



## **QED Statistics 1.1**

Copyright 2007, Pisces Conservation Ltd

# QED Statistics 1.1

Get to the heart of your data

---

*by Richard Seaby, Peter Henderson, John Prendergast & Robin Somes*

*QED Statistics offers a comprehensive range of statistics, chosen to meet the needs of all students, researchers and post-grads wanting to analyse quantitative data. The program holds your hand, from data input, through single sample stats, right up to General Linear Models.*

*QED Statistics is designed to give you confidence right from the start, and is unique in the level of help it offers. We help you:*

- \* To choose the right method.*
- \* To enter the data.*
- \* To explain and expand your results.*
- \* By showing step-by-step calculations of many of the methods.*
- \* By giving you extensive built-in help, demo data sets, worked examples, and animated guides.*

# Table of Contents

Foreword	I
<b>Part I Introduction</b>	<b>2</b>
1 System requirements and installation .....	2
2 Creating and opening a data set .....	3
Importing from Excel .....	4
Directly entering data .....	6
Data Entry Wizard .....	10
Opening a data set .....	10
No Working Data error .....	12
Entering Contingency table data .....	12
Maximum size of your data set .....	13
Saving edited data - Save and Save As .....	13
Editing existing data .....	13
3 Obtaining help .....	14
4 Common error messages .....	15
5 Checklist of data problems .....	17
6 Citation .....	17
7 References .....	18
<b>Part II Demonstration data sets</b>	<b>21</b>
<b>Part III The main window</b>	<b>24</b>
1 File drop-down menu .....	24
Export dialog .....	25
Print dialog .....	26
Printer setup dialog .....	26
2 Edit drop-down menu .....	27
Preferences - setup dialog .....	27
3 Simulation drop-down menu .....	28
Explore Distributions window .....	28
4 Single Sample drop-down menu .....	29
Mean .....	30
Median .....	31
Variance .....	31
Standard Deviation .....	31
Skewness .....	31
Kurtosis .....	32
Probability plot .....	32
Box and Whisker plot .....	33
Histogram plot .....	33
Testing for normality .....	34
Chi-squared test for normality - setup dialog.....	35
Chi-squared test for normality - results.....	36
Shapiro-Wilk test for non-normality - setup dialog.....	37
Shapiro-Wilk test - results.....	37
Lilliefors test for normality - setup dialog.....	38
Lilliefors test for normality - results.....	38
Single sample t-Test - setup dialog .....	38
Single sample t-Test - results .....	39

z Test - setup dialog .....	40
z Test - results.....	41
<b>5 Analysis of Frequency drop-down menu .....</b>	<b>42</b>
Fisher's Exact - setup dialog .....	42
Fisher's Exact - results.....	43
2 x 2 Contingency table - setup dialog .....	44
2 x 2 Contingency table - results.....	45
R x C Contingency table - setup dialog .....	46
R x C Contingency table - results.....	47
G-Test - setup dialog .....	48
G-Test - results.....	49
<b>6 Two samples drop-down menu .....</b>	<b>50</b>
t-Test (equal variance - balanced) - setup dialog .....	51
t-Test (equal variance - balanced) - results.....	51
t-Test (equal variance - unbalanced) - setup dialog .....	52
t-Test (equal variance - unbalanced) - results.....	53
t-Test (unequal variance) - setup dialog .....	54
t-Test (unequal variance) - results.....	55
Mann-Whitney two sample test - setup dialog .....	55
Mann-Whitney U - results.....	56
Two sample F Test - setup dialog .....	57
Two sample F Test - results.....	57
Paired t-Test - setup dialog .....	58
Paired t-Test - results.....	58
Wilcoxon matched pairs - setup dialog .....	59
Wilcoxon matched pairs - results.....	60
Two sample Chi-squared - setup dialog .....	60
Two sample Chi-squared - results.....	61
<b>7 Regression drop-down menu .....</b>	<b>62</b>
Pearson Correlation - setup dialog .....	62
Pearson Correlation - results.....	63
Kendall Correlation - setup dialog .....	64
Kendall Correlation - results.....	64
Spearman Rank Correlation - setup dialog .....	65
Spearman Rank Correlation - results.....	65
Linear Regression - setup dialog .....	66
Linear Regression - results.....	66
Multiple Linear Regression - setup dialog .....	68
Multiple Linear Regression - results.....	69
<b>8 ANOVAs drop-down menu .....</b>	<b>71</b>
1 way ANOVA - setup dialog .....	72
1 way ANOVA - results.....	73
1 way ANOVA repeated measures setup dialog .....	74
1 way ANOVA repeated measures results.....	76
2 way ANOVA - setup dialog .....	76
2 way ANOVA - results.....	77
Kruskal-Wallis - setup dialog .....	78
Kruskal-Wallis - results.....	79
<b>9 GLM drop-down menu .....</b>	<b>80</b>
GLM - setup dialog .....	80
GLM - results.....	82
<b>10 Planning drop-down menu .....</b>	<b>83</b>
Power Fisher's Exact - setup dialog .....	84
Power t-Test - unequal variances .....	85
Power t-Test - equal variances .....	86
Power of Correlation - setup dialog .....	86

11 Raw Data grid .....	87
12 Working Data grid .....	89
Data transformations and manipulations .....	90
Data transformations.....	91
Relativisations.....	92
Handling zeros.....	93
Transposing data.....	93
Saving the working data .....	94
13 Results tab .....	95
14 Expand tab .....	95
15 Explore tab .....	96
16 Summary tab .....	97
17 Help drop-down menu and Guides .....	98

## Part IV Single sample tests 100

1 Median .....	100
2 Mean .....	100
3 Variance .....	101
4 Standard Deviation .....	101
5 Skewness .....	102
6 Kurtosis .....	103
7 Probability plot .....	105
8 Box and Whisker plot .....	106
9 Histogram plot .....	107
10 Normality testing .....	107
Shapiro-Wilk test .....	108
Lilliefors test .....	108
Chi-squared test for normality .....	110
Normal distribution.....	111
Binomial distribution.....	111
Poisson distribution.....	112
Exponential distribution.....	113
11 t-Test : Comparing observations with a known mean .....	114
12 z Test : Comparing observations with a known mean .....	115

## Part V Analysis of Frequency 117

1 Fisher's Exact test .....	117
2 Contingency table Chi-squared test .....	118
Cramer's V .....	119
Contingency coefficient .....	119
3 Contingency table G-Test .....	120
4 Contingency table .....	120
5 Calculation of expected frequencies .....	121

## Part VI Two sample tests 123

1 t-Test: Comparing means of paired samples .....	123
2 t-Test: Comparing means of samples of the same size - equal variance .....	124
3 t-Test: Comparing means of samples of unequal size - equal variance .....	125

4	t-Test: Comparing means from samples with unequal variances .....	126
5	Testing for difference between two variances .....	127
6	Chi-squared two sample test .....	127
7	Mann-Whitney unpaired test .....	128
8	Wilcoxon paired-sample test .....	129
9	One- and two-tailed t-test .....	129

## **Part VII Regression and correlation 132**

1	Correlation coefficients .....	132
	Pearson Correlation .....	133
	Kendall's Correlation .....	133
	Spearman Rank Correlation .....	134
2	Linear Regression .....	135
	Is Linear Regression appropriate? .....	136
3	Multiple Linear Regression .....	137
	Stepwise Linear Regression .....	138
	Forward Stepwise Linear Regression.....	138
	Backward Stepwise Linear Regression.....	138
	Multicollinearity .....	139

## **Part VIII Analysis of Variance (ANOVA) 141**

1	One-way ANOVA .....	141
	One-way repeated measurements ANOVA .....	143
2	Homogeneity of variances test .....	143
3	Multiple comparison tests .....	144
	Tukey .....	144
	Scheffe .....	145
	Newman-Keuls test .....	145
	Tukey-Kramer .....	145
	Bonferroni .....	145
4	Two-way ANOVA .....	145
5	Fixed and random effects .....	147
6	Kruskal-Wallis test .....	147
7	Omega squared .....	148
8	An example one-way ANOVA .....	148
9	An example two-way ANOVA .....	150

## **Part IX General Linear Model 154**

1	Random and fixed effects .....	155
2	A simple ANOVA using a GLM .....	155
3	A simple linear regression using a GLM .....	156
4	Using more than 1 explanatory variable in a GLM .....	158
5	Using sequential and adjusted sums of squares .....	159
6	Combining continuous and categorical variables in a GLM .....	160
7	Studying interactions in a GLM .....	161
8	Fixed or categorical variables .....	163
9	Covariate variables .....	164
10	Coding categorical variables .....	164

---

Dummy coding .....	165
Effect coding .....	166
Orthogonal coding .....	167
<b>Part X Printing and saving results</b>	<b>169</b>
1 Exporting charts .....	169
2 Printing charts .....	169
3 Printing and exporting text and grid output .....	169
4 Preparing charts for output .....	169
<b>Index</b>	<b>172</b>

## Licence Agreement

### PISCES LICENSE AGREEMENT

This is a legal agreement between you the end user and PISCES Conservation Ltd. Lymington (PISCES). BY OPENING THIS PACKAGE YOU ARE AGREEING TO BE BOUND BY THE TERMS OF THIS AGREEMENT. IF YOU DO NOT AGREE TO THE TERMS OF THIS AGREEMENT PROMPTLY RETURN THE UNOPENED PACKAGE AND ALL ACCOMPANYING ITEMS (including written material) TO THE PLACE YOU OBTAINED THEM FOR A FULL REFUND.

1. GRANT OF LICENSE - This PISCES License Agreement ('License') permits you to use one copy of the PISCES software product acquired with this License (SOFTWARE) on any single computer, provided the SOFTWARE is in use on only one computer at any time. If you have multiple Licenses for the SOFTWARE then at any time, you may have as many copies of the SOFTWARE in use as you have Licenses. The SOFTWARE is 'in use' on a computer when it is loaded into the temporary memory (i.e. RAM) or installed into the permanent memory (e.g. hard disk, CD ROM, or other storage device) of that computer, except that a copy installed on a network server for the sole purpose of distribution to other computers is not 'in use'. If the anticipated number of users of the SOFTWARE will exceed the number of applicable Licenses then you must have a reasonable mechanism or process in place to assure that the number of persons using the SOFTWARE concurrently does not exceed the number of Licenses. If the SOFTWARE is permanently installed on the hard disk or other storage device of a computer (other than network server) and one person uses that computer more than 80% of the time it is in use then that person may also use the SOFTWARE on a portable or home computer.

2. COPYRIGHT - The SOFTWARE is owned by PISCES or its suppliers and is protected by all applicable national laws. Therefore, you must treat the SOFTWARE like any other copyrighted material (e.g. a book) except that if the software is not copy protected you may either (a) make one copy of the of the SOFTWARE solely for backup or archival purposes, or (b) transfer the SOFTWARE to a single hard disk provided you keep the original solely for backup or archival purpose. You may not copy the Product manual(s) or written materials accompanying the SOFTWARE.

3. DUAL MEDIA SOFTWARE - If the SOFTWARE package contains both 3-1/2" and 5-1/4" disks, then you may use only one set (either the 3-1/2" or 5-1/4") of the disks provided. You may not use the other disks on another computer or computer network, or lend, rent, lease, or transfer them to another user except as part of a transfer or other use expressly permitted by this PISCES License Agreement.

4. OTHER RESTRICTIONS - You may not rent or lease the SOFTWARE, but you may transfer your rights under this PISCES License Agreement on a permanent basis provided you transfer all copies of the SOFTWARE and all written materials, and the recipient agrees to the terms of this Agreement. You may not reverse engineer, decompile or disassemble the SOFTWARE. Any transfer must include the most recent update and all prior versions.

LIMITED WARRANTY - PISCES warrants that (a) the SOFTWARE will perform substantially in accordance with the accompanying Product Manual(s) for a period of 90 days from the date of receipt; and (b) any PISCES supplied hardware accompanying the SOFTWARE will be free from defects in materials and workmanship under nominal use and service for a period of one year from the date of receipt. Any implied warranties on the SOFTWARE and hardware are limited to 90 days and one (1) year respectively or the shortest period permitted by applicable law, whichever is greater. CUSTOMER REMEDIES - PISCES'S entire liability and your exclusive remedy shall be, at PISCES option, either (a) return of the price paid or (b) repair or replacement of the SOFTWARE or hardware that does not meet PISCES'S Limited Warranty, and which is returned to PISCES with a copy of your receipt. This Limited Warranty is void if failure of the SOFTWARE or hardware has resulted from accident, abuse or misapplication. Any replacement SOFTWARE will be warranted for the remainder of the original warranty period or 30 days, whichever is longer.

NO OTHER WARRANTIES - TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, PISCES DISCLAIMS ALL OTHER WARRANTIES. EITHER EXPRESS OR IMPLIED, INCLUDING BUT LIMITED NOT TO IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WITH RESPECT TO THE SOFTWARE, THE ACCOMPANYING PRODUCT MANUAL (S) AND WRITTEN MATERIALS, AND ANY ACCOMPANYING HARDWARE. THE LIMITED WARRANTY CONTAINED HEREIN GIVES YOU SPECIFIC LEGAL RIGHTS.

NO LIABILITY FOR CONSEQUENTIAL DAMAGES - TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW PISCES AND ITS SUPPLIERS SHALL NOT BE LIABLE FOR ANY OTHER DAMAGES WHATSOEVER (INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF BUSINESS PROFITS, BUSINESS INTERRUPTION, LOSS OF BUSINESS INFORMATION, OR OTHER PECUNIARY LOSS) ARISING OUT OF THE USE OF OR INABILITY TO USE THIS PISCES PRODUCT, EVEN IF PISCES HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. IN ANY CASE, PISCES'S ENTIRE LIABILITY UNDER ANY PROVISION OF THIS AGREEMENT SHALL BE LIMITED TO THE AMOUNT ACTUALLY PAID BY YOU FOR THE SOFTWARE.

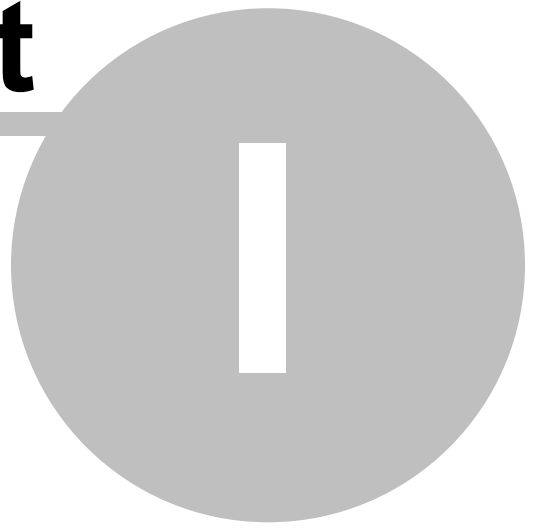
This Agreement is governed by the laws of England.

Should you have any questions concerning this Agreement, or if you desire to contact PISCES for any reason, please use the address information enclosed in this product to contact PISCES or write:

PISCES Conservation Ltd  
IRC House, The Square  
Pennington, Lymington  
Hampshire, England  
SO41 8GN  
Tel 01590 676622

# Part

---



# 1 Introduction

**QED Statistics** is a Windows program that offers all the standard statistical methods used in science and the social sciences. There are many statistical packages available, but most have not been designed for students with little statistical knowledge.

**QED Statistics** has been designed to be used by the novice, and particular care has been taken to fully explain the methods used and the meaning of the results obtained. The aim has been to produce a program which will meet most users' computational requirements in a highly supportive environment.

**QED Statistics** can act as a teaching aid for A level, high school and undergraduate students. It is also a powerful statistical system capable of undertaking the statistical analyses of most scientists.

**QED Statistics 1.0** was designed, developed and coded by Drs Peter Henderson and Richard Seaby. Testing was undertaken by Robin Somes and Claire Henderson.

To make using the program easy, several guides have been developed - see under **Help|Guides**.

For more information on [entering data](#)<sup>[3]</sup>, [obtaining Help](#)<sup>[14]</sup>, and the [main window](#)<sup>[24]</sup>.



## 1.1 System requirements and installation

### System requirements:

1. A PC running Windows XP or Vista.
2. 45 MB of hard disk space

QED Statistics does not limit the size of your data set, however your hardware will. To use QED Statistics with very large data sets (1000 or more species or samples) you will need a fast modern machine with 512 MB of RAM or more.

### Installation:

1. Place the QED Statistics CD in your CD drive: the installation process should begin automatically - follow the on-screen instructions.
2. If the CD does not auto-play, browse the CD in Windows Explorer or My Computer and click the file named Setup.exe in the root directory.
3. When installation is complete, there will be a QED entry under Start: Programs. An uninstall facility will also be created, in case you wish to remove the program. The program folder will be created by default in C:\Program Files\QED Statistics. A range of [demonstration data sets](#)<sup>[21]</sup> are installed with the program; they can be found in C:\My Documents\QED Statistics Data.

## 1.2 Creating and opening a data set

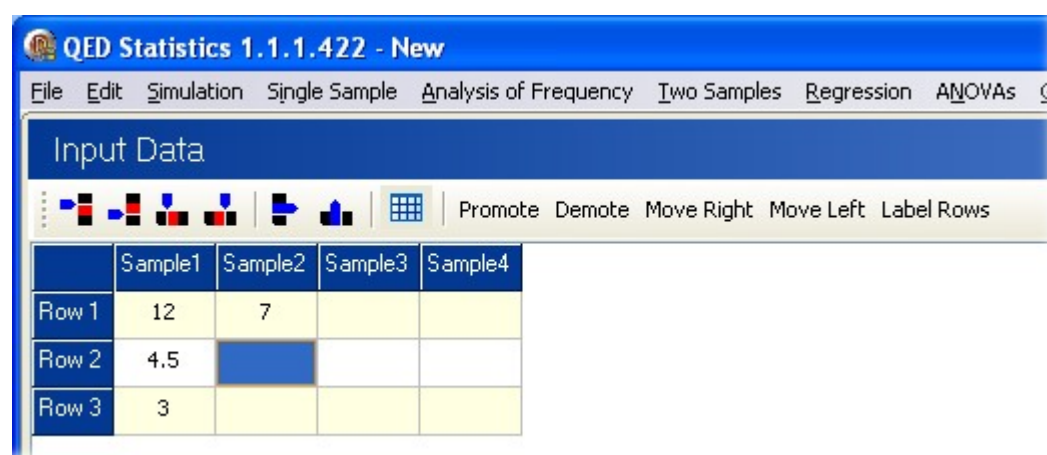
There are several ways you can enter your data into QED. The best method will depend on the size of your data set and the form in which it is presently held.

For large data sets it is best to organise your data using a spreadsheet program such as Excel. QED can directly import data from Excel files. See [Importing from Excel](#)<sup>[4]</sup>.

Small data sets can be conveniently typed directly into the data grid - see [Directly entering data](#)<sup>[6]</sup>, or created using the [Data Entry Wizard](#)<sup>[10]</sup>.

Data can also be copied and pasted into the data grid. Remember to click on the **Load the Data** button after the data is placed in the grid to make the data available to QED for analysis. When this button is activated the raw data is copied to the working data grid.

When data has been pasted in, the data grid tool bar is useful to promote the first row and column to become titles or to move the data across.



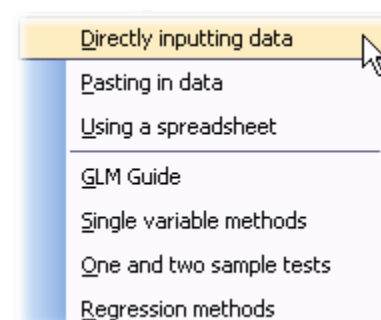
(For more details on the data grid tool bar, see [Directly entering data](#)<sup>[6]</sup>).

QED comes with many [demonstration data sets](#)<sup>[21]</sup> to show you how to organise your data. See [Entering contingency table data](#)<sup>[12]</sup> if you wish to enter frequency data.

There are no effective limits on the size of your data set - see [Maximum size of your data set](#)<sup>[13]</sup>.

See also [Editing existing data](#)<sup>[13]</sup>.

Remember that **Help|Guides** offers a range of guides to show you how to enter and organize your data.

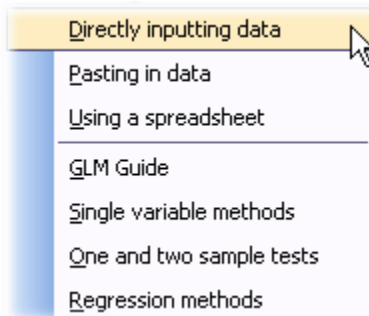


### 1.2.1 Importing from Excel

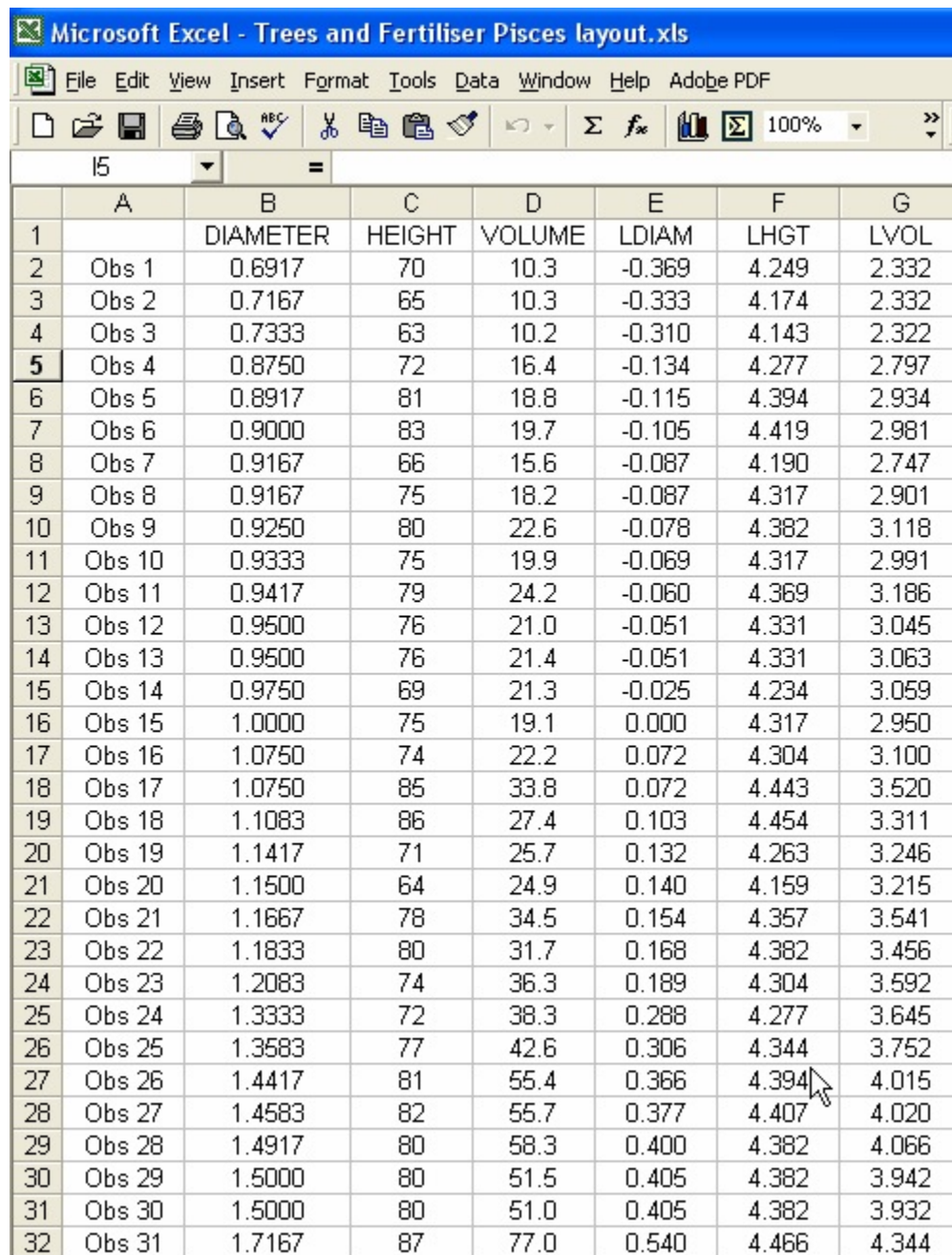
QED Statistics offers the ability to import data directly from a Microsoft Excel spreadsheet. It is important, however, that a number of points are observed:

1. If you are using a spreadsheet with multiple worksheets, the data will be imported from the worksheet that was open when the spreadsheet was last saved.
2. The data should be present as a contiguous rectangular block, starting at Cell A1.
3. Cell A1 itself should be empty, with names present in Row 1 and labels present in Column 1.
4. The import procedure ignores formulae in cells and imports the visible values.
5. Ensure that all cells rightwards and downwards from cell B2 contain numerical data.
6. QED Statistics can import directly from Excel whether Excel is open or not, provided the worksheet has been saved.

Watch **Help|Guides - Using a spreadsheet** to see how to enter data using a spreadsheet.

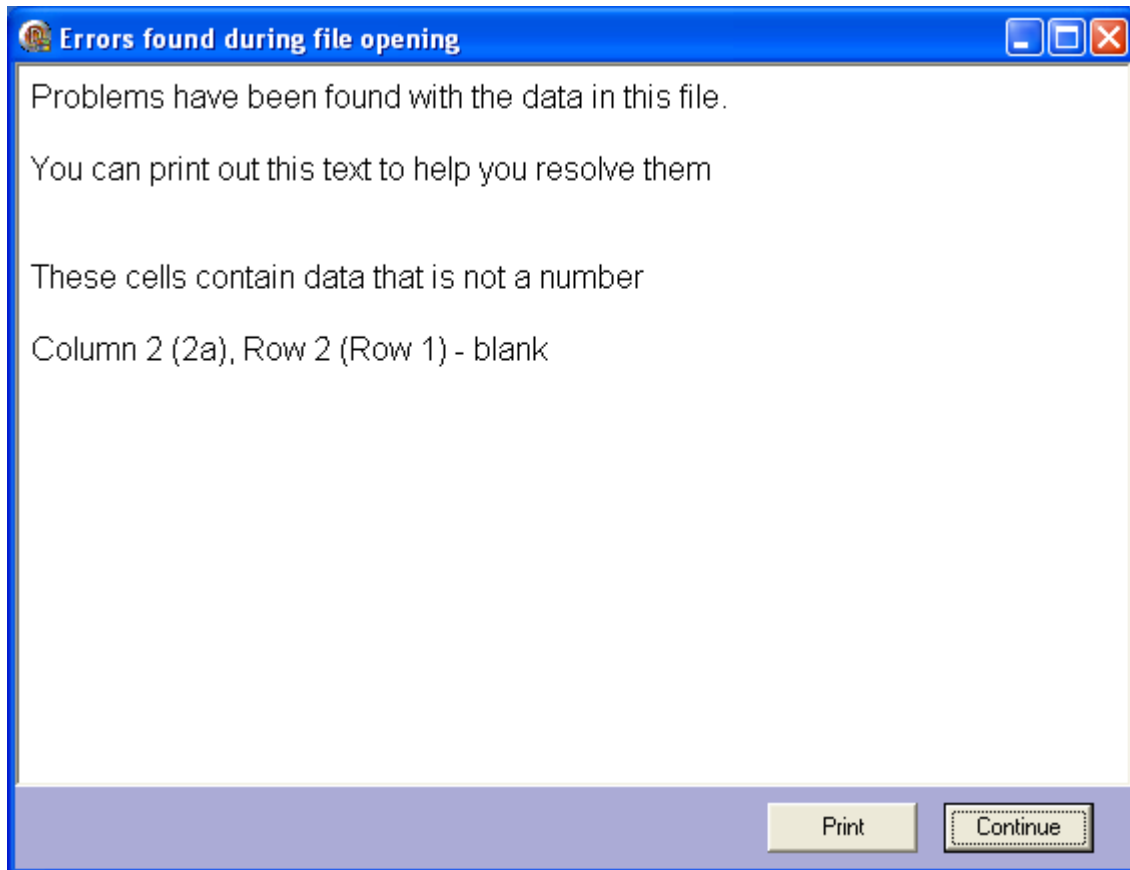


Here is an example of a data set in Excel ready to be imported into QED Statistics. There are 6 variables and 31 sets of observations. Note that cell A1 is empty



	A	B	C	D	E	F	G
1		DIAMETER	HEIGHT	VOLUME	LDIAM	LHGT	LVOL
2	Obs 1	0.6917	70	10.3	-0.369	4.249	2.332
3	Obs 2	0.7167	65	10.3	-0.333	4.174	2.332
4	Obs 3	0.7333	63	10.2	-0.310	4.143	2.322
5	Obs 4	0.8750	72	16.4	-0.134	4.277	2.797
6	Obs 5	0.8917	81	18.8	-0.115	4.394	2.934
7	Obs 6	0.9000	83	19.7	-0.105	4.419	2.981
8	Obs 7	0.9167	66	15.6	-0.087	4.190	2.747
9	Obs 8	0.9167	75	18.2	-0.087	4.317	2.901
10	Obs 9	0.9250	80	22.6	-0.078	4.382	3.118
11	Obs 10	0.9333	75	19.9	-0.069	4.317	2.991
12	Obs 11	0.9417	79	24.2	-0.060	4.369	3.186
13	Obs 12	0.9500	76	21.0	-0.051	4.331	3.045
14	Obs 13	0.9500	76	21.4	-0.051	4.331	3.063
15	Obs 14	0.9750	69	21.3	-0.025	4.234	3.059
16	Obs 15	1.0000	75	19.1	0.000	4.317	2.950
17	Obs 16	1.0750	74	22.2	0.072	4.304	3.100
18	Obs 17	1.0750	85	33.8	0.072	4.443	3.520
19	Obs 18	1.1083	86	27.4	0.103	4.454	3.311
20	Obs 19	1.1417	71	25.7	0.132	4.263	3.246
21	Obs 20	1.1500	64	24.9	0.140	4.159	3.215
22	Obs 21	1.1667	78	34.5	0.154	4.357	3.541
23	Obs 22	1.1833	80	31.7	0.168	4.382	3.456
24	Obs 23	1.2083	74	36.3	0.189	4.304	3.592
25	Obs 24	1.3333	72	38.3	0.288	4.277	3.645
26	Obs 25	1.3583	77	42.6	0.306	4.344	3.752
27	Obs 26	1.4417	81	55.4	0.366	4.394	4.015
28	Obs 27	1.4583	82	55.7	0.377	4.407	4.020
29	Obs 28	1.4917	80	58.3	0.400	4.382	4.066
30	Obs 29	1.5000	80	51.5	0.405	4.382	3.942
31	Obs 30	1.5000	80	51.0	0.405	4.382	3.932
32	Obs 31	1.7167	87	77.0	0.540	4.466	4.344

If there are no problems with the data set, QED Statistics will load the data into both the Raw Data and Working Data grids, and display the Working grid. If there are any problems - such as rows or columns that sum to zero - which might cause the calculations to malfunction, QED will normally alert you, and only load the data into the Raw Data grid. At this point, you will normally see an alert box, informing you of the problems found:



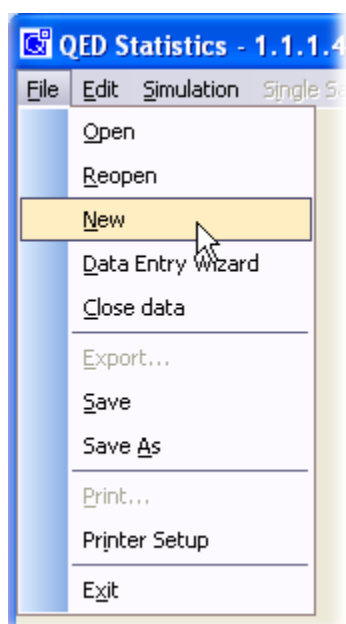
Sometimes it will require you to take some action, such as replacing a cell's contents, in which case, the cell(s) will be highlighted; otherwise it will make the required changes automatically. [Editing cell contents](#)<sup>[13]</sup> must be done on the Raw Data grid; when you have finished editing your data, you must then load the amended data set into the Working Data grid, by pressing the '**Load the Data**' button in the bottom right hand corner of the page. At this point the highlighting of the offending cells will disappear.

On the [Working Data grid](#)<sup>[89]</sup>, you can then use a wide range of [data transformations](#)<sup>[91]</sup>. QED Statistics will only make these changes to the [Working Data](#)<sup>[89]</sup> grid; the Raw Data will always contain the complete data array. The stored data file will not be altered unless you use **File|Save** to save the Raw Data grid, or **File|Export** to save the Working Data grid.

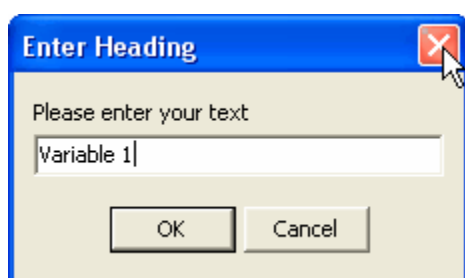
### 1.2.2 Directly entering data

Data sets can be created and edited within QED Statistics.

To create a new data set, select **File|New** from the drop-down menus (or alternatively, use the [Data Entry Wizard](#)<sup>[10]</sup> - refer to separate section):



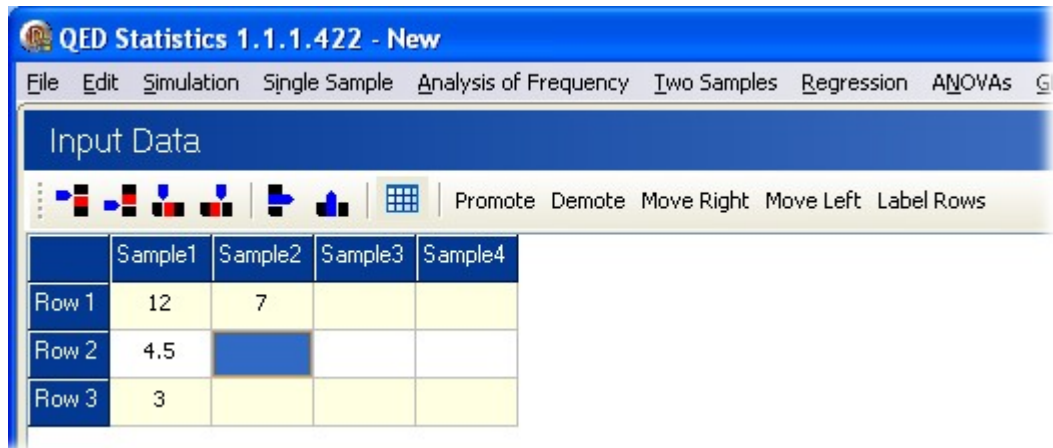
You are presented with a 3 x 3 grid in which the column and row headers (in dark blue) are designated to hold labels. It is almost always essential to have the columns labeled. It is not always essential to label the rows, but it can be useful. To enter column and row titles double-click on these cells and type into the dialog box that appears.



Leave the upper left-hand cell (Cell A1) empty. To input text or data into the grid, click into a cell and begin typing. Numbers can be either integer or real; some methods may require integers, but in most such cases the program will run with real data which will be automatically rounded.

The return key moves you sequentially through the grid. To type in a column of values just type a number into the top data cell of the column and then press **Enter** to move into the next cell down. To add a new row, select a cell in the bottom row of the data set, and press the Down arrow on your keyboard. To remove a row, click on a cell in that row, and press the Delete key on your keyboard. You should note that once a row has been deleted, the Undo function will not restore that row.

You can also add and delete rows and columns using the tool bar above the data grid. Just hover the cursor over the icons and QED Statistics will tell you the function.



From left to right, the tool bar buttons are:

**Insert row above selected**

**Insert row below selected**

**Insert column to left of selected**

**Insert column to right of selected**

**Delete selected row**

**Delete selected column**

**Resize grid**

**Promote** first data row to title row

**Demote** title row to first data row

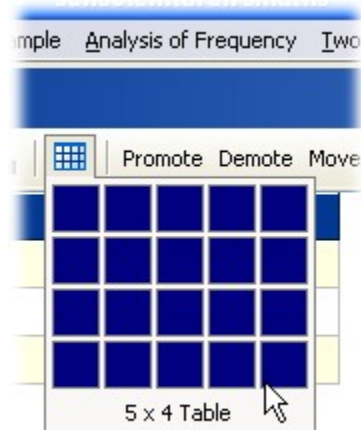
**Move Right** - Insert column to left of entire block of data (i.e. add a column of row header cells).

**Move Left** - Remove row header column from left of entire block of data

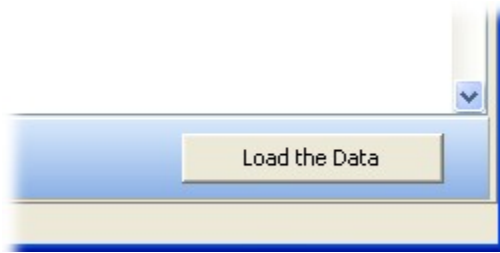
**Label Rows** - add labels (Row1, Row2, etc) to the row headers column.

**Label Columns** - add labels (Column1, Column2, etc) to the column headers row.

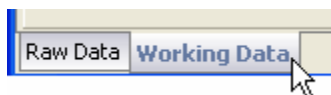
To set up a data grid of the required number of columns and rows, use the Resize Grid button (don't forget to add an extra column and row for the header cells); click and drag the cursor down and to the right to select the number of columns/rows:



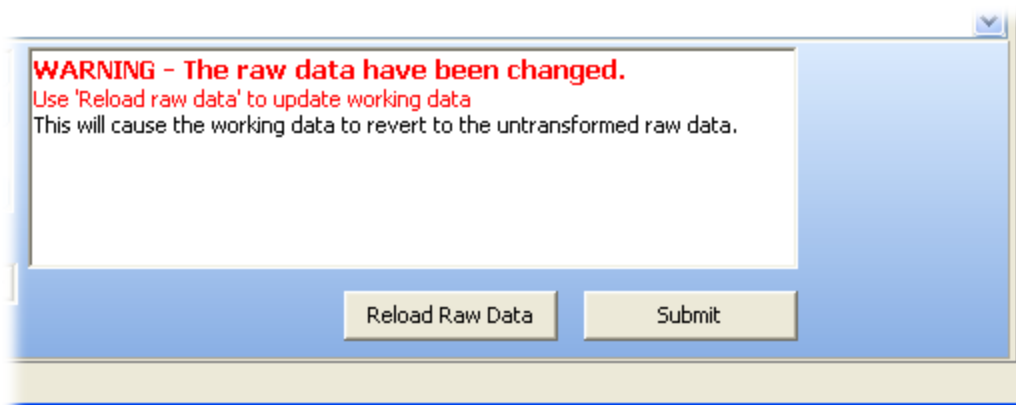
When you have entered your data, to make it available for analysis it must now be loaded into the working data grid. In the bottom right hand corner of the Raw Data tab, press the **'Load the Data'** button.



Alternatively, switch to the **Working Data** tab:

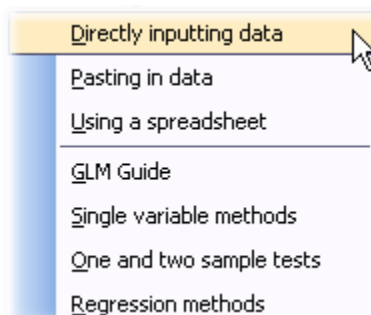


then load the data into the working grid by left clicking on the **Reload Raw Data** button.



When you have finished creating the data set, you can use File: Save or Save As to save it as a data file: see [Saving edited data or creating a new data file](#)<sup>[13]</sup>.

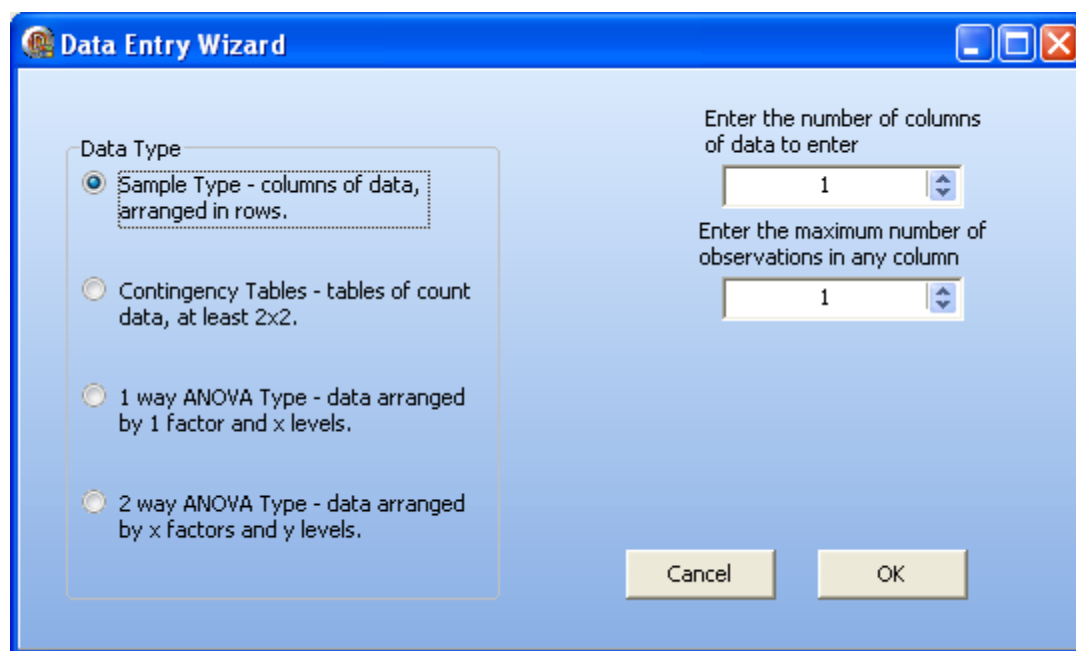
Run **Help|Guides - Directly inputting data** to watch how to enter data.



