860
Classification threshold probability, 687, 688, 694, 696, 700
Clipboard, 42, 44, 53, 64, 89, 91, 95, 96, 100, 103, 118, 119, 146, 150, 151, 199
Cloglog link function, 663, 669, 681, 682
Clopper-Pearson interval, 435, 439, 457, 688, 696
Cluster analysis, 825, 826, 827, 828, 836, 839, 840, 842, 844, 868
  Centroids, 834
  Cluster graph, 835, 841
  Cluster membership, 834, 837, 841, 842
  Cluster table, 834, 840, 842
  Dendrogram, 828, 833, 834, 836
  Distance between centroids, 834
  Distance matrix, 830
  Distance measures, 829
  Final cluster centres, 841, 842
  Hierarchical, 836, 840
  History, 833
  Linkage methods, 828, 831, 839
  Nonhierarchical, 840
Cluster bar chart, 243
Cluster graph, 835, 841
Cochran, 507, 509, 510
Cochran's C, 785, 787
Cochran-Armitage test, 506
Cochran's Q, 486, 487, 555, 575, 577
Code book procedure, 268
Coefficient of variation, 270, 288, 292
Cohen, J. A., 489, 1180
Cohen, L., 366, 369, 386, 388, 417, 428, 456, 470, 485, 486, 1180
Cohort studies, 447, 448, 510, 511
Colditz at al., 583, 1180
Colour
  Background, 55, 56, 107
  Border, 55, 59, 199
  Fill, 55, 56, 59, 71, 199
  Schemes, 108
Column
  Format procedure, 136, 171, 186
  Frozen, 105, 106
  Label, 155
  Sort, 164, 165, 185
  Width, 105, 106, 130
Complete linkage, 828, 831, 839
Compute matrix, 153, 176, 177, 196
Concordant, 503, 505
Conditional functions, 95, 104, 195, 609
Confidence interval, 98, 204, 275, 283, 671, 690, 713, 721, 1066
Conover, W. J, 468, 475, 477, 481, 1180
Constant term, 39, 212, 214, 215, 216, 228, 229, 230, 605, 607, 610, 611, 684, 690, 711, 719, 731, 849
Contingency coefficient, 502
Contingency table, 349, 357, 360, 432, 443, 449, 453, 454, 456, 459, 461, 491, 492, 493, 494, 499, 508, 697, 1056
Continuity correction, 411, 412, 416, 417, 419, 420, 422, 430, 431, 432, 433, 434, 439, 453, 501, 510
Continuous distributions, 259, 1169
Continuous group data, 356, 358
Contour plot, 85, 224, 225, 230
Contrasts, 39, 1095, 1099, 1122, 1125
Control charts, 932, 936, 937, 939, 942, 945, 952, 955, 956, 957, 958, 960, 963, 965
Cook's Distance, 616, 694
Correlation
  Box-Cox plot, 734
  Canonical, 825, 826, 851, 883, 884, 885
  Coefficient, 393, 456
  Intraclass, 394, 395, 488
  Kendall's rank, 15, 385, 390
  Matrix, 46, 392, 601, 603, 612, 621, 642, 646, 650, 671, 692, 714, 721, 826, 854, 873, 888, 890, 911, 1068, 1075
  Multiple, 632, 633, 634, 635, 872, 874
  Partial, 15, 392, 393, 633
  Pearson, 382, 387, 390, 502
  Point biserial, 15, 382, 387, 388, 389, 564, 565
  Sample size, 538
  Spearman's rank, 384, 390
  Tetrachoric, 456, 508
Covariance matrix, 379, 380, 601, 603, 612, 621, 646, 650, 671, 690, 692, 713, 714, 721, 722, 826, 849, 854, 869, 875, 876, 877, 888, 890, 911, 971, 973, 1066, 1068, 1069, 1070
Covariate, 690, 692, 713, 722, 759, 760, 765, 766, 767, 768, 817, 1060, 1061, 1062,

1063, 1064, 1065, 1067, 1068, 1069, 1070, 1072
Coverage probability, 282
Covratio, 617
Cox & Snell, 691, 1067
Cox regression, 36, 37, 598, 1030, 1060, 1062, 1063, 1064, 1065, 1073, 1074, 1075, 1076, 1077, 1078
Prediction, 1061, 1071, 1072
Cox-Snell residuals, 1069, 1078
Cramer’s V coefficient, 502
Cramer-von Mises test, 397, 405, 408, 1093, 1129
Critical values, 299, 301, 303, 304, 305, 429, 485, 525, 530, 531, 535, 787, 796
Cronbach's alpha, 888, 889, 890, 891
Crossover designs, 745, 757, 759, 1087, 1089, 1091, 1098, 1121, 1124
Triple, 1098, 1124
Twin, 1098, 1121, 1137
Cross-tabulation, 349, 357, 358, 359, 360, 432, 443, 449, 452, 461, 488, 491, 492, 496, 497, 499, 508, 509, 512, 697
CSV files. See Delimited text files
Cubic spline, 204, 245
Cumul(), 185, 186
Cumulative
Density function, 257, 258
Frequency, 294
Hazard function, 1069, 1070, 1072, 1078
Histogram, 326
Probability, 19, 26, 281, 301, 302, 304, 305, 331
Cumulative meta analysis, 555, 571
Curve fitting, 193, 202, 203, 205, 211, 213, 255
CUSUM chart, 936, 938, 939, 952
Cut-point, 28, 352, 356, 358, 359

D

Dallal, G.E., 402, 408, 1180
Data conversion functions, 101, 122, 123, 144, 160, 176, 177, 188
Data export / import, 92, 100, 103, 124, 128, 130, 131, 133, 134, 135, 136, 139, 140, 141, 144, 151
Data transformation, 1005, 1008, 1014
Database, 132, 140, 141, 145, 146, 147
Access, 140, 142
dBase, 140, 141, 142, 156, 158
FoxPro, 142
ODBC, 140, 141, 145, 146
Paradox, 143
SQL, 140, 141, 145, 146, 147
Date variable, 29, 125, 129
Date(), 190
Date-time data, 15, 102, 125, 126, 202
Default output medium, 42, 44, 91, 95, 96, 111
Delete, 152
Columns, 152
Rows, 153
Deleted residuals. See Press residuals
Delimited text (CSV) files, 132, 133, 134
Delimiter, 103, 133, 151
DeLong et al., 699, 1180
Delta-Beta, 617, 695, 1069, 1070, 1077, 1078
Dendrogram, 828, 833, 834, 836
Density function, 257, 258, 259, 300, 306, 1038, 1039
DerSimonian at al., 575, 1180
Descriptive statistics, 21, 33, 194, 208, 263, 265, 268, 271, 272, 273, 277, 283, 290, 292, 294, 295, 296, 297, 298, 596, 600, 602, 791
Deviance residuals, 1069, 1078
DfFit, 617
Diagnostic (case) statistics, 596, 612, 613, 614, 621, 663, 722, 772, 1061, 1064, 1068, 1077, 1078
Diagnostic odds ratio, 458, 698
Diagnostic tests, 15, 457, 508, 696, 700
DIF files, 140, 142
Discordant, 503, 505
Discrete distribution, 256, 326, 331, 332, 1173
Discrete uniform distribution, 299
Discriminant analysis, 709, 825, 826, 827, 844, 846, 849, 850, 853, 854, 855, 857, 859, 860
Canonical, 850, 853, 855, 860
Canonical statistics, 851, 855
Classification by case, 852, 856, 860
Classification by group, 689, 714, 852, 856, 860
Discriminant coefficients, 851
Discriminant Functions, 849, 851, 855, 856
Discriminant scores, 851, 856
Distance between centroids, 852, 856
K-th neighbour, 859, 860