### 1.2.3 Data Entry Wizard

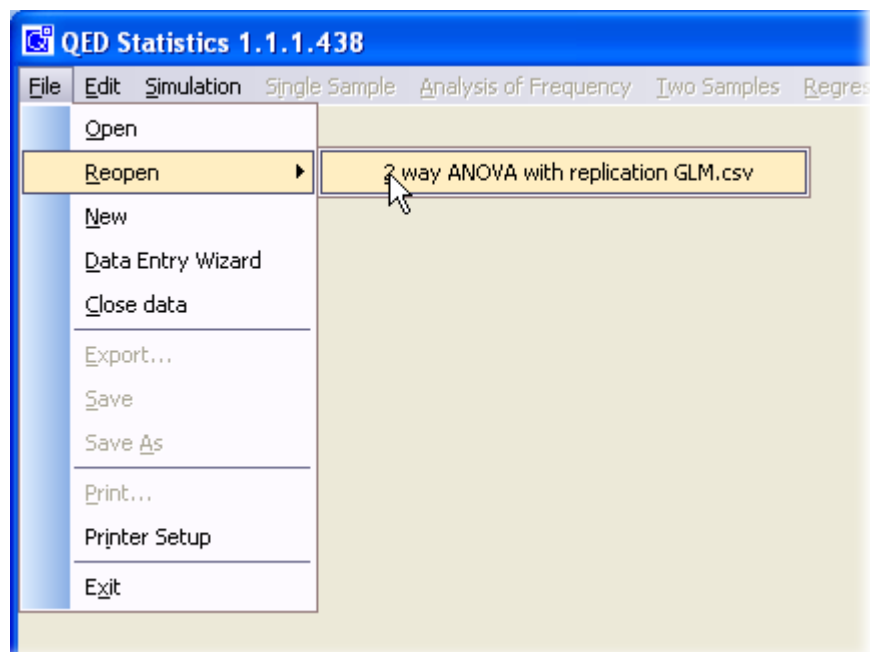
The Data Entry Wizard automatically creates a new data grid in the Raw Data grid of the proportions you specify, for 4 different types of data set:

- x by y array of columns and rows
- Contingency table, 2 x 2 or greater
- 1 way ANOVA, with one factor and x levels
- 2 way ANOVA with x factors and y levels



### 1.2.4 Opening a data set

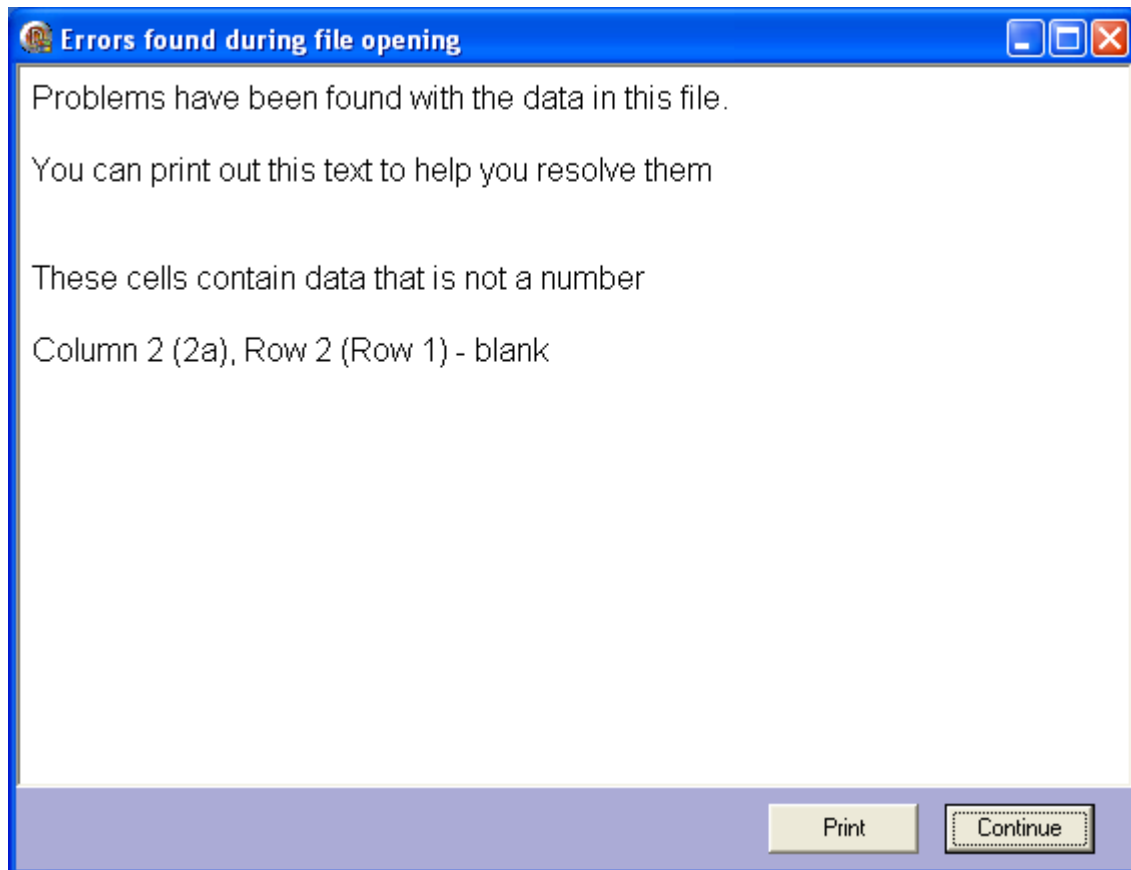
Use **File|Open** to start the file dialog to select a data file for analysis. To open a previously-used file, use **File|Reopen** to select the file to open:



QED Statistics's data files are in the Comma-Delimited Text format, with the file extension \*.csv. These are simple text files where the data in columns are separated by commas. This makes the data files easy to edit in a wide variety of spreadsheets and text editors, such as Excel, Lotus 1-2-3, Quattro Pro, MS Word, Wordpad or Notepad.

QED will also open Excel spreadsheet files directly - see [Importing from Excel](#)<sup>[4]</sup>.

If there are no problems with the data set, QED Statistics will load the data into both the [Raw Data](#)<sup>[87]</sup> and [Working Data](#)<sup>[89]</sup> grids, and display the Working grid. If there are any problems - such as rows or columns that sum to zero - which might cause the calculations to malfunction, QED will normally alert you, and only load the data into the Raw Data grid. At this point, you will normally see an alert box, informing you of the problems found:

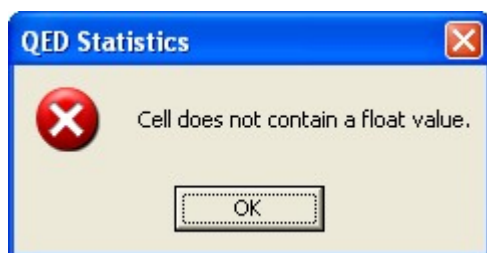


Sometimes it will require you to take some action, such as replacing a cell's contents, in which case, the cell(s) will be highlighted; otherwise it will make the required changes automatically. [Editing cell contents](#)<sup>[13]</sup> must be done on the Raw Data grid; when you have finished editing your data, you must then load the amended data set into the Working Data grid, by pressing the '**Load the Data**' button in the bottom right hand corner of the page. At this point the highlighting of the offending cells will disappear.

On the Working Data grid, you can then use a wide range of [data transformations](#)<sup>[91]</sup>. QED Statistics will only make these changes to the Working Data grid; the Raw Data will always contain the complete data array. The stored data file will not be altered unless you use **File|Save** to save the Raw Data grid, or **File|Export** to save the Working Data grid.

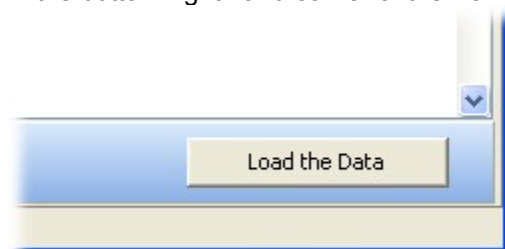
### 1.2.5 No Working Data error

If you create a new data set but forget to update the Working Data grid before trying to run an analysis, you will get the following dialogue box:

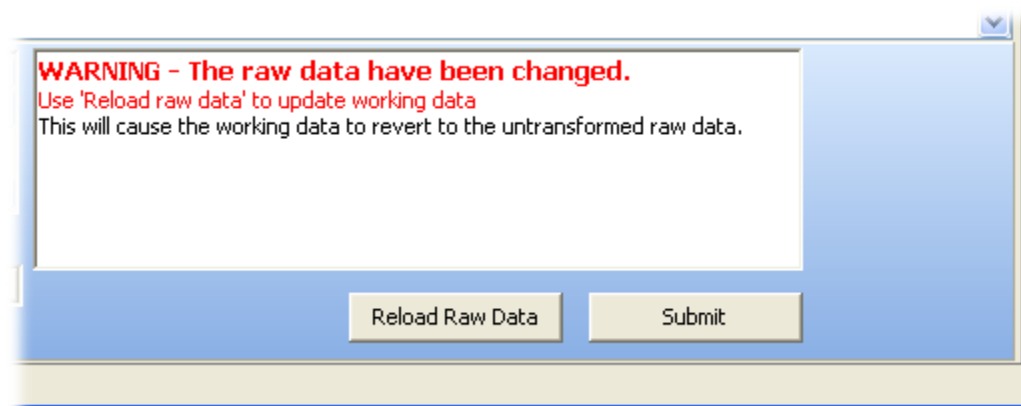


QED Statistics is telling you that it could not find a value in a cell where it was expected.

In the bottom right hand corner of the Raw Data tab, press the '**Load the Data**' button:



Alternatively, click on the Working Data tab and press on the **Reload Raw Data** button.



### 1.2.6 Entering Contingency table data

Data for analysis using a contingency table are entered as a 2 dimensional table. For example a standard 2 x 2 table will look like this:

Input Data		
	a	b
aa	1	4
bb	2	6

For an example of a 2 x 4 table of data see **2x4 contingency.csv**; more details under [Demonstration data sets](#)<sup>[21]</sup>.

Use the [Data Entry Wizard](#)<sup>[10]</sup> to create a contingency table of 2 x 2 or greater.

### 1.2.7 Maximum size of your data set

QED is programmed using dynamic data arrays. The program does not therefore set an upper limit on the size of the data sets it will handle. However, you will be limited by the memory of your computer and also by our ability to test the accuracy of the program.

There is no limit on the number of columns of data that can be entered. However, the program has not been tested rigorously with more than 256 columns of data.

All methods where it is appropriate will accept data sets of at least 1000 observations per column. Many will accept 2 columns with more than 3000 observations - for example linear regression. The program has only been rigorously test with data sets of up to 1000 observations per variable or treatment. It is known that most methods will run with 64,000 observations.

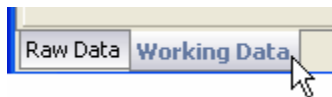
### 1.2.8 Saving edited data - Save and Save As

A data set can be copied and saved under a different name, from the [Raw Data grid](#)<sup>[87]</sup>, by selecting **File | Save As**.

Clicking **File|Save** will save your raw data and any changes you have made, over the original data file.

To save the [Working Data grid](#)<sup>[89]</sup>, which you would wish to do if the data set has been transformed, transposed or edited in some other way, do the following.

1. Click on the Working Data tab:



2. Select [File|Export](#)<sup>[25]</sup>, and choose the format you wish to save the data in.

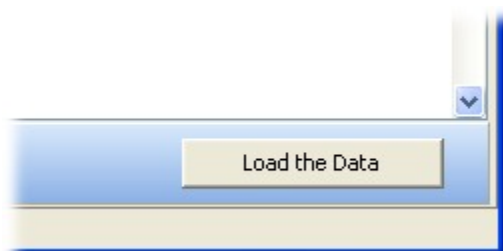
When **File | Exit** is selected to close QED Statistics, if a data set has been altered in any way, but not saved, you will be asked if you would like to save the data.

### 1.2.9 Editing existing data

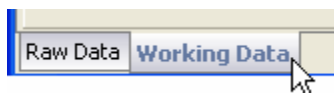
Data in the [Raw Data grid](#)<sup>[87]</sup> can be edited by using the mouse to click into a cell to select it, and typing in a new value.

Note: If you press the Delete key while a cell is selected, the entire row will be deleted. If you do wish to use the Delete key, then make sure that only the value in the cell is selected.

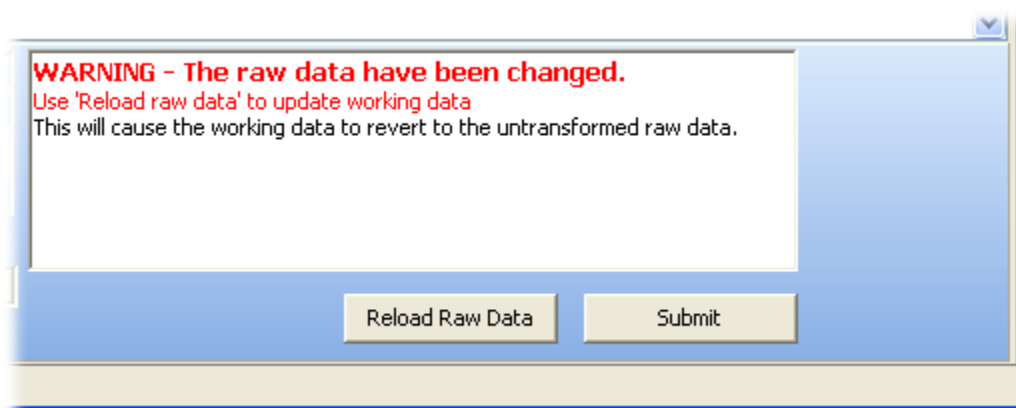
Changes made to the raw data will not alter a saved file until [File: Save](#)<sup>[13]</sup> is used. When you have finished editing your data, to make it available for analysis it must now be loaded into the [Working Data grid](#)<sup>[89]</sup>. In the bottom right hand corner of the Raw Data page, press the '**Load the Data**' button.



Alternatively, switch to the **Working Data** tab:



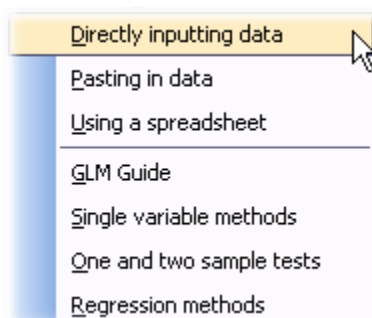
then load the data into the working grid by left clicking on the **Reload Raw Data** button:



QED Statistics will use the working data set thus created for all subsequent calculations. Note that output windows will not show calculations using the edited data until the methods have been re-run by selecting them in the normal fashion. The Working Data grid can also be edited using the [Transform](#)<sup>[91]</sup> functions. Changes made to the Working Data grid will not be transferred to the Raw Data grid. Transformed or otherwise changed working data can be saved from the Working Data grid as a new data set, using [File: Export](#)<sup>[25]</sup>.

## 1.3 Obtaining help

- For most active windows, context-sensitive help can be obtained by pressing **F1**, clicking on the Help button or selecting the Help drop-down menu. or clicking on the right-hand mouse button and choosing Help from the pop-up menu. If pressing F1, make sure that the window that you are seeking help for is the active one.
- To find out how to input data, choose the right method, and run most of the important methods, run a demonstration from the Help menu.
- Watch the **Help|Guides** to ensure you are setting up your data in the correct manner.



- If the program has displayed an error message, check the list of [Common error messages](#) <sup>15</sup>
- Work through the [Checklist of data problems](#) <sup>17</sup>
- Many software problems are transient, so try closing the program down and re-starting, to see if it recurs.
- Check on the [QED Statistics website](#), to see if there are any FAQs or announcements there relating to common problems.
- If you have problems using the program or entering data which you cannot solve then contact Pisces Conservation by e-mailing [pisces@irchouse.demon.co.uk](mailto:pisces@irchouse.demon.co.uk) or by phone +44 (0)1590 674000 during office hours (09.00 to 17.00 UK time). It will greatly help us to solve your problem if you can send us the data set which is causing the problem, and an exact description of the problem, the steps you took leading up to it, and any error messages displayed.

PISCES Conservation Ltd,  
IRC House, The Square  
Pennington, Lymington  
Hants, SO41 8GN  
UK

Telephone +44 (0) 1590 674000  
Fax +44 (0) 1590 675599

For details of our other software and e-books, visit our web site at [www.pisces-conservation.com](http://www.pisces-conservation.com)  
To buy software online, go to the [Pisces Conservation Shop](#)  
For details about our consultancy and other work, visit <http://www.irchouse.demon.co.uk>

## 1.4 Common error messages

When things go wrong, you may see a number of different error messages displayed by QED, either in the Results window, or as a pop-up message. These messages are explained below. You may also find it useful to work through the [Checklist of data problems](#) <sup>17</sup>.

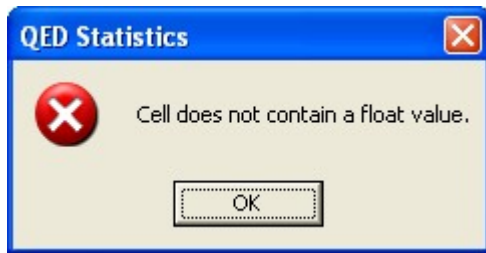
**NAN** - this stands for Not A Number, and will appear anywhere in the program where a calculation is impossible - for example, finding the square root of a negative number.

**+ve** and **-ve infinity**, which are when a number is infinitely large in either direction; usually caused by dividing by zero.

**"This file was not found"** - this error occurs when you have set the program to reload the last-used data file on start-up, and the file has subsequently been moved, deleted or renamed. To prevent this happening again, untick the "Always load last-used data file at startup" box on the

[Preferences dialog](#) <sup>27</sup>.

### "Cell does not contain a float value"



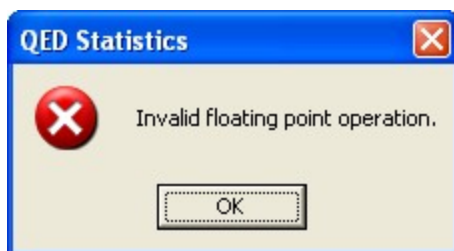
The most common cause is creating a new data set, and omitting to transfer the Raw Data into the Working Data Grid. Click OK to cancel the error message, then return to the Raw Data grid, check that you had finished entering the data, and press the 'Load the Data' button. See [No Working Data error](#) <sup>12</sup>.

### "Range check error"



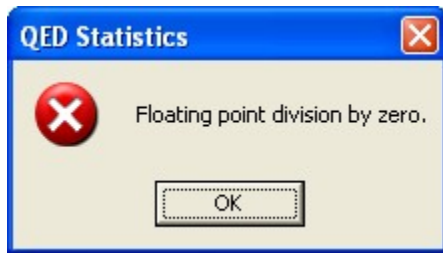
This can occur if you create a new, blank data set, and attempt to load it in to the Working Data grid. Ensure that the new data set is populated with data before transferring it to Working Data.

### "Invalid floating point operation"



This generally indicates a division by zero, or a similar mathematical impossibility; often caused by having a row or column in the data set which sums to 0.

### "Floating point division by zero"



Again, this is often caused by having a row or column in the data set which sums to 0.

## 1.5 Checklist of data problems

Here is a list of issues which can cause problems with the running of the program, or give unexpected results. We suggest that you run through the list to eliminate simple causes like these, before contacting us for support.

1. Many software problems are transient, so first, try closing the program and restarting it.
2. If you are opening an [Excel spreadsheet](#)<sup>[4]</sup>, have any recent changes to the data set been saved in Excel, and was your data set on the active worksheet (the visible one) when the file was saved?
3. Are all the columns labelled, with no duplicates? They should be.
4. Is cell A1 (the top left-hand corner of the data set) blank? It should be.
5. Have you got any columns or rows in the data set which you had not intended? For instance, was a row of column totals, means or standard deviations added in the original spreadsheet?
6. Are there any blank cells, or cells containing non-numerical data, in the data set? There should not be, apart from the column and row headers.
7. Are there any values (or even a non-numerical character, or a space) entered into cells outside the main block of data? The easiest way to do this is to open the data set in a spreadsheet, select the first 10 or so columns to the right of the data, and the first 10 rows below it, and press the Delete key, to clear any unwanted cell contents. Then save the data set and try again.
8. Do your data have no variability - i.e. are all the numbers the same?
9. Are the data sets perfectly correlated - either positively or negatively? This will occur if one variable is a simple factor of another.
10. Is your data set in the right format for the analysis you want to perform? Do you have the correct number of columns and rows? If in doubt, use the [Data Entry Wizard](#)<sup>[10]</sup>.
11. Have you got real numbers (i.e. 2.35, 1.796), when the analysis requires integer data (1, 2, 3)?
12. If the analysis depends upon having [fixed categorical](#)<sup>[163]</sup> or [covariate](#)<sup>[164]</sup> variables, are you sure that the range of data is consistent with this? For instance, a fixed categorical variable must run from 1 to  $n$ .
13. Is the data set excessively skewed, or in some other way out of the ordinary? Check [Skewness](#)<sup>[102]</sup>, [Kurtosis](#)<sup>[103]</sup> and the [Histogram plot](#)<sup>[107]</sup>.
14. If you are comparing two or more samples, have you considered the possibility that their [variances](#)<sup>[101]</sup> might be equal?
15. If you are performing a [Multiple Linear Regression](#)<sup>[137]</sup>, have you checked for [multicollinearity](#)<sup>[139]</sup>? This is when two (or more) variables are not truly independent, but can be expressed as a function of each other. One or more redundant variables from a set of directly-related variables should be removed.

## 1.6 Citation

For publication purposes this program should be cited as follows:

QED Statistics, Version 1.1, 2007, Pisces Conservation Ltd. Lymington, UK  
([www.pisces-conservation.com](http://www.pisces-conservation.com))

or alternatively, if you prefer...

Henderson, P.A. and Seaby R. H. M. (2007). QED Statistics 1.1, Pisces Conservation Ltd, Lymington, UK.

## 1.7 References

**The following title has been used for many of the example data sets used for QED Statistics.**

Grafen, A. and Hails, R. (2002) Modern Statistics for the life Sciences. Oxford University Press, Oxford.

**The following is a list of introductory statistics text books:**

**Introduction to the Practice of Statistics** Moore and McCabe; 1993; New York: W.H. Freeman and Company; 2nd Edition

**The Basic Practice of Statistics** Moore; 1995; New York: W.H. Freeman and Company

**Statistics** Freedman, Pisani, Purves and Adhikari; 1991; New York: W.W Norton and Company; 2nd Edition

**Understandable Statistics** Brase and Brase; 1995; Lexington, Massachusetts: D.C. Heath and Company; 5th edition

**Statistics: A First Course** Sanders; 1995; New York: McGraw-Hill, INC; 5th Edition

**Statistics in a World of Applications** Khazanie; 1996; New York: HarperCollins College Publishers; 4th Edition

**Statistics: A First Course** Freund and Simon; 1995; Englewood Cliffs, New Jersey: Prentice Hall; 6th Edition

**Statistics: Principles and Methods** Johnson and Bhattacharyya; 1996; New York: John Wiley and Sons, Inc.; 3rd Edition

**Introductory Statistics** Wonnacott and Wonnacott; 1990; New York: John Wiley and Sons, Inc.; 5th Edition

**The New Statistical Analysis of Data** Anderson and Finn; 1996; New York: Springer

**Statistics: An Introduction** Mason, Lind, and Marchal; 1994; Fort Worth: Saunders College Publishing; 4th Edition

**Statistics** McClave, Dietrich, and Sincich; 1997; Upper Saddle River, NJ: Prentice Hall; 7th Edition

**Introductory Statistics** Ross; 1996; New York: McGraw-Hill Companies

**Introduction to Probability and Statistics** Mendenhall and Beaver; 1994; Wadsworth Publishing; 9th Edition

**A Data-Based Approach to Statistics** Iman; 1994; Wadsworth Publishing

**Statistics: Learning in the Presence of Variation** Wardrop; 1993; Mosby-Year Book Inc.

**Statistical Methods** Freund and Wilson; 1996; Academic Press; Revised Edition

**Statistics: The Exploration and Analysis of Data** Devore and Peck; 1993; Wadsworth Publishing; 2nd Edition

**Contemporary Statistics: A Computer Approach** Gordon and Gordon; 1994; McGraw-Hill

**Statistics and Probability and their Applications** Brockett and Levine; 1985; Saunders Publishing

**Introductory Statistics** Devore and Peck; 1994; West Publishing; 2nd Edition  
**Statistics and Probability in Modern Life** Newmark; 1992; Fort Worth: Saunders College Publishing; 5th Edition

**Statistics: Concepts and Applications** Aczel; 1995; Chicago: Richard D. Irwin, Inc.

**Statistics and Data Analysis: An Introduction** Siegel and Morgan; 1996; New York: John Wiley and Sons; 2nd Edition

**A First Course in Statistics** McClave and Sincich; 1997; Upper Saddle River, NJ: Prentice Hall; 6th Edition

**An Introduction to Statistical Methods and Data Analysis** Ott; 1993; Belmont, CA: Duxbury Press; 4th Edition

**Statistics: The Conceptual Approach** Iversen and Gergen; 1997; New York: Springer-Verlag

**Introduction to Statistics** Milton, McTeer, and Corbet; 1997; New York: McGraw-Hill

**Introduction to Statistical Reasoning** Smith; 1998; Boston: WCB McGraw-Hill

**Understanding Data: Principles and Practice of Statistics** Griffiths, Stirling, and Weldon; 1998; Brisbane: John Wiley and Sons

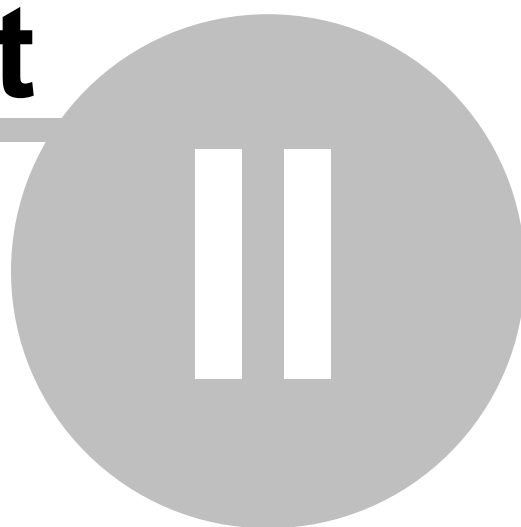
**Introductory Statistics** Mann; 1995; New York: John Wiley and Sons; 2nd Edition

**Applied Statistics: A First Course in Inference** Graybill, Iyer, and Burdick; 1998; Upper Saddle River, NJ: Prentice Hall

**Understanding Statistics** Naiman, Rosenfeld, and Zirkel; 1996; New York: McGraw-Hill; 4th Edition

# Part

---



## 2 Demonstration data sets

QED is supplied with a range of demonstration data sets which will show you how to organise your data for the different types of analysis. Demonstration data sets are stored by default in the folder C:\My Documents\QED Statistics Data.

### *Single and two sample tests*

The data set **1 way ANOVA rabbit ticks SFp208.csv** shows how data should be arranged if each column (representing a treatment or variable) is to be analysed.

### *Analysis of frequency*

**2x4 contingency.csv** holds a data set for a standard contingency table analysis. The data set describes the frequency of hair colour (Black, brown, blond and red) for boys and girls.

### *Regression Analysis*

**tree.csv** - This data set demonstrates a simple linear regression. Volume is the dependent variable and height the independent variable. This data set can also be analysed using a GLM; see [A simple linear regression using a GLM](#)<sup>[156]</sup>.

**school maths.csv** - This data set, with one dependent and 2 explanatory variables, demonstrates multivariate regression. The same data set can also be used by the General Linear Model method see [using more than 1 explanatory variable in a GLM](#)<sup>[158]</sup>.

### *One-way Analysis of Variance*

**1 way ANOVA rabbit ticks SFp208.csv** - This data set comprises the width of the scutum of larval ticks in samples taken from 4 cottontail rabbits. See [an example one-way ANOVA](#)<sup>[148]</sup>.

A one-way ANOVA can be done using either the ANOVA method or a General Linear Model method. The data is arranged differently for each method. The fertiliser example has been organised for both methods.

Open:

**1 way ANOVA fertiliser GH.csv** to run the analysis using a conventional ANOVA

**fertiliser GLM 1 factor.csv** to run the analysis using a GLM. The output is discussed under a [simple ANOVA using a GLM](#)<sup>[155]</sup>.

### *Two-way Analysis of Variance*

**2 way ANOVA with replication SFp302.csv** - See [an example two-way ANOVA](#)<sup>[150]</sup> for a discussion of this data set. The same data set coded for analysis using a General Linear model is in the file **2 way ANOVA with replication GLM.csv**

### *General Linear Model Examples*

For [A simple ANOVA using a GLM](#)<sup>[155]</sup> open **fertiliser GLM 1 factor.csv**.

For [A simple linear regression using a GLM](#)<sup>[156]</sup> open **tree.csv**.

For [Using more than 1 explanatory variable in a GLM](#)<sup>[158]</sup> open **school maths.csv**. This data set has 2 continuous explanatory variables.

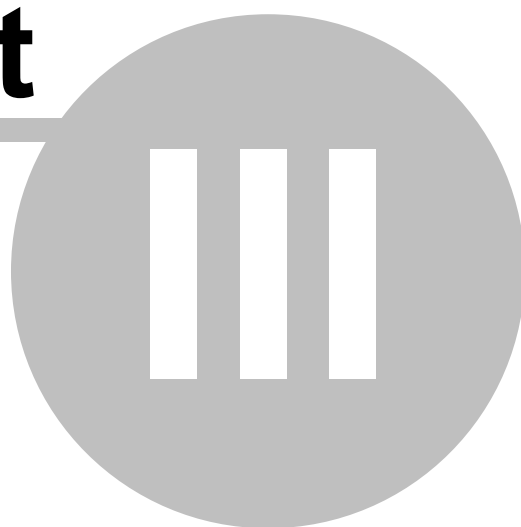
For [Combining continuous and categorical variables in a GLM](#)<sup>[160]</sup> open **fat.csv**.

For 3 fixed variables and [Studying interactions in a GLM](#)<sup>[161]</sup> open **tulip.csv**

A set of data for a Two-Way Analysis of Variance with replicates coded for use with a GLM is given in **2 way ANOVA with replication SFp302.csv**. The same data for use with the ANOVA method is available in **2 way ANOVA with replication SFp302.csv** - see [an example two-way ANOVA](#)<sup>[150]</sup> for a discussion of this data set.

# Part

---



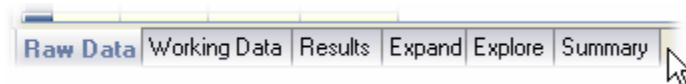
### 3 The main window

The top bar of the main window offers a number of drop-down menus.



[File](#)<sup>[24]</sup> - For opening, saving, printing and exporting data.  
[Edit](#)<sup>[27]</sup> - Copying, pasting and changing user preferences.  
[Simulation](#)<sup>[28]</sup> - To make and explore distributions.  
[Single sample](#)<sup>[29]</sup> - To examine, summarize and test a single variable.  
[Analysis of Frequency](#)<sup>[42]</sup> - Analysing contingency tables  
[Two Samples](#)<sup>[50]</sup> - Comparing two samples.  
[Regression](#)<sup>[62]</sup> - Fitting straight lines.  
[ANOVAs](#)<sup>[71]</sup> - One- and two-way analysis of variance  
[GLM](#)<sup>[80]</sup> - General linear models  
[Planning](#)<sup>[83]</sup> - Choosing the number of samples  
[Help](#)<sup>[98]</sup> - Run demos and get help.

The data and output of QED is organised under a number of tabbed sheets. The tabs which are displayed will depend on the methods you have used. Click on a tab to open the sheet.



The main tabs which are usually present, after a data set has been opened and a test run, are as follows.

[Raw Data](#)<sup>[87]</sup> - this sheet shows the current data set.  
[Working Data](#)<sup>[89]</sup> - shows the working data which may be an altered version of the working data.  
 Use this sheet to edit or transform the data.  
[Results](#)<sup>[95]</sup> - This sheet will present the results of your selected analysis in a grid.  
[Expand](#)<sup>[95]</sup> - Gives more details of how the calculation was carried out.  
[Explore](#)<sup>[96]</sup> - Takes you through the calculation steps.  
[Summary](#)<sup>[97]</sup> - shows the summary statistics of the data set; this tab is not normally shown by default, but can be select from the Single Sample menu; see [Summary tab](#)<sup>[97]</sup>.

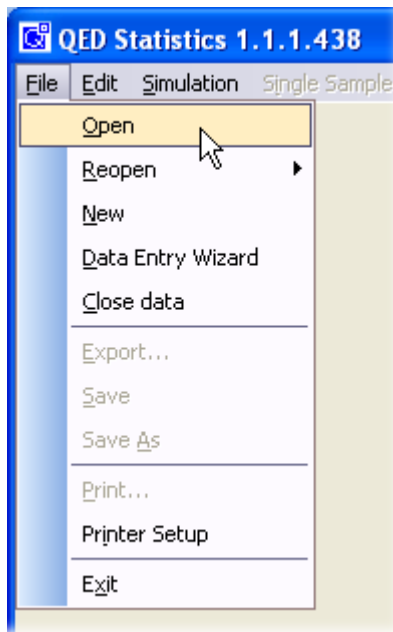
#### 3.1 File drop-down menu

This menu offers the standard Windows file menu. Choose:

**Open** - to open an existing file.  
**Reopen** - to open previously-used files  
**New** - to create a new data grid for data input.  
**Data Entry Wizard** - automatically creates a grid of the correct format for various types of data set  
**Close** - to close the active data set.  
[Export](#)<sup>[25]</sup> - to save the active grid in a variety of formats.  
**Save** - to save the open data set.  
**Save As** - to save the open data set under a new name.  
[Print](#)<sup>[26]</sup> - to print the active grid.  
[Printer Setup](#)<sup>[26]</sup> - Use this dialog to select a printer, choose settings and page orientation.

**Exit** - to close the program.

If you wish to change the number of recently-used files shown on the Reopen menu, you can do so under Edit: [Preferences](#) <sup>27</sup>.



### 3.1.1 Export dialog

The Export dialog offers a number of different formats in which to save the active grid. The active grid is the table you were looking at when you selected **File|Export**. It therefore could be the results of a test, or the working data.



Choose:

**CSV** to save as a standard comma delimited file. This format can be opened by QED Statistics, as well as many spreadsheets, word processors etc.

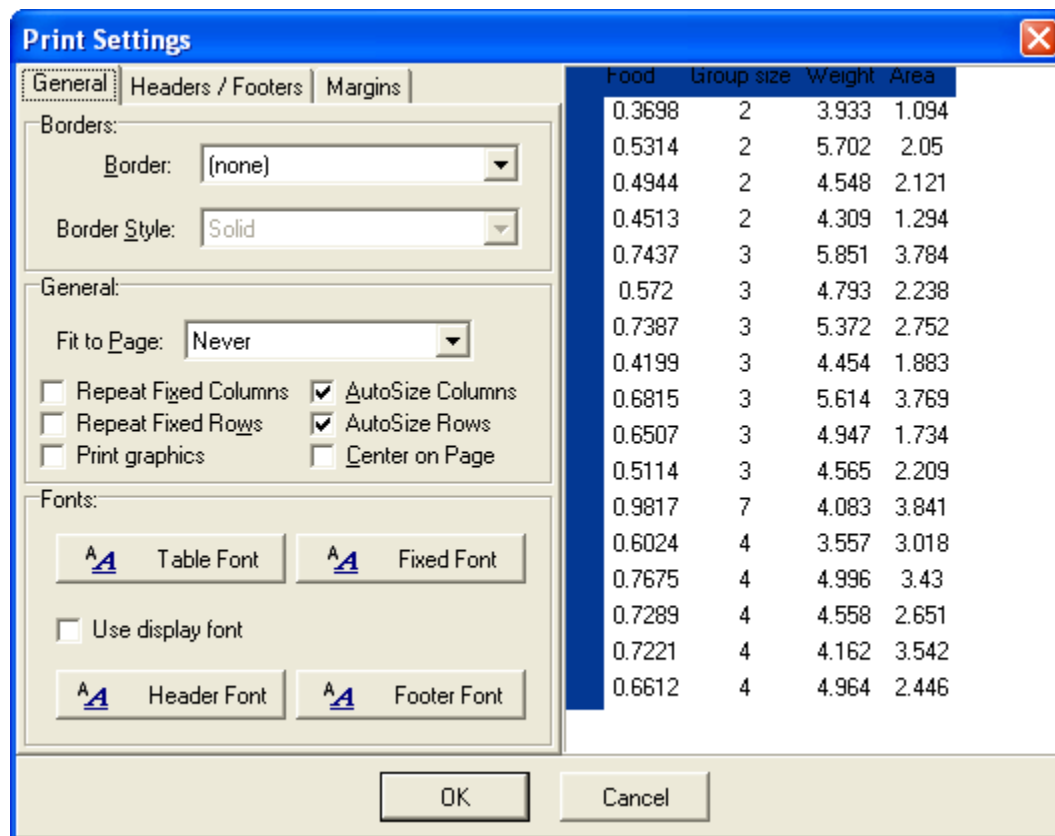
**ASCII** to save as a simple text file. This format can be opened by almost all text editors, as well as spreadsheets and many other programs.

**XLS** to save as an Excel file

**HTML** to save as an HTML file. This will save the data in an HTML-formatted table for use on websites etc.

### 3.1.2 Print dialog

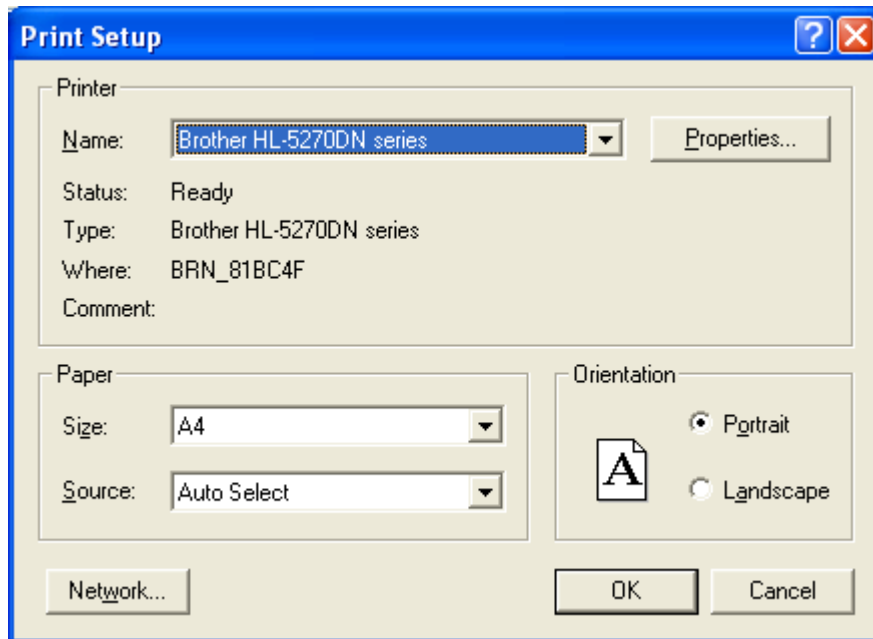
When **File|Print** is selected a Print Settings dialog window opens. This offers a wide range of options to format and print the data in the active grid. The panel on the right shows the data that will be printed. Select tabs and buttons to change margins and fonts, alter margins and add headers and footers.



See also [Printer setup dialog](#) <sup>26</sup>.

### 3.1.3 Printer setup dialog

Use **File|Printer Setup** to select your printer and its properties. The size of paper and its orientation are chosen here.



## 3.2 Edit drop-down menu

This menu offers the standard Windows edit menu plus preferences. Choose:

**Copy** - to copy the active grid or selected text to the Windows clipboard.

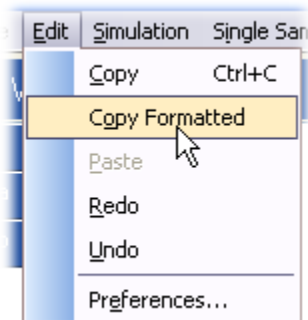
**Copy formatted** - copies to the Windows clipboard keeping the format to allow the active grid to be exported to Excel or Word.

**Paste** - to paste from the clipboard to the selected data grid.

**Redo** - to redo an action that was undone

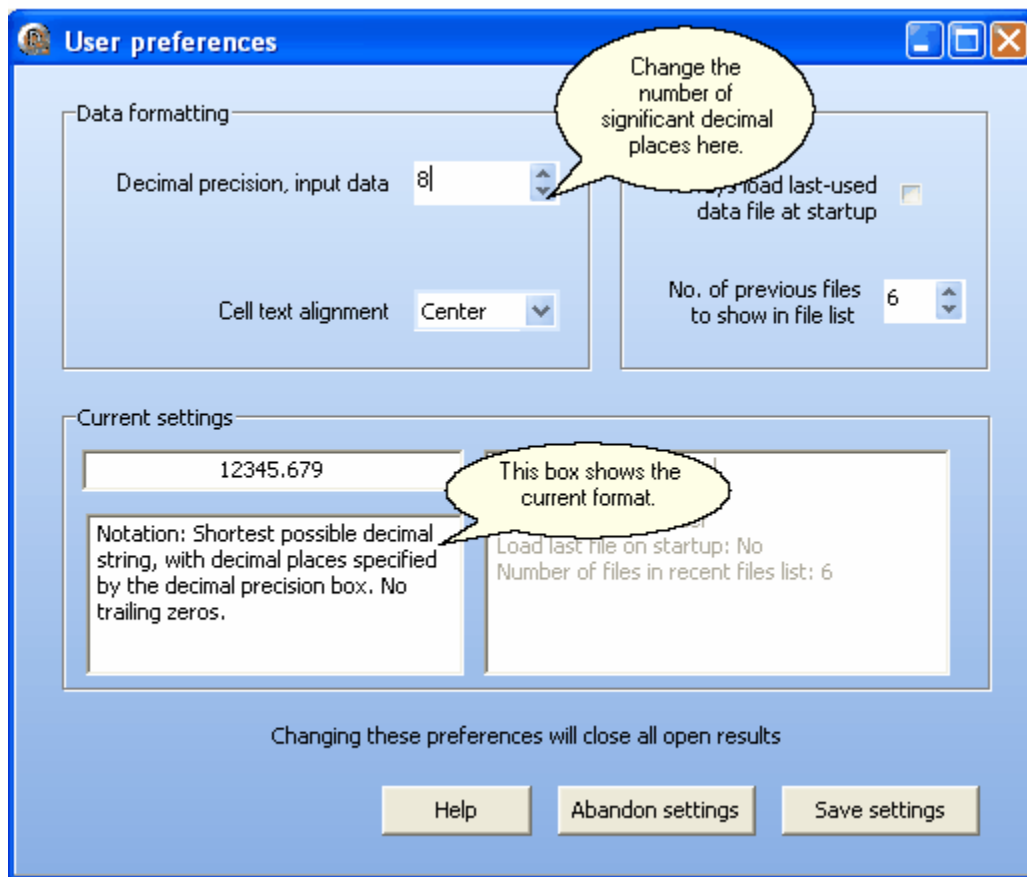
**Undo** - to reverse the last editing action

**Preferences** <sup>[27]</sup> - to change the look and output of the program.



### 3.2.1 Preferences - setup dialog

When **Edit|Preferences** is selected, a User Preferences dialog window opens. This offers a range of options to format your output. You can also change the number of recently used files listed under **File**.



Note that changing the decimal precision does not change the results until a new analysis has been run.

If you have "Always load last used data file at startup" ticked, and the last-used file is renamed, deleted or moved, then you will see an error message saying "This file was not found" on starting the program. To prevent seeing this error, disable this option in Preferences.

### 3.3 Simulation drop-down menu

This drop-down menu only has one item - [Explore Distributions](#)<sup>[28]</sup> opens a window in which you can generate distributions and explore their properties.

#### 3.3.1 Explore Distributions window

This option allows the generation of 1 or 2 normal distributions and displays them graphically. The data sets generated can then be analysed using any of the appropriate statistical tests. Because most data sets are not perfectly normal, there is also the option to add varying amounts of [skew](#)<sup>[102]</sup> and [kurtosis](#)<sup>[103]</sup> to each of the distributions generated.

The graphical presentation of the simulated data shows the [probability plot](#)<sup>[105]</sup>, a histogram of the distribution and a [box and whisker plot](#)<sup>[106]</sup>.

**Clear Data** - Click on this button to clear the data.

**Add Data** - Click on this button to add additional data points; the number added is given in the **No. to Add** box.

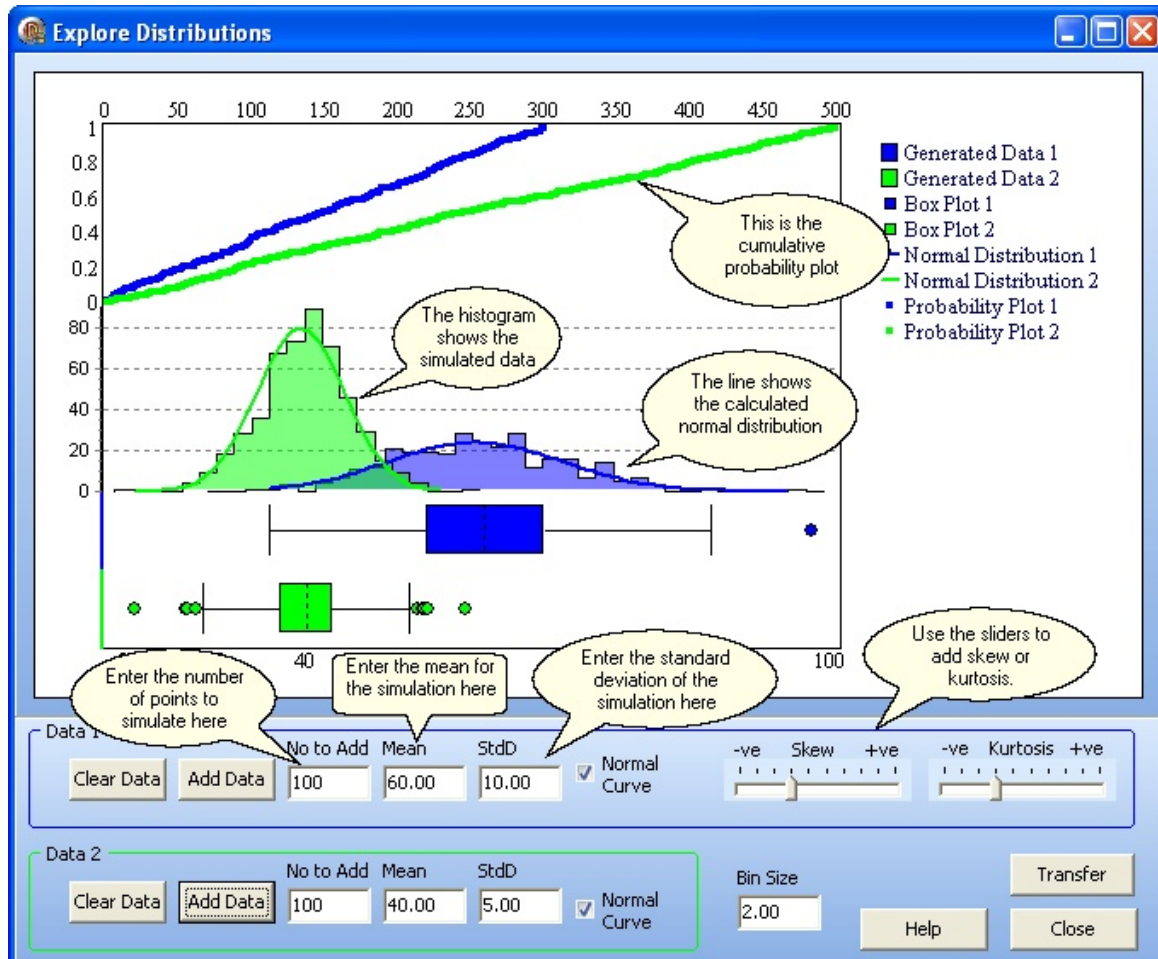
**No. to Add** - this box sets the number of simulated points to be added.

**Mean** - this is the [mean](#) of the simulated data.

**StdD** - this is the [standard deviation](#) of the simulated data.

The [skew](#) and [kurtosis](#) of the distribution can be varied using the sliders at the lower right corner of the window.

Use the **Transfer** button to move your simulated data to the working data grid for analysis using QED, or to be saved as a data file.



### 3.4 Single Sample drop-down menu

The **Single Sample** drop-down menu offers a range of methods for studying a single variable. Choose

[Mean](#) - to calculate the arithmetic mean of a list of numbers.

[Median](#) - to calculate the median of a list of numbers.

[Variance](#) - to calculate the variance of a list of numbers.

[Standard deviation](#) - to calculate the standard deviation of a list of numbers.

[Skewness](#) - to calculate the skew of a list of numbers.

[Kurtosis](#) - to calculate the kurtosis of a list of numbers.

[Probability Plot](#) - to examine the cumulative frequency distribution and investigate normality.

[Box and Whisker](#) - To create a box and whisker plot for a variable.

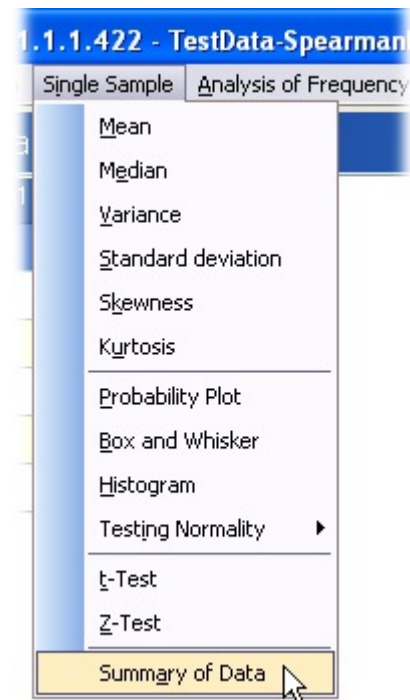
[Histogram](#) - Plots a histogram (bar chart) of a variable.

[Testing Normality](#)<sup>[34]</sup> - To test if the variable is normally distributed.

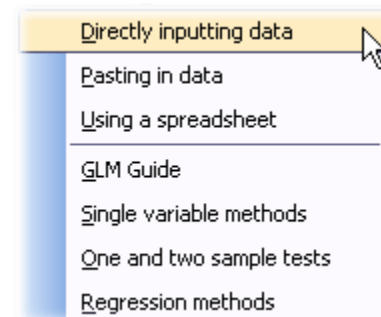
[t-Test](#)<sup>[38]</sup> - to test the mean of the variable for significant difference from a defined value.

[z Test](#)<sup>[40]</sup> - to test the mean of the variable for significant difference from a defined value.

[Summary of Data](#)<sup>[97]</sup> - to show or hide the page holding the data set statistics.



Watch **Help|Guides - Single variable methods** to see how to use these methods.



### 3.4.1 Mean

The mean of each variable in your data set is presented in a single grid.

Results: Mean				
	Bed	Water	Shade	Blooms
Mean	2	2	2	129

See [calculating the mean](#)<sup>[100]</sup> for further information.

### 3.4.2 Median

The median of each variable in your data set is presented in a single grid.

Results: Median				
	Bed	Water	Shade	Blooms
Median	2	2	2	111

See [calculating the median](#)<sup>[100]</sup> for further information.

### 3.4.3 Variance

The variance of each variable in your data set is presented in a single grid.

Results: Variance				
	Bed	Water	Shade	Blooms
Variance	0.6923	0.6923	0.6923	8590

See [calculating the variance](#)<sup>[101]</sup> for further information.

### 3.4.4 Standard Deviation

The standard deviation of each variable in your data set is presented in a single grid.

Results: Standard Deviation				
	Bed	Water	Shade	Blooms
Standard Deviation	0.8321	0.8321	0.8321	92.68

See [calculating the standard deviation](#)<sup>[101]</sup> for further information.

### 3.4.5 Skewness

The skew of each variable in your data set is presented in a single grid.

Results: Skewness				
	Bed	Water	Shade	Blooms
Skewness	0	0	0	0.7508

See [calculating the skew](#)<sup>[102]</sup> for further information.

### 3.4.6 Kurtosis

The kurtosis of each variable in your data set is presented in a single grid.

Results: Kurtosis				
	Bed	Water	Shade	Blooms
Kurtosis	-1.56	-1.56	-1.56	0.1959

See [calculating the kurtosis](#)<sup>[103]</sup> for further information.

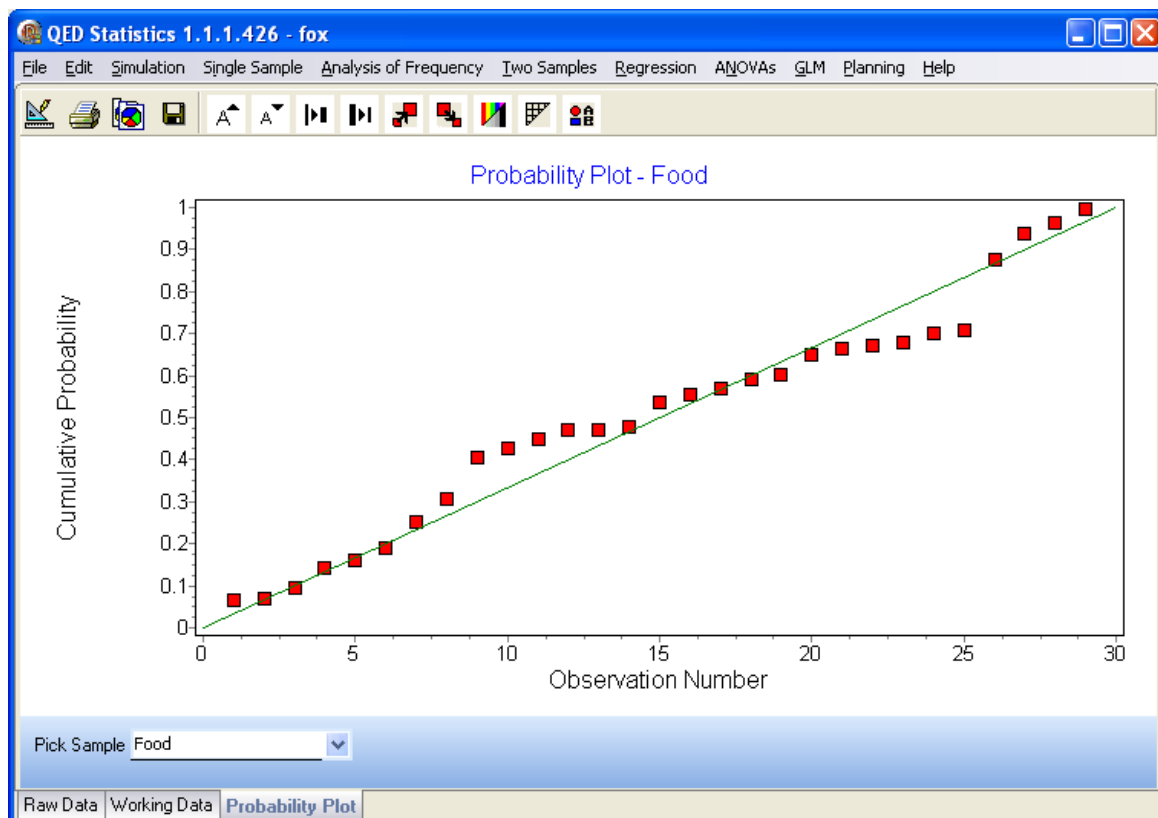
Two error messages will possibly occur during a kurtosis calculation; [NAN](#)<sup>[15]</sup> (Not A Number), and [positive or negative infinity](#)<sup>[15]</sup>.

### 3.4.7 Probability plot

Selecting **Single Sample|Probability** Plot displays the [cumulative normal probability plot](#)<sup>[105]</sup> for the selected variable.

The variable to plot is selected from the **Pick Sample** drop-down menu at the bottom left of the window.

All other aspects of the plot can be edited [using the graphics tool bar](#)<sup>[169]</sup> above the chart window.

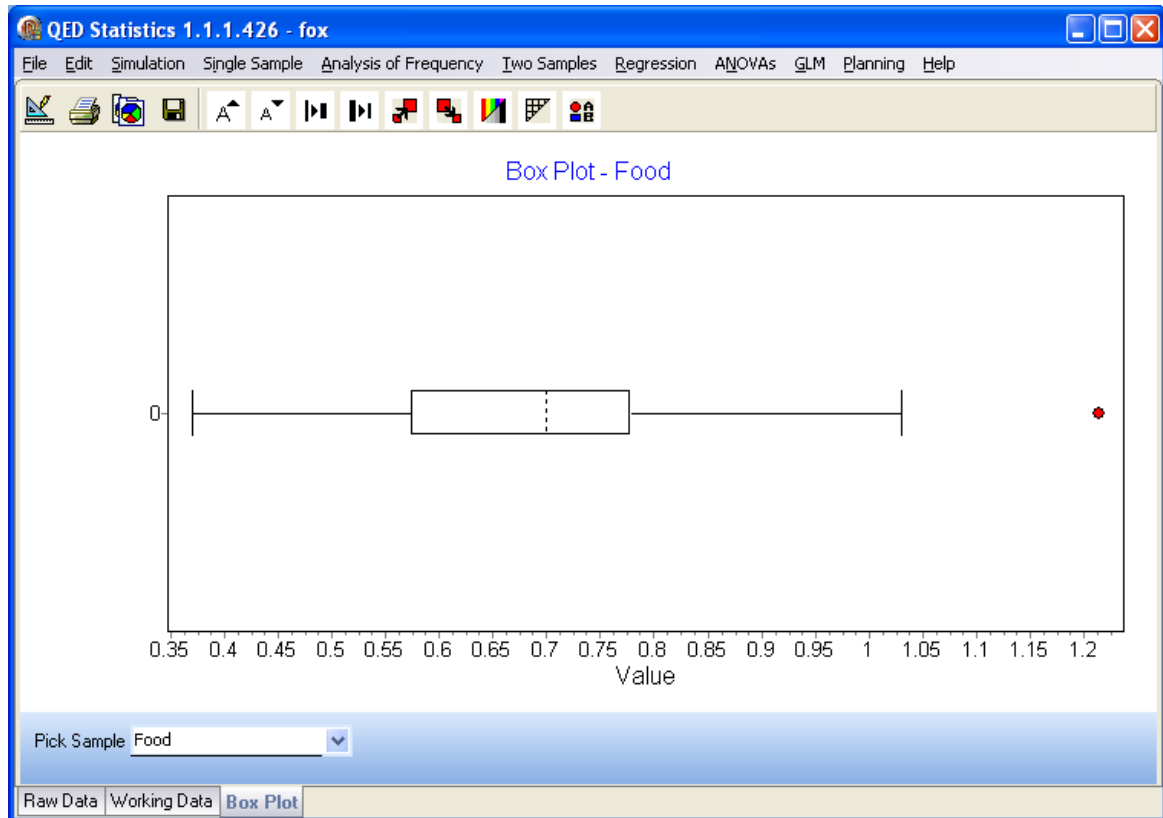


### 3.4.8 Box and Whisker plot

Selecting **Single Sample|Box and Whisker** displays a [box and whisker plot](#)<sup>[106]</sup> for the selected variable.

The variable to plot is selected from the **Pick Sample** drop-down menu at the bottom left of the window.

All other aspects of the plot can be edited [using the graphics tool bar](#)<sup>[169]</sup> above the chart window.

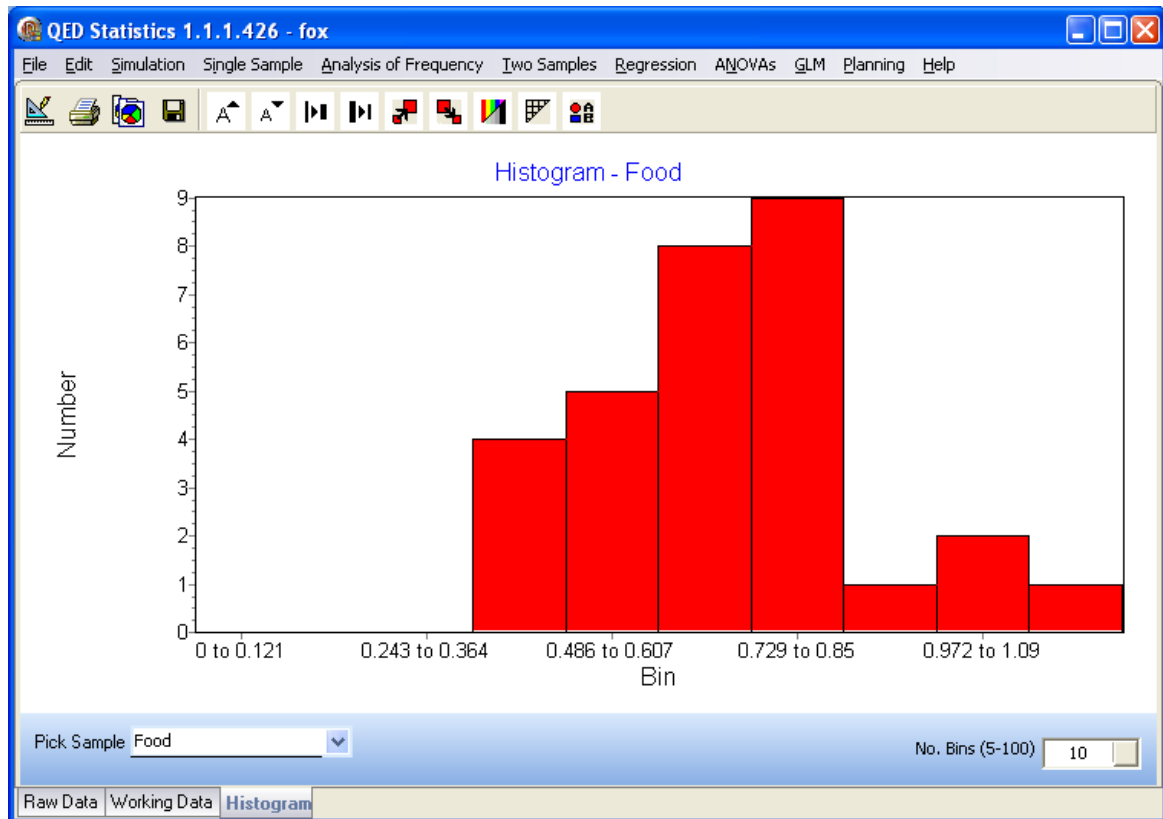


### 3.4.9 Histogram plot

Selecting **Single Sample|Histogram** displays a [histogram plot](#)<sup>[107]</sup> of the binned-up frequency of the selected variable data.

The variable to plot is selected from the **Pick Sample** drop-down menu at the bottom left of the window. Use the **No. Bins** box to specify the number of bins to class your data into. Press the button after then number of bins to activate your choice.

All other aspects of the graph can be edited [using the graphics tool bar](#)<sup>[169]</sup> above the plot.



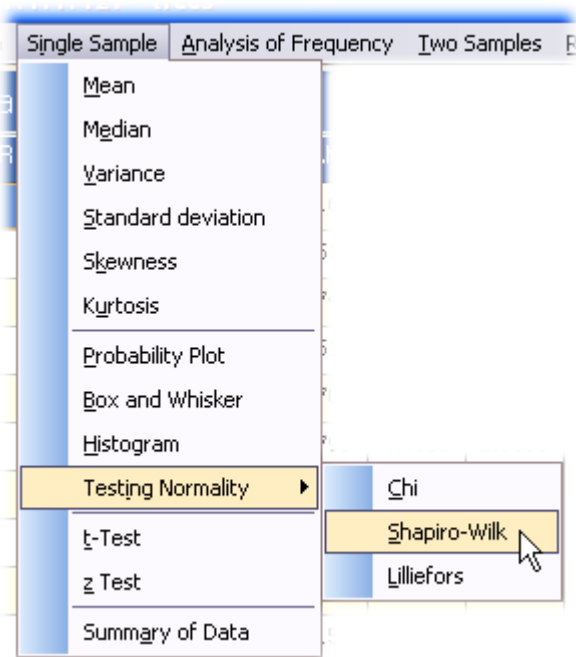
### 3.4.10 Testing for normality

There are three [tests for normality](#)<sup>[107]</sup>.

**Chi**<sup>[35]</sup> - compares the observed and expected distribution see [Chi-squared test](#)<sup>[110]</sup> for details - this method should not generally be used as the Shapiro-Wilk and Lilliefors are superior.

**Shapiro-Wilk**<sup>[37]</sup> - correlates the data with the corresponding normal scores see [Shapiro-Wilk test](#)<sup>[108]</sup> for details.

**Lilliefors**<sup>[38]</sup> - this is a generalisation of the Kolmogorov-Smirnov test; see [Lilliefors test](#)<sup>[108]</sup> for details.



#### 3.4.10.1 Chi-squared test for normality - setup dialog

To use a Chi-squared test for normality, the data must be binned to form a frequency distribution.

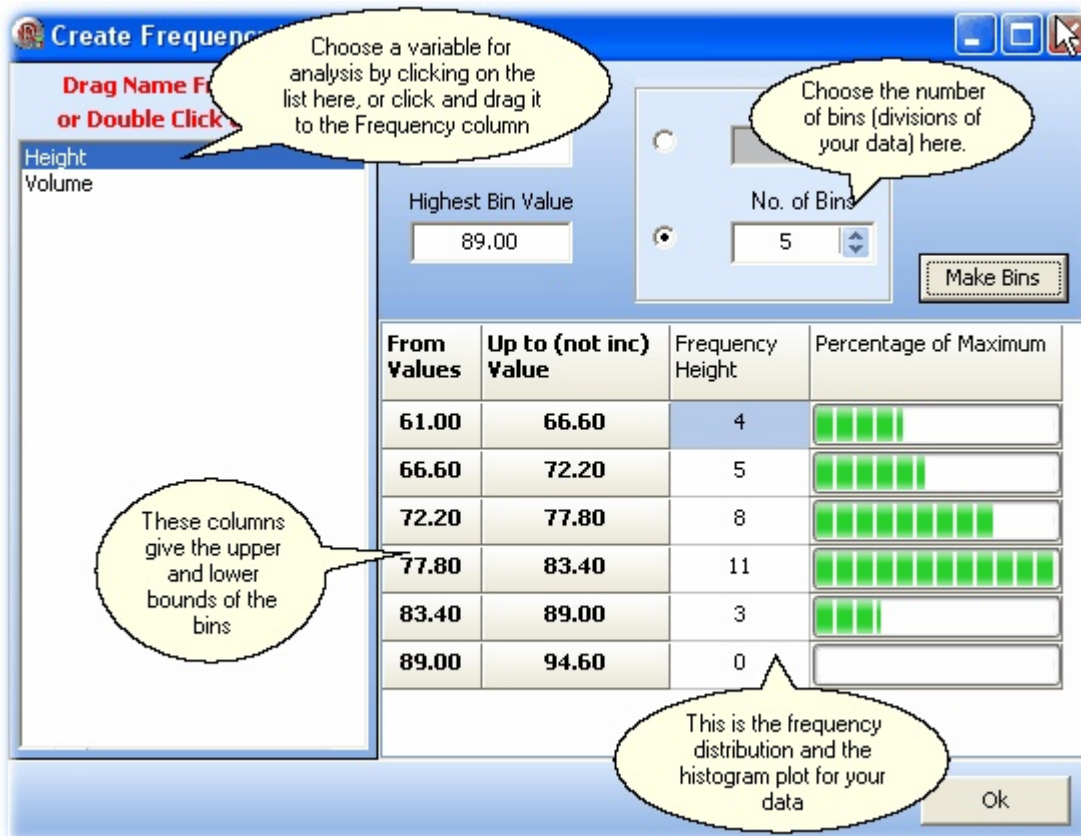
The form shown below will open when you select **Test for Normality|Chi**.

From the list of variables in the panel on the left, select one by dragging to the frequency column on the right, or alternatively just double click on the name.

Use **No. of Bins** to select an appropriate number of bins to ensure that most frequency intervals have greater than 5 observations.

You can also use the radio button to chose the **Step Size** for each bin rather than the number of bins.

Use the **Make Bins** button to allocate your data into the number of bins selected.



Once you have checked that the data is binned as desired, by looking at the frequency column and the plot, click OK to see your results.

#### 3.4.10.1.1 Chi-squared test for normality - results

The result is presented in a single grid.

**Chi-squared** is the test statistic.

**n** is the degrees of freedom.

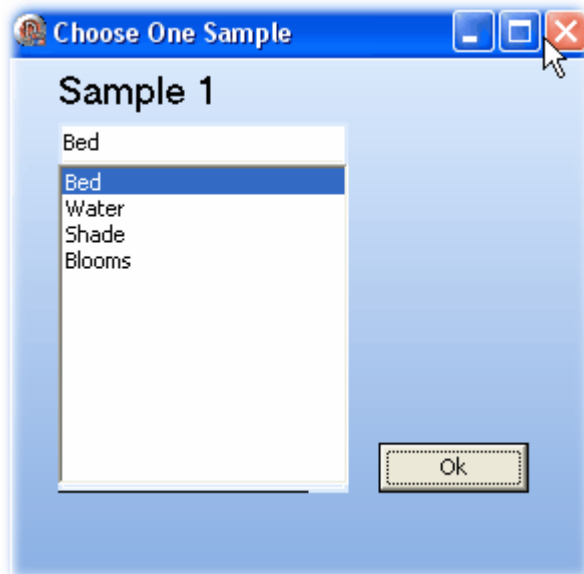
**Prob.** is the probability that the observed distribution is not significantly different from that expected for a normal distribution.

Results: Chi Test	
	Blooms
n	3
Chi squared	4.109
Prob.	0.7501

See [Chi-squared test for normality](#) for further information.

### 3.4.10.2 Shapiro-Wilk test for non-normality - setup dialog

First select a variable by clicking on the name. Then click OK to run the test for the selected variable.



#### 3.4.10.2.1 Shapiro-Wilk test - results

The results are presented in a single grid.

**n** is the number of observations.

**W** is the test statistic.

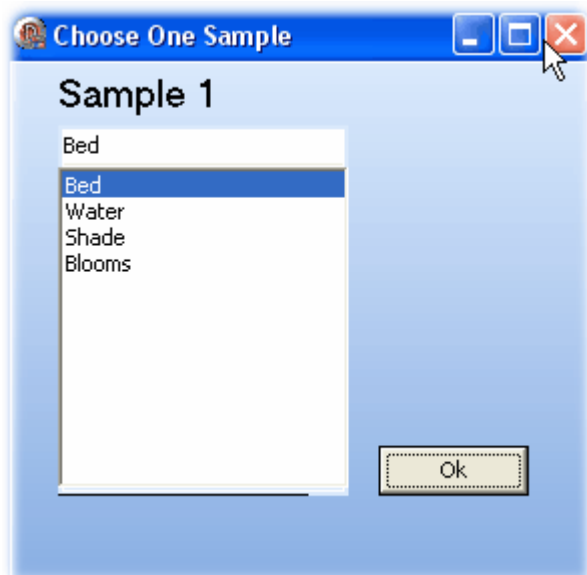
**Prob.** is the probability that the variable is normally distributed and that the null hypothesis cannot be rejected.

Results: Shapiro-Wilk Test for Normality	
	DIAMETER
n	31.000
W	0.941
Prob.	0.089
This means	The distribution of the variable DIAMETER does not deviate significantly from a normal distribution (Shapiro-Wilk test, $W = 0.941$ , $n = 31$ , $P = >0.05$ )

See [Shapiro-Wilk test](#) for further information.

### 3.4.10.3 Lilliefors test for normality - setup dialog

First select a variable by clicking on the name. Then click OK to run the test for the selected variable.



#### 3.4.10.3.1 Lilliefors test for normality - results

The results are presented in a single grid.

**Skewness** - is the calculated [skew](#)<sup>[102]</sup>.

**Kurtosis** - is the calculated kurtosis.

**Test Statistic** - is the test statistic.

Results: Lilliefors Test for Normality	
Blooms	
Skewness	0.7508
Kurtosis	0.1959
Test Stat	0.1739
This means	Sufficient evidence against normality.

See [Lilliefors test](#)<sup>[108]</sup> for further information.

### 3.4.11 Single sample t-Test - setup dialog

This is a [t-Test to compare a distribution against a known value](#)<sup>[114]</sup>.

Selecting **Single Sample t-Test** displays the t-Test template window which allows you to select data for analysis and input the value against which it is to be tested.

The left-hand side of the window is used to select or enter data for analysis. The drop-down menu in the upper left quarter is used to select one of variables from the working data grid. Alternatively, by selecting **Enter Here** you can enter a mean, standard deviation and number of observations for a distribution to test against a known value.

On the right-hand side use

**Value to Test** to enter the value to test the mean against.

**Tails** to select a one- or two-tailed test.

**Calculate** to undertake the test and show the results in the lower right hand quarter.

**t-Test Template**

**Known Distribution**

☒ From Column

Bed

☐ Enter Here

**Values**

Mean: 2.0000

Standard Deviation: 0.8321

Number of Obs: 27

**Enter Test Value**

Value to Test: 0.00

**Tails**

☒ One Tailed

☐ Two Tailed

Calculate

t Value: 12.4893

Probability: 0.0000

Df: 26

OK

**t Value** is the test statistic

**Probability** is the probability that the mean of the selected data is significantly different from the test value.

**Df** is the degrees of freedom.

Click OK to put the results in an output grid.

#### 3.4.11.1 Single sample t-Test - results

The result is presented in a single grid.

The name of the variable is shown at the top of the column ("Bed" in the example below). The value against which the data was compared is in the next column under "Known Value".

**Mean** is the mean of the selected variable or the mean you entered.

**StdD** is the standard deviation of the selected variable or the value you entered.

**N** is the number of observations.

**t Value** is the test statistic.

**df** is the degrees of freedom.

**Prob.** is the probability that the mean of the selected data is significantly different from the test value.

**Number of tails** - records if the result is for a one- or two-tailed test.

Results: T Test		
	Bed	Known Value
Mean	2.0000	0.00
StdD	0.8321	
N	27	
t	12.4893	
df	26	
Prob.	0.0000	
Number of tails	One	
This means	The mean of the known variable is significantly smaller than the selected value (t = 12.49 One tail , n = 26, P = <0.05)	

See [t-Test to compare a distribution against a known value](#)<sup>[114]</sup> for more information.

### 3.4.12 z Test - setup dialog

This is a [z Test to compare a distribution against a known value](#)<sup>[115]</sup>.

Selecting **Single Sample|z Test** displays the t-Test template window, which allows you to select data for analysis and input the value against which it is to be tested.

The left-hand side of the window is used to select or enter data for analysis. The drop-down menu in the upper left quarter is used to select one of the variables from the working data grid. Alternatively, by selecting **Enter Here** you can enter a mean, standard deviation and number of observations for a distribution to test against a known value.

On the right-hand side use

**Value to Test** to enter the value to test the mean against.

**Tails** to select a one- or two-tailed test.

**Calculate** to undertake the test and show the results in the lower right hand quarter.

**z Test Template**

**Known Distribution**

☒ From Column

Bed

☐ Enter Here

**Values**

Mean: 2.0000

Standard Deviation: 0.8321

Number of Obs: 27

**Enter Test Value**

Value to Test: 0.00

**Tails**

☒ One Tailed

☐ Two Tailed

Calculate

**Results**

z Value: 12.4893

Probability: 0.0000

Df: 0

OK

**z Value** is the test statistic

**Probability** is the probability that the mean of the selected data is significantly different from the test value.

**Df** is the degrees of freedom.

Click OK to put the results in an output grid.

#### 3.4.12.1 z Test - results

The result is presented in a single grid.

The name of the variable is shown at the top of the column ("Bed" in the example below). The value against which the data was compared is in the next column under "Known Value".

**Mean** is the mean of the selected variable or the mean you entered.

**StdD** is the standard deviation of the selected variable or the value you entered.

**N** is the number of observations.

**z** is the test statistic.

**Prob.** is the probability that the mean of the selected data is significantly different from the test value.

**Number of tails** - records if the result is for a one- or two-tailed test.

Results: z Test		
	Bed	Known Value
Mean	2.0000	0.00
StdD	0.8321	
N	27	
z	12.4893	
Prob.	0.0000	
Number of tails	One	
This means	The mean of the known variable is significantly smaller than the selected value (z = 12.49 One tail , n = 27, P = <0.05)	

See [z Test - Comparing observations with a known mean](#)<sup>[115]</sup> for more information.

### 3.5 Analysis of Frequency drop-down menu

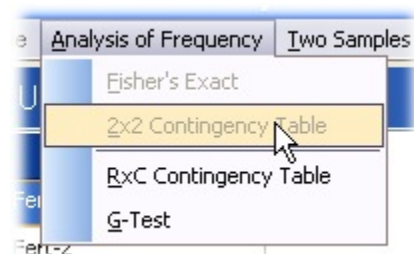
The Analysis of Frequency drop-down menu offers 4 tests for contingency tables. Choose:

[Fisher's Exact](#)<sup>[42]</sup> - to undertake Fisher's Exact test on a 2 x 2 contingency table.

[2 x 2 Contingency Table](#)<sup>[44]</sup> - to undertake a standard Chi-squared test on a 2 x 2 contingency table.

[R x C Contingency Table](#)<sup>[46]</sup> - to undertake a Chi-squared test with more than 2 rows or columns.

[G-Test](#)<sup>[48]</sup> - to undertake a G-Test rather than a Chi-squared test on a contingency table.



For instructions for entering data see [Entering Contingency table data](#)<sup>[12]</sup>.

#### 3.5.1 Fisher's Exact - setup dialog

If **Analysis of Frequency|Fisher's Exact** is selected, the following contingency table is opened. In the upper part of the window is displayed your original data with the row and column sums added. In the lower part of the window a second table of the calculated expected number of observations is displayed. See [Fisher's Exact test](#)<sup>[117]</sup> to find out how these values are calculated. Finally at the bottom of the window are the radio buttons to select a [one- or two-tailed test](#)<sup>[120]</sup>.

To see how to organise your data for this analysis see [Entering Contingency table data](#)<sup>[12]</sup>.

**Contingency Tables**

### Fisher's Exact

**Observed**

	a	b	Sum
aa	1	4	5
bb	2	6	8
Sum	3	10	Total Sum: 13

**Expected**

	a	b	Sum
aa	1.15	3.85	5
bb	1.85	6.15	8
Sum	3	10	

Type of Test

☒ One-Tailed ☐ Two-Tailed

Ok

Once you have checked that the data have been entered correctly, click OK to see your [results](#)<sup>[43]</sup>.

For more information about the test see [Fisher's Exact test](#)<sup>[117]</sup>.

#### 3.5.1.1 Fisher's Exact - results

The result of Fisher's Exact test is presented in a single grid.

Prob. is the probability that the two variables are independent of each other and that the null hypothesis is correct.

Results: Fisher's Exact	
Prob.	0.8042
No. Columns	2
No. Rows	2
This means	The null hypothesis is accepted. The two variables are independent of each other (Fisher's Exact = 0.8042, P = >0.05)

There is no Expand and Explore for this test.

For more information see [Fisher's Exact test](#)<sup>[117]</sup>.

Results can be [exported](#)<sup>[25]</sup> or [printed](#)<sup>[26]</sup> from this grid.

### 3.5.2 2 x 2 Contingency table - setup dialog

If **Analysis of Frequency|2x2 Contingency table** is selected the following contingency table is opened. In the upper part of the window is displayed your original data with the row and column sums added. In the lower part of the window a second table of the calculated expected number of observations is displayed. See [Calculation of expected frequencies](#)<sup>[127]</sup> to find out how these values are calculated. To see how to organise your data for this analysis see [Entering Contingency table data](#)<sup>[12]</sup>.

The screenshot shows a software window titled 'Contingency Tables' with a subtitle '2 x 2 Contingency'. It contains two sections: 'Observed' and 'Expected', each with a table. The 'Observed' table has values: (aa, a)=1, (aa, b)=4, (bb, a)=2, (bb, b)=6, (Sum, a)=3, (Sum, b)=10, and Total Sum=13. The 'Expected' table has values: (aa, a)=1.15, (aa, b)=3.85, (bb, a)=1.85, (bb, b)=6.15, (Sum, a)=3, (Sum, b)=10, and Total Sum=13. An 'Ok' button is at the bottom right.

	a	b	Sum
aa	1	4	5
bb	2	6	8
Sum	3	10	Total Sum: 13

	a	b	Sum
aa	1.15	3.85	5
bb	1.85	6.15	8
Sum	3	10	Total Sum: 13

Ok

Once you have checked that the data has been entered correctly click OK to see your [results](#)<sup>[45]</sup>.

For more information about the test, see [Contingency table Chi-squared test](#)<sup>[118]</sup>.

### 3.5.2.1 2 x 2 Contingency table - results

The results of 2 x 2 Chi-squared test are presented in a single grid.

[Chi-squared](#)<sup>[118]</sup> is the test statistic.

**df** is the degrees of freedom.

**Prob.** is the probability that the two variables are independent of each other and that the null hypothesis is correct.

[Cramer v](#)<sup>[119]</sup> is a measure of association between the two variables

[Contingency coefficient](#)<sup>[119]</sup> is a measure of relationship between the two variables.

Results: Contingency Table	
Chi squared	0.04333
df	1
Prob.	0.8351
Cramer v	0.05774
Contingency coefficient C	0.05764
No. Columns	2
No. Rows	2
This means	The null hypothesis is accepted. The two variables are independent of each other (Chi squared = 0.04333, df = 1, P = >0.05)

Results can be [exported](#)<sup>[25]</sup> or [printed](#)<sup>[26]</sup> from this grid.

### 3.5.3 R x C Contingency table - setup dialog

If **Analysis of Frequency|RxC Contingency table** is selected, a contingency table is opened. In the upper part of the window is displayed your original data with the row and column sums added. In the lower part of the window a second table of the calculated expected number of observations is displayed. See [Calculation of expected frequencies](#)<sup>[12]</sup> to find out how these values are calculated. To see how to organise your data for this analysis see [Entering Contingency table data](#)<sup>[12]</sup>.

Contingency Tables					
Row x Column Contingency					
Observed					
	Black	Brown	Blond	Red	Sum
Male	32	43	16	9	100
Female	55	65	64	16	200
Sum	87	108	80	25	Total Sum: 300
Expected					
	Black	Brown	Blond	Red	Sum
Male	29.00	36.00	26.67	8.33	100
Female	58.00	72.00	53.33	16.67	200
Sum	87	108	80	25	Total Sum: 300
Ok					

Once you have checked that the data has been entered correctly click OK to see your [results](#)<sup>[47]</sup>.

This data set is available as **2x4 contingency.csv**

For more information about the test see [Contingency table Chi-squared test](#)<sup>[118]</sup>.

### 3.5.3.1 R x C Contingency table - results

The results of R x C Chi-squared test are presented in a single grid.

[Chi-squared](#)<sup>[118]</sup> is the test statistic.

**df** is the degrees of freedom.

**Prob.** is the probability that the two variables are independent of each other and that the null hypothesis is correct.

[Cramer v](#)<sup>[119]</sup> is a measure of association between the two variables

[Contingency coefficient](#)<sup>[119]</sup> is a measure of relationship between the two variables.

Results: Contingency Table	
Chi squared	8.987
df	3
Prob.	0.02946
Cramer v	0.1731
Contingency coefficient C	0.1705
No. Columns	4
No. Rows	2
This means	The null hypothesis is rejected. The two variables are not independent of each other (Chi squared = 8.987, df = 3, P = <0.05)

Results can be [exported](#)<sup>[25]</sup> or [printed](#)<sup>[26]</sup> from this grid.

### 3.5.4 G-Test - setup dialog

If **Analysis of Frequency|G-Test** is selected a contingency table is opened. In the upper part of the window is displayed your original data with the row and column sums added. In the lower part of the window a second table of the calculated expected number of observations is displayed. See [Calculation of expected frequencies](#)<sup>[12h]</sup> to find out how these values are calculated. To see how to organise your data for this analysis see [Entering Contingency table data](#)<sup>[12]</sup>.

**G-Test**

Observed

	Black	Brown	Blond	Red	Sum
Male	32	43	16	9	100
Female	55	65	64	16	200
Sum	87	108	80	25	Total Sum: 300

Expected

	Black	Brown	Blond	Red	Sum
Male	29.00	36.00	26.67	8.33	100
Female	58.00	72.00	53.33	16.67	200
Sum	87	108	80	25	Total Sum: 300

Ok

Once you have checked that the data has been entered correctly click OK to see your [results](#)<sup>[49]</sup>.

For more information about the test see [Contingency table G-Test](#)<sup>[120]</sup>.

### 3.5.4.1 G-Test - results

The results of a G-Test are presented in a single grid.

**G** is the test statistic.

**df** is the degrees of freedom.

**Prob.** is the probability that the two variables are independent of each other and that the null hypothesis is correct.

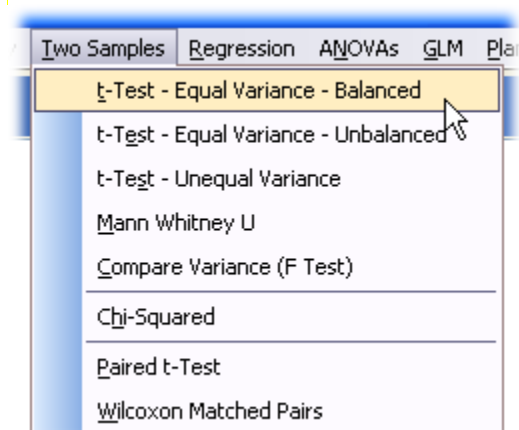
See [Contingency table G-Test](#)<sup>[120]</sup> for more information.

Results: G-Test	
G	4.641
DF	18.000
Prob.	0.999
No. Columns	3.000
No. Rows	10.000
This means	The null hypothesis is accepted. The two variables are independent of each other (G = 4.641, DF = 18, P = >0.05)

Results can be [exported](#)<sup>[25]</sup> or [printed](#)<sup>[26]</sup> from this grid.

### 3.6 Two samples drop-down menu

The **Two Samples** drop-down menu offers a range of methods for comparing two samples.



**t-Test - Equal Variance - Balanced**<sup>[51]</sup> to compare two [means](#)<sup>[100]</sup> when both samples are assumed to have the same variance and both samples have the same number of observations.

**t-Test - Equal Variance - Unbalanced**<sup>[52]</sup> to compare two [means](#)<sup>[100]</sup> when both samples are assumed to have the same variance and the samples have different numbers of observations.

**t-Test - Unequal Variance**<sup>[54]</sup> - If the samples cannot be assumed to have the same variances.

**Mann Whitney**<sup>[55]</sup> - a non-parametric test to determine if there is a significant difference between the [medians](#)<sup>[100]</sup> of two samples.

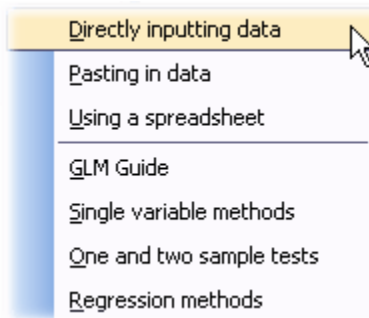
**Compare Variance (F Test)**<sup>[57]</sup> - to test if the [variances](#)<sup>[101]</sup> of two samples are the same.

**Chi-squared**<sup>[60]</sup> - to test two sets of frequencies for a significant difference.

**Paired t-Test**<sup>[58]</sup> - to compare the [means](#)<sup>[100]</sup> of paired samples.

**Wilcoxon Matched Pairs (Signed rank)**<sup>[59]</sup> - the non-parametric test for difference between matched pairs.

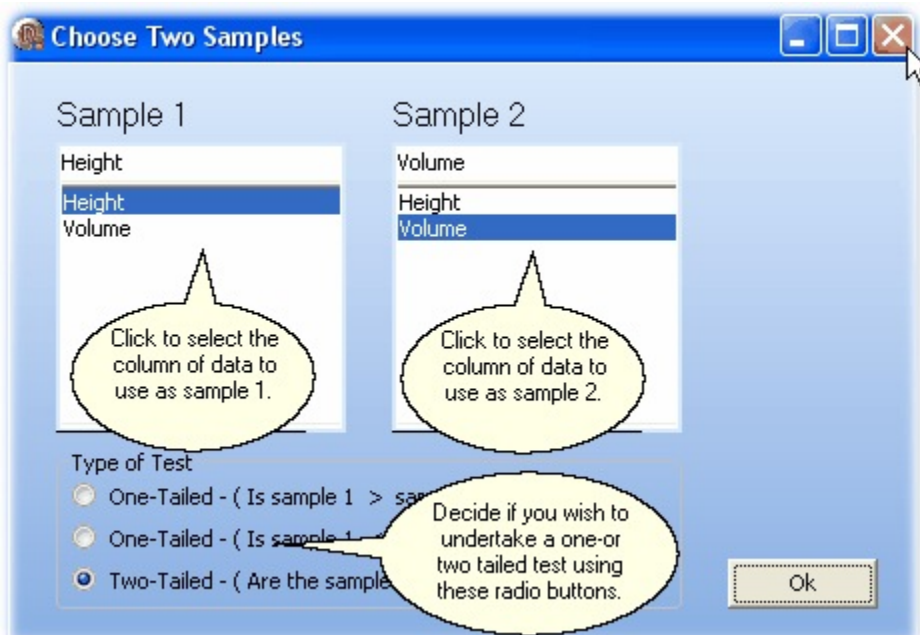
Watch **Help|Guides - One and two sample tests** to see how to use these methods.



### 3.6.1 t-Test (equal variance - balanced) - setup dialog

If **Two samples** **t-Test-Equal Variance - Balanced** is selected, the following dialog box is opened, which you use to select the two samples to compare, and whether you require a [one- or two-tailed test](#)<sup>[129]</sup>. For details of the method see [Comparing means of samples of the same size - equal variance](#)<sup>[124]</sup>.

Under Sample 1 and Sample 2, left click to choose the samples to compare. A one- or two-tailed test is selected using the radio buttons at the bottom of the dialog. Use a two-tailed test if you simply wish to test for a difference between the two means, which may be larger or smaller. Use a one-tailed test if one sample mean is larger than the other.



Once your selections have been made click OK to see the [results](#)<sup>[51]</sup>.

#### 3.6.1.1 t-Test (equal variance - balanced) - results

The results of t-Test are presented in a single grid.