Linear, 846, 849, 853, 854
Linear discriminant Functions, 849, 855
Prediction, 844
Stepwise, 844, 846, 847
Distance
Between centroids, 834, 852, 856
Block, 826, 830
Chebychev, 826, 829
Cook, 616, 694
Cosine, 829
Euclid, 829
Mahalanobis, 616
Power, 830
Squared Euclid, 829
Welsch, 616
Distribution, 1169
Bernoulli, 299, 1173
Beta, 299, 1170
Binomial, 281, 282, 299, 303, 330, 427, 450, 662, 962, 964
Chi-square, 299
Continuous, 259, 416, 1169
Discrete, 256, 326, 331, 332, 1173
Discrete uniform, 299, 1175
Erlang, 299, 1172
Exponential, 299, 1172
F, 299, 303, 475, 481, 533, 549, 550, 553, 741
Gamma, 299, 301, 312, 313, 1170
Geometric, 299, 1174
Hypergeometric, 299, 300, 309, 311, 331, 1175
Lognormal, 299, 1171
Negative binomial, 299, 300, 309, 330, 1174
Noncentral F-, 533, 549, 550, 553, 1183
Noncentral t-, 282, 991
Normal, 79, 277, 281, 299, 301, 302, 303, 305, 318, 324, 331, 381, 386, 397, 402, 405, 407, 408, 470, 503, 538, 541, 546, 656, 1169
Phi, 553, 554
Poisson, 299, 400, 957, 960, 1173
Student's t, 299, 310, 382, 385, 1169
Triangular, 299, 1171
Uniform, 299, 401, 402
Weibull, 79, 299, 300, 309, 977, 978, 979, 980, 1172
Distribution functions, 57, 81, 194, 256, 265, 299, 300, 302, 303, 307, 308, 309, 310, 312, 326, 330, 331, 336, 631
Dixon-Grubbs-Neumann, 930
Dot plot, 33, 263, 264, 283, 314, 315, 320
Dragging, 31, 56, 59
Dummy variable, 37, 38, 39, 187, 599, 608, 626, 630, 663, 668, 685, 712, 720, 772, 1062, 1063, 1065
Dummy(), 38, 39, 187, 602, 1074
Duncan, 795, 802
Dunnett test, 464, 476, 778, 788, 789, 792, 795, 796, 797, 807, 808, 809, 817, 1099
Durbin-Watson statistic, 610, 612, 620, 624, 625, 646, 650, 652, 654, 658, 781

E

e constant, 72, 73, 77, 180, 196, 251
ED50, 656, 657, 658, 659, 660, 1087, 1145, 1149, 1158
EDF tests, 405, See Normality tests
Effective dose, 1087, 1145, 1149, 1150, 1158
Egger at al., 578, 1181
Egger regression, 578
Eigenvalues, 827, 850, 851, 855, 863, 869, 870, 872, 873, 874, 875, 878, 879, 884, 885
Eigenvectors, 827, 850, 851, 863, 868, 869, 870, 873
Elliot, M. A., 623, 628, 1181
Elliott, D. F., 1079, 1181
Equimax rotation, 876
Equivalence test
for means, 34, 375, 440
Equivalence Test
for Binomial proportion, 440, 508
Erlang distribution, 299, 1172
Error bar plot, 315, 321
Error bars, 206, 207, 208, 223, 224, 227, 228, 241, 245, 618, 724, 1037, 1038, 1039
Eta, 503, 505
Euclid distance, 832
Excel, 21, 22, 42, 43, 44, 51, 53, 89, 91, 94, 96, 97, 111, 119, 173
Add-in mode, 21, 42, 43, 44, 61, 74, 92, 94, 109, 112, 113, 173, 207, 208, 209, 228, 307, 312, 583, 588, 618, 724, 734, 988, 1008, 1013, 1083, 1107, 1137
Files, 19, 114, 130, 131, 132, 142
Output to, 44, 96
Expected frequency, 256, 265, 300, 309, 311, 312, 398, 399, 400, 499, 500, 501, 673, 718, 1148

Exponential
  Decay, 636, 639, 651, 652, 654
  Distribution, 1172
  Fit, 193, 212, 215
  Weights, 916, 917, 933, 947, 950
Exponential distribution, 299
Exposure, 718, 720, 725