**Mean** is the arithmetic [mean](#)<sup>[100]</sup> or average of each sample.

**StdD** is the [standard deviation](#)<sup>[101]</sup>.

**N** is the number of observations in each sample.

**t** is the test statistic.

**DF** is the degrees of freedom.

**Prob.** is the probability that the two variables have the same mean.

**Number of tails** states if a [one- or two-tailed test](#)<sup>[129]</sup> was requested.

**Type** is the type of [t-Test](#)<sup>[123]</sup> undertaken

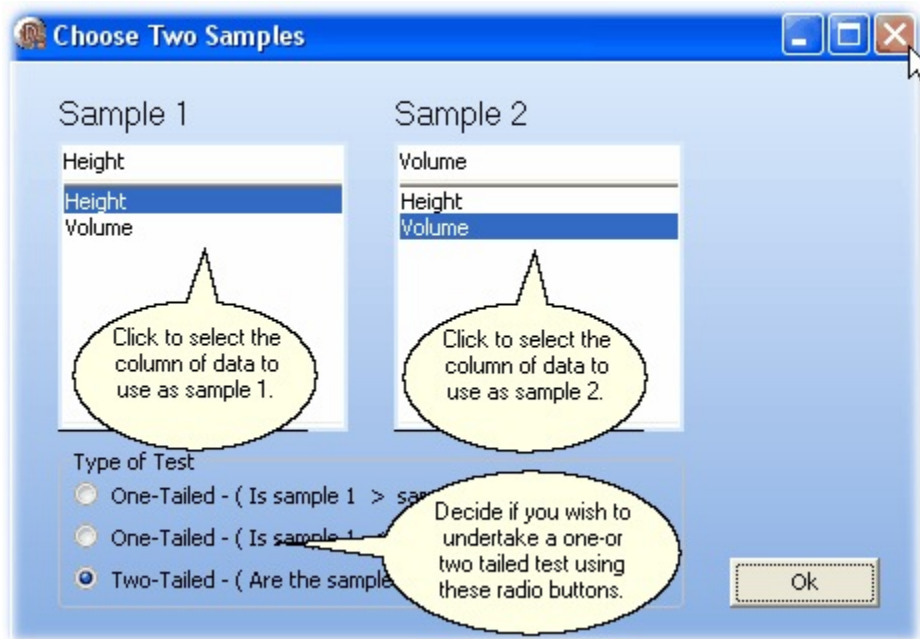
Results: 2 Sample t-Test: Equal Variances		
	Height	Volume
Mean	76.00	30.17
StdD	6.37	16.44
N	31.00	31.00
t	14.47	
DF	60.00	
Prob.	0.00	
Number of tails	Two	
Type	Balanced: Equal Variances	
This means	The null hypothesis is rejected. The two means are significantly different Height is larger than Volume (t = 14.47, DF = 60, P = <0.05)	

See [t-Test: Comparing means of samples of the same size - equal variance](#)<sup>[124]</sup> for further information.

### 3.6.2 t-Test (equal variance - unbalanced) - setup dialog

If **Two samples|t-Test-Equal Variance - Unbalanced** is selected the following dialog box is opened, which you use to select the two samples to compare, and whether you require a [one- or two-tailed test](#)<sup>[129]</sup>. For details of the method see [Comparing means of samples of unequal size - equal variance](#)<sup>[125]</sup>.

Under Sample 1 and Sample 2, left click to choose the samples to compare. A one- or two-tailed test is selected using the radio buttons at the bottom of the dialog. Use a two-tailed test if you simply wish to test for a difference between the two means, which may be larger or smaller. Use a one-tailed test if one sample mean is larger than the other.



Once your selections have been made click OK to see the [results](#)<sup>[53]</sup>.

### 3.6.2.1 t-Test (equal variance - unbalanced) - results

The results of t-Test are presented in a single grid.

**Mean** is the arithmetic [mean](#)<sup>[100]</sup> or average of each sample.

**StdD** is the [standard deviation](#)<sup>[101]</sup>.

**N** is the number of observations in each sample.

**t** is the test statistic.

**DF** is the degrees of freedom.

**Prob.** is the probability that the two variables have the same mean.

**Number of tails** states if a [one- or two-tailed test](#)<sup>[129]</sup> was requested.

**Type** is the type of [t-Test](#)<sup>[123]</sup> undertaken

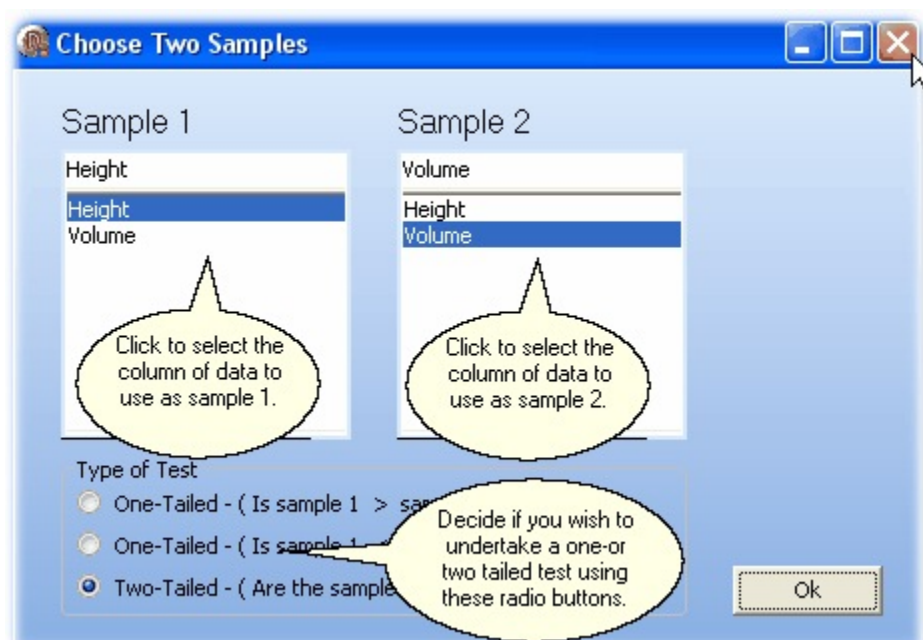
Results: 2 Sample t-Test: Equal Variances		
	Height	Volume
Mean	76.00	30.17
StdD	6.37	16.44
N	31.00	31.00
t	14.47	
DF	60.00	
Prob.	0.00	
Number of tails	Two	
Type	Unbalanced: Equal Variances	
This means	The null hypothesis is rejected. The two means are significantly different Height is larger than Volume (t = 14.47, DF = 60, P = <0.05)	

See [t-Test: Comparing means of samples of unequal size - equal variance](#)<sup>[125]</sup> for further information.

### 3.6.3 t-Test (unequal variance) - setup dialog

If **Two samples|t-Test- Unequal Variance** is selected the following dialog box is opened, which you use to select the two samples to compare, and whether you require a [one- or two-tailed test](#)<sup>[129]</sup>. For details of the method see [Comparing means from samples with unequal variances](#)<sup>[126]</sup>.

Under Sample 1 and Sample 2, left click to choose the samples to compare. A one- or two-tailed test is selected using the radio buttons at the bottom of the dialog. Use a two-tailed test if you simply wish to test for a difference between the two means, which may be larger or smaller. Use a one-tailed test if one sample mean is larger than the other.



Once your selections have been made click OK to see the [results](#)<sup>[55]</sup>.

### 3.6.3.1 t-Test (unequal variance) - results

The results of t-Test are presented in a single grid.

**Mean** is the arithmetic [mean](#)<sup>[100]</sup> or average of each sample.

**StdD** is the [standard deviation](#)<sup>[101]</sup>.

**N** is the number of observations in each sample.

**t** is the test statistic.

**DF** is the degrees of freedom.

**Prob.** is the probability that the two variables have the same mean.

**Number of tails** states if a [one- or two-tailed test](#)<sup>[129]</sup> was requested.

**Type** is the type of [t-Test](#)<sup>[123]</sup> undertaken

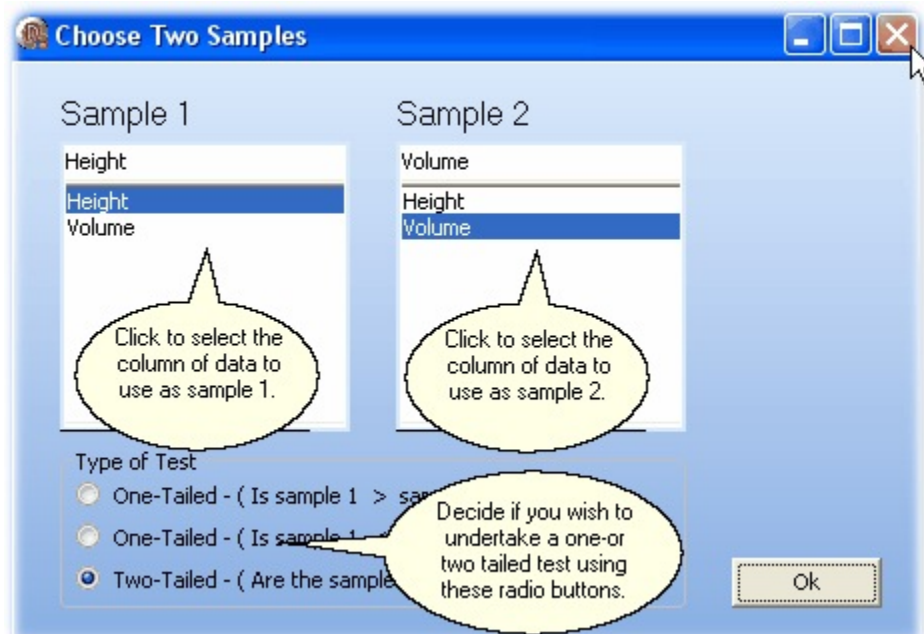
Results: 2 Sample t-Test: : Unequal Variances		
	Height	Volume
Mean	75.63	27.84
StdD	6.14	13.79
N	30.00	29.00
t	17.10	
DF	38.00	
Prob.	0.00	
Number of tails	Two	
Type	Unbalanced: Unequal Variances	
This means	The null hypothesis is rejected. The two means are significantly different Height is larger than Volume (t = 17.10, DF = 38, P = <0.05)	

See [t-Test: Comparing means from samples with unequal variances](#)<sup>[126]</sup> for further information.

### 3.6.4 Mann-Whitney two sample test - setup dialog

If **Two samples|Mann Whitney U** is selected the following dialog box is opened, which you use to select the two samples to compare, and whether you require a one- or two-tailed test.

Under Sample 1 and Sample 2, click to select the samples to compare. A one- or two-tailed test is selected using the radio buttons at the bottom of the dialog. Use a two-tailed test if you simply wish to test for a difference between the two means, which may be larger or smaller. Use a one-tailed test if one sample mean is larger than the other.



Once your selections have been made click OK to see the [results](#)<sup>[56]</sup>.

See [Mann-Whitney unpaired test](#)<sup>[128]</sup> for information about this method.

### 3.6.4.1 Mann-Whitney U - results

The results of the Mann Whitney test are presented in a single grid.

**Sample 1** Gives the name of the first sample.

**Sample 2** gives the name of the second sample.

**StdD** is the [standard deviation](#)<sup>[107]</sup>.

**U** is the test statistic.

**df** is the degrees of freedom.

**Prob.** is the probability that the two variables have the same median.

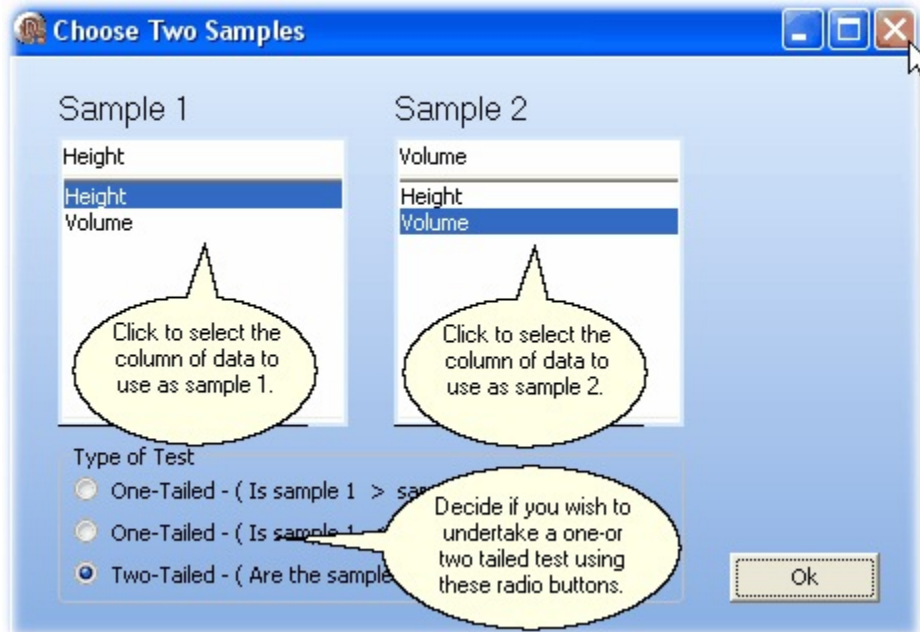
Results: Mann Whitney U	
Sample 1	Type I
Sample 2	Type II
U	18
Prob.	0.2031
This means	The null hypothesis is accepted. The two medians are not significantly different (U = 18, U1 = 7, U2 = 7, P = >0.05)

See [Mann-Whitney unpaired test](#)<sup>[128]</sup> for information about this method.

### 3.6.5 Two sample F Test - setup dialog

If **Two samples|Compare variance (F Test)** is selected the following dialog box is opened which you use to select the two samples to compare.

Under Sample 1 and Sample 2, click to select the samples to compare.



Once your selections have been made, click OK to see the [results](#)<sup>[57]</sup>.

For further information see [Testing for difference between two variances](#)<sup>[127]</sup>.

#### 3.6.5.1 Two sample F Test - results

The results of F Test are presented in a single grid.

**Sample 1** Gives the name of the first sample.

**Sample 2** gives the name of the second sample.

**F** is the test statistic.

**Prob.** is the probability that the two variables have the same variance.

Results: F Test	
Sample 1	Type I
Sample 2	Type II
F	1.233
Prob.	0.8061
This means	The null hypothesis is accepted. The two variances are not significantly different (F = 1.233, DF1 = 6, DF2 = 6, P = >0.05)

See [Testing for difference between two variances](#)<sup>[127]</sup> for further information

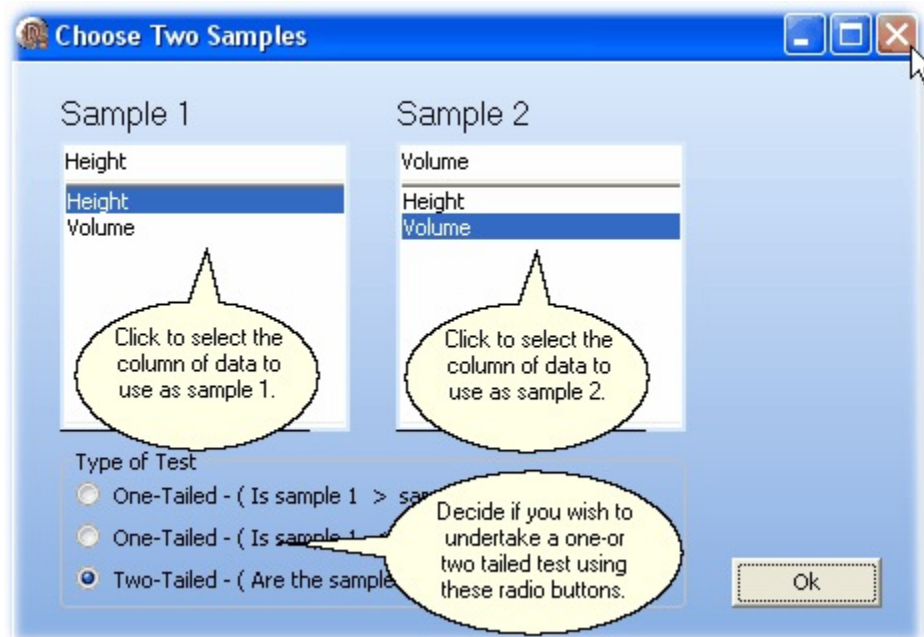
### 3.6.6 Paired t-Test - setup dialog

If **Two samples|Paired t-Test** is selected the following dialog box is opened, which you use to select the two samples to compare, and whether you require a one- or two-tailed test.

Under Sample 1 and Sample 2, click to select the samples to compare. A one- or two-tailed test is selected using the radio buttons at the bottom of the dialog. Use a two-tailed test if you simply wish to test for a difference between the two means, which may be larger or smaller. Use a one-tailed test if one sample mean is larger than the other.

For further information see [Comparing means of paired samples](#)<sup>[123]</sup>.

If you have more than 2 repeated measures use a [One-way repeated measurements ANOVA](#)<sup>[143]</sup>.



Once your selections have been made click OK to see the [results](#)<sup>[58]</sup>.

#### 3.6.6.1 Paired t-Test - results

The results of the t-Test are presented in a single grid.

**Mean** is the arithmetic [mean](#)<sup>[100]</sup> or average of each sample.

**StdD** is the [standard deviation](#)<sup>[101]</sup>.

**N** is the number of observations in each sample.

**t** is the test statistic.

**DF** is the degrees of freedom.

**Prob.** is the probability that the two variables have the same mean.

**Number of tails** states if a one- or two-tailed test was requested.

**Type** is the type of t-Test undertaken.

Results: Paired t-Test			
	Fert-1	Fert-2	Difference
Mean	5.445	3.999	1.446
StdD	0.976	0.972	1.658
N	10.000	10.000	10.000
t	2.758		
DF	9.000		
Prob.	0.022		
Number of tails	Two		
Type	Paired		
This means	The null hypothesis is rejected. The two means are significantly different Fert-1 is larger than Fert-2 (t = 2.758, DF = 9, P = <0.05)		

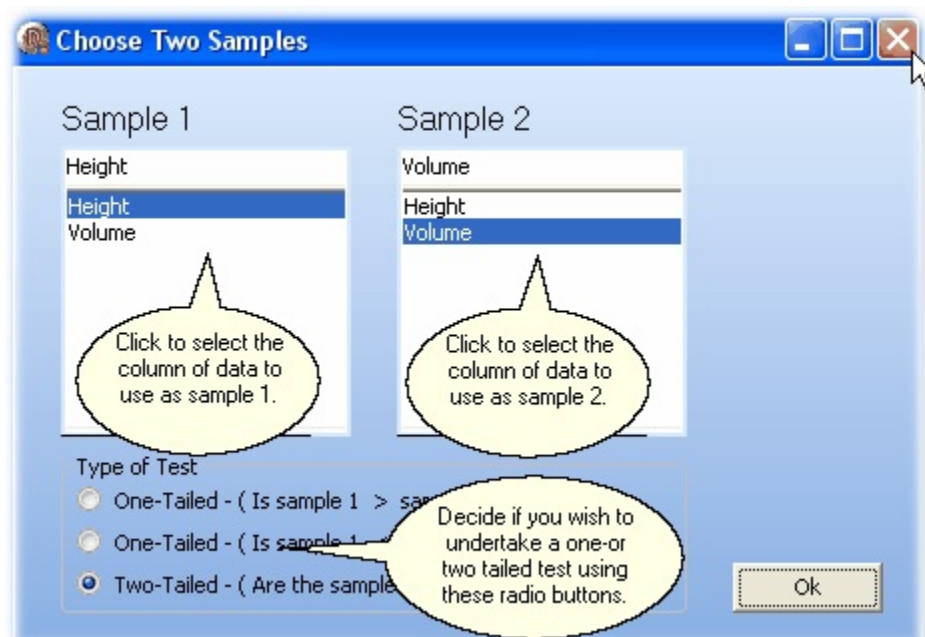
See [Paired t-Test](#)<sup>[123]</sup> for further information

### 3.6.7 Wilcoxon matched pairs - setup dialog

If **Two samples|Paired t-Test** is selected the following dialog box is opened, which you use to select the two samples to compare, and whether you require a one- or two-tailed test.

Under Sample 1 and Sample 2, click to select the samples to compare. A one- or two-tailed test is selected using the radio buttons at the bottom of the dialog. Use a two-tailed test if you simply wish to test for a difference between the two means, which may be larger or smaller. Use a one-tailed test if one sample mean is larger than the other.

For further information see [Wilcoxon paired-sample test](#)<sup>[129]</sup>



Once your selections have been made click OK to see the [results](#)<sup>[58]</sup>.

### 3.6.7.1 Wilcoxon matched pairs - results

The results of the Wilcoxon test are presented in a single grid.

**Sample 1** Gives the name of the first sample.

**Sample 2** gives the name of the second sample.

**t** is the test statistic.

**Prob.** is the probability that the two variables have the same median.

**Number of tails** states if a one- or two-tailed test was requested.

Results: Wilcoxon Matched Pairs	
Sample 1	Type I
Sample 2	Type II
t	0.169031
Prob.	>0.05
Number of tails	Two
This means	The null hypothesis is accepted. The two means are not significantly different (t = 0.169031, P = >0.05)

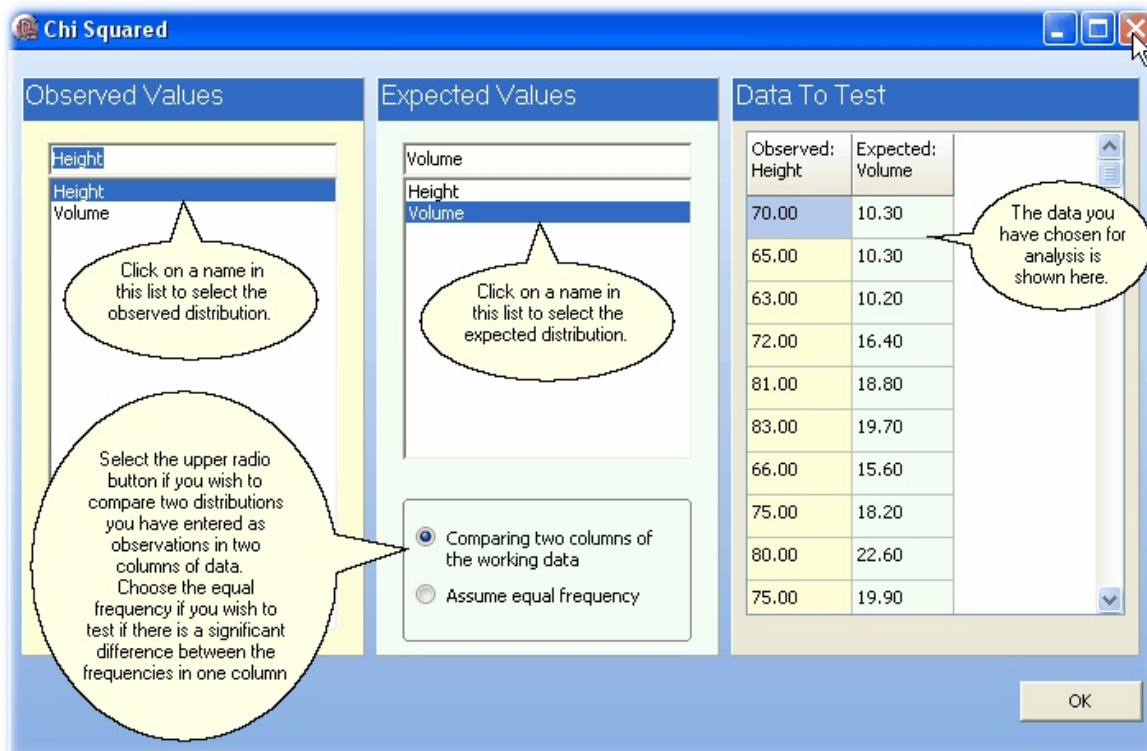
See [Wilcoxon matched pairs](#)<sup>[129]</sup> for further information

### 3.6.8 Two sample Chi-squared - setup dialog

If **Two samples|Chi-squared** is selected the following dialog box is opened, which you use to select the two samples to compare.

Under Sample 1 and Sample 2, click to select the samples to compare. If, instead of a list of expected values, you wish to compare the observed distribution against an even distribution select the **Assume equally likely** radio button.

For further information see [Chi-squared two sample test](#)<sup>[127]</sup>.



Once your selections have been made click OK to see the [results](#)<sup>[51]</sup>.

### 3.6.8.1 Two sample Chi-squared - results

The results of the Chi-squared test are presented in a single grid.

**Observed** Gives the name of the first sample.

**Expected** gives the name of the second sample.

**Chi-squared** is the test statistic.

**DF** is the degrees of freedom.

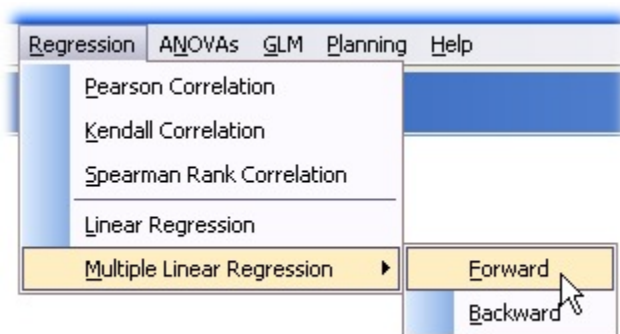
**Prob.** is the probability that the observed and expected come from the same distribution.

Results: Chi Squared	
Observed	Teacher1
Expected	Teacher2
Chi squared	0.741
DF	5.000
Prob.	0.981
This means	The null hypothesis is accepted. The observed and expected means are not significantly different (Chi Squared = 0.741, DF = 5, P = >0.05)

See [Two-sample Chi-squared](#)<sup>[127]</sup> for further information.

## 3.7 Regression drop-down menu

The **Regression** drop-down menu offers a range of methods for comparing two samples.



[Pearson Correlation](#)<sup>[133]</sup> - to calculate the Pearson Correlation between two variables.

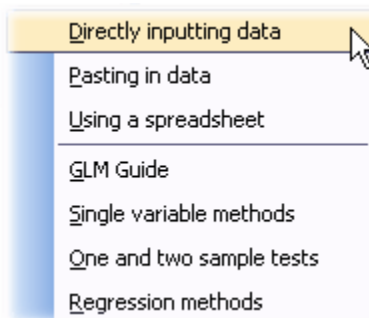
[Kendall Correlation](#)<sup>[64]</sup> - to calculate the Kendall Correlation between two variables.

[Spearman Rank Correlation](#)<sup>[65]</sup> - to calculate the Spearman Correlation between two variables.

[Linear Regression](#)<sup>[66]</sup> - to fit a linear equation to two variables.

[Multiple Linear Regression](#)<sup>[68]</sup> - to fit a linear equation with two or more independent variables.

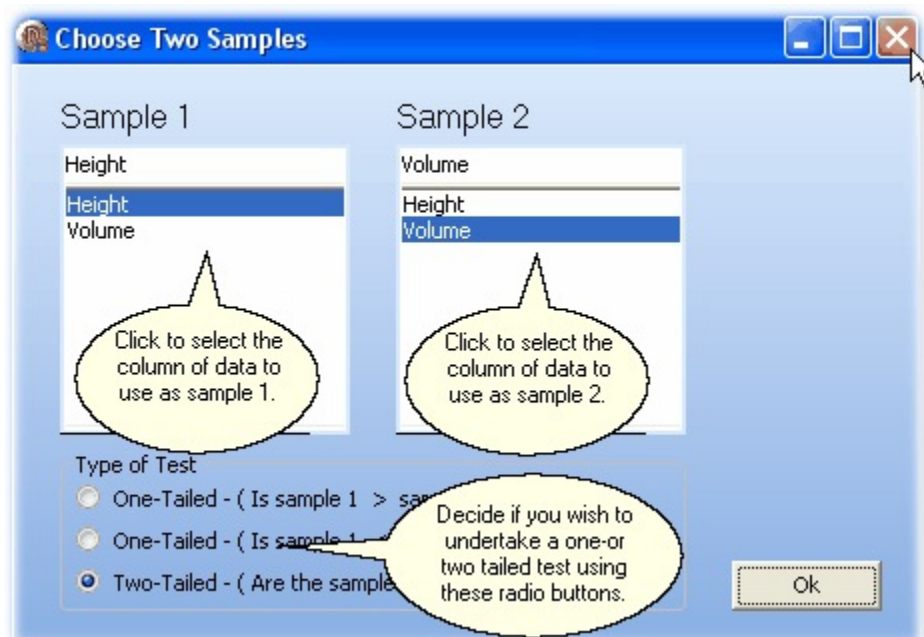
Watch **Help|Guides - Regression methods** to see how to use these options.



### 3.7.1 Pearson Correlation - setup dialog

If **Regression|Pearson Correlation** is selected, the following dialog box is opened which you use to select the two samples to correlate.

In the lists under Sample 1 and Sample 2 left-click to choose the samples to correlate.



Once your selections have been made click OK to see the [results](#)<sup>[63]</sup>.

### 3.7.1.1 Pearson Correlation - results

The results are presented in a single grid.

Results: Pearson Correlation	
Samples	Type I vs. Type II
Correlation Coef. r	0.056558
N	7
DF	5
t	0.126669
Prob.	0.904139
This means	There is no significant correlation between the two variables ( $r = 0.056558$ , $t = 0.126669$ , $DF = 5$ , $P = >0.05$ )

**Correlation Coef, r** is the [Pearson correlation coefficient](#)<sup>[133]</sup>.

**N** is the number of paired observations.

**t** is the test statistic.

**DF** is the degrees of freedom.

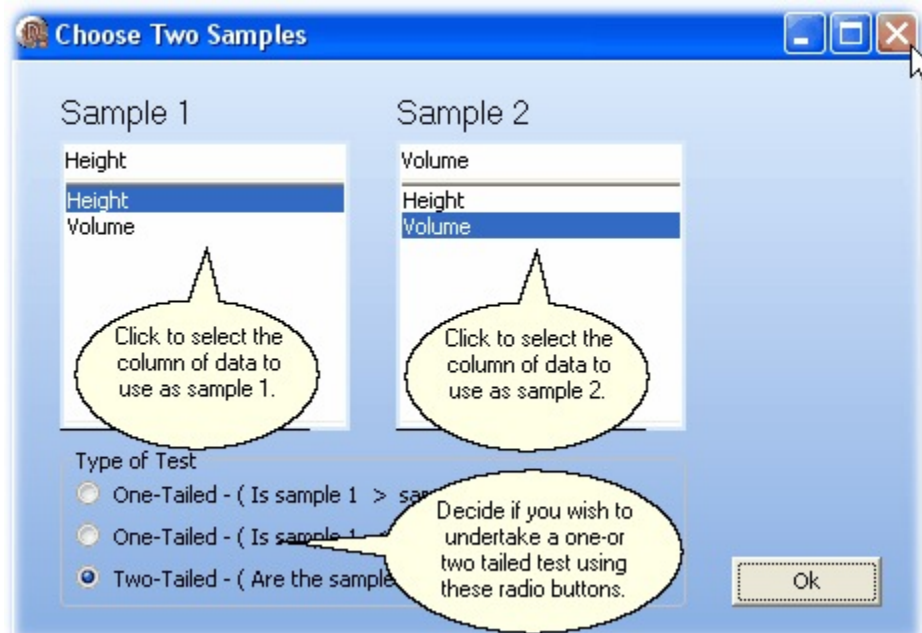
**Prob.** is the probability that the observed level of correlation could have occurred by chance.

See [Pearson correlation coefficient](#)<sup>[133]</sup> for more information.

### 3.7.2 Kendall Correlation - setup dialog

If **Regression|Kendall Correlation** is selected, the following dialog box is opened which you use to select the two samples to correlate.

In the lists under Sample 1 and Sample 2 left-click to choose the samples to correlate.



Once your selections have been made click OK to see the [results](#) <sup>64</sup>.

#### 3.7.2.1 Kendall Correlation - results

The results from the Kendall correlation analysis are presented in the Results grid.

**Tau** is the Kendall correlation coefficient.

**z** is the standardised deviate used to calculate the probability **Prob.** that the observed level of correlation could occur by random chance alone.

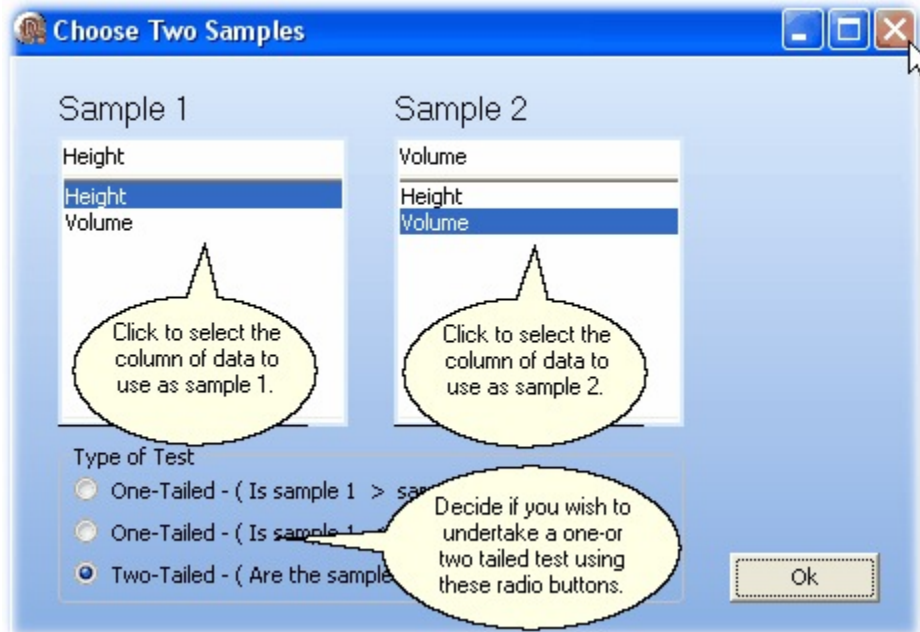
Results: Kendall Correlation	
Samples	DIAMETER vs. HEIGHT
Tau	0.317
z	2.504
Prob.	0.012
This means	There is a significant correlation between the two variables (Tau = 0.317, z = 2.504, P = <0.05)

See [Kendall Correlation](#) <sup>133</sup> for more information.

### 3.7.3 Spearman Rank Correlation - setup dialog

If **Regression|Spearman Rank Correlation** is selected, the following dialog box is opened which you use to select the two samples to correlate.

In the lists under Sample 1 and Sample 2 left-click to choose the samples to correlate.



Once your selections have been made click OK to see the [results](#)<sup>65</sup>.

#### 3.7.3.1 Spearman Rank Correlation - results

The results from the Spearman Rank Correlation are presented in the Results grid.

**r uncorrected** is the correlation coefficient.

The corrected coefficient, **r corrected**, allows for ties in the data.

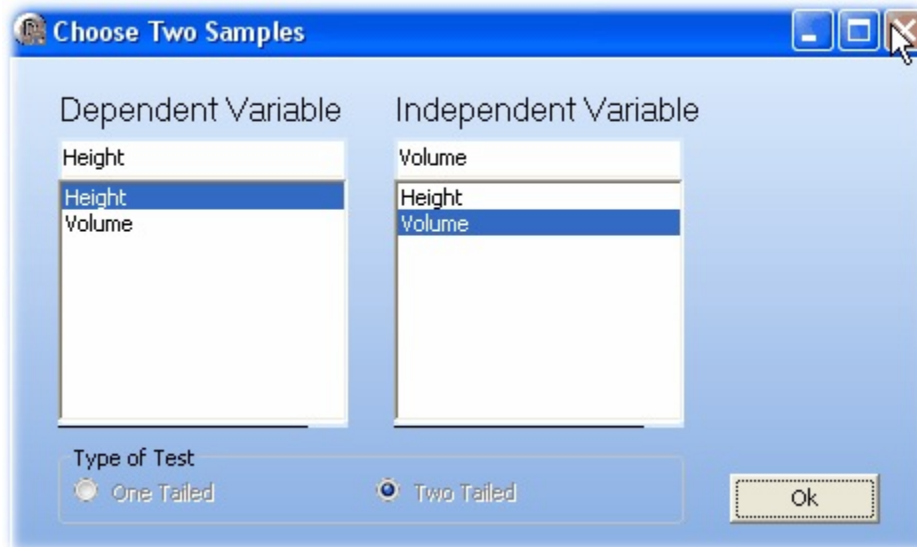
Results: Spearman Rank Correlation	
Samples	DIAMETER vs. HEIGHT
r uncorrected	0.443
r corrected	0.441
N	31
Prob.	0.013
This means	There is a significant correlation between the two variables (r = 0.441, DF = 31, P = <0.05)

See [Spearman Rank Correlation](#)<sup>134</sup> for more information.

### 3.7.4 Linear Regression - setup dialog

If **Regression|Linear Regression** is selected, the following dialog box is opened, which you use to select the dependent and independent variables for the regression. See [Linear Regression](#)<sup>[135]</sup> for information about the method. Expand and Explore is available for this method.

In the lists under Dependent Variable and Independent Variable left-click to select a variable.



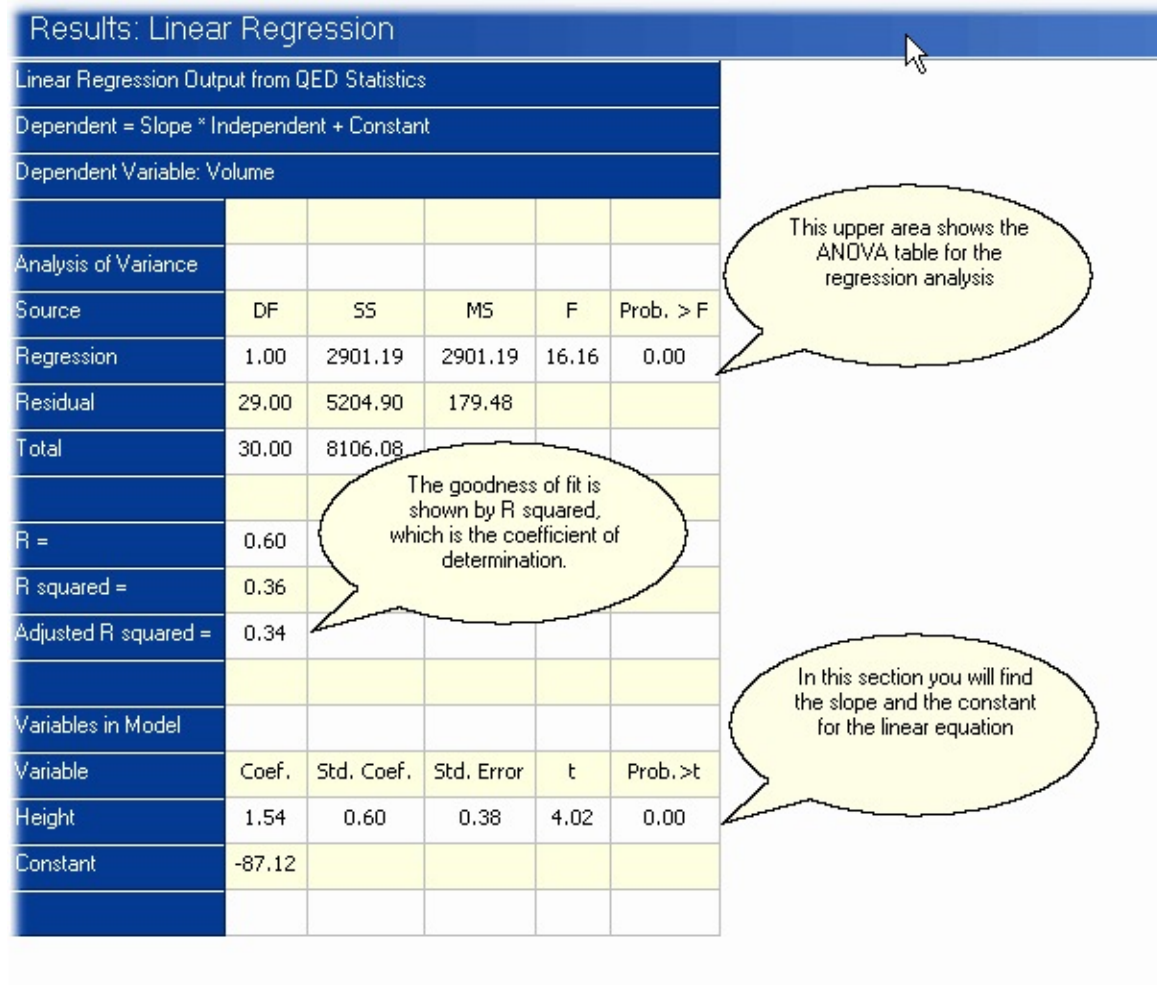
Once your selections have been made click OK to see the [results](#)<sup>[66]</sup>.

#### 3.7.4.1 Linear Regression - results

There are a number of components to the Linear Regression output. The main results are given in **Results** tab.



This presents the results in a grid, as follows:



The Analysis of Variance Table shows how much of the variability in the dependent variable is explained by the linear model. If the F value is significant, then the model explains more of the variability than would be expected by random chance.

**R** is the [correlation coefficient](#)<sup>[133]</sup>.

**R squared** is the coefficient of determination.

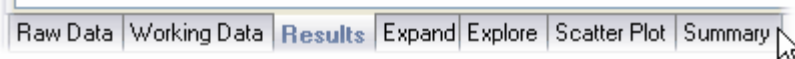
**Adjusted R squared** is the adjusted coefficient of determination.

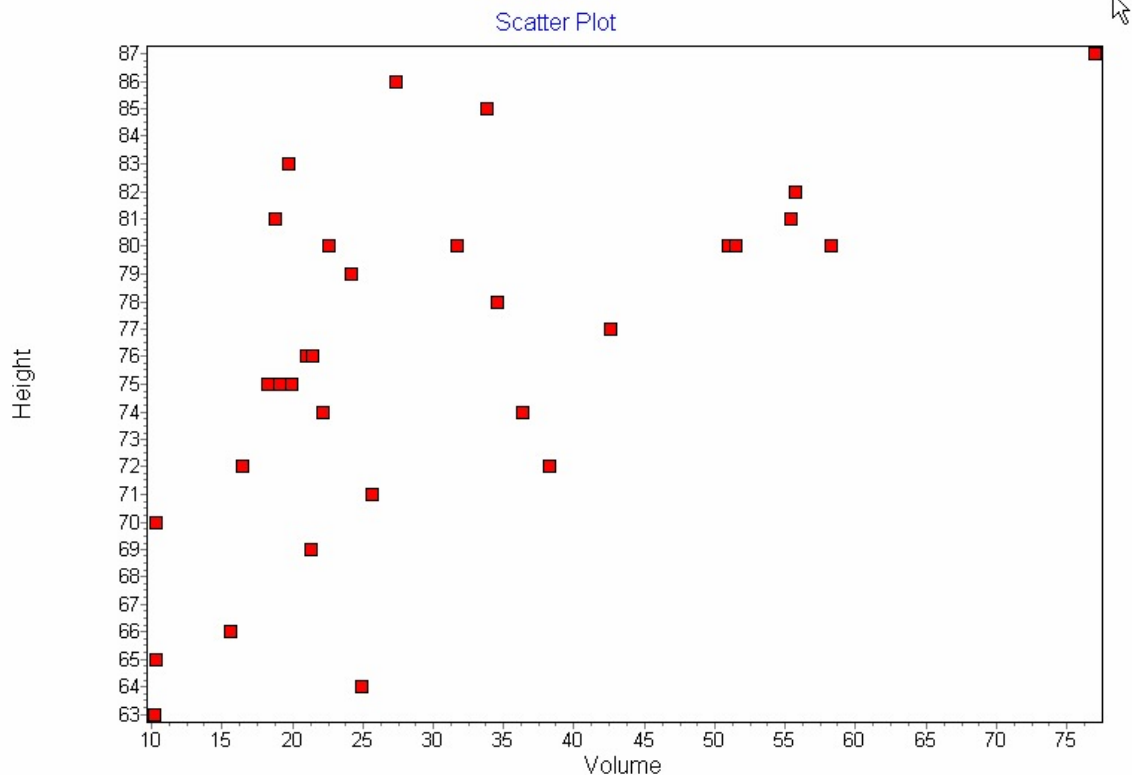
See [Linear Regression](#)<sup>[135]</sup> for more details.

Variables in the Model gives the estimated model parameters and their errors and significance. In the example above Volume was independent variable so the final equation is estimated to be

$$\text{Volume} = -87.12 + 1.54 \times \text{Height}.$$

This equation can viewed fitted to the data by clicking on the Scatter Plot tab.





To add your regression line or the line  $y = x$  to your scatter plot using the radio buttons below the scatter plot.

☐ 1/1 line  
☐ Regression line

Almost all features of your plot can be changed using the chart tool bar above the scatter plot. See [Preparing charts for output](#) <sup>169</sup>



### 3.7.5 Multiple Linear Regression - setup dialog

If **Regression**|**Multiple Linear Regression** is selected the menu offers [forward or backward stepwise](#) <sup>138</sup> multiple regression.

See [Multiple Linear Regression](#) <sup>137</sup> <sup>135</sup> for information about the method.

After the direction of the steps has been chosen the following dialog box is opened, which you use to select the dependent and independent variables for the regression.

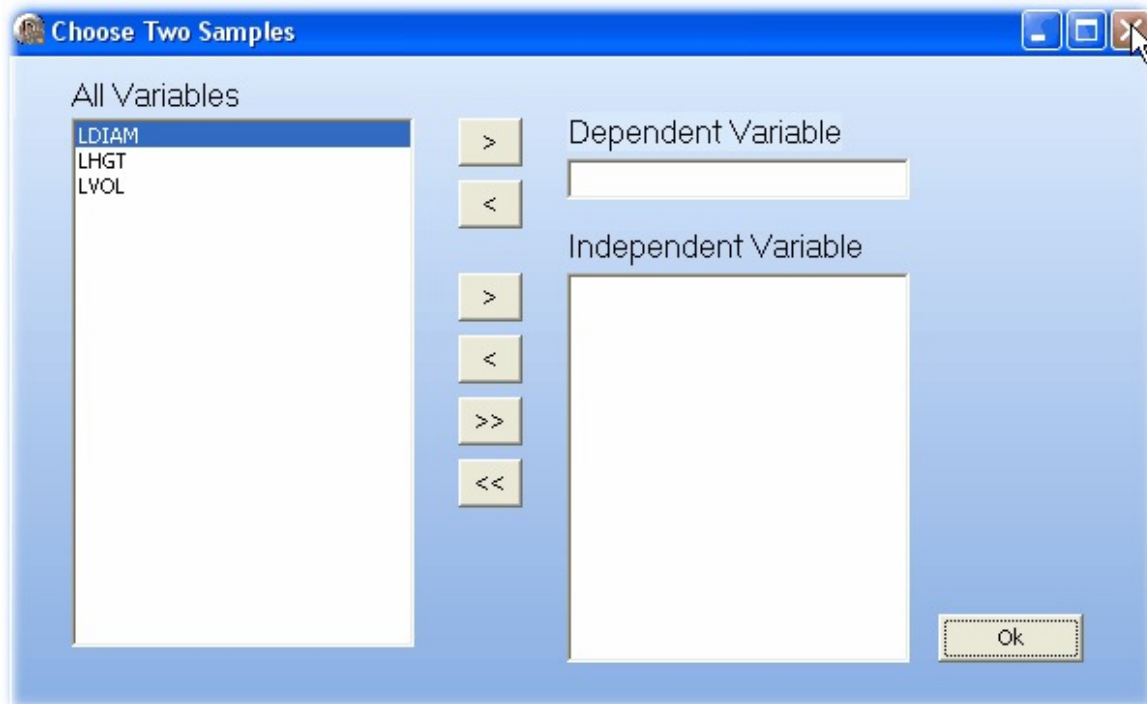
In the left-hand box, labelled "All variables", is a list of all the variables available in the open data set.

To choose a dependent variable

1. Click on a variable in the **All Variables** list so that it is highlighted in blue.
2. Then click on the > button for the **Dependent Variable** box.

Independent variables are selected in a similar fashion.

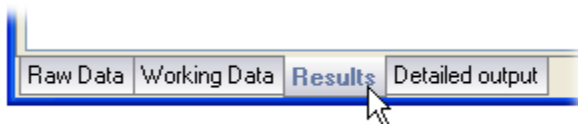
To remove or change a variable use the **< buttons** to return them to the **All Variables** list.



Once your selections have been made, click OK to see the [results](#) <sup>69</sup>.

#### 3.7.5.1 Multiple Linear Regression - results

There are two components to the Multiple Linear Regression output. The main results are given in **Results** tab.



These present the results in a grid as follows:

Results: Forward Stepwise Regression					
Stepwise Multiple Regression Output from QED Statistics					
Forward method in which parameters are added sequentially					
Dependent Variable: LVOL					
Analysis of Variance					
Source	DF	SS	MS	F	Prob. > F
Regression	2.000	8.123	4.062	613.187	0.000
Residual	28.000	0.185	0.007		
Total	30.000	8.309			
R =					
R squared =	0.989				
Adjusted R squared =	0.978				
Variables in Model					
Variable	Coef.	Std. Coef.	Std. Error	t	Prob. > t
LDIAM	1.983	0.880	0.075	26.432	0.000
LHGT	1.117	0.182	0.204	5.465	0.000
Constant	-6.632				

The Analysis of Variance Table shows how much of the variability in the dependent variable is explained by the linear model. If the F value is significant, then the model explains more of the variability than would be expected by random chance.

**R** is the [correlation coefficient](#)<sup>[133]</sup>.

**R squared** is the coefficient of determination.

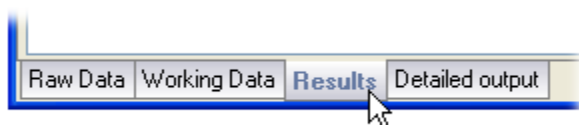
**Adjusted R squared** is the adjusted coefficient of determination.

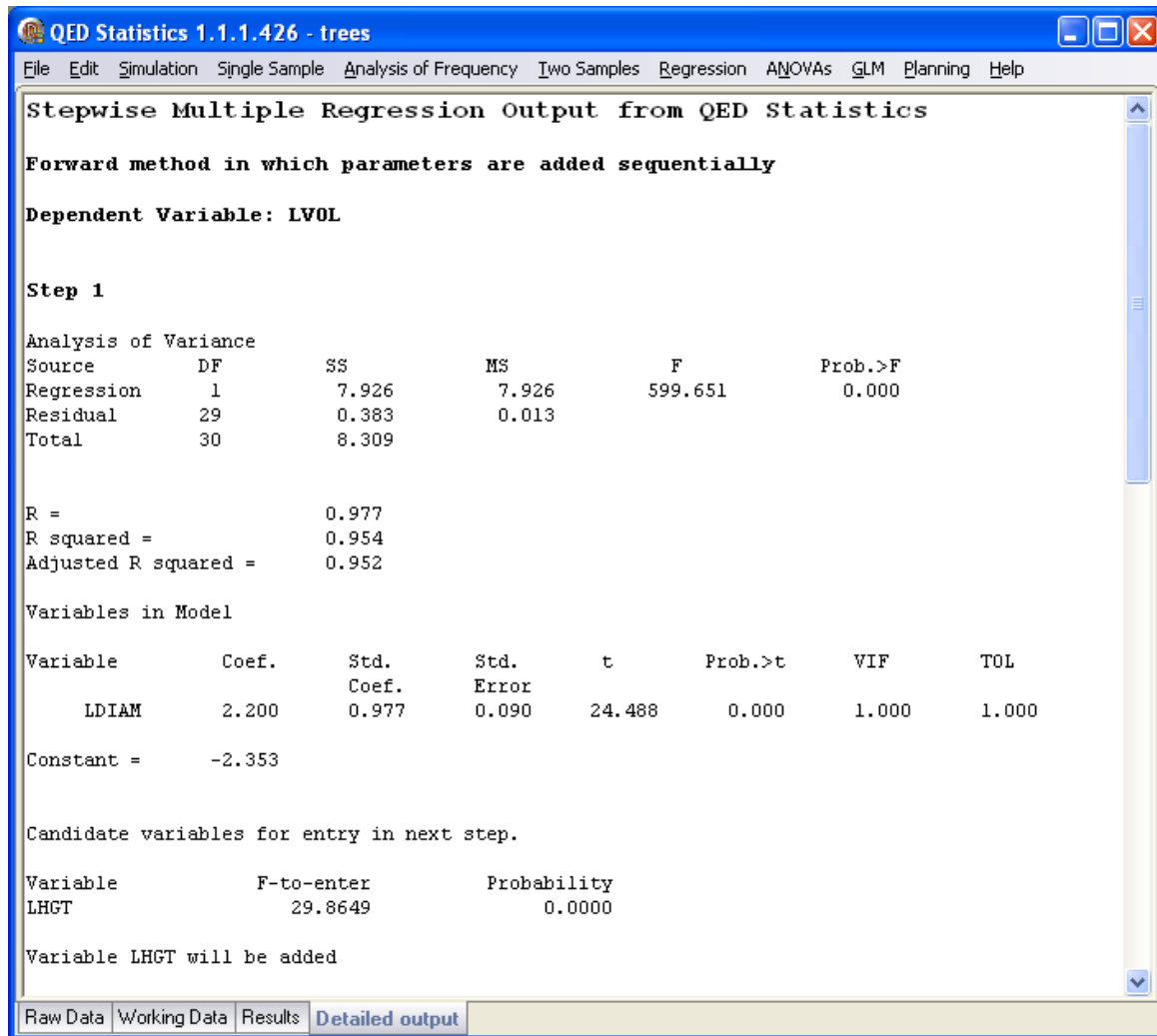
See [Multiple Linear Regression](#)<sup>[137]</sup> for more information.

Variables in the Model gives the estimated model parameters and their errors and significance. In the example above LDIAM and LHGT were independent variables so the final equation is estimated to be

$$LVOL = -6.63 + 1.98 \text{ LDIAM} + 1.12 \text{ LHGT}.$$

Detailed output from the steps during independent variable addition and removal can be found under the **Detailed output** tab





For more information on the Detailed Results tab, see [Multiple Linear Regression](#)<sup>[137]</sup> and the examples under [Stepwise Linear Regression](#)<sup>[138]</sup>.

### 3.8 ANOVAs drop-down menu

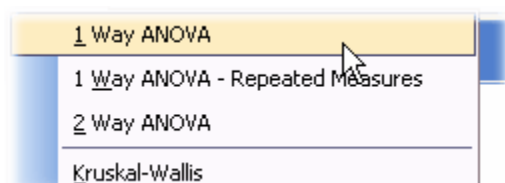
The ANOVAs drop-down menu offers one- and two-way [Analysis of Variance](#)<sup>[141]</sup> plus a non-parametric test. For more complex designs, for example a one-way nested design, QED offers [General Linear Model](#)<sup>[154]</sup> methods.

[1 Way ANOVA](#)<sup>[72]</sup> - select to undertake a one-way (single classification) Analysis of Variance.

[1 Way ANOVA - Repeated Measures](#)<sup>[74]</sup> - select to undertake a one-way (single classification) Analysis of Variance with repeated measurements of the same subjects.

[2 Way ANOVA](#)<sup>[76]</sup> - select to undertake a two-way

[Kruskal-Wallis](#)<sup>[78]</sup> - select for a one-way non-parametric test.



See also: [an example one-way ANOVA](#)<sup>[148]</sup>

If you wish, you can use the [Data Entry Wizard](#)<sup>[10]</sup> to create a new 1 Way or 2 Way ANOVA data set.

### 3.8.1 1 way ANOVA - setup dialog

If **ANOVAs|1 way ANOVA** is selected, a dialog window is opened which allows the selection of variables for analysis.

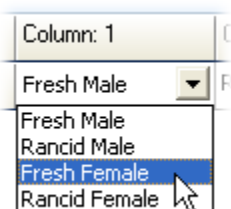
The data for each of the treatments are assumed to be arranged in columns in the [working data grid](#)<sup>[89]</sup>.

At the top of the window is a box to select the number of levels for the treatment (in the example below it is 4).

Radio buttons allow the choice between [fixed and random effects](#)<sup>[155]</sup>. The default is a fixed effects model.

The upper grid is used to select which columns hold the data for the various treatments. In our example, there were 4 different rabbits sampled for their ticks which were individually measured. Each variable comprises the measurements of ticks from one rabbit.

To select a different variable from that initially present, click on the drop-down menu and select from the variable list.



To help you in the selection of the variables the working data is shown below in the **Available data** table.

The output can also include a multiple comparison test, which is selected in the radio panel below the data selection panel. The default is **None**.

When the variables have all been selected click **OK** to run the analysis and see your [results](#)<sup>[73]</sup>.

**One way ANOVA**

Column Effect Type: ☒ Fixed ☐ Random

Rows Effect Type: ☒ Fixed ☐ Random

Test level: 0.050

Multiple Comparisons: ☒ None ☐ Tukey ☐ Scheffe ☐ Newman-Keuls ☐ Tukey-Kramer ☐ Bonferroni

Available data:

	Rabbit1	Rabbit2	Rabbit3	Rabbit4
	380	350	354	376
	376	356	360	344
	360	358	362	342
	368	376	352	372
	372	338	366	374
	366	342	372	360
	374	366	362	
	382	350	344	

Ok

See also -

[an example one-way ANOVA](#) <sup>148</sup>  
[one-way ANOVA](#) <sup>147</sup>

### 3.8.1.1 1 way ANOVA - results

The results of a one-way ANOVA are presented in a single grid.

**DF** is the degrees of freedom.

**SS** is the Sums of Squares.

**MS** is the Mean Squares

**F** is the test statistic

**Prob.** is the probability that the difference in the means of the treatments could have arisen by chance.

**Omega<sup>2</sup>** is [Omega squared](#) <sup>148</sup>, a measure of the amount of variability explained by the treatments.

**Between** gives results between treatments.

**Within** gives results within treatments

**Total** is the total variance etc.

Results: One way ANOVA						
One Way ANOVA						
	DF	SS	MS	F	Prob. >F	Omega <sup>2</sup>
Between	3	1807.73	602.58	5.26	0.00	0.26
Within	33	3778.00	114.48			
Total	36	5585.73	359.70			
This means	There is a significant difference in location between the treatments (F = 5.26, DF1 = 3, DF2 = 33, P = <0.05)					

If a multiple comparisons test has also been requested the results will be shown below the ANOVA table.

The test significance level is shown followed by the means of each level for the treatment.

This is followed by the results for each pair-wise comparison.

Newman Keuls					
Selected Significance Level	0.05				
Samples	Mean				
Rabbit2	354.40				
Rabbit3	355.31				
Rabbit4	361.33				
Rabbit1	372.25				
Samples	Difference	Statistic	DF	Prob.	Conclusion
Rabbit2 vs. Rabbit3	0.91	0.21	2.00	0.88	Means Same
Rabbit2 vs. Rabbit4	6.93	1.59	3.00	0.51	Means Same
Rabbit2 vs. Rabbit1	17.85	4.09	4.00	0.03	Means Different
Rabbit3 vs. Rabbit4	6.03	1.38	2.00	0.34	Means Same
Rabbit3 vs. Rabbit1	16.94	3.88	3.00	0.03	Means Different
Rabbit4 vs. Rabbit1	10.92	2.50	2.00	0.09	Means Same

See [One-way ANOVA](#)<sup>[147]</sup> and [an example one-way ANOVA](#)<sup>[148]</sup> for more information.

### 3.8.2 1 way ANOVA repeated measures setup dialog

If **ANOVAs|1 way ANOVA - Repeated Measures** is selected, a dialog window is opened which allows the selection of variables for analysis.

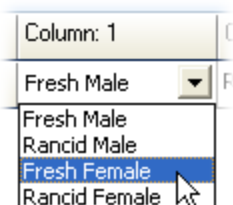
The data for each of the treatments are assumed to be arranged in columns in the [working data grid](#)<sup>[89]</sup>.

At the top of the window is a box to select the treatments (in the example below it is 4). Radio buttons allow the choice between [fixed and random effects](#)<sup>[150]</sup>. The default is a fixed effects

model.

The upper grid is used to select which columns hold the data for the various treatments. In our example, there were 4 repeated tests undertaken on 7 subjects.

To select a different variable from that initially present, click on the drop-down menu and select from the variable list.



To help you in the selection of the variables the working data is shown below in the **Available data** table.

When the variables have all been selected click **OK** to run the analysis and see your [results](#) <sup>76</sup>.

**Analysis**

Select the number of treatments (columns) here.

Column Effect Type  
☒ Fixed ☐ Random

Rows Effect Type  
☒ Fixed ☐ Random

	Column: 1	Column: 2	Column: 3	Column: 4
	test 1	test 2	test 3	rest 4

Multiple Comparisons  
 Test level: 0.050  
☒ None ☐ Tukey ☐ Tukey-Kramer ☐ Bonferroni

**Available data**

	test 1	test 2	test 3	rest 4
subject 1	14	17	14	8
subject 2	12	15	11	6
subject 3	10	12	10	5
subject 4	10	9	10	4
subject 5	9	9	8	2
subject 6	6	7	7	2
subject 7	5	7	7	2

OK

See also -

[one-way repeated measurements ANOVA](#) <sup>143</sup>

### 3.8.2.1 1 way ANOVA repeated measures results

The results of a one-way ANOVA are presented in a single grid.

**DF** is the degrees of freedom.

**SS** is the Sums of Squares.

**MS** is the Mean Squares

**F** is the test statistic

**Prob.** is the probability that the difference in the means of the treatments could have arisen by chance.

**Between** gives results between treatments.

**Within subjects** gives the results between the subjects.

**Treatments** gives the variability due to the treatments on the subjects.

**Residual** is the variance that cannot be explained by the model.

**Total** is the total variance.

Results: One way ANOVA - Repeated Measures					
Treatments by Subjects (A x S) ANOVA Results					
	DF	SS	MS	F	Prob. > F
Subjects	6	205	34.166667		
Within subjects	21	204	9.714286		
Treatments	3	185.857143	61.952381	61.464567	0.000000
Residuals	18	18.142857	1.007937		
Total	27	409	15.148148		
This means	There is a significant difference in means between the treatments (F = 61.464565, DF1 = 3, DF2 = 18, P = <0.05)				

See [one-way repeated measurements ANOVA](#)<sup>[143]</sup> for more information.

### 3.8.3 2 way ANOVA - setup dialog

If **ANOVAs|2 way ANOVA** is selected, a dialog window is opened which allows the selection of variables for analysis.

The data for each of the treatment cells is assumed to be arranged in columns in the [working data grid](#)<sup>[89]</sup>.

At the top of the window is a box to select the number of levels for treatment 1 (in the example below it is 2 - fresh or rancid).

At the left is a box to select the number of levels for treatment 2 (in the example below it is 2 male or female).

Radio buttons allow the choice between [Fixed and random effects](#)<sup>[155]</sup>. The default is a fixed effects model.

The upper grid is used to select which columns hold the data for the various treatments and levels. In our example, there were 3 observations in a 2 x 2 table.

To select a different variable from the one initially selected, click on the drop-down menu and select from the variable list.



To help you in the selection of the variables the working data is shown below in the **Available data**

table.

When the variables have all been selected click **OK** to run the analysis and see your [results](#) <sup>[77]</sup>.

The screenshot shows the 'Two way ANOVA' dialog box. It includes a 'Column Effect Type' section with 'Fixed' (selected) and 'Random' radio buttons. A callout explains: 'Select if the treatment in the columns is a fixed or random effect. Fixed is more common.' Below this is a table with two columns: 'Column: 1' and 'Column: 2'. The first column has 'Fresh Male' and 'Fresh Female' as options, and the second column has 'Rancid Male' and 'Rancid Female'. A callout states: 'The drop down menus can be used to select the column of data to put in each cell of the two-way table.' To the left, the 'Rows Effect Type' section also has 'Fixed' (selected) and 'Random' radio buttons, with a callout: 'Select if the treatment in the rows is a fixed or random effect. Fixed is more common.' Below the effect types is the 'Multiple Comparisons' section with radio buttons for 'None', 'Scheffe', 'Tukey' (selected), 'Newman-Keuls', and 'Bonferroni'. A callout says: 'Use the radio buttons to select a multiple comparison test.' At the bottom is an 'Available data' section with a grid of data. A callout points to this grid: 'This grid shows you the columns of data you have available for the analysis.' The grid contains the following data:

	Fresh Male	Rancid Male	Fresh Female	Rancid Female
Obs 1	709	592	657	508
Obs 2	679	538	594	505
Obs 3	699	476	677	539

An 'Ok' button is located at the bottom right of the dialog box.

Note that multiple comparisons tests are not offered for a two-way ANOVA. Please use a [General Linear Model](#) <sup>[154]</sup> for more detailed analysis.

See also -

[two-way ANOVA](#) <sup>[145]</sup>

[an example two-way ANOVA](#) <sup>[150]</sup>

### 3.8.3.1 2 way ANOVA - results

The results of a two-way ANOVA are presented in a single grid.

**DF** is the degrees of freedom.

**SS** is the Sums of Squares.

**MS** is the Mean Squares

**F** is the test statistic

**Prob.** is the probability that the difference in the means of the treatments could have arisen by chance.

**Omega<sup>2</sup>** is [Omega squared](#) <sup>[148]</sup> a measure of the amount of variability explained by the treatments.

**Between Columns** gives results between the levels of treatment 1.

**Between Rows** gives results between the levels of treatment 2.

**Within Groups / Error** gives results within treatments.

**Total** is the total SS etc.

Results: Two way ANOVA						
Two Way ANOVA	DF	SS	MS	F	Prob. >F	Omega <sup>2</sup>
Between Columns	1	61204.08	61204.08	41.97	0.00	0.76
Between Rows	1	3780.75	3780.75	2.59	0.15	0.03
Interaction	1	918.75	918.75	0.63	0.45	0
Within Groups/Error	8	11666.67	1458.33			
Total	11	77570.25	7051.84			
Omega <sup>2</sup>	0.78					

See [Two-way ANOVA](#)<sup>[145]</sup> and [an example two-way ANOVA](#)<sup>[150]</sup> for more information.

### 3.8.4 Kruskal-Wallis - setup dialog

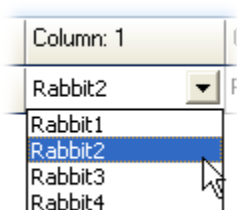
If **ANOVAs|Kruskal-Wallis** is selected, a dialog window is opened which allows the selection of variables for analysis.

The data for each of the treatments is assumed to be arranged in columns in the [working data grid](#)<sup>[89]</sup>.

At the top of the window is a box to select the number of levels for the treatment (in the example below it is 4).

The upper grid is used to select which columns hold the data for the various treatments. In our example, there were 4 different rabbits sampled for their ticks, which were individually measured. Each variable comprises the measurements of ticks from one rabbit.

To select a different variable from that initially selected, click on the drop-down menu and select from the variable list.



To help you in the selection of the variables the working data is shown below in the **Available data** table.

When the variables have all been selected click **OK** to run the analysis and see your [results](#)<sup>[79]</sup>.

**Kruskal-Wallis**

4

Column Effect Type  
☒ Fixed ☐ Random

1

Rows Effect Type  
☒ Fixed ☐ Random

Column: 1	Column: 2	Column: 3	Column: 4
Rabbit1	Rabbit2	Rabbit3	Rabbit4

Select the columns that hold the observations for the different treatments using the drop down menus here.

Multiple Comparisons Test level  
 0.050

Multiple Comparisons  
☒ None ☐ Scheffe ☐ Tukey-Kramer  
☐ Tukey ☐ Newman-Keuls ☐ Bonferroni

Available data

Rabbit1	Rabbit2	Rabbit3	Rabbit4
380	350	354	376
376	356	360	344
360	358	362	342
368	376	352	372
372	338	366	374
366	342	372	360
374	366	362	
382	350	344	

The columns of data available for selection for your analysis are shown here.

OK

See [Kruskal-Wallis test](#)<sup>[147]</sup> for further information on this method.

### 3.8.4.1 Kruskal-Wallis - results

The results of a one-way [Kruskal-Wallis test](#)<sup>[147]</sup> are presented in a single grid.

**h** is the test statistic

**DF** is the degrees of freedom.

**Prob.** is the probability that the difference in the means of the treatment levels could have arisen by chance.

**QED Statistics 1.1.1.425 - 1 way ANOVA rabbit ticks SFp208**

File Edit Simulation Single Sample Analysis of Frequency Two Samples Regression

Results: Kruskal-Wallis

H	11.500
DF	3
Prob.	0.009

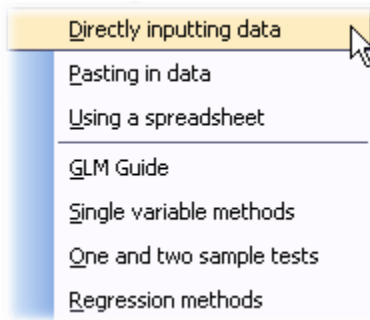
This means There is a significant difference in location between the treatments (H = 11.500, DF = 3, P = <0.05)

See [Kruskal-Wallis test](#)<sup>[147]</sup> for further information on this method.

## 3.9 GLM drop-down menu

This drop-down menu only has one item, [GLM](#)<sup>[80]</sup>, which opens a dialog window to select data and criteria for a [General Linear Model](#)<sup>[154]</sup>.

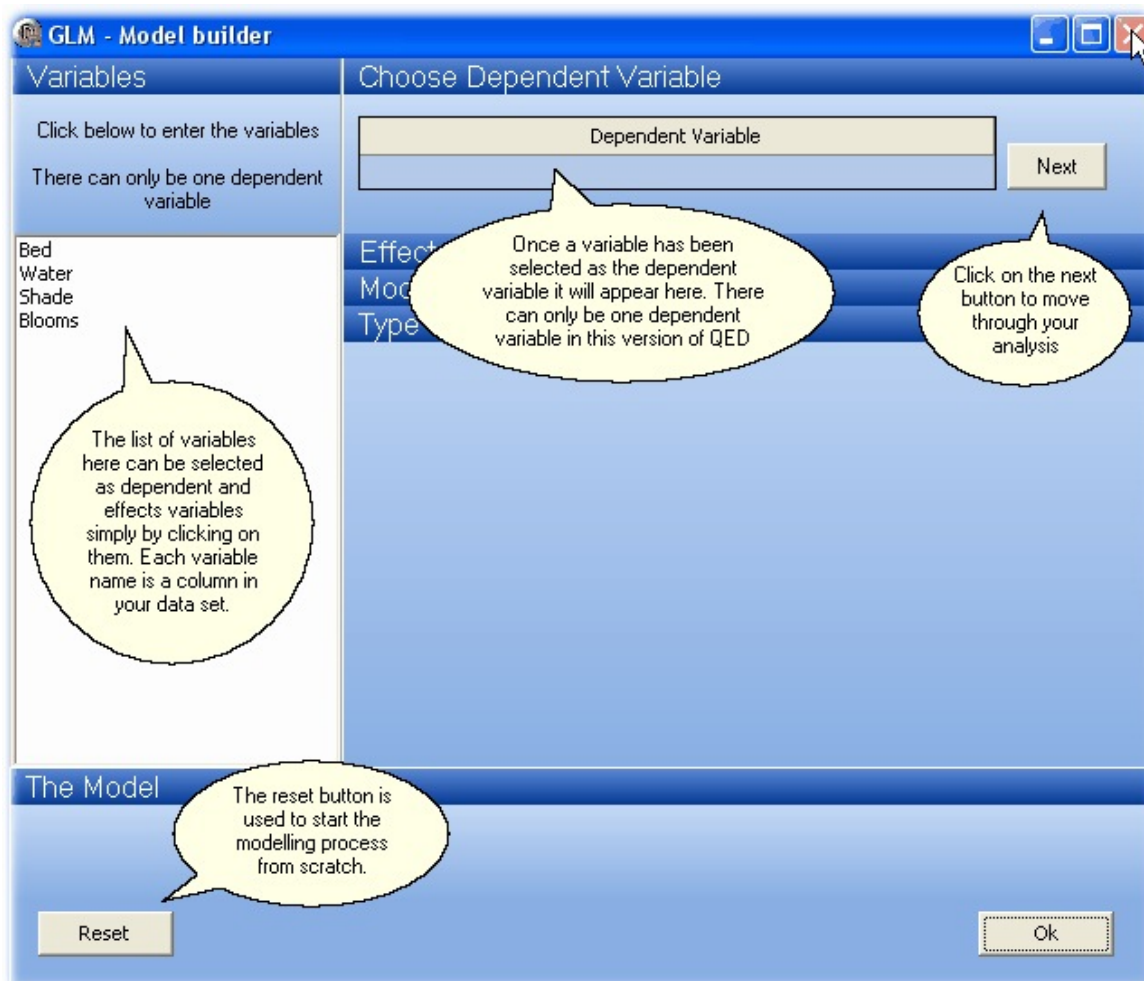
Watch **Help|Guides - GLM Guide** to see how to use this method.



### 3.9.1 GLM - setup dialog

The GLM dialog window takes you through the process of model building. See [General Linear Model](#)<sup>[154]</sup> for background information on this procedure.

Using the Variables listed in the left hand panel this dialog takes you through the model building steps.



1. Click on the dependent variable to select it, then click **Next**. If you wish to choose a different dependent variable, simply click on a different one from the list of variables, and it will be replaced.
2. Click on each variable selected as an explanatory or independent variable. The program will automatically select its choice of Variable Type. Use the **Delete** button to remove a variable. The **Back** button takes you back to the previous step.
3. If you wish to change the Variable Type that QED has automatically selected, click on the chosen Type - a drop down menu will appear to allow the selection of the variable as [Fixed](#)<sup>[163]</sup>, [Random](#)<sup>[155]</sup> or [Covariate](#)<sup>[164]</sup>. Select the appropriate type then click **Next**.

Variable Name	Variable Type
Shade	Fixed Categorical
Water	Fixed Categorical
Bed	Fixed Categorical

4. Select the interactions you want included in your model. The radio buttons allow you to decide which interactions are available for selection. Your model in words will be shown at the bottom of the dialog window.

### The Model

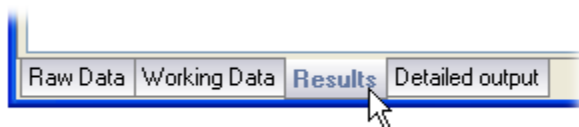
Blooms = Bed + Water + Shade + Water\*Shade

Then Click **Next**.

5. Select the [type of coding](#)<sup>[164]</sup> to be used as [Dummy](#)<sup>[165]</sup>, [Effect](#)<sup>[166]</sup> or [Orthogonal](#)<sup>[167]</sup>.
6. Click **Run** and your model will be run and the [results](#)<sup>[82]</sup> presented in a number of windows.

#### 3.9.1.1 GLM - results

There are two components to the GLM output. The main results are given in **Results** tab.



These present the results in a grid as follows:

Results: General Linear Model							
SOURCE	DF1	DF2	Inc. SS	Adj. SS	MS	F	
Bed	2.000000	16.000000	13811.349609	13811.349609	6905.674805	3.880024	0.042285
Water	2.000000	16.000000	103625.781250	6365.334961	3182.667480	1.788214	0.199110
Shade	2.000000	16.000000	36375.937500	42548.472656	21274.236328	11.953146	0.000668
Water*Shade	4.000000	16.000000	41058.140625	41058.140625	10264.535156	5.767233	0.004529
Error	16.000000		28476.835938		1779.802246		
Total	26.000000		223348.046875				

The grid presents the Analysis of Variance Table .

**DF1** and **DF2** are the degrees of freedom.

**Inc. SS** is the [incremental sums of squares](#)<sup>[159]</sup>.

**Adj. SS** is the [adjusted sums of squares](#)<sup>[159]</sup>.

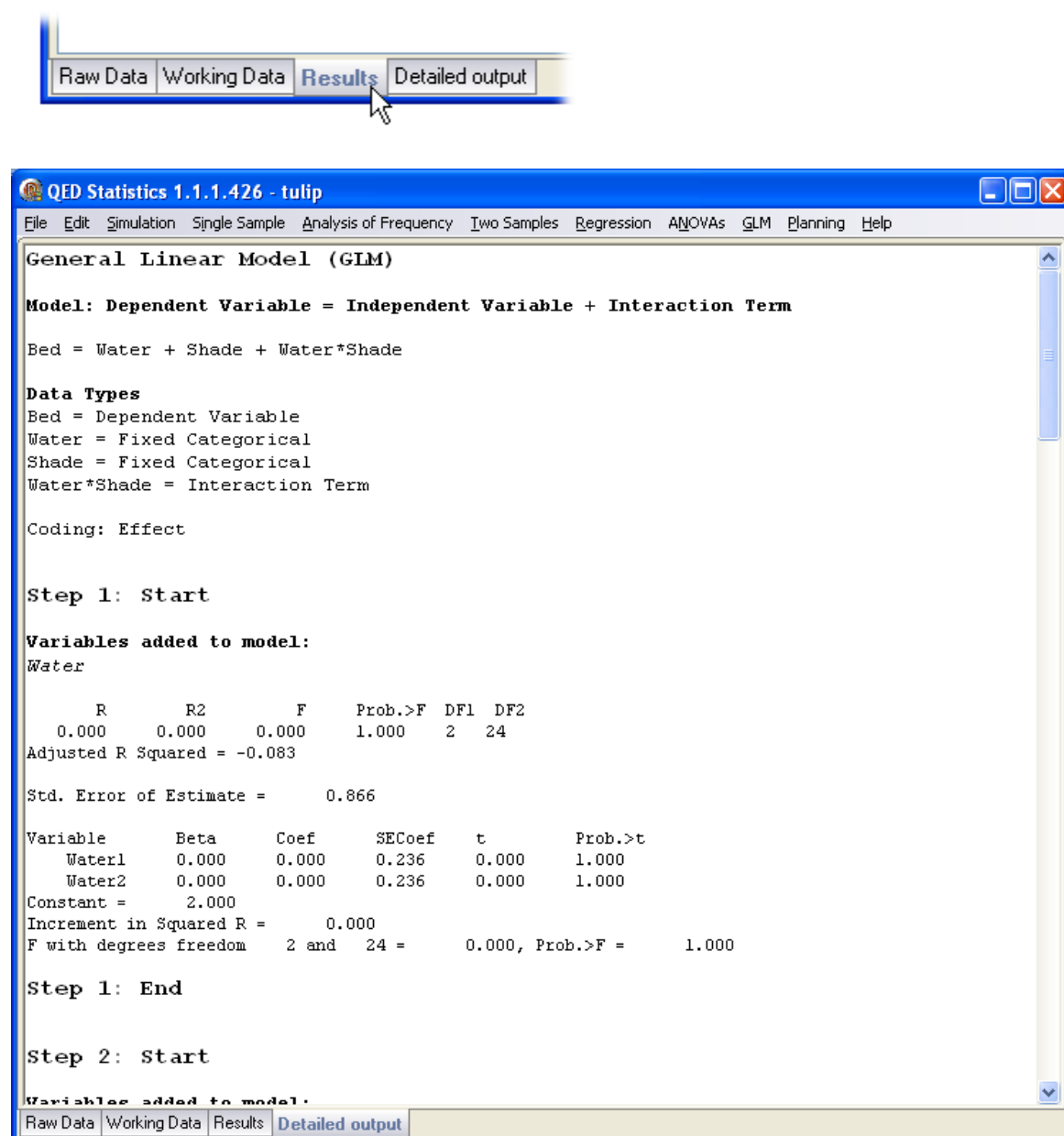
**MS** is the mean squares

**F** is the F statistic

**Prob.** is the probability that the observed effect could have occurred by chance.

See [General Linear Model](#)<sup>[154]</sup> for more information.

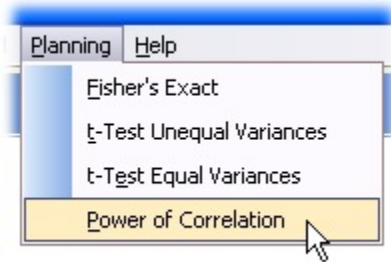
Detailed output, including stepped addition and removal of variables can be found under the **Detailed output** tab



For more information on the Detailed Results tab, see the examples under [General Linear Model](#).

### 3.10 Planning drop-down menu

The **Planning** drop-down menu offers 4 types of calculation for the statistical power of an analysis. You can use these methods to decide how many samples are required to have a good chance of proving that a difference is significant.



**Fisher's Exact** <sup>[84]</sup> - calculates the power of a 2 x 2 contingency table analysed using Fisher's exact test.

**t-Test Unequal Variances** <sup>[85]</sup> - calculates the power for detecting the differences between two means taken from distributions with different variances.

**t-Test Equal Variances** <sup>[86]</sup> - calculates the power for detecting the differences between two means taken from distributions with similar variances.

**Power of Correlation** <sup>[86]</sup> - calculates the power to detect a correlation between two variables.

### 3.10.1 Power Fisher's Exact - setup dialog

This dialog window will calculate the power of Fisher's Exact test or a Chi-squared test for a 2 x 2 [contingency table](#) <sup>[120]</sup>. The data can be arranged into a table as follows:

	Female	Male
Alive	5	8
Dead	2	1

Fill in suitable values as follows and then click **Calculate Power** to obtain the power of the test.

**Probability of Events in Group 1** - in our example the probability of female - a value between 0 and 1.

**Probability of Events in Group 2** - in our example the probability of being alive - a value between 0 and 1.

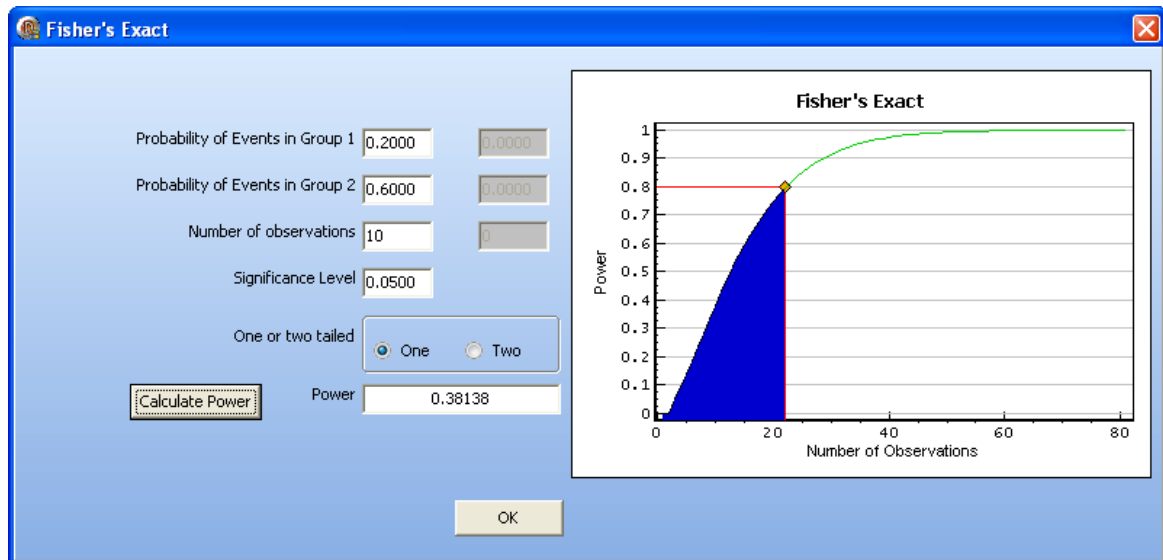
**Number of observations** - the total number of observations in the table.

**Significance level** - the significance level at which the two variables (sex and death in our example) are dependent upon each other.

**One- or two-tailed** - if you want a one-tailed or two-tailed test - choose a two-tailed test if the result would be significant if, for instance, females had either a greater or less chance of death.

**Power** - this is the calculated power for the chosen values.

The plot of the power curve showing the change in power with the number of observations is also shown. the number of observations required for a power of 0.8 (which is commonly considered a suitable value to plan for) is shown on this plot.



### 3.10.2 Power t-Test - unequal variances

This dialog window will calculate the power of a t-Test for comparing two means when the variances are assumed unequal.

Fill in suitable values as follows and then click **Calculate Power** to obtain the power of the test.

**Mean Group 1 Group 2** - enter the means of the two groups of observations.

**Standard deviation Group 1 Group 2** - enter the standard deviations for each group of observations.

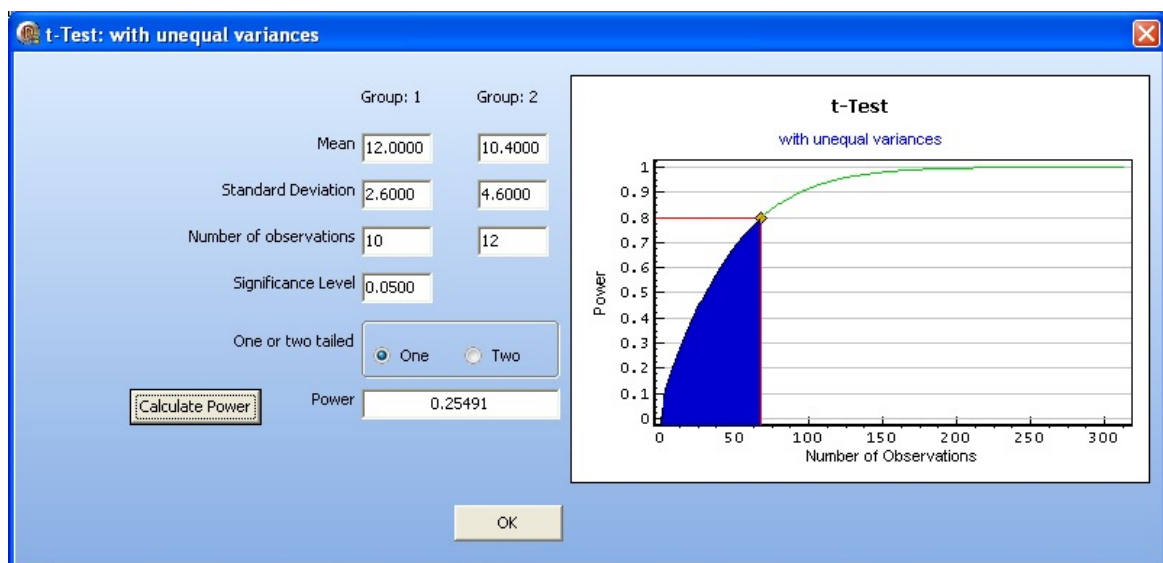
**Number of observations Group 1 Group 2** - the total number of observations in each group.

**Significance level** - the significance level at which the two variables (sex and death in our example) are dependent upon each other.

**One- or two-tailed** - if you want a one-tailed or two-tailed test.

**Power** - this is the calculated power for the chosen values.

The plot of the power curve showing the change in power with the number of observations is also shown. the number of observations required for a power of 0.8 (which is commonly considered a suitable value to plan for) is shown on this plot.



### 3.10.3 Power t-Test - equal variances

This dialog window will calculate the power of a t-Test for comparing two means when the variances are assumed equal.

Fill in suitable values as follows and then click **Calculate Power** to obtain the power of the test.

**Mean Group 1 / Group 2** - enter the means of the two groups of observations.

**Standard deviation** - enter the standard deviation for each group of observations.

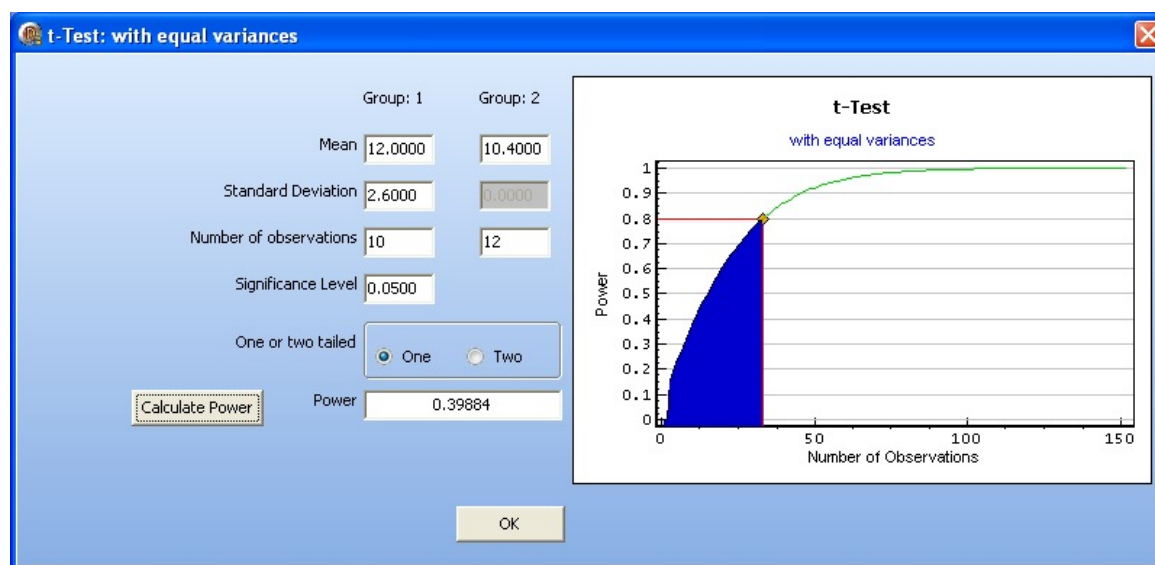
**Number of observations Group 1 / Group 2** - the total number of observations in each group.

**Significance level** - the significance level at which the two variables (Sex and death in our example) are dependent upon each other.

**One- or two-tailed** - if you want a one-tailed or two-tailed test.

**Power** - this is the calculated power for the chosen values.

The plot of the power curve showing the change in power with the number of observations is also shown. the number of observations required for a power of 0.8 (which is commonly considered a suitable value to plan for) is shown on this plot.



### 3.10.4 Power of Correlation - setup dialog

This dialog window will calculate the power to detect a certain difference in correlation between two variables.