F

F distribution, 299
Factor analysis, 825, 826, 872, 878
  Factor matrix, 874, 876, 878, 879, 880
  Factor score coefficients, 875, 876, 877, 879, 881
  Factor scores, 875, 876, 877, 879, 881
  Principal axis, 872, 873, 874, 878, 879, 880
  Principal components, 872, 874
  Rotations, 876
  Variance table, 869, 874
Factor scores, 875, 876, 877, 879, 881
Factor variables, 32, 40, 614, 693
Fan grid plot, 223
Fct(), 180, 251
F-distribution, 303, 475, 481, 533, 549, 550, 553, 741, 1170
Field delimiter, 103, 133, 135, 151
Fieller's Theorem, 1101, 1135
File formats, 94, 128, 131, 143, 144, 145, 156
  Access, 140, 142
  dBase, 140, 141, 142, 156, 158
  Delimited (CSV), 132, 133, 134
  DIF, 140, 142
  Excel, 19, 114, 130, 131, 142, 156, 158
  Fixed format text, 137, 138, 139, 140, 155
  FoxPro, 142
  Free format text, 103, 133, 134, 136, 137, 155, 157
  HTML, 143
  Lotus, 132, 142, 156, 158
  Paradox, 143
  Sylk, 140, 142, 156, 158
  UNISTAT, 63, 104, 105, 106, 114, 118, 130, 142, 145
File output, 97
Finney, D. J., 1101, 1104, 1124, 1128, 1135, 1142, 1148, 1157, 1181
Fisher, 381, 420, 500, 538, 542, 739, 849
  Exact test, 420, 452, 502, 503, 508
  Linear discriminant function, 849
  Z transformation, 381, 538, 542
Fitting distribution functions, 256, 310, 330
Fixed factor, 769
Fixed format text, 137, 138, 139, 140, 155
Fleiss, J. L., 488, 489, 1181
Flexible linkage, 828, 832
Fonts
  as graphics symbols, 83
  Default, 99
  Graphics, 70, 74, 82, 83
Forecasting, 913, 914, 915, 916, 919, 921, 924, 927, 931, 1179
  ARIMA, 913
  Brown's exponential, 918, 919, 920
Forest plot, 571, 576, 578, 579, 591
Format column, 136, 171, 186
Formula, 174, 176
Formula editor, 118, 174, 178, 252, 253
Fourier analysis, 1079, 1081, 1083
Free format text, 103, 133, 134, 136, 137, 155, 157
Freeman-Halton test, 502
Freq(), 187, 397
Frequency, 293
  Counts, 265, 266, 283, 294, 311, 326, 397, 488, 491
  Distributions, 265, 268, 293, 294, 295, 312, 328, 410
  Histogram, 331
  Table, 293, 294, 488
Friedman two-way ANOVA, 474, 475, 476, 477, 478, 480, 485, 797
Frozen columns, 105, 106
F-tests, 34, 349, 353, 362, 364, 365, 366, 367, 369, 370, 371, 372, 373, 786, 971
  Levene's, 365, 373, 374, 1094
Functions, 178
  Conditional, 95, 104, 195, 609
  Data conversion, 101, 122, 123, 144, 160, 176, 177, 188
  Date and time, 125, 191
  Plot of, 19, 26, 178, 251, 253, 256
  Scalar, 162, 167, 168, 177, 183, 196
  Statistical, 39, 164, 171, 184, 397, 742, 744, 746, 749, 752, 753, 755, 756, 757, 789, 798, 812, 821, 976, 979, 1017, 1089
  Trigonometric, 181, 182, 251
  UNISTAT, 173, 193, 212, 215, 229, 230, 614, 693, 718
Funnel plot, 578, 580, 581

G

Gallery. See Chart gallery
Gamma, 180, 301, 312, 313, 504, 1170
  Distribution, 301, 312, 313, 1170
  Function, 180
Gamma distribution, 299
Gardner, M. J., 412, 446, 1181
Gauge R&R, 1017, 1021, 1023, 1024
  ANOVA Method, 1022
  Average and Range Method, 1020
Gehan test, 1055, 1056, 1058, 1059
General linear model, 35, 41, 606, 739, 741, 746, 747, 748, 750, 759, 760, 764, 766, 767, 776, 795
  Coefficients, 770
  Fitted values, 770
  F-ratio selection, 769
  Multiple comparisons, 770
  Residuals, 770
  Table of means, 770
  Variable selection, 760
Geometric
  Distribution, 1174
  Fit, 216
Geometric distribution, 299
Geometric mean, 270, 276, 286, 292, 322, 339
Gompit, 1145, 1147, 1154, 1155
  Regression, 34, 36, 595, 596, 663, 665, 667, 673, 676, 678, 684, 709, 1087
Gompit link function, 664, 665
Gompit scale, 79, 978, 1152
Goodman-Kruskal's gamma, 503
Goodman-Kruskal's lambda, 505
Goodness of fit tests, 34, 312, 397, 403, 404, 407
Goto, 148, 149
Graeco-Latin squares designs, 742, 753, 759
Graphics objects, 45, 56, 71, 91, 199
Graphics template files, 53, 62, 63, 64, 71, 109, 199
Greene, W.H., 672, 678, 726, 1181
Grid lines, 68, 69, 148, 223
Grouped data, 265, 266, 268, 285, 311, 326

H

Harmonic mean, 270, 276, 287, 292, 322, 472, 530, 532, 533, 534, 792, 794
Hartley's F test, 785, 787, 791
Hazard function, 256, 259, 1028, 1031, 1033, 1037, 1038, 1040, 1044, 1048, 1049, 1055, 1056, 1060, 1061, 1064, 1069, 1070, 1072, 1078
Hazard ratio, 571, 1067
Hedges g, 565, 571
Heterogeneity of regression tests, 782, 797, 815, 816, 817, 818, 821
Hierarchical approach ANOVA, 764
Higgins at al., 577, 1181
Histogram
  2D, 15, 33, 256, 263, 266, 283, 309, 326, 333, 336, 1000
  3D, 15, 88, 267, 283, 314, 335, 336, 345
  Cumulative, 326
  Frequency, 331
Hodges-Lehmann estimator, 34, 414, 415, 425, 426
Holt's linear forecasting, 920, 921, 923, 926
Homogeneity of variance tests, 782, 784, 785, 789, 791, 797, 801, 1092, 1122, 1125, 1130, 1159, 1161, 1162
Hosmer-Lemeshow test, 690, 692
Hotelling's T-Squared
  Analysis, 971, 972
  Chart, 972, 973
  Test, 362, 379, 380, 971
HTML, 42, 44, 48, 51, 52, 91, 96, 143
Hypergeometric distribution, 299, 300, 309, 311, 331, 452, 1175
Hypothesis matrix, 771

I

Icon plots, 892, 894, 898
Incidence rate ratio, 718, 722, 728
  Poisson regression, 718, 722, 728
Information, 162
Input mode, 120
Insert, 153
  Columns, 154
  Rows, 154
Installation, 63, 114
Interaction plot, 774
Interaction selection, 759, 765
Interaction term, 37, 38, 39, 187, 598, 601, 607, 608, 626, 630, 663, 668, 684, 685, 711, 719, 731, 744, 752, 753, 755, 758, 765, 767, 776, 780, 782, 783, 1062, 1063

Interactions, 37, 38, 187, 599, 608, 739, 747, 752, 755, 764, 765, 766, 767, 780, 782, 783, 1062, 1063
Interactive data points, 59, 173, 201, 221, 224, 235
Intercept, 605, 782, 796, 797, 815, 817, 818, 819, 821
Interpolation, 193, 204, 212, 215, 229, 230, 614, 647, 693, 709, 718
Intraclass correlation coefficients, 394, 395, 488
I-square, 555, 577

J

Jackknife residuals. See Studentised residuals
Jeffreys interval, 435
Johnson transformation, 956, 985, 1001, 1005, 1006, 1007, 1008
Jonckheere's trend test, 470

K

k x n chi-square test. See Tests for tables
Kaplan-Meier analysis, 1043, 1044, 1048, 1049, 1051, 1065
Kappa test, 394, 488, 489, 507
Karber. See Spearman-Karber
Kendall’s rank correlation, 15, 385
Kendall's concordance coefficient, 483, 484
Kendall's rank coreelation, 578
Kendall's tau b and tau c, 503
Key sort, 164, 165
Kolmogorov-Smirnov
  One sample test, 402
  Two sample test, 403
Kolmogorov-Smirnov test, 34, 397, 401, 402, 403, 405, 408, 1129, 1182
Kruskal-Wallis one-way ANOVA, 34, 461, 462, 464, 465, 466, 467, 468, 796
Kurtosis, 271, 290, 292