**The null hypothesis correlation** - this will often be 0 (no correlation), but can range between -1 and +1.

**The alternative hypothesis correlation** - the level of correlation you wish to detect.

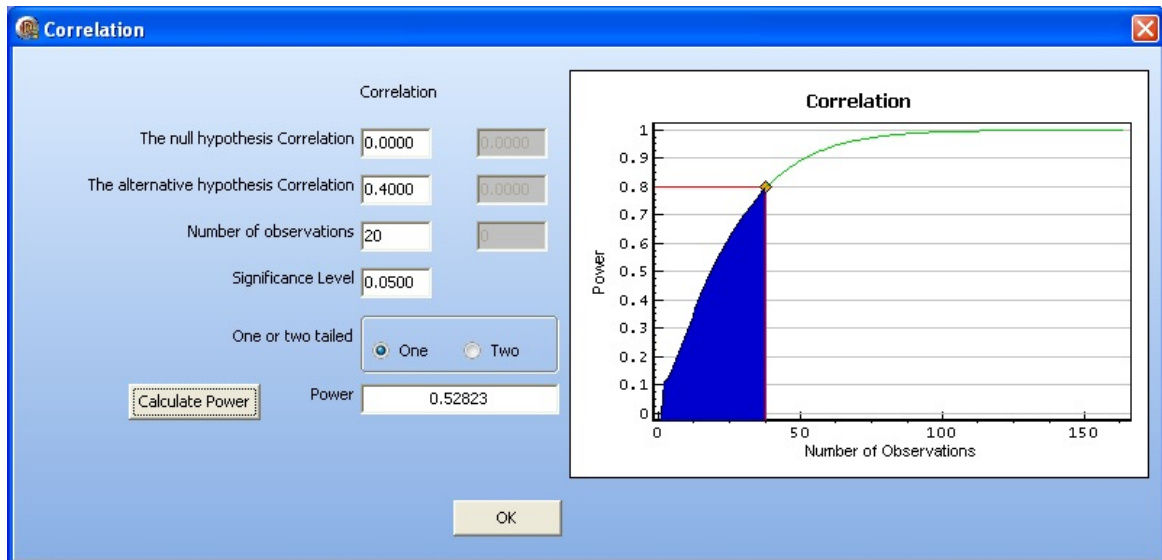
**Number of observations** - the total number of observations in the table.

**Significance level** - the significance level at which the two variables (sex and death in our example) are dependent upon each other.

**One- or two-tailed** - if you want a one-tailed or two-tailed test.

**Power** - this is the calculated power for the chosen values.

The plot of the power curve showing the change in power with the number of observations is also shown. the number of observations required for a power of 0.8 (which is commonly considered a suitable value to plan for) is shown on this plot.



### 3.11 Raw Data grid

The Raw Data grid displays the currently open data set in a grid. You also use the Raw Data grid to [directly enter new data](#)<sup>[6]</sup>, and to edit individual cells in a data set. See [Working Data](#)<sup>[89]</sup> to make [transformations and manipulations](#)<sup>[90]</sup> to your data set. If you wish, you can use the [Data Entry Wizard](#)<sup>[10]</sup> to create a new data set, in various different formats, in the Raw Data grid.

The image below shows a typical example of data. The blue row and column headers hold titles. In this example we have data for 4 variables, Bed, Water, Shade and Blooms. The rows have not been given titles. These data are in the format for a [General Linear Model](#)<sup>[154]</sup>.

Input Data				
	Bed	Water	Shade	Blooms
	1	1	1	0
	1	1	2	0
	1	1	3	111.04
	1	2	1	183.47
	1	2	2	59.16
	1	2	3	76.75
	1	3	1	224.97
	1	3	2	83.77
	1	3	3	134.95
	2	1	1	80.1
	2	1	2	85.95
	2	1	3	19.87
	2	2	1	213.13
	2	2	2	124.99
	2	2	3	65.48
	2	3	1	361.66
	2	3	2	197.13
	2	3	3	134.93
	3	1	1	10.02
	3	1	2	47.69
	3	1	3	106.75
	3	2	1	246

Raw Data Working Data Results Expa

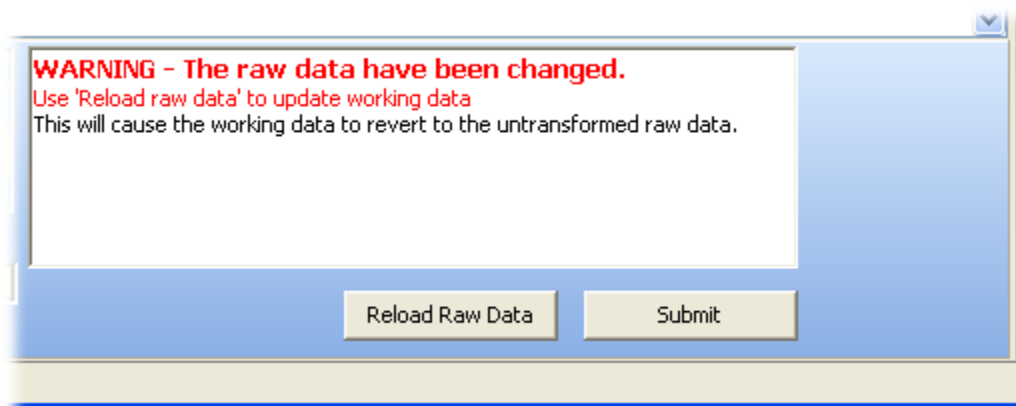
Below is another set of Data arranged for a contingency table analysis. It shows the frequency of hair colour observed in a group of boys and girls.

Input Data				
	Black	Brown	Blond	Red
Male	32	43	16	9
Female	55	65	64	16

### 3.12 Working Data grid

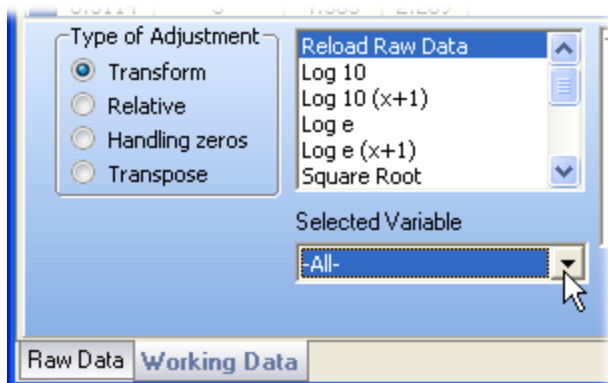
This window allows you to make a variety of changes to the [raw data](#)<sup>[87]</sup> prior to undertaking an analysis. Any changes undertaken to the working data will not change your raw data nor the saved file. If you create modified working data which you wish to save, choose **Export** from the File menu - see [Saving the Working Data](#)<sup>[94]</sup>.

If you have just entered raw data into the [Raw Data grid](#)<sup>[87]</sup> or just edited the Raw Data grid, you will be warned that the working data needs to be updated as follows. Load the data into the working grid by clicking the **Reload Raw Data** button.

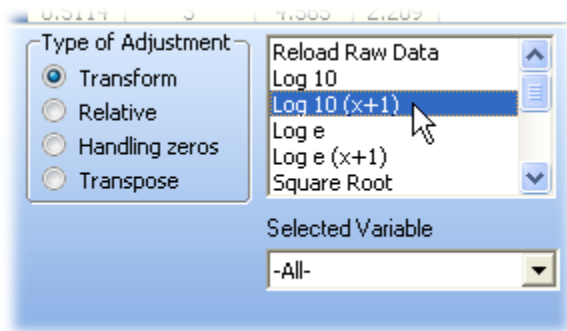


If you want to revert to the original raw data at any time simply click the **Reload Raw Data** button.

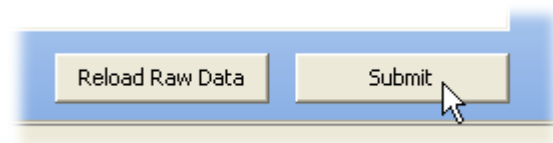
Initially, you will be presented with a grid filled with the raw data; this can be adjusted using the options in the panel below the data grid.



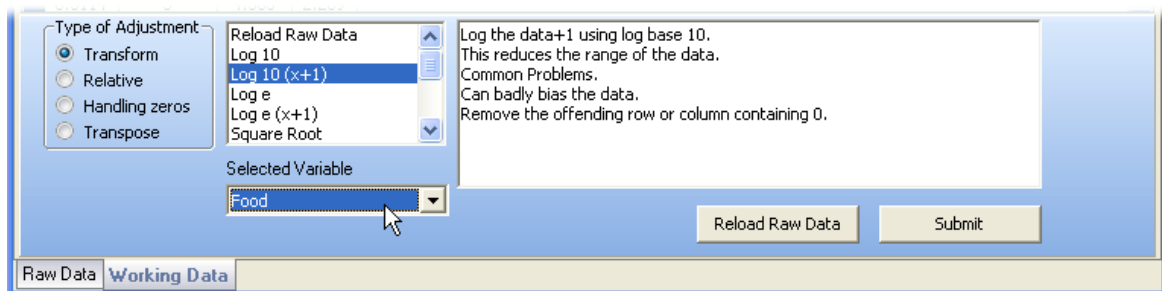
Use the radio button to select from the different types of adjustment then select from the list to the right



and click the Submit button to undertake the transformation or adjustment to your data.



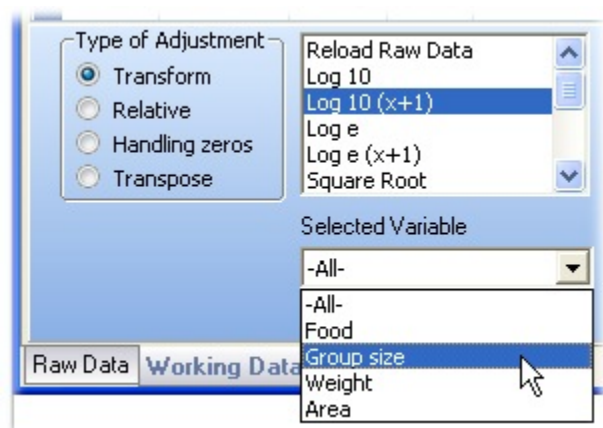
Any transformation or data manipulation you choose can be applied to all or only one variable, using the drop-down menu below the list of options.



For details about the transformations and other data manipulations available, see [Data transformations](#) <sup>90</sup>

### 3.12.1 Data transformations and manipulations

The values in the Working Data grid can be transformed and adjusted in a wide variety of ways. These changes can be useful if you wish to use a test that requires normally distributed data. They can also be used to remove zeros or unwanted observations. On the Working Data window select Transform in the Type of Adjustment panel.



See the topics below for the details on the transformations available

[Transform](#)<sup>91</sup>  
[Relative](#)<sup>92</sup>  
[Handling zeros](#)<sup>93</sup>  
[Transposing data](#)<sup>93</sup>

### 3.12.1.1 Data transformations

The Working data set can be altered without having any effect on the Raw Data. If you wish to save the Working Data under a new name use [File|Export](#)<sup>25</sup>.

The transformation options within QED Statistics are itemised below.

**Reload raw data** - This will cause the working data to revert to the raw data.

**Log(10)** - Each value is transformed to the log to base 10. This cannot be done for numbers  $\leq 0$ .

**Log10(x+1)** - Each value is transformed by adding 1 and then calculating the log to base 10. This is used when the data contains zero values.

**Log e** - Each value is transformed to the log to base e (natural logs). This cannot be done for numbers  $\leq 0$ .

**Log e (x+1)** - Each value is transformed by adding 1 and then calculating the log to base e. This is used when the data contains zero values.

**Square root** - the square root of each number is calculated. This cannot be done for negative numbers.

**Arcsin** - The Arcsin of each value is calculated. A transformation often used for percentage data.

**Arcsin root** - The Arcsin of the square root of each number is calculated.

**Power** - Each value, x, is transformed to x to the power a, where a is chosen by the user.

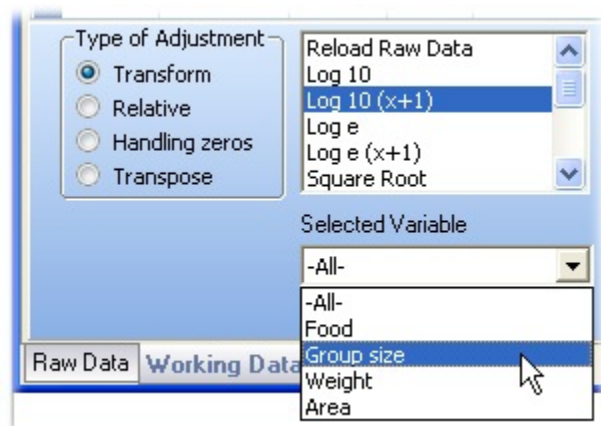
**Add constant** - A constant value, chosen by the user, is added to each value.

**Subtract constant** - A constant value, chosen by the user is subtracted from each value.

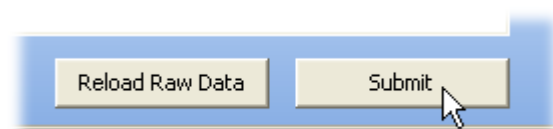
**Multiply by constant** - Each value is multiplied by a constant value chosen by the user.

**Divide by constant** - Each value is divided by a constant value chosen by the user.

Use the radio button to choose from the different types of adjustment available, then select the adjustment from the list box to the right. If you wish to apply the adjustment to only one of the columns in your data set, use the Selected Variable drop-down menu to choose the one you want. Otherwise, to apply the change to the whole of your data set, leave Selected Variable set to All.



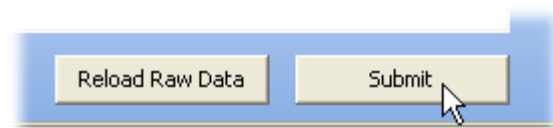
Then, click the Submit button to undertake the transformation or adjustment to your data.



Note that if a log transformation is attempted, with a 0 or negative number in the grid, the calculation will cease at that cell, which will be highlighted. You should either reload the raw data and use  $\text{Log}_{10}(x+1)$ , or edit the data set on the Raw Data grid to correct the offending cells, then reload the data into the Working Data grid.

### 3.12.1.2 Relativisations

The values in each column of the working data can be transformed so that their magnitudes are expressed relative to a variety of statistical measures. On the Working Data page select "Relative" in the Type of Adjustment panel, then click the Submit button to undertake the adjustment to your data.



The possible relative measures available within QED Statistics are given below. In each case, where by row or column is not stated you can select which will be used. Select the adjustment to be made and click Submit to make the change.

**By Maximum value** - For each column the maximum value is found and all values are divided by the maximum.

**By Mean** - For each column the mean value is found and all values are subtracted from the mean.

**By SD** - For each column the standard deviation value is found and all values are divided by the standard deviation.

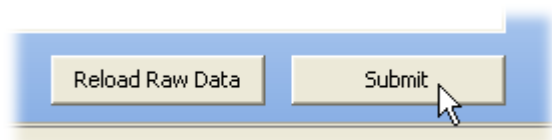
**Binary using Mean** - For each column the mean is found and all values above the mean are given the value 1 and all values below the mean zero.

**Binary using Median** - For each column the median is found and all values above the median are given the value 1 and all values below the median zero.

### 3.12.1.3 Handling zeros

Rows or columns in the working data holding zero values or missing values can be removed. On the Working Data page select Handling zeros in the Type of Adjustment panel.

Select the adjustment to be made and click on Submit to make the change.



The possible options are as follows.

**Close up Data:** Compresses a column of data ignoring any blanks.

**Missing to zero:** Puts a zero in every empty cell - this allows you to quickly enter sparse data by not entering all the zeros.

**Remove rows with missing values:** Removes observations with missing values or any variable from the analysis.

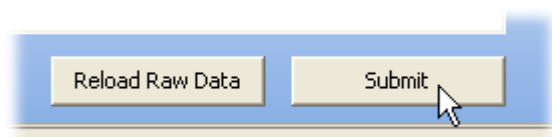
**Deselect column:** Removes the selected column (variable) from the analysis.

**Delete 0 columns** - Every column in the data set that only contains zeros is removed.

**Remove sparse columns** - Every column in the data set which contains < x non zero elements is removed. The value of x is entered by the user in the "At least x non zero value" text box.

### 3.12.1.4 Transposing data

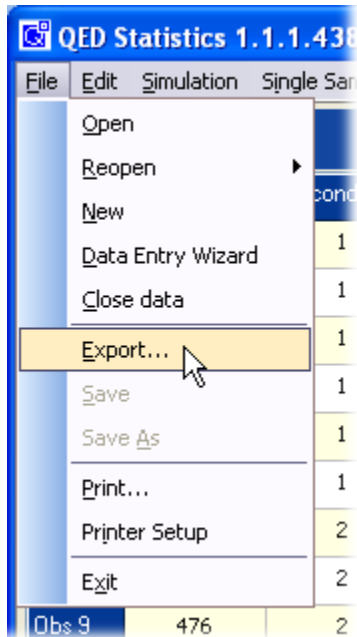
Use this option to switch the rows and columns of the data set. Like all the other adjustments it is applied to the working data set. Select transpose in the Type of Adjustment radio box and click the Submit button.



The required arrangement of data within QED Statistics is to have the variables as columns and the individual observations as the rows. If the data has been entered with the observations as columns use Transpose to switch them round.

### 3.12.2 Saving the working data

To save the working data, select **File|Export**.



This will open a dialog window offering a number of file formats. If you want the data set to be easily reopened in QED Statistics then choose the CSV file option.



### 3.13 Results tab

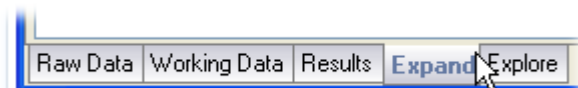
This sheet will show the main results in a grid as shown below.

Results: Pearson Correlation	
Samples	Type I vs. Type II
Correlation Coef. r	0.056558
N	7
DF	5
t	0.126669
Prob.	0.904139
This means	There is no significant correlation between the two variables (r = 0.056558, t = 0.126669, DF = 5, P = >0.05)

You can [Export](#)<sup>[25]</sup>, [Copy](#)<sup>[27]</sup> to the clipboard or [Print](#)<sup>[26]</sup> the contents of this grid.

### 3.14 Expand tab

The **Expand** tab displays a grid which shows the calculation with some intermediate steps.



For example, if Single **Sample|Mean** is selected, the mean of every variable present in the data set is displayed in the **Results** grid. The output under Expand is laid out as follows:

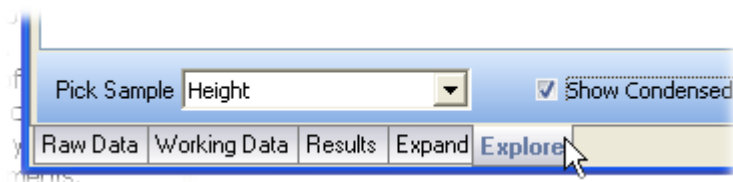
Explain the Statistic: Mean					
DIAMETER	HEIGHT	VOLUME	LDIAM	LHGT	LVOL
8.3000002	70	10.3	2.1163001	4.2484999	2.3320999
8.6000004	65	10.3	2.1517999	4.1743999	2.3320999
8.8000002	63	10.2	2.1747999	4.1430998	2.3224001
...	...	...	...	...	...
18	80	51.5	2.8903999	4.382	3.9416001
18	80	51	2.8903999	4.382	3.9317999
20.6	87	77	3.0253	4.4658999	4.3438001
Sum = 410.69995	Sum = 2356	Sum = 935.29999	Sum = 79.277679	Sum = 134.14441	Sum = 101.4549
N = 31	N = 31	N = 31	N = 31	N = 31	N = 31
Mean = 13.248385	Mean = 76	Mean = 30.170967	Mean = 2.5573444	Mean = 4.327239	Mean = 3.2727387
$\mu_x = \frac{\sum_{i=1}^n X_i}{n}$					

The variable names, Diameter, Height, Volume etc. in this case, are shown in the first blue row. The first 3 and the last 3 of each data set is then displayed. Below the data, in yellow, are shown the intermediate stages in the calculation. **Sum** is the sum of the observations, **N** is the number of observations. In green are the results, and in some calculations, the equation used to derive the result is shown.

Even further details of the calculation are available in [Explore](#)<sup>96</sup>.

### 3.15 Explore tab

The **Explore** tab displays a further series of tabbed pages which step through the calculation.



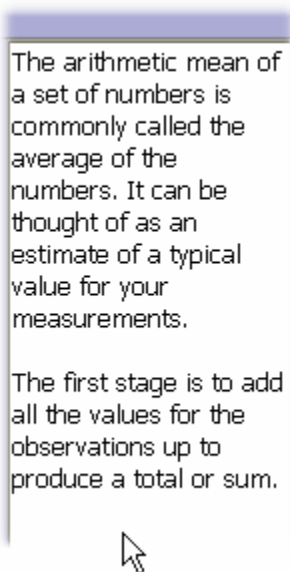
For example, if **Single Sample|Mean** is selected, the output under Explore offers the following tabs on the top row of the window:



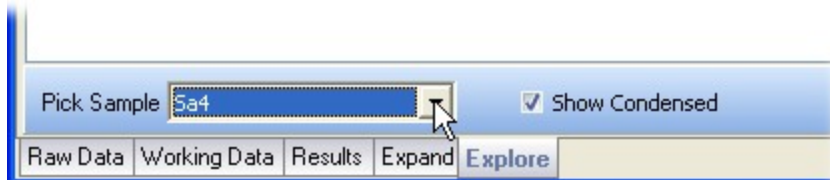
Each tabbed sheet shows a step in the calculation. For example, the first stage when calculating the mean is to find the sum of the values.

Values	Cumulative
70.000000	70.000000
65.000000	135.000000
63.000000	198.000000
...	...
80.000000	2189.000000
80.000000	2269.000000
87.000000	2356.000000
	Sum = 2356.000000

Over to the right there is a text box which describes the calculations undertaken during each step.



At the bottom of the page are the Pick Sample drop-down menu, and the Show Condensed tick box:



When you are using the Single Sample analyses on a data set with more than one sample or column, use Pick Sample to select which sample to display the analysis for.

If your data set has many rows, causing the display of the calculations to run off the bottom of the window, tick Show Condensed to display just the top and bottom 3 rows of the data set.

### 3.16 Summary tab

Summary statistics for both the raw and working data sets are displayed by clicking Summary of Data from the [Single Sample](#) <sup>[29]</sup> menu. Summary Data is not displayed until a data set has been loaded.

You can choose summary statistics for either the raw or working data sheets. Statistics can be generated for individual columns and general summary statistics for the entire data set.

When first activated the data grid will display the following column statistics for the working data.

Summary of Data													
Row	Mean	Median	Max	Min	Zeros	Non-zeros	% zeros	Sum	Sum Sqr	Total Variance	Sample Variance	Skewness	Kurtosis
Height	76.00	76.00	87.00	63.00	0	31	0.00	2356.00	180274.00	1218.00	40.60	-0.39	-0.45
Volume	30.17	24.20	77.00	10.20	0	31	0.00	935.30	36324.99	8106.08	270.20	1.12	0.77

The column statistics for the columns in the data set (Height and Volume in our example) are calculated are as follows:

**Mean** <sup>[100]</sup> - This is the mean of all the values in the data matrix.

**Median** <sup>[100]</sup> - This is the median of all the values in the data matrix.

**Max** - This is the maximum value in the data matrix.

**Min** - This is the minimum value in the data matrix.

**Zeros** - This is the number of zero entries in the data matrix.

**Non-zeros** - This is the number of non-zero entries in the data matrix.

**% Zeros** - (Number of zeros/Total number of cells) \* 100

**Sum** - This is the sum of all the values in each row or column in the data matrix.

**SumSqr** - This is the sums of squares of all the values in each row or column in the data matrix.

**Total Variance** - this is the variance of each column.

**Sample Variance**<sup>[101]</sup> - This is the estimated variance of all the values in each row or column in the data matrix.

**Skewness**<sup>[102]</sup> - This is the skewness of all the values in each row or column in the data matrix.

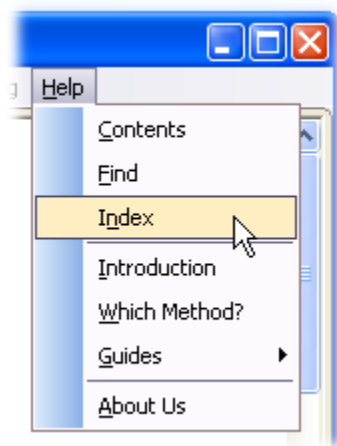
**Kurtosis**<sup>[103]</sup> - This is the kurtosis of all the values in each row or column in the data matrix.

To obtain general statistics for all columns of data select **General** in the panel at the bottom of the window.



### 3.17 Help drop-down menu and Guides

Use the Help drop-down menu to obtain help on how to use QED and the methods it uses.



There is also a help system to guide you to the best method for your purpose. Finally, Help offers a number of demonstrations which take you through the steps required for a variety of tasks. These demonstrations even include recorded guided talks.

To ensure that the guides and tutorials work correctly, please make sure that both your default internet browser, and Internet Explorer, will allow popup windows (i.e., any popup-blocking utility is disabled). Also, playing the demonstrations requires that both Internet Explorer and your default browser have the Macromedia Flash player installed as a plugin. If the plugin is not installed, the program will seek to download and install it from the internet. If your PC is connected to the internet, this process will occur automatically. If your PC is not connected to the internet, or is blocked by a firewall, this may interfere with the playing of the guides.

**Part**

---

**IV**

## 4 Single sample tests

The statistical literature offers a wide range of methods for summarising and testing a set of values collected for a single variable. QED offers the following:

[Mean](#) <sup>[100]</sup>

[Median](#) <sup>[100]</sup>

[Variance](#) <sup>[101]</sup>

[Standard deviation](#) <sup>[101]</sup>

[Skewness](#) <sup>[102]</sup>

[Kurtosis](#) <sup>[103]</sup>

[Probability Plot](#) <sup>[105]</sup>

[Box and Whisker](#) <sup>[106]</sup>

[Histogram](#) <sup>[107]</sup>

[Testing Normality](#) <sup>[107]</sup>

[t-Test](#) <sup>[114]</sup>

[z Test](#) <sup>[115]</sup>

### 4.1 Median

The median is the value that comes in the middle of a list of values which is ordered from smallest to largest.

For example, the median of 1,2,3,4,5 is 3.

It is a measure of central tendency and can be used to summarize the magnitude of a distribution of values.

The median is less sensitive to extreme scores than the mean, and this makes it a better measure than the mean for expressing the central magnitude in highly skewed distributions. For example, median income is usually more informative than mean income because the mean is increased greatly by those few individuals who earn millions of pounds or dollars per year.

To compare the medians of two samples see [Mann-Whitney Test](#) <sup>[128]</sup>.

### 4.2 Mean

The mean or average is the most common measure of the general magnitude of a range of values. The mean is calculated as the sum of the values divided by the number of values.

For example the mean of 7,8,9 is

$7 + 8 + 9 = 24$ , divided by 3, giving a mean of 8.

This is expressed mathematically as:

$$\mu_X = \frac{\sum_{i=1}^n X_i}{n}$$

It is standard practice to use the Greek letter  $\mu$  (pronounced mu) for the mean.

The mean and the [median](#) <sup>[100]</sup> are the same for symmetric distributions. In general, the mean will be higher than the median for positively [skewed](#) <sup>[102]</sup> distributions, and less than the median for negatively skewed distributions. The mean is more affected by extreme scores than the median, and is therefore not a good measure of central tendency for extremely skewed distributions.

We are describing above the arithmetic mean. There are other types of mean in occasional use including the geometric and harmonic. These are not calculated by QED Statistics.

### 4.3 Variance

The variance is a measure of the spread of a distribution. For a series of observations it is calculated as the average squared deviation of each observation from the mean of the observations. For example, for the numbers 1, 2, and 3, the mean is 2 and the variance is:

$$(1-2)^2 + (2-2)^2 + (3-2)^2 = 1 + 0 + 1 = 2, \text{ divided by 3 observations, giving a variance of } 0.667.$$

Because we are usually estimating the variance of a distribution from a subsample of observations, the estimated variance is calculated using the equation:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \mu_x)^2}{n-1}$$

where n is the number of observations and n-1 is termed the degrees of freedom. By using n-1 rather than n, a less biased estimate is produced.

The variance gives the average variability of the values about the mean expressed as squared deviations. The larger the variance the larger the spread in the data.

Note that the variance is measured as squared deviations so if the observations were lengths measured in meters, the variance is expressed in square meters.

It is not possible to mark the variance on a frequency distribution. However, we can mark the position of the square root of the variance which is called the standard deviation.

Other measures of the spread of the distribution which can be used are the range, and the first and third quartiles.

Related topics: [Comparing the variances of two populations](#)<sup>[127]</sup>

### 4.4 Standard Deviation

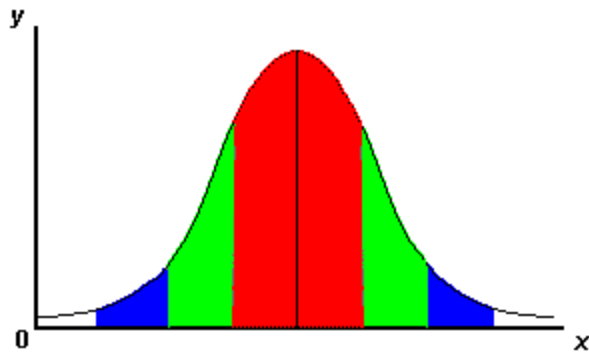
The standard deviation, s, is a measure of the spread of a distribution which can be more useful than the [variance](#)<sup>[107]</sup>. It is defined as the square root of the variance and is calculated for a sample of measurements using the equation:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu_x)^2}{n-1}}$$

where n is the number of observations and n-1 is termed the degrees of freedom. By using n-1 rather than n, a less biased estimate is produced.

The standard deviation is particularly useful when your data are [normally distributed](#)<sup>[117]</sup>. For any normal distribution one standard deviation distance away from the mean in either direction contains about 68.26 % of the total population (the red area in the graph below). 1.96 standard

deviation units away from the mean in either direction contains 95 % of the population (the green and red area below). Finally, 3 standard deviation units away from the mean in either direction contains 99.73 % of the population (the red, green and blue areas).



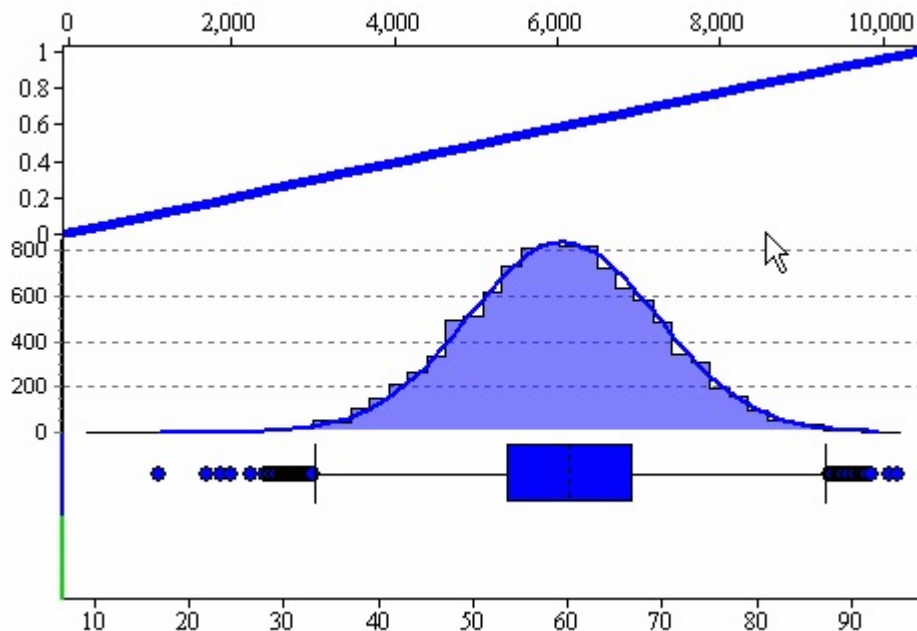
## 4.5 Skewness

Skewness is a measure of the lack of symmetry of a distribution. A distribution is termed skewed if one of the tails is longer than the other. The unbiased estimate of skewness is calculated for a series of observations  $X$  using the equation:

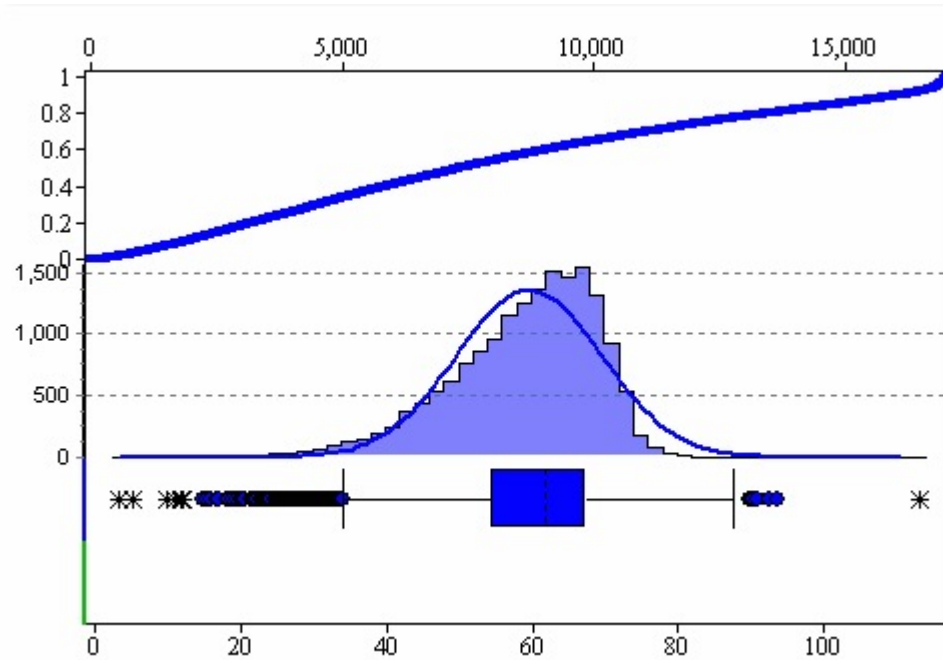
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (X_i - \mu_X)^3}{s^3}$$

where  $s$  is the standard deviation,  $n$  the number of observations and  $\mu$  mean of the distribution.

The figure below shows simulated data from a normal distribution with no skew.



In comparison the next figure shows a simulation of a skewed distribution.



Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail holds more of the distribution than the right tail. Similarly, skewed right means that the right tail is heavier than the left tail. The distribution above shows a left skew.

When summarising skewed data, the [median](#)<sup>[101]</sup> may be a better measure of the central tendency than the mean.

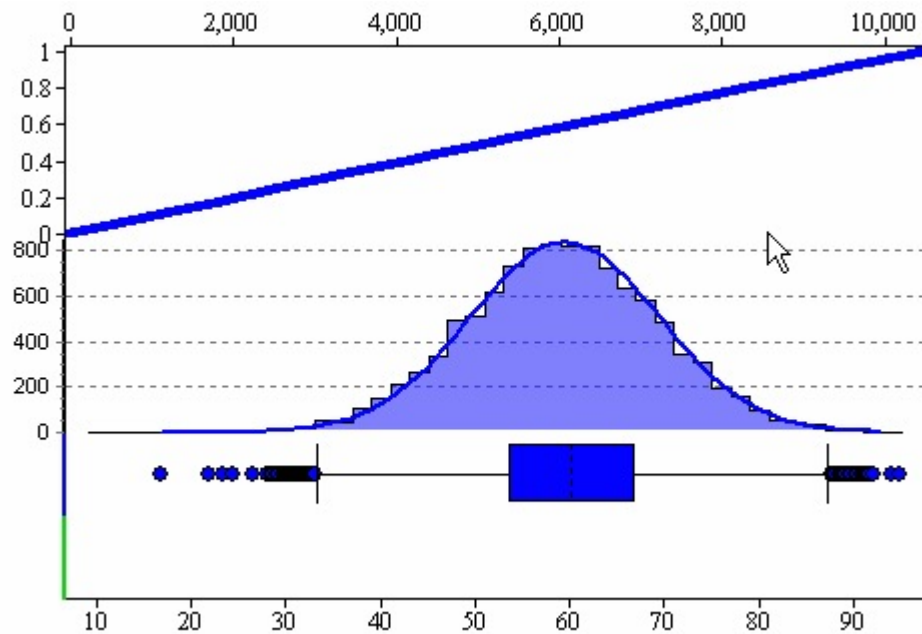
## 4.6 Kurtosis

Kurtosis is a measure of how peaked a distribution is. A high kurtosis distribution has a sharper peak than a normal distribution. The unbiased estimator for kurtosis is given by the equation:

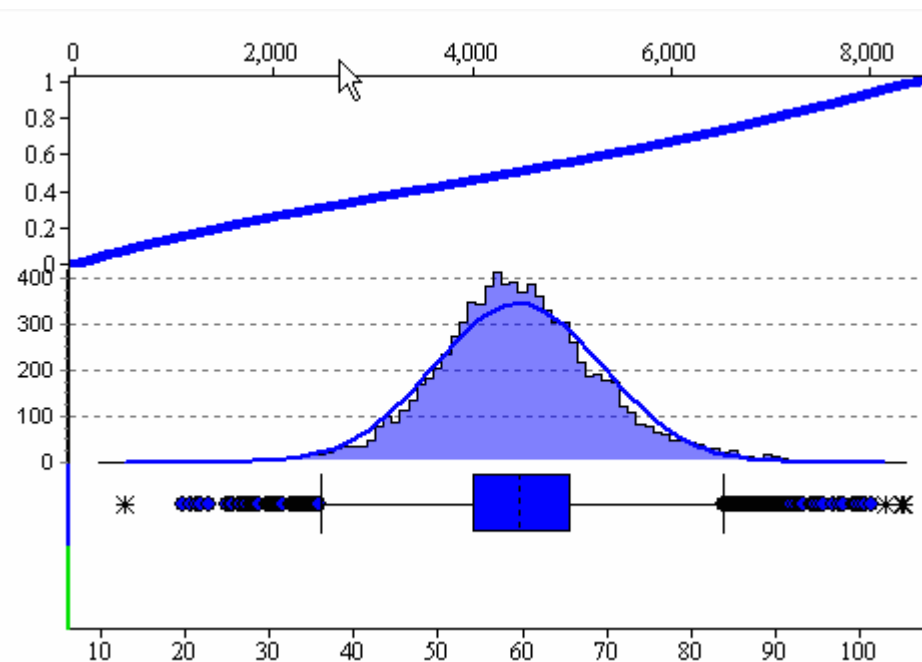
$$Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (X_i - \mu_x)^4}{s^4} - 3 \frac{(n-1)^2}{(n-2)(n-3)}$$

where  $s$  is the [standard deviation](#)<sup>[101]</sup>,  $n$  the number of observations and  $\mu$  mean of the distribution.

The figure below shows simulated data from a normal distribution with no skew.



In comparison, the next figure shows a simulation of a distribution with positive kurtosis. Note there are more observations towards the centre than would be the case for a normal distribution (plotted as a curve).



Distributions with zero kurtosis are called mesokurtic. The best example is a normal distribution.

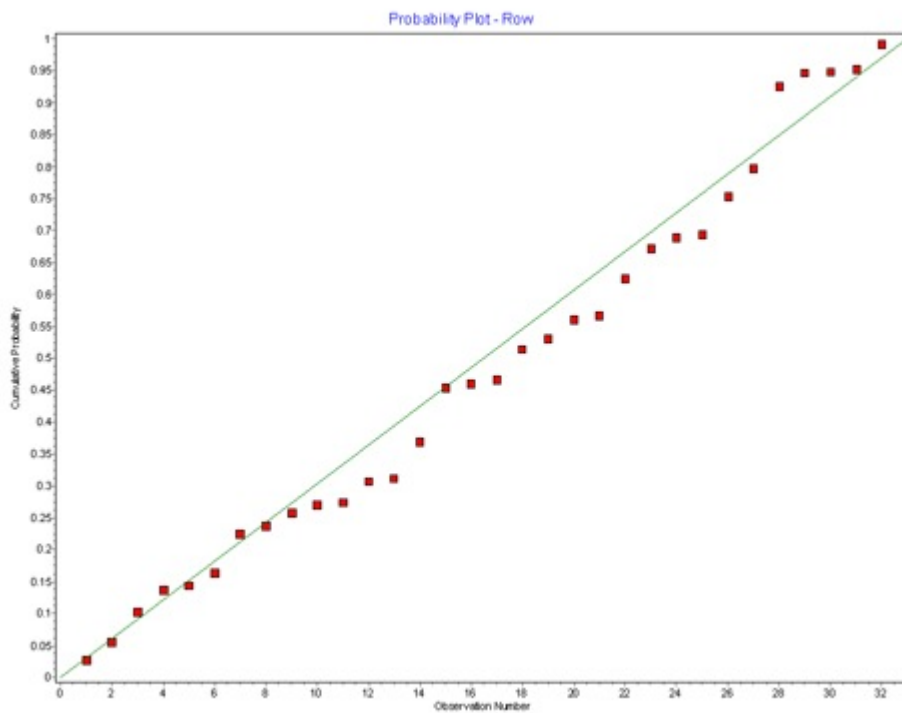
A distribution with positive kurtosis is called leptokurtic. A leptokurtic distribution has a more acute peak around the mean than a normal distribution.

A distribution with negative kurtosis is called platykurtic. A platykurtic distribution has a lower peak around the mean than a normal distribution.

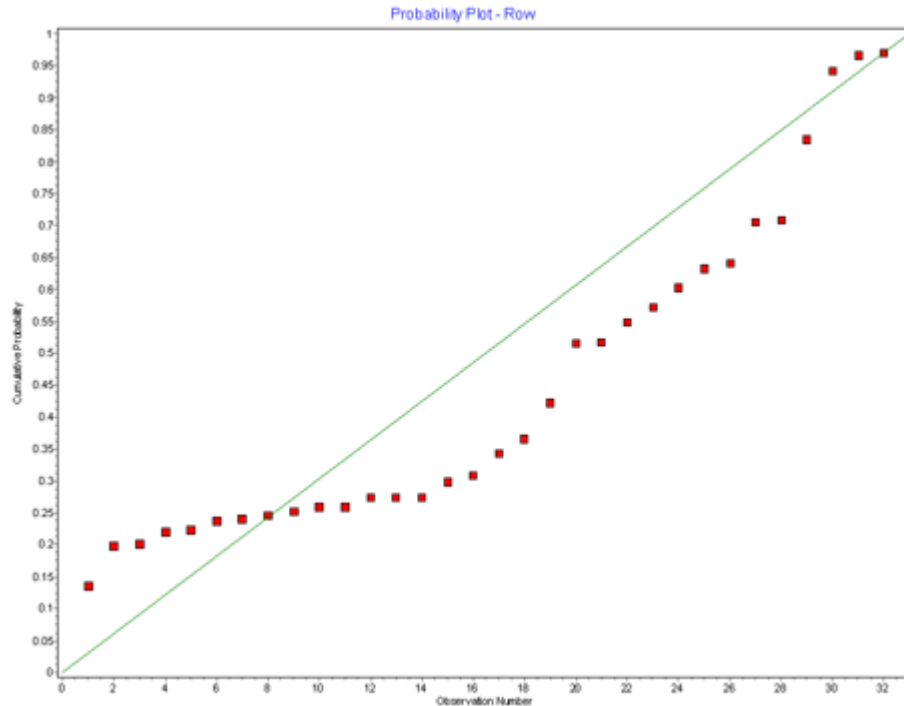
## 4.7 Probability plot

A Probability Plot provides a simple graphical method for assessing whether your observed values fit a normal distribution. The plot compares your data with what would be expected if the data were perfectly normally distributed. If the data perfectly fits a normal distribution then the probability plot will be a straight line. If not, the plot will be some sort of curve.

If the data are normal then the plot will look like the plot below with little deviation from the straight line.



If it is not normally distributed and there is considerable skewness and kurtosis it will look like the plot below, which is clearly not straight.



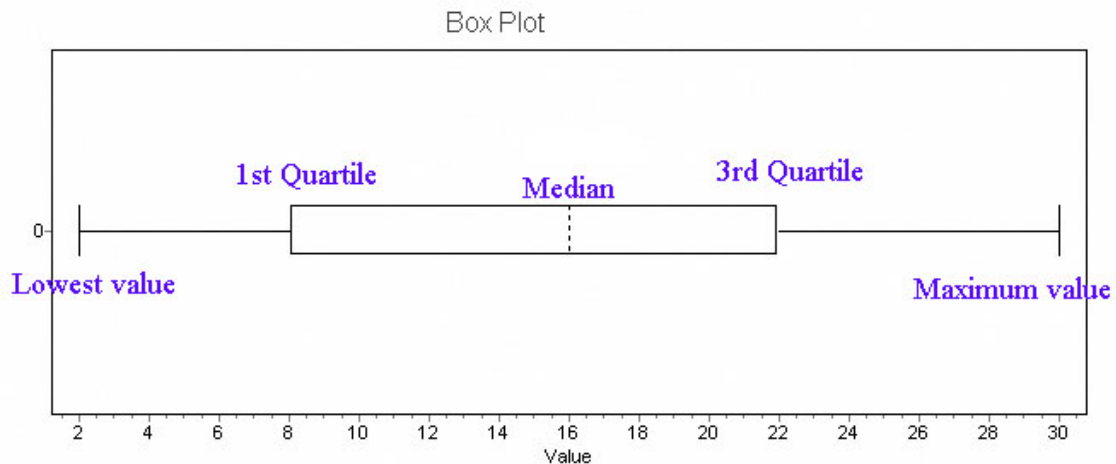
The method works by calculating the standardised normal deviate (z-value) for each observation and then using this z-value to calculate the cumulative probability distribution expected for a normal distribution. First the data is sorted from smallest to largest. Then the cumulative probabilities are calculated for each point. This expected value is then plotted against the observed cumulative probability distribution, which is simply the sample number divided by the total number of observations.

Also see [Normality testing](#) <sup>[107]</sup>

## 4.8 Box and Whisker plot

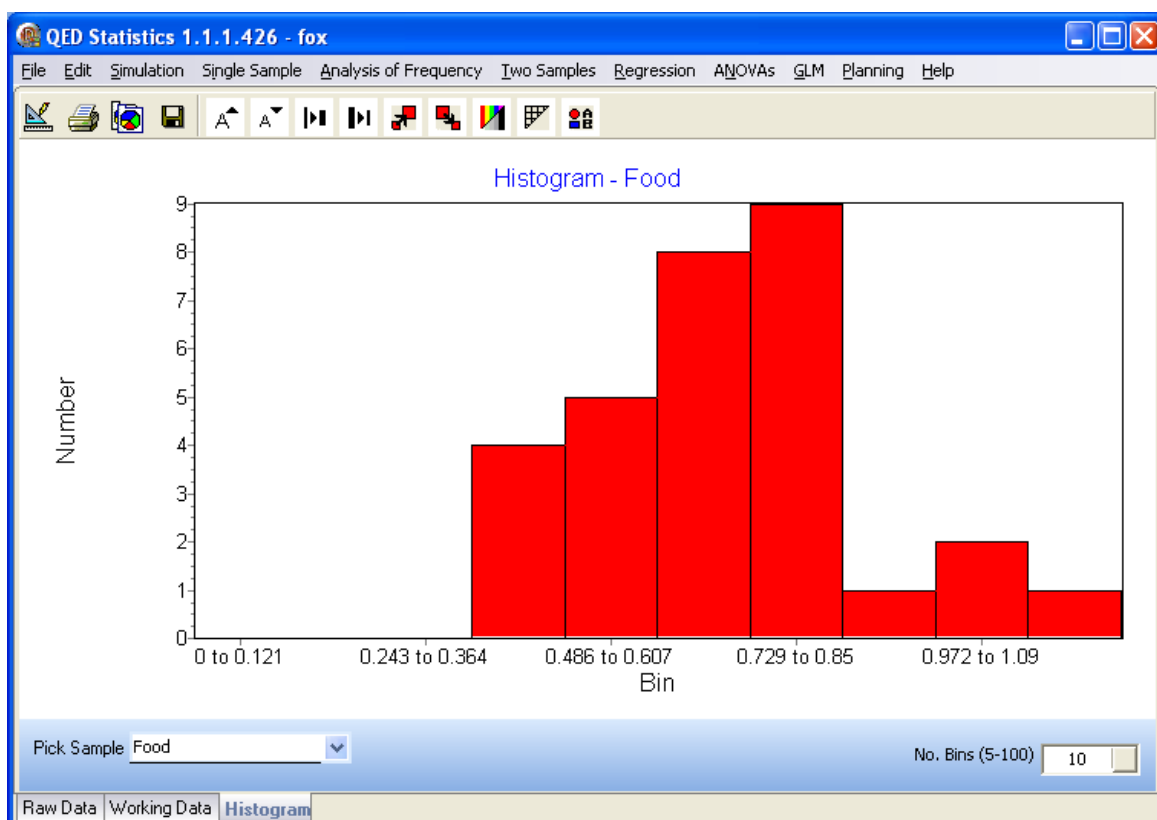
A box and whisker plot is an excellent way to summarise the characteristics of a set of data.

As shown below, the box of the diagram encloses the range from the first to the third quartile. The median value (which is also the second quartile) is shown as a dotted line within this box. The smallest and largest values in the distribution are represented as the tips of the whiskers. However, if the smallest or largest value exceeds 1.5 times the interquartile range (the difference between the 1st and 3rd quartiles) then the upper and lower adjacent values are plotted. The upper adjacent value is the largest observation that does not exceed the upper quartile plus 1.5 times the interquartile range. The lower adjacent value is the smallest observation that is not less than the lower quartile minus 1.5 times the interquartile range. When this is the case, the maximum and minimum values are plotted as a red dot (moderate outlier) or a star (extreme outlier), to show that they are outliers - see [this example](#) <sup>[33]</sup>.



## 4.9 Histogram plot

The [histogram plot](#)<sup>[33]</sup> displays the binned-up frequency of the selected variable data. This is useful for looking at the distribution of your data to check normality, outlying data points, etc. This method can be sensitive to the size and / or number of bins.



## 4.10 Normality testing

Many statistical procedures are based on the assumption that errors are normally distributed. It is therefore important to test if a variable (or its transformation) is normally distributed.

There are a number of ways that you can test whether a set of data is normally distributed. None of

these methods is entirely satisfactory. The simplest way for assessing normality is to examine the frequency distribution. If normally distributed it should form a symmetrical curve. However, it is impossible to use this approach with small numbers of observations. The other methods are described below.

[Probability plots](#)<sup>[105]</sup>

[Chi-squared test for normality](#)<sup>[110]</sup>

[Shapiro-Wilk test for non-normality](#)<sup>[108]</sup>

[Lilliefors test for normality](#)<sup>[108]</sup>

#### 4.10.1 Shapiro-Wilk test

This is a standard test for non-normality. The test statistic,  $W$ , acts like a measure of correlation between your data and their corresponding normal scores. If  $W = 1$  then your data has a perfect fit to a normal distribution. When  $W$  is significantly smaller than 1, the assumption of normality is not met.

The Shapiro-Wilk test statistic,  $W$ , is defined as:

$$W = \frac{\left(\sum_{i=1}^n a_i x_i'\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where  $n$  is the number of observations,  $x_i$  the original data,  $x_i'$  the original data sorted by magnitude,  $a_i$  is a variable calculated from the expected values for a normal distribution and  $\bar{x}$  the mean of the observations.

$W$  is the squared correlation coefficient between the sorted samples and  $a_i$  which are proportional to the normal scores and gives a measure of the straightness of the normal probability plot.

see also [Lilliefors test](#)<sup>[108]</sup>

#### 4.10.2 Lilliefors test

Lilliefors test is a test for normality which is a more applicable generalisation of the Kolmogorov-Smirnov test. The main difference between these tests is that Lilliefors test does not assume that the mean and standard deviation are known.

The Lilliefors test statistic is calculated as the difference between the observed and expected cumulative distribution function (cdf) as follows.

- The observations ( $x_i$ ) are converted to Z-scores by subtracting the observed mean ( $\bar{x}$ ) and dividing by the observed standard deviation ( $S_x$ ).

$$Z = \frac{x_i - \bar{x}}{S_x}$$

- The empirical cdf of this Z-score series is computed. To do this the Z-scores are arranged from smallest to largest and the proportion of scores less than or equal to each score is calculated.
- The cdf of the standard normal distribution is then calculated at the same probability points

using:

$$N(Z_i) = \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}Z_i^2\right]$$

- The maximum difference of the two cdfs at any point is then calculated. This is the test statistic. The critical values for this statistic are given in the tables below.

see also [Shapiro-Wilk test](#)<sup>[108]</sup>

<i>N</i>	$\alpha = .20$	$\alpha = .15$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
4	.3027	.3216	.3456	.3754	.4129
5	.2893	.3027	.3188	.3427	.3959
6	.2694	.2816	.2982	.3245	.3728
7	.2521	.2641	.2802	.3041	.3504
8	.2387	.2502	.2649	.2875	.3331
9	.2273	.2382	.2522	.2744	.3162
10	.2171	.2273	.2410	.2616	.3037
11	.2080	.2179	.2306	.2506	.2905
12	.2004	.2101	.2228	.2426	.2812
13	.1932	.2025	.2147	.2337	.2714
14	.1869	.1959	.2077	.2257	.2627
15	.1811	.1899	.2016	.2196	.2545
16	.1758	.1843	.1956	.2128	.2477
17	.1711	.1794	.1902	.2071	.2408
18	.1666	.1747	.1852	.2018	.2345
19	.1624	.1700	.1803	.1965	.2285
20	.1589	.1666	.1764	.1920	.2226
21	.1553	.1629	.1726	.1881	.2190
22	.1517	.1592	.1690	.1840	.2141
23	.1484	.1555	.1650	.1798	.2090
24	.1458	.1527	.1619	.1766	.2053
25	.1429	.1498	.1589	.1726	.2010
26	.1406	.1472	.1562	.1699	.1985
27	.1381	.1448	.1533	.1665	.1941
28	.1358	.1423	.1509	.1641	.1911

$N$	$\alpha = .20$	$\alpha = .15$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
29	.1334	.1398	.1483	.1614	.1886
30	.1315	.1378	.1460	.1590	.1848
31	.1291	.1353	.1432	.1559	.1820
32	.1274	.1336	.1415	.1542	.1798
33	.1254	.1314	.1392	.1518	.1770
34	.1236	.1295	.1373	.1497	.1747
35	.1220	.1278	.1356	.1478	.1720
36	.1203	.1260	.1336	.1454	.1695
37	.1188	.1245	.1320	.1436	.1677
38	.1174	.1230	.1303	.1421	.1653
39	.1159	.1214	.1288	.1402	.1634
40	.1147	.1204	.1275	.1386	.1616
41	.1131	.1186	.1258	.1373	.1599
42	.1119	.1172	.1244	.1353	.1573
43	.1106	.1159	.1228	.1339	.1556
44	.1095	.1148	.1216	.1322	.1542
45	.1083	.1134	.1204	.1309	.1525
46	.1071	.1123	.1189	.1293	.1512
47	.1062	.1113	.1180	.1282	.1499
48	.1047	.1098	.1165	.1269	.1476
49	.1040	.1089	.1153	.1256	.1463
50	.1030	.1079	.1142	.1246	.1457
$> 50$	$\frac{0.741}{f_N}$	$\frac{0.775}{f_N}$	$\frac{0.819}{f_N}$	$\frac{0.895}{f_N}$	$\frac{1.035}{f_N}$

### 4.10.3 Chi-squared test for normality

One approach that can be taken to test for normality is to compare the observed frequency distribution against that which would be expected if the data were [normally distributed](#)<sup>[111]</sup>. A standard Goodness-of-fit test for frequency data is the Chi-squared test.

The steps in the calculation are as follows.

- The observations are allocated to predetermined class intervals and the number of observations in each class counted. It is desirable to have at least 10 class intervals.
- The mean and standard deviation of the observations are calculated and used to calculate the expected frequency within each class interval for a normal distribution.
- To reduce bias in the test, classes are combined to ensure that no frequency class has an expected frequency of observations fewer than 5.
- The Chi-squared test statistic is calculated using the equation:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

where

$f_i$  is the observed frequency or number of counts in class,  $i$

$\hat{f}_i$  is the predicted frequency or number of counts in class,  $i$

and k is the number of classes.

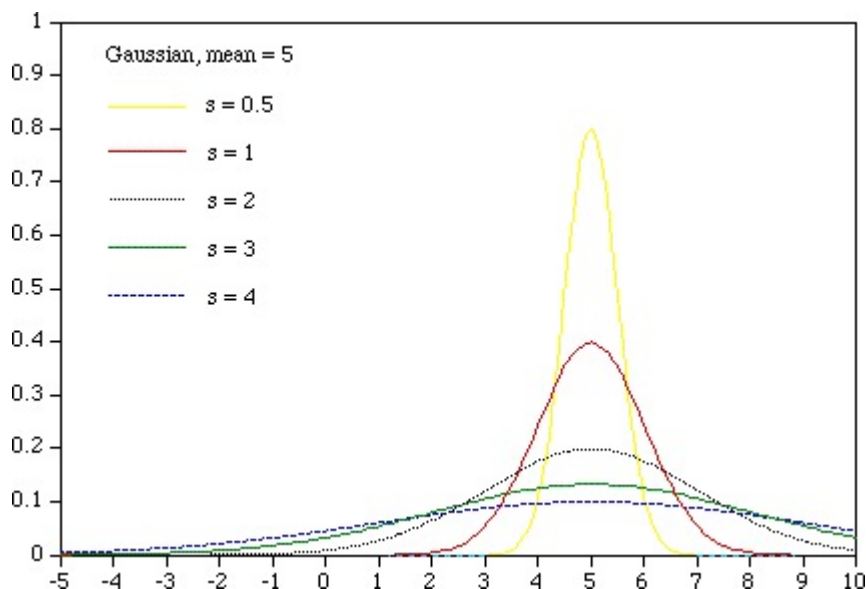
For small data sets this method is unreliable, and other methods should be considered - please see [Normality testing](#)<sup>[107]</sup>.

#### 4.10.3.1 Normal distribution

The Normal or Gaussian distribution plays a central role in statistics and has been found to be a useful model for many continuous distributions. The Normal Distribution function with [mean](#)<sup>[100]</sup> (m) and [standard deviation](#)<sup>[101]</sup> (s) is defined by the equation:

$$f(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2}$$

This equation describes a bell shaped curve as shown in the examples below.



Other distributions you may encounter include [Poisson](#)<sup>[112]</sup>, [Binomial](#)<sup>[111]</sup> and [Exponential](#)<sup>[113]</sup>.

#### 4.10.3.2 Binomial distribution

The binomial distribution describes the possible number of times that a particular event will occur in a sequence of observations. The binomial distribution is used when we are only interested in the frequency of occurrence of an event and not in the magnitude.

The binomial distribution is specified by the number of observations, n, and the probability of occurrence of the event per trial, p.

The classic example used to illustrate the binomial theory is the tossing of a coin.

If a coin is tossed 4 times, then we may obtain 0, 1, 2, 3, or 4 heads. We may also obtain 4, 3, 2, 1, or 0 tails, but these outcomes are equivalent to 0, 1, 2, 3, or 4 heads.

The likelihood of obtaining 0, 1, 2, 3, or 4 heads with a fair coin for which the probability of a head on any one toss is p = 0.5 is, respectively, 1/16, 4/16, 6/16, 4/16, and 1/16.

These probabilities are calculated using the equation:

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

where  $p(X=k)$  is the probability of  $k$  events (For example if the events is heads,  $k = 3$  and  $n = 4$ , then  $p(X=3)$  is the probability of 3 heads after 4 tosses).

If the probability of an event,  $p$ , is small then the distribution is approximately [Poisson distributed](#) [112].

Return to [normal distribution](#) [111].

#### 4.10.3.3 Poisson distribution

The Poisson distribution is used to describe rare discrete events. It was first derived by the French mathematician Poisson in 1837, and the earliest application was the description of the number of deaths from being accidentally kicked by a horse in the Prussian cavalry! More modern phenomena which might follow a Poisson distribution include child deaths, book misprints or the incidence of advantageous mutations.

The only fact you need to specify for the Poisson distribution is the mean number of occurrences for which the symbol lambda ( $\lambda$ ) is usually used.

For the Poisson it is assumed that:

- the counts are for rare events
- all events are independent
- average rate of occurrence does not change over the period of interest

The terms of the Poisson distribution are given by:

$$P(X) = \frac{e^{-\lambda} \lambda^x}{X!}$$

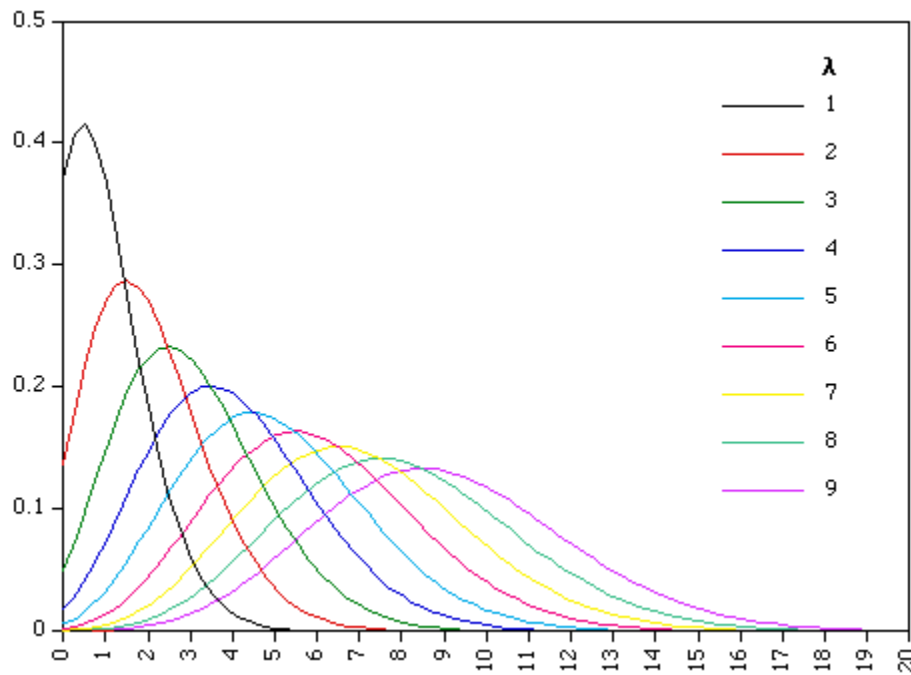
where  $P(X)$  is the probability of observing  $X$  events.

Thus, for example, the probability of zero events is

$$P(X = 0) = e^{-\lambda}$$

as the factorial of zero is 1 and  $\lambda$  to the power of zero is 1.

The graph below shows examples of the Poisson distribution.



Return to [normal distribution](#) .

#### 4.10.3.4 Exponential distribution

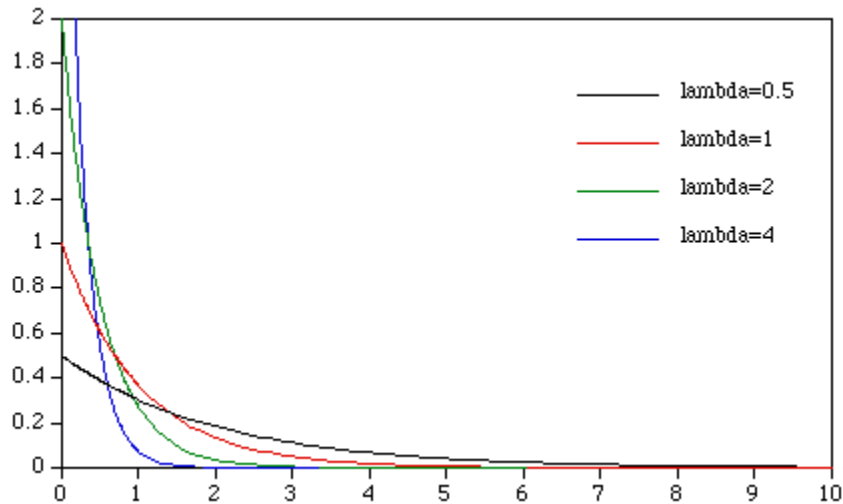
The exponential distribution is used to model the failure of components when the failure rate is a constant. The distribution describes the distribution of time between events which occur at a constant average rate.

The probability distribution function for the exponential distribution is given by the equation:

$$f(x) = \lambda e^{-\lambda x}$$

The mean of the distribution is given by  $1/\lambda$ .

The graph below shows examples of exponential distributions.



Return to [normal distribution](#)<sup>[111]</sup>.

## 4.11 t-Test : Comparing observations with a known mean

A t-Test can be used to compare a set of observations against a known mean.

Suppose a particular species is known to have a mean length of 10 mm. You could use this test to decide if a set of measurements with a mean of 11 and a standard deviation of 0.6 was likely to have come from a population with a mean of 10 mm.

The test statistic,  $t$ , is defined as:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

where

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

$\bar{x}$  is the mean of the observations,  
 $\mu$  is the hypothesized mean to be tested against,  
 $s$  is the estimated [standard deviation](#)<sup>[101]</sup> of the observations and  
 $n$  is the number of observations.

You should use a two-tailed test when you simply want to know if there is a significant difference from the mean and you are not concerned if the observed mean is larger or smaller than the known mean.

Frequently the hypothesized mean value is assumed to be zero. For example, you might look at the growth of a group of fish and wish to determine if their length had changed over the study period. The change in length of each individual could be recorded and these values, which could be positive or negative, tested against a hypothesis that there had been no growth ( $\mu = 0$ ).

If you have a large number of observations ( $n$ ) then a [z Test](#)<sup>[115]</sup> can also be used.

## 4.12 z Test : Comparing observations with a known mean

The z Test gives the probability of obtaining a random sample of observations with a calculated mean from a population with a known or hypothesized mean.

The normal deviate for the normal distribution of mean values is given by:

$$Z = \frac{\bar{X} - \mu}{\sigma_X}$$

$\bar{x}$  is the mean of the observations,

$\mu$  is the hypothesized mean to be tested against and

$\sigma$  is the [standard deviation](#) of the mean.

The standard deviation of the mean (often called the standard error of the mean) is:

$$\sigma_X = \frac{\sigma}{\sqrt{n}}$$

where

$\sigma$  is the [standard deviation](#) of the observed values and  
 $n$  is the number of observations.

We do not know the standard deviation of the population so it must be estimated from the observed [variance](#) of the observations. This estimate is only reliable if you have a large number of observations ( $n$ ). If this is not the case you should use a [t-Test](#).

You should use a two-tailed test when you simply want to know if there is a significant difference from the mean and you are not concerned if the observed mean is larger or smaller than the known mean.

# Part

---



## 5 Analysis of Frequency

QED offers 3 methods for analysing frequency of occurrence data arranged in a [contingency table](#) <sup>[120]</sup>.

[Fisher's Exact](#) <sup>[117]</sup> - to undertake Fisher's exact test on a 2 x 2 contingency table - suitable for small numbers of observations.

[Chi-squared test](#) <sup>[118]</sup> - the conventional method for comparing observed and expected frequencies.

[G-Test](#) <sup>[120]</sup> - a similar test to Chi-squared, but considered to have superior properties.

### 5.1 Fisher's Exact test

Fisher's exact test is used in the analysis of categorical data where sample sizes are small. It is named after R. A. Fisher, who devised the test (Fisher, 1922).

The test is used to examine the significance of the association between two variables in a 2 x 2 contingency table. Such tables are used when the observed frequencies are divided into two categories in two separate ways. For example, in a study of the outcome of a surgical procedure we might divide a group of patients into male or female and dead or alive

	Male	Female	Total
Dead	2	3	5
Alive	6	4	10
Total	8	7	15

The aim of the test is to investigate if the observed uneven distribution could have been generated by chance and that sex and survival are independent. To be more specific, the question we can address about our example data is knowing that 10 of these 15 patients are alive, what is the probability that these 10 would be so unevenly distributed between males and females?

We represent the frequencies in each cell by the letters A, B, C and D and call the totals across rows and columns marginal totals, and represent the grand total by N. So the table is:

	Male	Female	Total
Dead	A	B	A+B
Alive	C	D	C+D
Total	A+C	B+D	N

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{A!B!C!D!N!}$$

where the symbol ! indicates the factorial operator.

This formula gives the exact probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that the odds ratio between male and female among dead and alive equals to 1 in the population from which the sample was drawn. Fisher's exact test computes the probability, given the observed marginal totals, of obtaining exactly the frequencies observed and any configuration more extreme. By "more extreme," we mean any configuration (given observed marginal totals) with a smaller probability of occurrence in the same direction (one-tailed) or in both directions (two-tailed).

For our example data set all configurations with the same marginal frequencies include:

0	5	1	4	2	3	3	2	4	1	5	0
8	2	7	3	6	4	5	5	4	6	3	7

with corresponding probabilities:

.007      .093      .326      .392      .163      .019

Those tables outlined in yellow constitute the configurations more extreme than the observed configuration in the same direction. More extreme configurations in the same direction are identified by locating the smallest frequency in the table, subtracting 1, and then computing the remaining items given the observed marginal frequencies. Those tables outlined in green are the configurations more extreme in the opposite direction. Extremity is defined in terms of probability, so the probability of any configuration to the right of the table of observed frequencies with probability less than or equal to that of the observed configuration are added to the total probability of more extreme configurations.

Thus, the one-tailed probability for this table would be:

$$.326 + .093 + .007 = .426$$

whereas the two-tailed probability would be:

$$.326 + .093 + .007 + .163 + .019 = .608$$

The probability for the fourth configuration is not included because it is less extreme (more probable) than the observed frequency configuration.

With large samples, a Chi-squared test can be used instead of the Exact Test. Fisher's Exact Test is best when the expected values in any of the cells of the table is below 10, and there is only one degree of freedom. The Fisher test is exact, and it can therefore be used regardless of the sample characteristics. It becomes difficult to calculate with large samples or well-balanced tables, but fortunately these are exactly the conditions where the chi-squared test is appropriate.

## 5.2 Contingency table Chi-squared test

Use a Chi-squared test to determine if the observed frequency of observations in the different categories in a [contingency table](#)<sup>[120]</sup> is likely to be due to random chance.

The first stage is to [calculate the expected frequencies](#)<sup>[121]</sup> for each cell under the assumption that the variables vary independently, which is the null hypothesis.

The observed and expected frequencies are then used to calculate the test statistic with the

general formula for Pearson's chi-squared test statistic:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where  $O_i$  is the frequency observed in a cell,  $E_i$  is the frequency expected on the null hypothesis, over all cells in the contingency table.

The association between the variables in a contingency table can be measured using Chi-squared based measures - see [Contingency coefficient](#)<sup>[119]</sup> and [Cramer's V](#)<sup>[119]</sup>.

### 5.2.1 Cramer's V

Cramer's V is a popular measure of the association in a contingency table larger than 2x2. It is based on Chi-squared.

It is calculated using the formula

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where  $N$  is the total number of observations and  $k$  is the smaller of the number of rows or columns.

Cramer's V can range from 0 to 1.

For 2x 2 tables,  $k = 2$  so the  $k-1$  term becomes 1. Consequently, for 2 x 2 tables Cramer's V is equal to another association measure, phi.

See also:

[Contingency coefficient](#)<sup>[119]</sup>

[Chi-squared test](#)<sup>[118]</sup>

### 5.2.2 Contingency coefficient

The coefficient of contingency is a Chi-squared-based measure of the relation between two categorical variables.

It is calculated using:

$$cc = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

where  $N$  is the total number of observations.

The theoretical range of  $cc$  is 0 to 1 (where 0 is complete independence). However the upper limit is constrained by the size of the table, so the upper value of 1 can only be achieved with an unlimited number of rows and columns. Because it can only approach 1 for large tables, some recommend that it is only used for 5 x 5 contingency tables or larger.

See also:

[Cramer's V](#)<sup>[119]</sup>

[Chi-squared test](#)<sup>[118]</sup>

### 5.3 Contingency table G-Test

G-Tests are likelihood-ratio or maximum likelihood statistical significance tests that are increasingly being used for the analysis of [contingency tables](#)<sup>[120]</sup> where [Chi-squared tests](#)<sup>[118]</sup> were previously recommended. G-Tests have come into increasing use.

The first stage is to [calculate the expected frequencies](#)<sup>[121]</sup> for each cell under the assumption that the variables vary independently, which is the null hypothesis.

The observed and expected frequencies are then used to calculate the test statistic, G, using:

$$G = 2 \sum_i O_i \ln \left( \frac{O_i}{E_i} \right)$$

where  $\ln$  denotes the natural logarithm (log to the base e) and the sum is again taken over all cells in the contingency table.

Given the null hypothesis that the observed frequencies result from random sampling from a distribution with the given expected frequencies, the distribution of G is approximately that of [chi-squared](#)<sup>[118]</sup>, with the same number of degrees of freedom as in the corresponding chi-squared test.

For samples of a reasonable size, the G-Test and the chi-squared test will lead to the same conclusions. However, the approximation to the theoretical chi-squared distribution for the G-Test is better than for the Pearson [chi-squared tests](#)<sup>[118]</sup> in cases where for any cell  $|O_i - E_i| > E_i$ , and in any such case the G-Test should always be used.

For very small samples, [Fisher's Exact test](#)<sup>[117]</sup> is preferred to either the [Chi-squared test](#)<sup>[118]</sup> or the G-Test.

### 5.4 Contingency table

Contingency tables are used to record and analyse the relationship between two or more variables, most usually categorical variables.

As an example consider the presentation of data on handedness in men and women. The values of both variables in a random sample of 100 people can be presented in a contingency table as follows:

	right-handed	left-handed	TOTAL
male	43	9	52
female	44	4	48
TOTAL	87	13	100

The figures in the right-hand column and the bottom row are called marginal totals, and the figure in the bottom right-hand corner is the grand total. These totals are used for the [calculation of expected frequencies](#)<sup>[121]</sup>.

The table allows us to see at a glance that the proportion of men who are right-handed is about the same as the proportion of women who are. The two proportions are not identical, and the statistical significance of this difference can be tested with a [Chi-square test](#)<sup>[118]</sup>, a [G-Test](#)<sup>[120]</sup> or [Fisher's Exact test](#)<sup>[117]</sup>. Use the Exact test when the expected frequency in any cell is less than 10. The G-Test is now considered superior to the Chi-squared test although in practice there is often little difference.

While we have presented as an example a 2 x 2 contingency table, tables with many more rows and columns can be constructed and tested for independence.

## 5.5 Calculation of expected frequencies

As an example consider the presentation of data on handedness in men and women. The values of both variables in a random sample of 100 people can be presented in a [contingency table](#)<sup>[120]</sup> as follows:

	right-handed	left-handed	TOTAL
male	43	9	52
female	44	4	48
TOTAL	87	13	100

The figures in the right-hand column and the bottom row are called marginal totals and the figure in the bottom right-hand corner is the grand total.

If handedness and sex are independent then the expected frequencies are determined from the marginal totals using the equation:

$$e_{ij} = (n_{i+})(n_{+j})/n_{++}$$

where  $n_{i+}$  is the frequency for the  $i$ th row,  $n_{+j}$  is the frequency for the  $j$ th column, and  $n_{++}$  is the total frequency for the entire table.

For example, the expected number of right-handed males =  $52 * 87/100 = 45.24$

# Part

---



VI

## 6 Two sample tests

A variety of two sample tests are available within QED.

To decide if the mean or median of two samples are the same, QED Statistics offers parametric and nonparametric tests. If your samples are approximately normally distributed, or can be transformed into normally distributed variables, then use a t-Test. A number of forms of the t-Test are available to handle different types of data.

[Comparing the means of samples with related pairs of observations](#)<sup>[123]</sup>

[Comparing the means of samples with the same numbers of observations equal variance](#)<sup>[124]</sup>, <sup>[125]</sup>

[Comparing the means of samples with different numbers of observations equal variance.](#)<sup>[125]</sup>

[Comparing the means of samples with different numbers of observations unequal variance.](#)<sup>[126]</sup>

A t-Test may not be applicable if the variances of the two samples are significantly different. An [F Test](#)<sup>[127]</sup> can be used to test for differences in the variances.

If your data is non-normal then there are nonparametric tests for paired and unpaired samples:

[Mann-Whitney unpaired test](#)<sup>[128]</sup>

[Wilcoxon paired-sample test](#)<sup>[129]</sup>

If you are comparing the goodness of fit to a distribution, or your aim is to compare two distributions to determine if they are both derived from the same population, a [Chi-squared test](#)<sup>[127]</sup> can be used.

### 6.1 t-Test: Comparing means of paired samples

If the two samples to be compared were not randomly selected, and the second sample is the same as the first after some treatment has been applied, a paired test should be used. This would be the case if pairs of measurements had been taken from the same animal, or the same plot of land. In the example below, the number of leaves afflicted with rust is counted on the same tree for two years, and we wish to determine if the number has changed between years. In this case the between-tree variation in rust incidence is so great that an unpaired t-Test would not have found any significant difference.

The basis equation for the calculation of the test statistic t is:

$$t = \frac{\sum d}{\sqrt{\frac{n \sum d^2 - (\sum d)^2}{(n-1)}}}$$

where d is the difference between the pairs of values and n is the number of matched pairs.

#### Step by step description of the method

**Step 1.** Tabulate the data in pairs and calculate the difference between the paired observations. For example the following data is for the incidence of rust on apple trees:

tree	number of rusted leaves: year 1	number of rusted leaves: year 2	difference
1	38	32	6
2	10	16	-6
3	84	57	27

4	36	28	8
5	50	55	-5
6	35	12	23
7	73	61	12
8	48	29	19

**Step 2.** Calculate the mean and standard deviation of the difference. For the above example the mean = 10.5 and the Standard deviation = 12

**Step 3.** Calculate the test statistic  $t = \text{mean divided by the standard deviation} / \sqrt{n}$ , where  $n$  is the number of paired samples.

**Step 4.** Choose the level of significance required (normally  $p = 0.05$ ) and read the tabulated  $t$  value in a table with  $n-1$  degrees of freedom where  $n$  is the number of paired samples. In our example the degrees of freedom =  $8 - 1 = 7$ .

**Step 5.** If the calculated  $t$  value exceeds the tabulated value then the means are significantly different.

For a non-parametric paired test see [Wilcoxon paired-sample test](#)<sup>[129]</sup>.

If you wish to compare more than 2 measurements taken over the same subjects use a [one-way repeated measurement ANOVA](#)<sup>[143]</sup>.

See also:

[Comparing the means of samples with different numbers of observations equal variance.](#)<sup>[125]</sup>

[Comparing the means of samples with the same numbers of observations equal variance](#)<sup>[124]</sup><sup>[126]</sup>

[Comparing the means of samples with different numbers of observations unequal variance.](#)<sup>[126]</sup>

[Mann-Whitney test.](#)<sup>[128]</sup>

## 6.2 t-Test: Comparing means of samples of the same size - equal variance

A t-Test is used to compare the means of two treatments. The two treatments are assumed to have equal numbers of observations (replicates) and the same variance. The  $t$ -test compares the difference between two means in relation to the variation in the data (expressed as the standard deviation of the difference between the means)

### Step by step description of the method

**Step 1.** Record the number ( $n$ ) of observations for each treatment (the number of observations for treatment 1 is  $n_1$  and the number for treatment 2,  $n_2$ )

**Step 2.** Calculate [mean](#)<sup>[100]</sup> of each treatment.

**Step 3.** Calculate [variance](#)<sup>[101]</sup> ( $s^2$ ) for each treatment.

**Step 4.** Calculate the  $t$  value using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

**Step 5.** Calculate the degrees of freedom as:

$$\text{degrees of freedom} = (n_1 + n_2 - 2).$$

**Step 6.** Find the value for  $t$  in a table with the appropriate degrees of freedom; choose the level of significance required (normally  $p = 0.05$ ) and read the tabulated  $t$  value.

**Step 7.** If the *calculated*  $t$  value *exceeds* the tabulated value then the means are *significantly different*.

See also:

[Comparing the difference between paired samples.](#) <sup>[123]</sup>

[Comparing the means of samples with different numbers of observations equal variance.](#) <sup>[125]</sup>

[Comparing the means of samples with different numbers of observations unequal variance.](#) <sup>[126]</sup>

[Mann-Whitney test.](#) <sup>[128]</sup>

### 6.3 t-Test: Comparing means of samples of unequal size - equal variance

A t-Test is used to compare the means of two treatments. The two treatments are assumed to have different numbers of observations (replicates) and the same variance. The  $t$ -test compares the difference between two means in relation to the variation in the data (expressed as the standard deviation of the difference between the means)

#### Step by step description of the method

**Step 1.** Record the number ( $n$ ) of observations for each treatment (the number of observations for treatment 1 is  $n_1$  and the number for treatment 2,  $n_2$ )

**Step 2.** Calculate [mean](#) <sup>[100]</sup> of each treatment.

**Step 3.** Calculate [variance](#) <sup>[101]</sup> ( $s^2$ ) for each treatment.

**Step 4.** Calculate the  $t$  value using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[ \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \left( \frac{n_1 + n_2}{n_1 n_2} \right)}}$$

**Step 5.** Calculate the degrees of freedom as:

$$\text{degrees of freedom} = (n_1 + n_2 - 2).$$

**Step 6.** Find the value for  $t$  in a table with the appropriate degrees of freedom; choose the level of significance required (normally  $p = 0.05$ ) and read the tabulated  $t$  value.

**Step 7.** If the *calculated*  $t$  value *exceeds* the tabulated value then the means are *significantly different*.

See also:

[Comparing the difference between paired samples.](#)<sup>[123]</sup>

[Comparing the means of samples with the same numbers of observations equal variance](#)<sup>[124]</sup>,<sup>[126]</sup>

[Comparing the means of samples with different numbers of observations unequal variance.](#)<sup>[126]</sup>

[Mann-Whitney test.](#)<sup>[128]</sup>

## 6.4 t-Test: Comparing means from samples with unequal variances

A t-Test is used to compare the means of two treatments. The two treatments are assumed to have different numbers of observations (replicates) and that this in turn is likely to result in different variances. The  $t$ -test compares the difference between two means in relation to the variation in the data (expressed as the standard deviation of the difference between the means). This test is also known as Welch's t-Test.

### Step by step description of the method

**Step 1.** Record the number ( $n$ ) of observations for each treatment (the number of observations for treatment 1 is  $n_1$  and the number for treatment 2,  $n_2$ )

**Step 2.** Calculate [mean](#)<sup>[100]</sup> of each treatment.

**Step 3.** Calculate [variance](#)<sup>[101]</sup> ( $s^2$ ) for each treatment.

**Step 4.** Calculate the  $t$  value using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

**Step 5.** Calculate the degrees of freedom. We assume variances are unequal the degrees of freedom is estimated as as the integer part of the equation:

$$\nu = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left( \frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left( \frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

**Step 6.** Find the value for  $t$  in a table with the appropriate degrees of freedom; choose the level of significance required (normally  $p = 0.05$ ) and read the tabulated  $t$  value.

**Step 7.** If the *calculated*  $t$  value *exceeds* the tabulated value then the means are *significantly different*.

See also:

[Comparing the difference between paired samples.](#) <sup>[123]</sup>

[Comparing the means of samples with different numbers of observations equal variance.](#) <sup>[125]</sup>

[Comparing the means of samples with the same numbers of observations equal variance](#) <sup>[124]</sup> <sup>[126]</sup>

[Mann-Whitney test.](#) <sup>[128]</sup>

## 6.5 Testing for difference between two variances

Given two samples taken at random from normal populations this test is used to determine if the [variances](#) <sup>[101]</sup> of the two populations are equal. The test is called a F Test or variance ratio test.

Step by step description of the method

**Step 1.** Calculate the [variance](#) <sup>[101]</sup> of each sample.

**Step 2.** Calculate the test statistic  $F = \text{variance sample 1} / \text{variance sample 2}$  or variance sample 2 / variance sample 1, whichever is the larger.

**Step 3.** Calculate the degrees of freedom as the number of observations in sample 1 - 1 ( $v_1$ ) and the number of observations in sample 2 - 1 ( $v_2$ ).

**Step 4.** Choose the level of significance required (normally  $p = 0.05$ ) and read the critical F value from tables with  $v_1$  and  $v_2$  degrees of freedom. If the calculated F value is greater than the critical value the variances are significantly different.

## 6.6 Chi-squared two sample test

The aim of this test is to compare two distributions to determine if they are both derived from the same population. Your data needs to be divided into a number of categories or bins for each of which you have recorded the number of observations.

Chi-square is calculated by finding the difference between each observed and expected frequency for each category, squaring them, dividing each by the expected frequency, and taking the sum of the results:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where:

$O_i$  = an observed frequency

$E_i$  = an expected (theoretical) frequency, asserted by the null hypothesis.

The degrees of freedom to be used when looking up the critical Chi-squared value will depend on how the expected values are derived. If they are derived independently of the observed data then the degrees of freedom is the number of categories.

### Step by step description of the method

**Step 1.** For each bin, subtract the number of observations for each distribution and square the result

**Step 2.** For each bin, divide the squared differences by the number of observations for the expected variable and find the sum over all the bins.

**Step 3.** Choose the level of significance required (normally  $p = 0.05$ ) and read the tabulated chi-squared value in a table with  $n$  degrees of freedom where  $n$  is the number of paired samples.

**Step 4.** If the calculated Chi-squared value exceeds the tabulated value then the distributions are significantly different.

## 6.7 Mann-Whitney unpaired test

The Mann-Whitney U test is a non-parametric test to determine if there is a significant difference between the [medians](#)<sup>[100]</sup> of two samples. It is the nonparametric equivalent of the [two sample t-Test](#)<sup>[126]</sup>.

### Step by step description of the method

**Step 1.** Arrange all the observations into a single ranked series. That is, rank all the observations without regard to which sample they are from.

**Step 2.** Find the sum of the ranks in each sample. Simply add up the ranks in sample 1. The sum of ranks in sample 2 can be calculated using  $N(N + 1) / 2$  where  $N$  is the total number of observations.

**Step 3.** The test statistic  $U$  is then given by:

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$

where  $n_1$  and  $n_2$  are the two sample sizes, and  $R_1$  is the sum of the ranks in sample 1.

**Step 4.** Choose the level of significance required (normally  $p = 0.05$ ) and consult a table of  $U$  values to determine the critical value for  $n_1$  and  $n_2$  sample sizes. If the calculated value of  $U$  is greater than the critical value the medians are significantly different.

See also:

[Comparing the difference between paired samples.](#)<sup>[123]</sup>

[Comparing the means of samples with the same numbers of observations equal variance.](#)<sup>[124]</sup><sup>[126]</sup>

[Comparing the means of samples with different numbers of observations equal variance.](#)<sup>[125]</sup>

[Comparing the means of samples with different numbers of observations unequal variance.](#)<sup>[126]</sup>

## 6.8 Wilcoxon paired-sample test

This test is a nonparametric equivalent of the [paired-sample t-Test](#)<sup>[123]</sup>. The Wilcoxon test is also known as the Signed-Rank test. If the two samples to be compared were not randomly selected and the second sample is the same as the first after some treatment has been applied a paired test should be used. This would be the case if pairs of measurements had been taken from the same animal or the same plot of land.

### Step by step description of the method

**Step 1.** Tabulate the data in pairs and calculate the difference between the paired observations.

**Step 2.** Rank these differences without regard to sign. Differences of zero are discarded.

**Step 3.** After ranking, restore the sign (plus or minus) to the ranks.

**Step 4.** Calculate the sums of the positive and negative ranks respectively (termed  $W^+$  and  $W^-$ ).

**Step 5.** Obtain critical values for  $W^+$  and  $W^-$  from tables.

For a parametric test see [Comparing the difference between paired samples](#).<sup>[123]</sup>

See also:

[Comparing the difference between paired samples](#).<sup>[123]</sup>

[Comparing the means of samples with the same numbers of observations equal variance](#).<sup>[124]</sup><sup>[126]</sup>

[Comparing the means of samples with different numbers of observations equal variance](#).<sup>[125]</sup>

[Comparing the means of samples with different numbers of observations unequal variance](#).<sup>[126]</sup>

## 6.9 One- and two-tailed t-test

One- and two-tailed tests are terms used in statistics when determining whether the observed difference would be expected by chance.

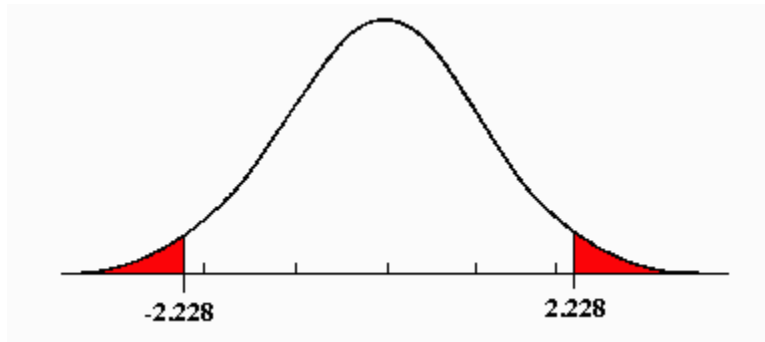
Generally, you should use a two-tailed test when you are simply interested in the existence of a relationship and do not care about the direction. For example, when comparing the mean of two samples use a two-tailed test if you wish to determine if sample 1 is different from sample 2. As another example, use a two-tailed test for a significant correlation if you are equally interested in a significant negative and positive correlation.