L

Labels
  Column, 155
  Point, 209, 226, 227, 239, 315, 934, 969
  Row, 157, 209, 226, 227, 239, 827
Ladder plot, 15, 88, 267, 314, 344
Lag/lead variable, 36, 37, 40, 596, 598, 601, 607, 608, 668, 669, 684, 685, 686, 711, 712, 719, 720, 731, 1062, 1063, 1065
Latin squares designs, 742, 751, 753, 759, 1087, 1089, 1091, 1098, 1118
LD50, 1145, 1149, 1158
Least significant difference, 795, 805, 806, 807
Least squares means, 770, 771, 772, 773, 774
Lee-Desu statistic, 1055, 1056, 1058, 1059
Legend, 53, 54, 55, 56, 57, 64, 68, 70, 71, 81, 82, 87, 199, 201, 210, 212, 229
Len(), 162, 163, 167, 168, 183, 184, 196
Lethal dose, 1145, 1149
Level(), 186, 187, 742, 744, 746, 749, 752, 753, 755, 756, 757, 789, 798, 812, 821, 1017, 1089
Levene's F test, 365, 368, 373, 374, 377, 785, 788, 790, 791, 1094
Leverage, 613, 616, 694
Levey-Jennings chart, 937, 955
Life table analysis, 1031, 1032, 1033, 1037, 1040, 1043, 1044, 1051, 1060, 1065
Likelihood
  Displacement, 1070, 1077, 1078
  Maximum, 663, 664, 710, 719, 977, 1060, 1061, 1065, 1148
  Ratio test, 501, 690, 733, 1008, 1013, 1067
Lilliefors test, 402, 403, 408, 1180, 1182
Line chart, 91
Line plot, 19, 615, 895, 896
Linear
  Regression, 36
Linear discriminant analysis, 846, 849, 853, 854
Linear regression, 205, 607, 627, 691, 730, 731, 735
Linear scale, 75, 78
Linear-by-linear test, 501, 502
Linearity test. See Regression with replicates
Link function, 663, 664, 665, 669, 673, 678, 681, 682, 1145, 1146
Ljung-Box statistic, 911
Log base 10 scale, 76, 200, 224, 314
Log base 2 scale, 77
Log base e scale, 77, 200, 224, 314
Log file, 113, 114
Logical condition, 104, 172, 195
Logistic regression, 15, 34, 35, 36, 595, 596, 598, 663, 684, 685, 687, 688, 689, 691, 694, 703, 709, 711, 718
Logit, 510, 511, 1145, 1147, 1149, 1154

Logit / Probit / Gompit regression, 34, 36, 595, 596, 663, 665, 667, 673, 676, 678, 684, 709, 1087
Logit link function, 663, 665
Logit residuals, 695
Logit scale, 78, 1152
Loglinear model, 718
Loglog, 1148
Loglog link function, 15, 663, 664, 665, 669
Loglog scale, 78, 79, 80
Lognormal distribution, 299, 1171
Long string table, 100, 101, 122, 123, 160, 161, 188, 189, 190, 739
Long strings, 101, 122, 123, 608
Long(), 101, 122, 160, 189, 190
Lotus 123 files, 132, 142, 156, 158
LSMeans. See Least squares means

M

Macros, 25
  Combining, 111
  Recording, 110
  Running, 111
Mahalanobis distance, 616
Mann-Whitney U test, 16, 34, 410, 412, 413, 699
Mantel-Haenszel, 570, 571, 572, 573, 574, 577, 588
Mantel-Haenszel test, 497, 501, 510, 1056, 1058, 1059
Marginal effects, 15, 671, 681
Marginal means, 770, 771, 772
Martingale residuals, 1069, 1078
Matrix, 893
Matrix plot, 892
Matrix plot, 892
Matrix sort, 165, 166, 167
Matrix statistics, 34, 36, 37, 596, 598, 600, 601, 602, 826
Max(), 162, 163, 167, 168, 183, 184, 196
McFadden, 691, 1066
McNemar test, 453, 454, 501, 508
McNemar-Bowker test, 501
McQuitty linkage, 828, 832
MdRk(), 188
MDS analysis, 861, See Multidimensional scaling
Mean
  Arithmetic, 162, 167, 802, 806, 809, 1174
  Geometric, 270, 276, 286, 292, 322, 339
  Harmonic, 270, 276, 287, 292, 322, 472, 530, 532, 533, 534, 792, 794
  Quadratic, 270, 288
Mean deviation, 289, 292
Mean failure time, 978, 980
Means plot, 201, 206, 216, 220, 241, 245, 774
Median, 269, 276, 287, 296, 316, 326, 332, 335, 412, 413, 414, 415, 416, 419, 420, 425, 426, 471, 472, 796, 1040, 1047, 1051
Median linkage, 828, 832
Median ranks, 188, 976, 978, 979
Median test, 34, 410, 419, 420, 471, 472, 796
  Multisample, 34, 471, 472, 796
  Two sample, 34, 410, 419, 420, 471
Mehta at al., 502, 1182
Merge, 144
Meta analysis, 15, 555, 556, 558, 576
  Cumulative, 555, 571
  One study removed, 555, 571
  Subgroups, 555, 586
Metafile, 43, 44, 45, 53, 64, 89, 90, 91, 96, 199
  Background, 91
  Copy, 90
  Orientation, 90, 579
Min(), 162, 163, 167, 168, 183, 184, 196
Minimum detectable difference, 522, 525, 526, 527, 531, 532, 533, 534, 548, 549, 550, 551, 552
Mis(), 162, 163, 167, 168, 183, 184, 196
Missing data code, 93, 94, 95, 127, 129, 131, 135
Modified LSD, 795, 806, 807
Modified ridit scores, 498
Moment matrix, 601, 604
Montgomery, D. C., 741, 742, 744, 752, 753, 756, 795, 802, 806, 809, 1182
Moses extreme reaction test, 34, 410, 418, 419
Mova(), 186
Moving average chart, 936, 938, 939, 947, 948, 950, 953
Moving averages, 186, 916
Multicollinearity, 187, 610, 611, 689, 772, 777, 780, 844, 1065
Multidimensional scaling, 826, 861, 862, 864, 865, 1182
  Classical, 862
  Ordinal, 861, 864

Multinomial regression, 34, 35, 36, 595, 596, 598, 709, 711, 712, 714
Multiple comparisons, 741, 770, 772, 792
  Between medians, 471
  Between variances, 788
  Bonferroni (Modified LSD), 795, 806, 807
  Duncan, 795, 802
  Dunnett, 464, 476, 778, 788, 789, 792, 795, 796, 797, 807, 808, 809, 817, 1099
  Friedman ANOVA, 475
  GLM, 772
  Kruskal-Wallis ANOVA, 462
  Least Significant Difference (LSD), 795, 805, 806, 807
  Nonparametric, 462, 463, 475, 481, 796, 797
  Quade ANOVA, 481
  Scheffe, 795, 804
  Student-Newman-Keuls, 463, 788, 789, 795, 797, 798, 799, 802, 804
  Tukey-B, 795, 802
  Tukey-HSD, 462, 463, 465, 471, 475, 481, 788, 789, 795, 799, 800, 801, 802, 817
Multiple range tests. See Multiple comparisons
Multivariate analysis, 825, 826, 868, 1182
Multi-way table, 491, 496, 509, 510

N

Nagelkerke, 691, 1067
Natural response rate, 663, 669
Negative binomial distribution, 299, 300, 309, 330, 1174
Negative likelihood ratio, 458, 698
Negative predictive value, 458, 697
Nested designs, 595, 756, 759, 762, 763, 771
Nested factors, 756, 761, 762, 763, 767
Neville's polynomial, 212, 213
No data code, 93, 94, 129, 135, 136
Noncentral F-distribution, 533, 549, 550, 553
Noncentral t-distribution, 282, 991
Noncentrality parameter, 549, 551, 991
Nonhierarchical cluster analysis, 840
Noninferiority Test, 438, 439, 440, 508
Nonlinear regression, 178, 596, 636, 637, 638, 639, 643, 645, 649, 651, 652, 654, 655, 658, 1183
Nonparametric tests, 169, 349, 410, 421, 461, 488
  1-2 samples, 34, 355, 357, 412, 413, 415, 417, 418, 420, 423, 424, 425, 426, 427, 429, 431, 436, 437, 441, 445, 447, 448, 450, 451, 454, 455, 456, 459
  Multisample, 34, 395, 464, 465, 466, 468, 470, 472, 476, 477, 478, 481, 483, 485, 486, 487, 488, 489, 796, 797
Normal
  Approximation, 416, 417, 427, 430, 431, 444, 450
  Data transformation, 1005
  Distribution, 79, 277, 281, 301, 302, 303, 305, 318, 324, 331, 381, 386, 397, 402, 405, 407, 470, 503, 538, 541, 546, 656, 1169
  Plot of residuals, 619, 774
  Probability plot, 15, 33, 59, 79, 263, 314, 325, 406, 619, 734, 1000, 1008, 1013, 1014
Normal distribution, 299
Normality tests, 34, 397, 403, 405, 407, 1092
  Anderson-Darling, 324, 397, 405, 406, 409, 734, 988, 1000, 1007, 1008, 1010, 1013, 1014, 1093, 1129
  Cramer-von Mises, 397, 405, 408, 1093, 1129
  Kolmogorov-Smirnov test, 397, 401, 402, 403, 405, 408, 1129, 1182
  Shapiro-Wilk, 397, 405, 406, 407, 1007, 1093, 1094, 1104, 1129
Np chart, 960, 962
Number formats, 74, 98, 136
Number of columns, 93
Number of rows, 93
Nyquist frequency, 1080

O

Oblimin rotation, 876, 877
OC curve, 522, 548, 549, 553, 934, 935
ODBC, 140, 141, 145, 146
Odds ratio, 79, 447, 448, 455, 508, 510, 560, 561, 566, 573, 591, 663, 692
  Logistic regression, 692
  Paired, 447, 455
  Unpaired, 455, 510

Unpaired proportions, 447
One sample test, 264, 349, 363, 397, 402, 403, 523, 524, 526, 527, 528, 544
Chi-square test, 398
Kolmogorov-Smirnov test, 401, 402, 403
Sample size and power estimation, 523
t-test, 363, 379
One study removed meta analysis, 555, 571
Operating characteristic plot. See OC plot
Optimality criterion, 687
Ordinal multidimensional scaling, 861, 864
Orthogonal polynomial contrasts, 15, 767
Outliers, 61, 173, 316, 383, 384, 607, 613, 859, 934, 935
Output medium, 37, 42, 91, 111, 119, 598, 601
Output medium toolbar, 22, 42, 43, 46, 54, 91, 96, 119, 152, 307, 734, 827, 988, 1008, 1013, 1066, 1083, 1113, 1117, 1120, 1154
Output options dialogue, 41

P

P chart, 960, 962, 963, 964, 965
Page's L trend, 485
Paired t-test, 369, 370
Parallel line method, 60, 1087, 1088, 1089, 1093, 1094, 1104, 1110, 1114, 1118, 1121, 1124, 1128, 1129
Parallelism test. See Heterogeneity of regression and Parallel line method
Parametric tests, 17, 33, 34, 362, 364, 366, 367, 369, 371, 372, 373, 375, 377, 380
Pareto chart, 968
Partial autocorrelations, 906, 907
Partial correlations, 393
Password protection, 105
Pasting clipboard data, 103, 118, 151
Pearson chi-square, 501
Pearson correlation, 382
Pearson's chi-square test, 493, 690
Percentiles, 33, 263, 268, 276, 279, 282, 287, 999
Perfect fit, 611
Perspective, 84, 85, 225
Peto odds ratio, 560
Pharmacopoeia, 1093, 1094, 1096, 1099, 1100, 1101, 1102, 1104, 1110, 1114, 1118, 1121, 1128, 1137, 1139, 1152, 1156, 1159, 1162, 1164, 1181
Phi coefficient, 502
Phi distribution, 553, 554
Phi statistic, 522, 533, 534
Pi constant, 72, 73, 196, 251
Pie chart, 33, 54, 238, 898
Plot of functions, 19, 26, 178, 251, 253, 256
Plot of residuals, 46, 618, 619, 648, 724, 773, 774, 911
PNG files, 44, 51, 52, 91
Point biserial correlation, 15, 382, 387, 388, 389, 564, 565
Point identification, 60, 201, See Interactive data points
Point labels, 209, 226, 227, 239, 315, 934, 969
Point plot, 234, 253
Poisson distribution, 299, 400, 957, 960, 1173
Poisson regression, 34, 35, 36, 595, 596, 598, 718, 719, 721, 725, 726
Polar plot, 33, 59, 201, 216, 220, 241
Polynomial
Fitting, 205, 214
Neville's, 212, 213
Plots, 214, 254
Rational, 213
Regression, 15, 34, 35, 36, 60, 193, 214, 605, 606, 626, 627, 628, 648
Roots, 214, 254
Polynomial contrasts, 15, 767
Polytomous regression, 709, See Multinomial regression
Pooled variance, 365, 445, 530, 533
Positive likelihood ratio, 458, 698
Positive predictive value, 458, 697
Post hoc tests. See Multiple comparisons
Potency, 1087, 1089, 1096, 1099, 1100, 1101, 1102, 1103, 1110, 1117, 1120, 1121, 1134, 1135, 1142, 1145, 1149, 1151, 1152, 1153, 1159, 1160, 1161, 1162, 1163
Assigned, 1100, 1101, 1110, 1116, 1117, 1120, 1121, 1123, 1126, 1134, 1135, 1142, 1152, 1153, 1158
Combined, 1162, 1163
Estimated, 1110
Relative, 1100, 1135, 1151, 1152
Unweighted mean, 1163
Weighted mean, 1159, 1161, 1162, 1163, 1164
Power distance, 826

Power of the test, 527, 533, 536, 539, 542, 545, 552
Precision plot, 578, 581
Prediction, 193, 614, 615, 627, 647, 693, 694, 709, 718, 844, 868, 1061, 1072
  Cox regression, 1071, 1072
  Discriminant analysis, 844
  Principal components analysis, 868
Press
  Residuals, 616, 617
  Statistic, 612, 620, 624, 625, 781
Prevalance, 458, 459, 697, 698
Principal axis factoring, 872, 873, 874, 878, 879, 880
Principal components analysis, 825, 868, 870, 872, 873, 874, 875, 883
  Eigenvectors, 869
  Plot of eigenvalues, 870, 875
  Plot of principal components, 870
  Predicition, 868
  Principal components, 869, 870
Principal components factoring, 872, 874
Print, 54, 63, 71, 90
Printer setup dialogue, 64, 147, 148
Probit, 1147, 1149, 1153, 1156, 1157
Probit analysis, 663, 665
Probit link function, 664, 665
Probit scale, 79, 325, 619, 774, 1145, 1152
Process capability analysis, 983, 984, 987, 999, 1001, 1005
Profile plot, 217, 774
Projection matrix, 614
Projection plot, 85, 224, 226
Proportional hazards model. See Cox regression
Proportions
  Binomial, 355, 430, 431, 436, 437, 441, 443, 449, 491, 508
  Paired, 357, 358, 359, 360, 443, 449, 450, 451, 454, 455, 456, 457, 459, 491, 503, 508, 697
  Sample size, 544
  Unpaired, 357, 358, 359, 360, 443, 445, 447, 448, 449, 491, 508
Proximity matrix, 825, 826, 827, 829, 830, 850
Pseudo R-squared, 691, 1066
Publication bias, 578

Q

Quade two-way ANOVA, 480, 481, 797
Quadratic mean, 270, 288
Quality control, 730, 731, 932, 934, 939, 942, 945, 948, 950, 953, 955, 956, 957, 958, 960, 963, 965, 969, 973, 980, 985, 1179
Quantal response method, 60, 1087, 1145, 1153, 1156, 1157
Quantile, 33, 263, 268, 276, 279, 280, 281, 282, 283, 287, 332, 999, 1046, 1047, 1053
Quartile, 270, 271, 283, 285, 287, 316, 319, 322, 326, 332, 335, 1046
Quartimax rotation, 876

R

R chart, 936, 938, 939, 941, 945, 983, 1020, 1023
R x C table, 461, 488, 500
Random numbers, 194, 299, 300, 309, 310, 312
Randomised block designs, 741, 744, 745, 759, 1087, 1090, 1098, 1104, 1110, 1114
Range statistics, 118, 162, 168, 207, 228
Rank scores, 498
Rank(), 185
Rational function, 213
Receiver operating characteristic plot. See ROC plot
Reciprocal scale, 78, 200, 224
Recode, 95, 104, 162, 166, 167, 353, 609
Rectangular plot, 892, 893
Redo, 150
Reg(), 193, 212, 215, 229, 230, 614, 693, 718, 976, 979
Regression analysis, 30, 31, 35, 36, 60, 61, 522, 596, 601, 605, 609, 610, 618, 620, 623, 624, 627, 628, 633, 643, 649, 651, 655, 673, 676, 678, 703, 714, 724, 725, 726, 735, 780, 883, 1001, 1060
  Bivariate, 214, 812, 815
  Box-Cox, 730, 731, 733, 735, 1005, 1011, 1012
  Cox, 36, 37, 598, 1030, 1060, 1062, 1063, 1064, 1065, 1073, 1074, 1075, 1076, 1077, 1078
  Gompit, 34, 36, 595, 596, 663, 665, 667, 673, 676, 678, 684, 709, 1087
  Linear, 36, 205, 607, 627, 691, 730, 731, 735

Logistic, 15, 34, 35, 36, 595, 596, 598, 663, 684, 685, 687, 688, 689, 691, 694, 703, 709, 711, 718
Logit, 34, 36, 595, 596, 663, 665, 667, 673, 676, 678, 684, 709, 1087
Multinomial, 34, 35, 36, 595, 596, 598, 709, 711, 712, 714
Noconstant, 214, 228, 605, 606, 607, 684, 711, 719, 731
Nonlinear, 178, 596, 636, 637, 638, 639, 643, 645, 649, 651, 652, 654, 655, 658, 1183
Plane, 224, 228, 229
Poisson, 34, 35, 36, 595, 596, 598, 718, 719, 721, 725, 726
Polynomial, 15, 34, 35, 36, 60, 193, 214, 605, 606, 626, 627, 628, 648
Probit, 34, 36, 595, 596, 663, 665, 667, 673, 676, 678, 684, 709, 1087
Stepwise, 34, 35, 36, 595, 596, 598, 605, 606, 630, 631, 633, 634, 847
Surface, 85, 86, 193, 228, 229, 230
Weighted, 596, 610, 654
with replicates, 782, 812, 813, 814
Relative risk ratio, 709, 710, 712, 713, 714
Cross-tabulation, 510
Multinomial regression, 709, 710, 712, 713, 714
Reliability analysis, 826, 887, 888, 889
Cronbach's alpha, 888, 889, 890, 891
Item total statistics, 888, 890
Split analysis, 888, 889, 891
Repeatability, 1017, 1021, 1022, 1023
Repeated measures ANOVA, 595, 745, 746, 747, 748, 749, 750, 755, 757, 758, 759, 761, 762, 763, 766, 888
Reproducibility, 1017, 1021, 1022, 1023
Residual bar, 211, 228
Residuals
Adjusted, 500
ARIMA, 911
Cox-Snell, 1069, 1078
Deviance, 1069, 1078
Logit, 695
Martingale, 1069, 1078
Modified Cox-Snell, 1069, 1078
Normal plot of, 619, 774
Plot of, 46, 618, 619, 724, 773, 774, 911
Press (Deleted), 616
Score (Partial), 1069, 1078
Standard Error of, 615
Standardised, 500, 615, 694, 695, 723
Studentised, 616, 695, 796, 797, 799, 802
Ribbon chart, 33, 201, 216, 241, 247
Rich text, 45, 53, 57, 72
Ridit scores, 498
Risk difference, 560, 571, 574
Risk ratio, 446, 447, 508, 555, 560, 574, 583, 588
Robust confidence interval, 318, 322
ROC analysis, 15, 684, 688, 699
ROC curve, 686, 688, 699, 701
ROC plot, 702
Root mean square, 270, 288, 292, 614, 646, 650, 652, 654, 658
ROUND, 171, 186
Row Labels, 157, 209, 226, 227, 239, 827
Row(), 194
R-squared, 620, 624, 625, 634, 635, 650, 652, 654, 658, 748, 776, 781, 813, 814
Runs test, 417, 430, 431, 662

S

S chart, 936, 938, 939, 941, 942, 943
Sachs, L., 931, 1183
Sample size and power estimation
ANOVA, 548
Correlation, 538
One sample, 523
Phi statistic, 522, 533, 534
Proportions, 544
Regression, 522
Two samples, 529
Variance, 535
Sample statistics, 265, 268, 269, 271, 283, 285, 290, 292
Satterthwaite's approximation, 368
Scalar functions, 162, 167, 168, 177, 183, 196
Scale transformations, 325
Scatter diagram, 82, 235, 236, 268, 298, 835, 851
Scheffe, 795, 804
Score (Partial) Residuals, 1069, 1078
Scores, 491, 497, 501, 504, 505, 507
Modified ridit, 498
Rank, 498
Ridit, 498
Table, 498
Scott at al., 458, 698, 1184
Scree plot, 870, 874, 875

Select case. See Select row
Select row, 31, 40, 61, 104, 162, 172, 173, 193, 609, 614, 618, 693, 724, 1107, 1110, 1137, 1138
Select(), 193
Sensitivity, 457, 459, 663, 686, 687, 688, 696, 697, 698, 699, 700, 701, 702, 712, 720, 732, 1011, 1064
Sequence diagram, 297
Ser(), 167, 168, 184, 196
Shapiro-Wilk test, 397, 405, 406, 407, 1007, 1093, 1094, 1104, 1129
Sheppard's correction, 285, 289, 290
Shewhart charts, 936
Shift parameter, 414, 425
Short String, 100, 101, 122, 123, 161, 188, 189, 190
Short(), 189
Sign test, 426, 427
Single linkage, 828, 831, 839
Skewness, 271, 290, 292, 416
Skewness coefficient
  Pearson, 290, 292
Slope ratio method, 60, 1096, 1128, 1137, 1139, 1142
Slope test. See Heterogeneity of regression or Slope ratio method
Smoothing, 916
  Double exponential, 920
  Level, 918, 920, 923, 926, 928
  Seasonal, 923, 926, 928
  Trend, 920, 923, 926, 928
Somer's delta, 503, 505
Sort(), 164, 185
Spearman's rank correlation, 384
Spearman-Karber method, 15, 1146, 1149, 1150
Special functions, 188
Specificity, 457, 459, 663, 687, 688, 696, 697, 698, 699, 700, 701, 702
Spin plot, 59, 235
Spline, 204, 245
Split-plot designs, 754, 757, 759, 762, 763
Sprent, P., 402, 403, 1184
SQL, 140, 141, 145, 146, 147
Ssq(), 162, 167, 168, 184, 196, 605, 606, 658
Stack columns, 168, 742, 744, 746, 749, 752, 753, 755, 756, 757, 789, 798, 812, 821
Stacked bar chart, 244, 246
Stan(), 186
Stand-alone mode, 18, 31, 37, 39, 42, 43, 44, 45, 60, 61, 74, 91, 92, 94, 103, 105, 107, 201, 202, 207, 209, 212, 215, 224, 228, 229, 235, 307, 312, 397, 598, 601, 611, 614, 618, 693, 724, 734, 827, 976, 979, 988, 1008, 1013, 1017, 1066, 1083, 1089, 1107, 1110, 1137, 1138
Standardised alpha, 888, 889
Standardised coefficients, 851, 855
Standardised Delta-Beta, 617, 1070, 1078
Standardised DfFit, 617
Standardised fitted values, 615
Standardised mean difference, 561, 563, 565, 566, 567
Standardised proximity matrix, 826, 850
Standardised residuals, 500, 615, 694, 695, 723
Standardised SSP matrix, 868
Statistical functions, 39, 164, 171, 184, 397, 742, 744, 746, 749, 752, 753, 755, 756, 757, 789, 798, 812, 821, 976, 979, 1017, 1089
Std(), 162, 167, 168, 183, 196
Stem and leaf plot, 268, 296
Stephens, M.A., 1183
Stepwise discriminant analysis, 844, 846
Stepwise regression, 34, 35, 36, 595, 596, 598, 605, 606, 630, 631, 633, 634, 847
Stratified analysis
  2 x 2 tables, 361, 491, 496, 497, 509
  Survival analysis, 1030, 1061, 1063
String data, 21, 100, 101, 120, 121, 122, 123, 130, 132, 135, 140, 151, 160, 161, 188, 189, 190, 195, 263, 332, 350, 352, 354, 496, 520, 609, 685, 711
String variable, 29, 100, 101, 122, 129, 130, 131, 133, 135, 136, 140, 141, 164, 165, 166, 179, 326, 332, 352, 356, 359, 491, 597, 844
Stuart's tau c, 504
Student's t distribution, 299
Studentised residuals, 616, 695, 796, 797, 799, 802
Student-Newman-Keuls, 788
Student's t distribution, 385
Subsample selection, 104, 596, 610
Subscript, 53, 57, 58, 72
Sum of squares decomposition, 780
Sum(), 162, 167, 168, 184, 196

Summary statistics, 21, 27, 31, 32, 33, 168, 208, 263, 265, 268, 271, 272, 273, 283, 290, 349, 598
Superiority test, 439, 440, 508
Superscript, 53, 57, 58, 72
Surface fitting, 85, 86, 193, 228, 230
Surface plot, 86, 233, 253
Survival analysis, 30, 36, 1028, 1030, 1031, 1040, 1043, 1049, 1051, 1054, 1057, 1058, 1060, 1063, 1064, 1073, 1074, 1077
  Comparison statistics, 1030, 1054, 1057, 1058, 1059
  Cox regression, 36, 37, 598, 1030, 1060, 1062, 1063, 1064, 1065, 1073, 1074, 1075, 1076, 1077, 1078
  Kaplan-Meier analysis, 1043, 1044, 1049, 1051, 1065
  Life table, 1031, 1032, 1033, 1037, 1040, 1043, 1044, 1051, 1060, 1065
Survival function, 1028, 1031, 1033, 1038, 1043, 1044, 1046, 1047, 1048, 1049, 1051, 1053, 1054, 1060, 1065, 1068, 1072, 1078
Survivor function, 256, 258, 259
Sylk files, 140, 142, 156, 158
Symbols, 82

T

Tabachnick, B. G., 620, 820, 853, 870, 878, 885, 1184
Table of means, 169, 520, 759, 766, 772, 777, 782, 783, 785, 972
Table probability, 452, 503
Table scores, 498
Tables, 491
Tarone test, 511
Tetrachoric correlation, 456, 508
Text files, 129
  By column, 135, 136
  By row, 133, 134, 135, 136, 137, 143
  Delimited (.CSV), 103, 132, 133, 137, 142, 143, 151
  Field delimiter, 103, 133, 135, 151
  Fixed format, 137, 138, 139, 140, 155
  Free format, 103, 133, 136, 137, 155, 157
Text qualifier, 103, 134, 135, 151
Time variable, 29, 125, 129, 163, 164, 165, 759, 766, 1040, 1049, 1051, 1053, 1063, 1068, 1070, 1073, 1075
Tolerance, 632, 641, 668, 686, 712, 720, 732, 1011, 1064, 1146
Tools menu, 92
Torgerson, W. S., 862, 1184
Transpose, 163, 164, 483, 828
Trend line, 203, 893, 894
Triangular distribution, 299, 1171
Trigonometric functions, 181, 182, 251
True time axis, 202
t-test, 28, 29, 34, 349, 353, 362, 363, 364, 366, 367, 369, 370, 371, 372, 373, 377, 739, 971
  One sample, 363, 379
  Paired, 369, 370
  Pooled variance, 365
  Separate variance, 368
Tukey mean difference plot, 338
Tukey scores, 325
Tukey-B, 795, 802
Tukey-HSD, 462, 463, 465, 471, 475, 481, 788, 789, 795, 799, 800, 801, 802, 817
Two sample tests
  Chi-square test, 312, 330, 399
  F-test, 365, 372, 786
  Kolmogorov-Smirnov test, 403
  Mann-Whitney U test, 16, 34, 410, 412, 413
  Median test, 34, 410, 419, 420, 471, 472, 796
  Moses extreme reaction test, 34, 410, 418, 419
  Pooled variance t-test, 365
  Sample size and power estimation, 529
  Separate variance t-test, 368
  Wald-Wolfowitz runs test, 34, 416, 417, 431
  Walsh test, 428, 429
  Wilcoxon signed rank test, 421, 423, 424, 426, 428
Type I error, 457, 522, 525, 531, 535, 541, 546, 549, 696, 739, 952, 953, 972
Type II error, 457, 522, 525, 531, 535, 541, 549, 696, 953

U

U chart, 960
Unbalanced design, 759, 764, 766
Uncertainty coefficient, 506
Undo, 45, 118, 150, 172
Uniform distribution, 299, 401, 402
Unistat65.ini, 33, 37, 46, 47, 50, 109, 114, 141, 143, 407, 412, 414, 418, 422, 423,

611, 631, 632, 689, 739, 771, 772, 785, 791, 792, 796, 797, 806, 809, 845, 869, 990, 1065, 1081, 1093, 1099, 1101
Unstack columns, 168, 170, 171
US Pharmacopoeia, 15, 1159, 1163
User-defined log base, 77

V

Validity of Assay, 1096, 1099, 1123, 1126, 1130, 1148, 1158
Van der Waerden scores, 325
Var(), 162, 167, 168, 184, 196, 605
Variable control charts, 932, 936, 937, 939, 942, 945, 955, 956
  CUSUM chart, 936, 938, 939, 952
  Exponential weights moving average chart, 950
  Geometric moving average. See Exponential weights moving average
  Moving average chart, 936, 938, 939, 947, 948, 950, 953
  R chart, 936, 938, 939, 941, 945, 983
  S chart, 936, 938, 939, 941, 942, 943
  Variable selection, 936
  Variance chart, 936, 938, 943
  X bar chart, 936, 938, 939, 945, 947, 983
  X chart (Levey-Jennings), 937, 955
Variable selection dialogue, 26
Variance chart, 936, 938, 943
Varimax rotation, 876, 878, 880
Vector graphics, 90
View file, 137, 139
View point, 233, 236

W

Wadsworth, H. M., 662, 1184
Wald interval, 432, 439, 457, 688, 696
Wald statistic, 671, 689, 690, 713, 721, 1065, 1066, 1074, 1075
Wald-Wolfowitz runs test, 34, 416, 417, 431
Walsh averages, 425
Walsh test, 428, 429
Ward linkage, 828
Ward's linkage, 832
Web browser, 42, 44, 51, 61, 89, 91, 95, 96, 111, 119
Weibull, 300, 309, 977, 978, 979, 980, 1172
  Chart, 80, 980
  Distribution, 1172
Weibull distribution, 299
Weibull paper, 80, 979
Weighted regression, 596, 610, 654
Weights, 30, 407, 415, 489, 520, 596, 600, 610, 630, 637, 639, 640, 652, 653, 654, 658, 685, 686, 711, 719, 731, 739, 916, 917, 933, 947, 950, 1054
Welsch distance, 616
Wilcoxon signed rank test, 421, 423, 424, 426, 428
Wilkinson, L., 402, 408, 1180
Wilson (score) interval, 434
Winer, B. J., 746, 749, 795, 1184
Winter's seasonal forecasting, 923, 926
Word, 42, 43, 44, 48, 51, 52, 53, 72, 89, 90, 91, 95, 96, 97, 100, 111, 119
Wordpad, 42, 43, 45, 46
Working folder, 41, 149

X

X bar chart, 935, 936, 938, 939, 945, 947, 983, 1020, 1023
X chart (Levey-Jennings), 937, 955
XLSX files, 130
X-Y Plots, 16, 19, 20, 22, 27, 30, 33, 56, 57, 59, 61, 68, 75, 81, 82, 200, 206, 208, 216, 220, 223, 226, 227, 238, 240, 241, 242, 248, 255, 325, 618, 724, 774, 979, 1137
X-Y-Z grid plot, 232, 253
X-Y-Z scatter plot, 59, 85, 224

Y

Yates correction, 494, 501
Youden's index, 458, 687, 697, 699

Z

Zar, J. H., 385, 389, 393, 413, 420, 424, 427, 431, 453, 454, 465, 466, 472, 477, 483, 487, 492, 493, 522, 524, 525, 526, 527, 530, 532, 533, 534, 536, 537, 539, 540, 542, 543, 544, 546, 549, 550, 551, 552, 554, 601, 624, 633, 785, 789, 797, 799, 801, 804, 812, 818, 819, 1185
Zhang at al., 992, 1185
Zoom / unzoom, 54, 65, 66