Use a one-tailed test when you are only interested in one direction. For example, when comparing the mean of two samples use a one-tailed test if you wish to determine if sample 1 is greater than sample 2. As another example, use a one-tailed test for a significant correlation if you are equally interested in a significant positive correlation. You should take care not to decide to use a one-tailed test after you have undertaken the experiment and have seen how the data looks. So for example, using a one-tailed test when you notice that the mean of sample one is higher than sample two would not be correct if you have only decided to do this test after you saw the result.

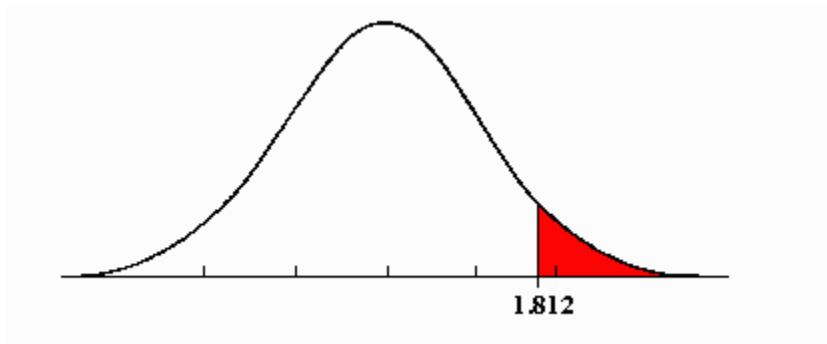
To show the difference between one- and two-tailed t-tests, consider the situation where we wish to test for a significant difference at the 5% or  $p = 0.05$  probability level. The probability of different values of  $t$  occurring by chance can be calculated to produce a curve as shown below. Using the area under these curves we can find the critical  $t$  value for which there is only a 5% probability that a greater value could occur by chance.

For the two-tailed test this 5 % is divided equally between the probability that the mean of sample 1 is greater than the mean of sample 2 and visa versa. Therefore we find the  $t$  values at the 0.025 probability level at each end of the distribution. For example, the critical value of  $t$  when there are

10 degrees of freedom ( $df = 10$ ) and  $p = .05$ , is  $t_{crit} = \pm 2.228$ .



For a one-tailed test we are only interested in one tail of the distribution. As an example we will consider a one-tailed test in the positive direction (mean of sample 1 > mean of sample 2). Therefore using the same example as above we look to find the value at which only one tail of the distribution holds 5% of the area under the curve probability curve. For example the critical value of  $t$  when there are 10 degrees of freedom ( $df = 10$ ) and  $p = .05$ , is  $t_{crit} = +1.812$ .



In the negative direction the critical  $t$  value would be  $-1.812$ .

The value  $t_{crit}$  would be negative. For example, when  $p = .05$  with ten degrees of freedom ( $df=10$ ),  $t_{crit}$  would be equal to  $-1.812$ .

**Part**

---

**VII**

## 7 Regression and correlation

QED Statistics offers three measures of correlation, two linear regression methods, and a General Linear Model (GLM) procedure. Calculate correlation when you want a measure of the association between two variables. Use linear regression when you wish to fit a straight line relationship between one independent and one or more dependent variables. For the fitting of complex models involving both categorical and continuous variables, use the General Linear Model procedure.

[Pearson Correlation](#)<sup>[133]</sup>

[Kendall Correlation](#)<sup>[133]</sup>

[Spearman Rank Correlation](#)<sup>[134]</sup>

[Simple Linear Regression](#)<sup>[135]</sup>

[Multiple Linear Regression](#)<sup>[137]</sup>

[General Linear Model](#)<sup>[154]</sup>

### 7.1 Correlation coefficients

QED Statistics calculates both Pearson and Kendall correlation coefficients. Correlation coefficients measure the amount of association between two variables and range in magnitude between -1 and +1. A value close to +1 indicates that the two variables are highly positively correlated so that as one variable increases, the second variable also increases. Conversely a value close to -1 indicates that the two variables are highly negatively correlated, as one variable increases the other tends to decrease. To calculate a correlation coefficient you will need data giving a series of pairs of observations.

For example you might have data on length and weight of fish as follows:

Length (cm)	Weight (g)
10	28
33	78
59	114
5	2
12	19

Several authors have offered guidelines for the interpretation of a correlation coefficient. Cohen (1988), for example, has suggested the following interpretations for correlations in psychological research:

Correlation	Negative	Positive
Small	-0.29 to -0.10	0.10 to 0.29
Medium	-0.49 to -0.30	0.30 to 0.49
Large	-1.00 to -0.50	0.50 to 1.00

As Cohen himself has observed, however, all such criteria are in some ways arbitrary and should not be observed too strictly. This is because the interpretation of a correlation coefficient depends on the context and purposes. A correlation of 0.9 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors.

Pearson's correlation coefficient is a parametric statistic, and it may be less useful if the underlying assumption of normality is violated. Non-parametric correlation methods, such as Kendall's  $\tau$  may be useful when distributions are not normal; they are a little less powerful than parametric methods if the assumptions underlying the latter are met, but are less likely to give distorted results when the assumptions fail.

[Calculating a Pearson correlation coefficient](#)<sup>[133]</sup>  
[Calculating a Kendall correlation coefficient](#)<sup>[133]</sup>

### 7.1.1 Pearson Correlation

The Pearson Product Moment Correlation Coefficient is the most widely used measure of correlation or association. It is named after Karl Pearson who developed the correlational method for agricultural research. The product moment part of the name comes from the way in which it is calculated, by summing up the products of the deviations of the scores from the mean.

The correlation coefficient  $r$  is defined by the equation:

$$r = \frac{\sum (X - \mu_X)(Y - \mu_Y)}{N\sigma_X\sigma_Y}$$

where  $X$  and  $Y$  are the two variables whose correlation is being calculated,  $\mu$  subscript  $X$  or  $Y$  is their respective [mean](#)<sup>[100]</sup> and  $\sigma$  subscript  $X$  or  $Y$  is their respective [standard deviation](#)<sup>[101]</sup>.  $N$  is the number of observations.

The numerator of this formula says that we sum up the products of the deviations of variable  $X$  from the mean of the  $X$ s and the deviation of the variable  $Y$  from the mean of the  $Y$ s. This summation of the product of the deviation scores is divided by the number of pairs of observations times the standard deviation of the  $X$  variable times the standard deviation of the  $Y$  variable.

If  $r$  is  $-1$ , there is a perfect negative correlation.

- " falls between  $-1$  and  $-0.5$ , there is a strong negative correlation.
- " falls between  $-0.5$  and  $0$ , there is a weak negative correlation.
- " is  $0$ , there is no correlation.
- " falls between  $0$  and  $0.5$ , there is a weak positive correlation.
- " falls between  $0.5$  and  $1$ , there is a strong positive correlation.
- " is  $1$ , there is a perfect positive correlation

### 7.1.2 Kendall's Correlation

Kendall's Rank Correlation is a non-parametric correlation measure.

To calculate Kendall's rank correlation,  $\tau$  the following steps are undertaken:

1. Express the pairs of data points  $(x,y)$  by their rank value.

For example the pairs  $(1.0, -1.5)$ ,  $(3.5, 5.0)$ ,  $(-1.0, 0)$ ,  $(2, -4)$  becomes  
 $(2, 2)$ ,  $(4, 4)$ ,  $(1, 3)$ ,  $(3, 1)$

2. Now sort the ranked pairs in terms of the rank of  $y$ . So that our example becomes:

1.  $(3, 1)$ ,
2.  $(2, 2)$ ,
3.  $(1, 3)$ ,
4.  $(4, 4)$

3. Now count the number of pairs of  $X$ s where the ranks are out of order,  $Q$ .

In our example the answer is  $Q = 3$ .  $X$  is out of order when comparing  $3$  &  $2$ ,  $2$  &  $1$  and  $3$  &  $1$

4. Finally calculate Kendall's rank correlation using:

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

If  $\tau$  is -1, there is a perfect negative correlation.

- " falls between -1 and -0.5, there is a strong negative correlation.
- " falls between -0.5 and 0, there is a weak negative correlation.
- " is 0, there is no correlation.
- " falls between 0 and 0.5, there is a weak positive correlation.
- " falls between 0.5 and 1, there is a strong positive correlation.
- " is 1, there is a perfect positive correlation

### 7.1.3 Spearman Rank Correlation

Spearman's Rank Correlation measures the direction and strength of the relationship between two variables. It provides a distribution free test of independence between two variables, but is insensitive to some types of dependence. Both corrected and uncorrected Spearman's Rank Correlations are calculated.

The correlation coefficient,  $\rho$ , can range between -1 and +1.

To calculate Spearman's Rank Correlation the following steps are undertaken

1. Rank both sets of data from the highest to the lowest. Make sure to check for tied ranks. The average rank is used for ties.
2. Subtract the two sets of ranks to get the difference  $d$ .
3. Square the values of  $d$ .
4. Add the squared values of  $d$  to get  $\phi$
5. Calculate  $\rho$  using the formula:

$$\rho = 1 - \left( \frac{6\phi}{n^3 - n} \right)$$

If  $\rho$  is -1, there is a perfect negative correlation.

- " falls between -1 and -0.5, there is a strong negative correlation.
- " falls between -0.5 and 0, there is a weak negative correlation.
- " is 0, there is no correlation.
- " falls between 0 and 0.5, there is a weak positive correlation.
- " falls between 0.5 and 1, there is a strong positive correlation.
- " is 1, there is a perfect positive correlation.

Spearman's rank correlation provides a distribution free test of independence between two variables. It is, however, insensitive to some types of dependence. [Kendall's rank correlation](#)<sup>[133]</sup> may give a better measure of correlation and is also a better two-sided test for independence.

The corrected Spearman's rank correlation coefficient ( $r$ ) is calculated as:

$$\rho = \frac{\sum_{i=1}^n R(x_i) R(y_i) - n\left(\frac{n+1}{2}\right)^2}{\left( R(x_i)^2 - n\left(\frac{n+1}{2}\right)^2 \right)^{0.5} \left( R(y_i)^2 - n\left(\frac{n+1}{2}\right)^2 \right)^{0.5}}$$

- where  $R(x)$  and  $R(y)$  are the ranks of a pair of variables ( $x$  and  $y$ ) each containing  $n$  observations.

## 7.2 Linear Regression

Linear regression is a method used to fit a straight line, for predicting a value for a dependent variable, given a value for an independent variable.

The regression equation is given by:

$$Y = a + bX$$

where X is the independent variable, Y is the dependent variable, a is the intercept and b is the slope of the line.

Regression analysis is most often used for prediction. The goal is to create a mathematical model that can be used to predict the values of a dependent variable, Y, based upon the values of an independent variable, X.

Regression analysis assumes that for a fixed value of X (the independent variable), the population of Y (the dependent variable) is normally distributed with equal variances across the Xs. Note that it is assumed that there is no error in the measurement of the independent variable. See [Is Linear Regression appropriate](#) <sup>[136]</sup>?

An example of the use of linear regression is the description of linear growth. In the data set below, the height of a seedling was measured daily for 10 days.

Day (X)	Height (cms) (Y)
1	0.1
2	0.2
3	0.5
4	0.8
5	1.3
6	1.5
7	1.9
8	2.2
9	2.3
10	2.7

While the day can be known accurately and cannot be changed by the observer, the height of the seedling cannot be measured with complete accuracy. The time axis is the independent variable and the height is the dependent variable. In other words, the size of the seedling is dependent on the time day since germination.

A regression line is fitted by the method of least squares. The steps in the calculation are as follows:

1. Arrange the data into X, Y pairs
2. Compute the [mean](#) <sup>[100]</sup> of all of the X (independent) values.
3. Compute the sum of the  $X^2$  by squaring each X value and adding up the squares.
4. Compute the sum of the  $Y^2$  in the same manner.
5. Compute the sum of each X value multiplied by its corresponding Y value.
6. Calculate the slope (b) of the line using

$$b = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}$$

7. Calculate the Y-intercept (a) where  $X = 0$  using

$$a = \bar{Y} - b\bar{X}$$

where  $\bar{Y}$  and  $\bar{X}$  are the means of the dependent and independent variables respectively.

Before using linear regression consider [if linear regression is appropriate](#) <sup>136</sup>?

### 7.2.1 Is Linear Regression appropriate?

To check that linear regression is an appropriate analysis for your data, ask yourself the following questions.

Can the relationship between X and Y be graphed as a straight line?	Examine the scatter plot. If the the relationship between X and Y is curved, linear regression will be inappropriate. You may be able to make the relationship linear by transformation. If this is impossible linear regression should not be used.
Is the scatter of data around the line approximately normal? Look for extreme outliers and check for unimodality.	Linear regression analysis assumes that the errors are normally distributed.
Is the variability the same everywhere?	Linear regression assumes that scatter of points around the best-fit line has the same standard deviation all along the curve. The assumption is violated if the points with high or low X values tend to be further from the best-fit line. The assumption that the standard deviation is the same everywhere is termed homoscedasticity.
Do you know the X values precisely?	The linear regression model assumes that X values are exactly correct, and that experimental error or biological variability only affects the Y values. This is rarely the case, but it is sufficient to assume that any imprecision in measuring X is very small compared to the variability in Y.
Are the data points independent?	Whether one point is above or below the line is a matter of chance, and does not influence whether another point is above or below the line.
Are the X and Y values intertwined?	If the value of X is used to calculate Y (or the value of Y is used to calculate X) then linear regression calculations are invalid. This would be the case, for example, if you fitted a linear regression to a plot of hydrogen ion concentration and pH, since one is calculated from the other.

## 7.3 Multiple Linear Regression

Simple [linear regression](#) fits the equation :

$$Y = a + bX$$

where  $X$  is the independent variable,  $Y$  is the dependent variable,  $a$  is the intercept and  $b$  is the slope of the line.

Multiple linear regression is an extension of this procedure for situations where you have multiple predictor variables. The linear equation then takes the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

where  $k$  is the number of predictors. The regression coefficients (or  $b_1 \dots b_k$  coefficients) represent the independent contributions of each independent variable to the prediction of the dependent variable.

The computations involved in solving a multiple regression problem are conveniently expressed using matrix notation. Assume there are  $n$  observed values of  $Y$  and  $n$  associated observed values for each of  $k$  different  $X$  variables. Then  $Y_i$ ,  $X_{ik}$ , and  $e_i$  can represent the  $i$ th observation of the  $Y$  variable, the  $i$ th observation of each of the  $X$  variables, and the  $i$ th unknown residual value, respectively. Collecting these terms into matrices we have

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 & \dots & \dots & x_{1k} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 1 & x_n & \dots & \dots & x_{nk} \end{bmatrix}, e = \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ \vdots \\ e_n \end{bmatrix}$$

The multiple regression model in matrix notation then can be expressed as

$$Y = Xb + e$$

where  $b$  is a column vector of 1 (for the intercept) +  $k$  unknown regression coefficients.

The regression coefficients that minimize the sum of the squared residuals are found by solving the set of normal equations:

$$X'Xb = X'Y$$

When the  $X$  variables are linearly independent there is a unique solution to the normal equations. Premultiplying both sides of the matrix formula for the normal equations by the inverse of  $X'X$  gives

$$(X'X)^{-1}X'Xb = (X'X)^{-1}X'Y$$

or

$$b = (X'X)^{-1}X'Y$$

An important assumption of multiple linear regression that frequently leads to problems is that the independent variables are linearly independent. This is often not the case and can result in failure

to find a solution because the inverse of  $X'X$  does not exist.

Variables can be added in a [stepwise fashion](#) <sup>[138]</sup>.

If the independent variables are correlated you can use a [general linear model](#) <sup>[154]</sup>.

### 7.3.1 Stepwise Linear Regression

Stepwise regression is a method to identify the set of independent variables that are the best predictors for the dependent variable. Independent variables are added or removed one at a time and the improvement of the fit to the observed data assessed to determine if the fit of the model has been improved. This addition of variables can be done in a forwards direction, in which variables are sequentially added, or a backwards direction, in which all the independent variables are initially used and then removed in a sequential fashion.

[Forward Stepwise Regression](#) <sup>[138]</sup>  
[Backward Stepwise Regression](#) <sup>[138]</sup>

See [Multiple Linear Regression](#) <sup>[137]</sup> for information about the method.

#### 7.3.1.1 Forward Stepwise Linear Regression

Forward stepwise regression starts with no variables in the model. It then adds the most significant explanatory variable, that is, the one with the lowest p-value, at each step, until all variables have been added. By scrutinising the overall fit of the model, variables will be automatically added (and, if they do not improve the fit, removed again) until the optimum model is found.

The results report shows the sequences of the procedure as steps:

Step 1 - Shows the effect of introducing the most explanatory variable into the model, and a list of candidate variables that will be entered next, providing they fit the selection criteria.

Step 2 - Shows the effect of introducing the next most explanatory variable into the model, and a list of candidate variables that will be entered next, providing they fit the selection criteria.

...and so on.

A common problem in Multiple Linear Regression is [multicollinearity](#) <sup>[139]</sup> between explanatory variables; one or more redundant variables from a set of directly-related variables should be removed from the data set to avoid problems in regression analysis.

#### 7.3.1.2 Backward Stepwise Linear Regression

Backward stepwise regression starts with all explanatory variables included the model. It then removes the least significant explanatory variable, that is, the one with the highest p-value, at each step, until all variables have been added. By scrutinising the overall fit of the model, variables will be automatically removed until the optimum model is found.

The results report shows the sequences of the procedure as steps:

Step 1 - Shows the effect of including all the explanatory variables into the model, with the individual p-values.

Step 2 - Shows the effect of removing the least explanatory variable from the model.

...and so on.

A common problem in Multiple Linear Regression is [multicollinearity](#)<sup>[139]</sup> between explanatory variables; one or more redundant variables from a set of directly-related variables should be removed from the data set to avoid problems in regression analysis.

### 7.3.1.3 Multicollinearity

Multicollinearity occurs when you have one or more variables which are directly related, where one variable can be derived from the others. A good example of this would be the composition of a river bed, expressed as a percentage of total sample mass. If you have 4 variables representing particle size, pebble, sand, silt and clay, then you only need an observation for 3 of the variables, since the fourth can be calculated from the other three:

$$\% \text{clay} = 100 - (\% \text{pebble} + \% \text{sand} + \% \text{silt})$$

In order to avoid problems in the regression analysis, one of a set of directly related variables should be removed from the data set.

**Part**



## 8 Analysis of Variance (ANOVA)

QED Statistics offers two main pathways for undertaking an Analysis of Variance. You can use a [General Linear Model](#)<sup>[154]</sup> to undertake even simple ANOVAs, and it is the only procedure offered for complex multi-factorial ANOVAs. Alternatively, a conventionally organised one- and two-way ANOVA can be selected.

These procedures are offered for those unfamiliar with the General Linear Model methodology and jargon, and also as a teaching aid to show the steps in the basic method.

[One-way ANOVA](#)<sup>[141]</sup>

[One-way ANOVA repeated measures](#)<sup>[143]</sup>

[Two-way ANOVA](#)<sup>[145]</sup>

[General Linear Model](#)<sup>[154]</sup>

If your data is far from normally distributed then a [Kruskal-Wallis](#)<sup>[147]</sup> test can be used instead of a one-way ANOVA.

For examples of how to undertake an ANOVA see:

[An example one-way ANOVA](#)<sup>[148]</sup>

[An example two-way ANOVA](#)<sup>[150]</sup>

### 8.1 One-way ANOVA

A one-way ANalysis of Variance (ANOVA) is used to compare the means of three or more samples or treatments. While you can compare the means between two treatments using a [t-Test](#)<sup>[126]</sup>, with more than two samples it is both inefficient and misleading to undertake all the individual pair-wise comparisons.

As a simple example, consider an experiment to determine if the mean number of ground beetles differed between 3 localities. At each locality 6 pitfall traps were set up for 1 night, and the number of beetles caught in each trap recorded. The results are shown below.

Obs	Site 1	Site 2	Site 3
Pitfall 1	5	5	2
Pitfall 2	4	3	6
Pitfall 3	1	1	1
Pitfall 4	2	6	1
Pitfall 5	3	3	0
Pitfall 6	0	6	5
mean	2.5	4	2.5
Variance	3.5	4	5.9

In this example we assume that all the pitfall traps were set up the same so we have 6 replicate observations at each site.

If measurements for the different treatments were taken a number of times on the same subjects this is a [repeated measurements ANOVA](#)<sup>[143]</sup>.

The mean number of beetles per trap was the same for sites 1 and 3, and highest at site 2. However, note that the variance at each site is different, and highest at site 3. An important assumption underlying Analysis of Variance is that all treatments have similar variance. If there are strong reasons to doubt this then the data might need to be transformed before the test can be done. See [Homogeneity of variances test](#)<sup>[143]</sup> to check if variances are similar. If your data cannot meet the assumptions required for an ANOVA you can use a [Kruskal-Wallis test](#)<sup>[147]</sup>.

### Step by step description of the method

**Step 1** With the data for each treatment (site in our case) arranged in columns calculate for each treatment  $\Sigma x$ ,  $n$ ,  $\Sigma x^2$  and  $(\Sigma x)^2 / n$ .

**Step 2** For each column calculate

$$\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}$$

**Step 3** Calculate the sum of squares for all the observations irrespective of treatment . Call this sum A.

**Step 4** Calculate for all the observations

$$\frac{(\sum_{i=1}^n X_i)^2}{n}$$

and call this sum B.

**Step 5** Calculate the sum of all the observations (the grand total). Square the grand total and divide it by total number of observations. Call this C.

**Step 6** Calculate the total sums of squares, A-C.

**Step 7** Calculate the between treatments sums of squares, B-C.

**Step 8** Calculate the residual sums of squares A-B

**Step 9** Construct an ANOVA results table as follows:

Source of variance	Sum of squares (S of S)	Degrees of freedom (df)	Mean square = S of S / df
Between treatments	B-C	u - 1	B-C/u - 1
Residual	A-B	u(v-1)	A-B/u(v-1)
Total	A-C	(uv)-1	

where  $u$  = number of treatments and  $v$  = number of replicates.

**Step 10** Using the mean squares in the final column of this table, do a variance ratio test to obtain an F value:

F = Between treatments mean square / Residual mean square  
and look up the significance of this F value in Tables.

If the value is significant it tells you that there is a significant difference between the means of the different treatments. However it does not indicate which treatments differ significantly from each other. To identify these differences use a [Multiple Range Test](#) <sup>[144]</sup>.

Ideally, for a one-way ANOVA you should would have the same number of replicates for each site - as we have in our example. However, this is not essential as there are methods for dealing with

missing values.

ANOVAs can be undertaken with [fixed or random effects](#)<sup>[147]</sup>.

### 8.1.1 One-way repeated measurements ANOVA

A one-way repeated measures ANOVA is used to compare the difference in the means from a number of treatments when the same subjects have been tested under each treatment.

For example, an experimenter might wish to measure the ability of a group of people under different levels of background noise. For the experiment 10 people are tested under 3 background noise levels, 10, 50 decibels and 100 decibels. There are therefore 3 measurements on each person.

You cannot use a standard one-way ANOVA in this case because it fails to model the correlation between the repeated measures on the same person.

If you have only taken two measurements on each subject then you can use [a paired t-test](#)<sup>[123]</sup>.

The data for a repeated measures ANOVA is laid out with the measurements as the columns and the subjects, which are repeatedly measured as the rows. For example, in an experiment on the changing attitudes, 7 peoples were tested 4 times after various levels of training. giving the following data set of test scores:

	test 1	test 2	test 3	test 4
subject 1	14	17	14	8
subject 2	12	15	11	6
subject 3	10	12	10	5
subject 4	10	9	10	4
subject 5	9	9	8	2
subject 6	6	7	7	2
subject 7	5	7	7	2

## 8.2 Homogeneity of variances test

To test that each treatment has a similar variance calculate the variance for each treatment. For example, using data on pitfall traps from 3 localities we have the following variances for sites 1, 2 and 3:

Obs	Site 1	Site 2	Site 3
Pitfall 1	5	5	2
Pitfall 2	4	3	6
Pitfall 3	1	1	1
Pitfall 4	2	6	1
Pitfall 5	3	3	0

<b>Pitfall 6</b>	0	6	5
<b>Variance</b>	<b>3.5</b>	<b>4</b>	<b>5.9</b>

Divide the highest variance value by the lowest to obtain a variance ratio (F). In our example above this is  $5.9/3.5 = 1.686$ . Then look up this value in a table of  $F_{\max}$  values for the number of treatments (3 sites in our example) and the degrees of freedom (number of replicates per treatment - 1 which is  $6-1=5$  in our example).

If our variance ratio *does not exceed* the tabulated  $F_{\max}$  value, the variances are sufficiently homogeneous for an ANOVA.

If not, you should consider [transforming your data](#)<sup>[91]</sup>.

## 8.3 Multiple comparison tests

QED offers a number of multiple comparison tests for use with a [one-way ANOVA](#)<sup>[141]</sup> to compare the differences between the treatments.

[Tukey](#)<sup>[144]</sup>

[Scheffe](#)<sup>[145]</sup>

[Newman-Keuls](#)<sup>[145]</sup>

[Tukey-Kramer](#)<sup>[145]</sup>

[Bonferroni](#)<sup>[145]</sup>

### 8.3.1 Tukey

The Tukey multiple comparisons test is also known as the "honestly significant difference test". If the single factor analysis of variance rejects the null hypothesis that all the means are identical a multiple comparison test is needed to tell which means are significantly different.

1. Order the sample means from largest to smallest.
2. Calculate and tabulate all pair-wise differences.
3. The q value is calculated by dividing the differences by the standard error

$$SE = \sqrt{\frac{s^2}{n}}$$

where  $s^2$  is the error mean square from the analysis of variance and n is the number of observations in the two groups.

4. If this critical value is greater than a critical q value the means are significantly different.

### 8.3.2 Scheffe

The Scheffe test computes a new critical value for an F Test conducted when comparing two groups from the ANOVA. The formula simply modifies the F-critical value by taking into account the number of groups being compared.

### 8.3.3 Newman-Keuls test

The Newman-Keuls test or Student-Newman-Keuls or SNK test is performed like the [Tukey Test](#) [144] with one exception. This is that the critical value is calculated differently depending on the number of means that are compared.

A multiple comparisons test with different critical values depending on the range is termed a multiple range test.

### 8.3.4 Tukey-Kramer

The Tukey-Kramer test is an extension of the Tukey test to unbalanced designs. Unlike Tukey test for balanced designs, it is not exact.

### 8.3.5 Bonferroni

In order to ensure that the probability is no greater than say 0.05 that a difference will appear to be statistically significant when there are no underlying difference, each of the 'm' individual comparisons is performed at the  $(0.05/m)$  level of significance. This is termed the Bonferroni adjustment. The major disadvantage to the Bonferroni adjustment is that it is not an exact procedure, and the adjusted P value is larger than the true P value. In other words this test is conservative.

## 8.4 Two-way ANOVA

Analysis of Variance (ANOVA) is frequently used to test for differences in experiments in which 2 or more factors are considered simultaneously. If there are more than 2 factors in the analysis see [General Linear Models](#) [154]. With two factors, this type of analysis is called a two-way ANOVA and is applied to data from experiments where combinations of treatments of two factors (e.g. pH and temperature) have been applied in all possible combinations. These are called factorial designs, and we can analyse them even if we do not have replicates for each combination. However, you will need replicate observations if you wish to investigate the interactions between the two factors.

As an example consider a situation in which the ability to catch fish of 3 different designs of fish trap were tested for 1 day and 1 night. First the counts of the number of fish caught are arranged in a table as follows:

	Trap 1	Trap 2	Trap 3
Day	25	14	38
Night	23	105	40

Note that in this first simple example we have no replicates - there is a single count for each combination of trap and Day/Night.

Now the following are calculated.

1.  $\sum x$ ,  $\sum x^2$ , and  $(\sum x)^2 / n$ , for each column in the table.

2.  $\sum x$ ,  $\sum x^2$ , and  $(\sum x)^2 / n$ , for each row.
3. Find the grand total by adding all  $\sum x$  for columns (it should be the same for rows). Square this grand total and then divide by  $uv$ , where  $u$  is the number of data entries in each row, and  $v$  is number of data entries in each column. **Call this value D.**
4. Find the sum of  $\sum x^2$  values for columns; **call this A.** It will be the same for  $\sum x^2$  of rows.
5. Find the sum of  $\sum x^2/n$  values for columns; **call this B.**
6. Find the sum of  $\sum x^2/n$  values for rows; **call this C.**
7. Set out a table of analysis of variance as follows:

Source of variance	Sum of squares	Degrees of freedom*	Mean square
Between columns (Traps)	<b>B - D</b>	$u - 1$ (=2)	<b>B-D</b> / $u - 1$
Between rows (Day/Night)	<b>C - D</b>	$v - 1$ (= 1)	<b>C-D</b> / $v - 1$
Residual	<b>(A-D)-(B-D)</b>	$(u-1)(v-1)$ (=2)	<b>(A-D)-(B-D)</b> / $(u-1)(v-1)$
Total	<b>A - D</b>	$(uv)-1$ (=5)	<b>A - D</b> / $(uv)-1$

\* Where  $u$  is the number of data entries in each row, and  $v$  is the number of data entries in each column; note that the total df is always one fewer than the total number of entries in the table of data.

Now do a variance ratio test to obtain F values:

(1) **For between columns** (Type of trap):  $F = \text{Between columns mean square} / \text{Residual mean square}$

(2) **For between rows** (Day/Night):  $F = \text{Between rows mean square} / \text{Residual mean square}$

In each case, consult a **table of F** ( $p = 0.05$  or  $p = 0.01$  or  $p = 0.001$ ) where  $u$  is the between-treatments df (columns or rows, as appropriate) and  $v$  is residual df. If the calculated F value exceeds the tabulated value then the treatment effect (trap or Day/Night) is significant.

By comparing the size of residual (error) mean square (MS) with that of the columns (traps) or rows (Day/Night) the this analysis you can deduce if there is a strong interaction between type of trap and Day/ Night). If the residual mean square is low compared with the Traps or Day/night values it indicates that most variation in the data is accounted for by the separate effects of traps and Day/Night.

If you wish to analyse for interactions then the experiment would need to be replicated so that we have more than 1 observation in each Trap - Day/Night combination.

For example, each combination was actually repeated over 3 nights so the full data set was as follows:

	Trap 1	Trap 2	Trap 3
Day	25, 7, 30	14, 8, 7	38, 56, 62
Night	23, 14, 22	105, 68, 77	40, 50, 53

Now when the Two-way ANOVA is undertaken we produce an ANOVA table which includes an interaction mean square value.

ANOVAs can be undertaken with [fixed or random effects](#) <sup>[147]</sup>.

## 8.5 Fixed and random effects

In statistical terminology a fixed variable is one that is assumed to be measured without error. In contrast, a random variable is assumed to be drawn from a larger population with its values representing a random sample of the possible values. Therefore, we expect to generalise the results obtained with a random variable to all other possible values of that random variable. Most of the time in ANOVAs and regression analysis we assume the independent variables are fixed.

The terms Random and Fixed Effects are used in ANOVA and regression models, and refer to the statistical model. Almost always, researchers use fixed effects regression or ANOVAs. A fixed effects ANOVA refers to assumptions about the independent variable and the error distribution for the variable. An example is the easiest example for illustrating the idea.

Consider a study in which mice are fed 0 mg, 1 mg, or 2 mg of an experimental drug each day and after 10 days the mice are weighed. To determine the effects on weight of this drug a fixed effects ANOVA would be appropriate. This is because we are interested in studying the effects of these particular doses of the drug. However, if you wanted to make inferences about the effects of other doses of the drug, say 1.5 mg, a random effects model should be used.

Random effects models are sometimes referred to as Model II or variance component models.

Analyses using both fixed and random effects are called mixed models.

## 8.6 Kruskal-Wallis test

The Kruskal-Wallis one-way analysis of variance by ranks (named after William Kruskal and Allen Wallis) is a non-parametric method for testing equality of population medians among groups. Intuitively, it is identical to a [one-way ANOVA](#) <sup>[147]</sup> with the data replaced by their ranks. As it is a non-parametric method, the Kruskal-Wallis test does not assume a normal population, unlike the analogous [one-way analysis of variance](#) <sup>[147]</sup>. However, it still assumes that population variabilities among groups are equal. If your data meets the assumptions required for a standard [ANOVA](#) <sup>[147]</sup>, you should not use a Kruskal-Wallis test.

Given  $k$  independent samples of sizes  $n_1$  to  $n_k$  the test is carried out as follows.

1. Rank all data from all groups together; i.e., rank the data from 1 to  $N$  ignoring group membership. Assign any tied values the average of the ranks they would have received had they not been tied.

2. Find

$$\bar{R}_i,$$

the average of the ranks of the observations in the  $i$ th sample.

3. The test statistic is then

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left( \bar{R}_i - \frac{N+1}{2} \right)^2$$

and reject the null hypothesis that all  $K$  distributions are the same if

$$H > \chi_{k-1}^2$$

## 8.7 Omega squared

Omega squared is one of the most frequently applied methods for estimating the proportion of the dependent variable variability accounted for by an independent variable.

Omega squared is an estimate of the dependent variance accounted for by the independent variable in the population for a [fixed effects model](#)<sup>[155]</sup>. The between-subjects, fixed effects, form of the  $\omega^2$  formula is:

$$\omega^2 = (SS_{\text{effect}} - (df_{\text{effect}})(MS_{\text{error}})) / MS_{\text{error}} + SS_{\text{total}}.$$

(Note: Do not use this formula for repeated measures designs)

Omega squared provides a relative measure of the strength of an independent variable ranging from 0.0 to 1.0, however, in many areas of research it is unlikely that high omega-squared values will be obtained because of the large relative magnitude of the error variance.

In some fields a value of 0.15 or greater is considered large, a medium effect is .06 to 0.15 and a small effect is 0.01.

Omega-squared can give useful insight when an F Test is not significant, because it is unaffected by sample size, whereas F ratios are affected by small sample sizes.

## 8.8 An example one-way ANOVA

This example uses a data set from Sokal and Rohlf (1969). The width of the scutum of tick larvae collected from 4 cottontail rabbits were measured. The objective of the analysis is to determine if the size of the ticks differed significantly between the rabbits.

The [demonstration data set](#)<sup>[21]</sup> is supplied as **1 way ANOVA rabbit ticks SFp208.csv**, and is also shown at the bottom of this page.

Open the data set using **File|Open**

Select a one-way ANOVA using **ANOVAs|1 way ANOVA**

The following dialog will open, which shows that the 4 different rabbits are the 4 treatments. Click **OK** to run the analysis

**One way ANOVA**

4

Column Effect Type  
☒ Fixed ☐ Random

Rows Effect Type  
☒ Fixed ☐ Random

Input the significance level for the multiple comparison test here.  
 0.050

Multiple Comparisons  
☒ None ☐ Scheffe ☐ Tukey-Kramer  
☐ Tukey ☐ Newman-Keuls ☐ Bonferroni

Column: 1 Column: 2 Column: 3 Column: 4  
 Rabbit1 Rabbit2 Rabbit3 Rabbit4

Available data

	Rabbit1	Rabbit2	Rabbit3	Rabbit4
	380	350	354	376
	376	356	360	344
	360	358	362	342
	368	376	352	372
	372	338	366	374
	366	342	372	360
	374	366	362	
	382	350	344	
		344	342	

Ok

The results of a one-way ANOVA are presented in a single grid.

**DF** is the degrees of freedom.

**SS** is the Sums of Squares.

**MS** is the Mean Squares

**F** is the test statistic

**Prob.** is the probability that the difference in the means of the treatments could have arisen by chance.

**Omega<sup>2</sup>** is Omega squared, a measure of the amount of variability explained by the treatments.

**Between** gives results between treatments.

**Within** gives results within treatments

**Total** is the total variance etc.

Results: One way ANOVA						
One Way ANOVA						
	DF	SS	MS	F	Prob. >F	Omega <sup>2</sup>
Between	3	1807.73	602.58	5.26	0.00	0.26
Within	33	3778.00	114.48			
Total	36	5585.73	359.70			
This means	There is a significant difference in location between the treatments (F = 5.26, DF1 = 3, DF2 = 33, P = <0.05)					

In this example it is clear that the mean size of the ticks is significantly different between treatments.

The data used for the above example is tabulated below.

Rabbit1	Rabbit2	Rabbit3	Rabbit4
380	350	354	376
376	356	360	344
360	358	362	342
368	376	352	372
372	338	366	374
366	342	372	360
374	366	362	
382	350	344	
	344	342	
	364	358	
		351	
		348	
		348	

## 8.9 An example two-way ANOVA

This example uses a data set from Sokal and Rohlf (1969). The data set measures the difference in food consumption when rancid lard was substituted for fresh lard in the diet of rats. The data is the food eaten over 73 days by 12 rats classified in two ways - fresh vs rancid lard and male vs female.

The [demonstration data set](#)<sup>[21]</sup> is supplied as **2 way ANOVA with replication SFp302.csv**, and is also shown at the bottom of this page.

Open the data set using **File|Open**

Select a one-way ANOVA using **ANOVAs|2 way ANOVA**

The following dialog will open which shows a 2 x 2 grid.

The data for each of the treatment cells is assumed to be arranged in columns in the [working data grid](#)<sup>[89]</sup>.

At the top of the window is a box to select the number of levels for treatment 1 (in the example below it is 2 fresh or rancid).

At the left is a box to select the number of levels for treatment 2 (in the example below it is 2 male or female).

Radio buttons allow the choice between [Fixed and Random effects](#)<sup>[155]</sup>. The default is a fixed effects model.

The upper grid is used to select which columns hold the data for the various treatments and levels.

In our example, there were 3 observations in a 2 x 2 table.

To select a different variable from that initially present, click on the drop-down menu and select from the variable list.

Row: 2	Fresh Female
	Fresh Male
	Rancid Male
	Fresh Female
	Rancid Female

To help you in the selection of the variables the working data is shown below in the **Available data** table.

When the variables have all been selected click **OK** to run the analysis and see your [results](#).

**Two way ANOVA**

Column Effect Type: ☒ Fixed ☐ Random

Rows Effect Type: ☒ Fixed ☐ Random

	Column: 1	Column: 2
Row: 1	Fresh Male	Rancid Male
Row: 2	Fresh Female	Rancid Female

Multiple Comparisons: ☒ None ☐ Scheffe ☐ Tukey ☐ Newman-Keuls ☐ Bonferroni

**Available data**

	Fresh Male	Rancid Male	Fresh Female	Rancid Female
Obs 1	709	592	657	508
Obs 2	679	538	594	505
Obs 3	699	476	677	539

Ok

The results of a two-way ANOVA are presented in a single grid.

**DF** is the degrees of freedom.

**SS** is the Sums of Squares.

**MS** is the Mean Squares

**F** is the test statistic

**Prob.** is the probability that the difference in the means of the treatments could have arisen by

chance.

**Omega<sup>2</sup>** is Omega squared, a measure of the amount of variability explained by the treatments.

**Between Columns** gives results between the levels of treatment 1.

**Between Rows** gives results between the levels of treatment 2.

**Within Groups / Error** gives results within treatments.

**Total** is the total SS etc.

Results: Two way ANOVA						
Two Way ANOVA						
	DF	SS	MS	F	Prob. >F	Omega <sup>2</sup>
Between Columns	1	61204.08	61204.08	41.97	0.00	0.76
Between Rows	1	3780.75	3780.75	2.59	0.15	0.03
Interaction	1	918.75	918.75	0.63	0.45	0
Within Groups/Error	8	11666.67	1458.33			
Total	11	77570.25	7051.84			
Omega <sup>2</sup>	0.78					

In this example there is a highly significant difference between the columns (fresh vs rancid fat) proving that the rats are responsive to fresh fat in their diet. There was no significant difference between rows, showing that the sexes did not differ in their response. Finally, there was no significant interaction between sex and fat consumption.

The data used for the above example is tabulated below.

	Fresh Male	Rancid Male	Fresh Female	Rancid Female
Obs 1	709	592	657	508
Obs 2	679	538	594	505
Obs 3	699	476	677	539

# Part

---



IX

## 9 General Linear Model

It was conventional to analyse the effects of categorical variables using an [ANOVA](#)<sup>[141]</sup>, and continuous variables using [regression](#)<sup>[132]</sup>. Both types of analysis are forms of a general linear model (GLM). The general linear model incorporates a range of different statistical models: ANOVA, ANCOVA, MANOVA, MANCOVA, ordinary linear regression, t-Test and F Test. If there is only one dependent variable and a number of independent variables, which are not correlated, then the model is a [multiple linear regression](#)<sup>[137]</sup>.

Within QED Statistics we have limited the available analyses to those possible with a single dependent variable. However, you can explore models with many independent or explanatory variables, and these can be categorical or continuous. The program can therefore be used to analyse multi-factorial analysis of variance models with interaction terms and situations where you have a mixture of categorical and continuous variables.

A general linear model (GLM) takes the form:

$$Y = BX + U$$

where **Y** is a matrix with a series of multivariate measurements,

**X** is a matrix that may hold indicator variables that indicate group membership and independent variables.

**B** is a matrix containing parameters that are usually to be estimated and

**U** is a matrix containing residuals (i.e., errors or noise).

The application of a GLM for different types of data is presented in the following examples:

[A simple ANOVA using a GLM](#)<sup>[155]</sup>

[A simple linear regression using a GLM](#)<sup>[156]</sup>

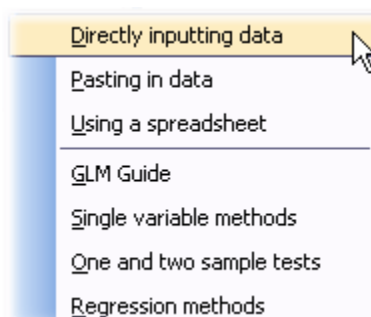
[Using more than 1 explanatory variable in a GLM](#)<sup>[158]</sup>

[Using sequential and adjusted sums of squares](#)<sup>[159]</sup>

[Combining continuous and categorical variables in a GLM](#)<sup>[160]</sup>

[Studying interactions in a GLM](#)<sup>[161]</sup>

Go to **Help|Guides - GLM Guide** to see how to use this method.



## 9.1 Random and fixed effects

Random and fixed factors are terms used in ANOVA, regression and General Linear Models.

Each factor, classification variable or independent variable in an ANOVA or General Linear Model must be classified as either a fixed or random factor. This is necessary in order to find the correct error term for each effect.

**Random factors** are categorical variables used when you wish to generalise to a larger population of factor levels which can have values different from the levels used in the analysis. Define a factor as random when the levels are viewed as a sample from a larger population. A factor is random if :

1. It is desired to make inferences about the wider population with exposure to other levels of the factor.
2. The levels of the factor used in the experiment were selected by a random procedure.
3. If the experiment were replicated, different levels of that factor would be used in the new experiment.

**Fixed factors** are categorical variables used when the selected levels are of interest and the objective is not to make inferences about a larger population with a far wider range of levels. A factor is fixed if

1. The results of the factor generalise only to the levels that were included in the experimental design.
2. Some method is used to select and define the levels used in the experiment.
3. If the experiment were replicated, the same levels of that factor would be used.

The difference between random and fixed factors can be illustrated with a simple experimental design. Consider a drug study using 0, 5, or 10 mg doses of an experimental drug on groups of experimental animals. If the researcher is only interested in the difference in response between the 3 levels of drug exposure, the level of the drug is fixed. This is the usual situation. However, if the researcher studies the effect of a compound in the natural diet of animals he might select animals from 3 populations which vary in their intake and define these as levels 1, 2, and 3. This will be a random effect factor, as the actual level is not fixed by the researcher and would change if the experiment were replicated at another time. Further, he will almost certainly wish to make inferences about the effect of the compound on other populations of the animal which may well be exposed to different quantities of the compound.

## 9.2 A simple ANOVA using a GLM

We use as an example the fertiliser yield data in [Grafen and Hails](#) <sup>[18]</sup> page 2. The [demo data set](#) <sup>[21]</sup> is included with QED, filename: **fertiliser GLM 1 factor.csv**, and is listed at the bottom of this page. You can also run this analysis using a conventional one-way ANOVA using the file **1 way ANOVA fertiliser GH.csv**.

The data set comprises the crop yield in tonnes from 10 field plots allocated to each of 3 fertilisers. The basic question is, does the fertiliser affect the yield? We can express this question using the word equation:

YIELD = FERTILISER

The variable in the right-hand column in the data set below, YIELD, is the dependent (response) variable we wish to explain. The variable on the left, FERTILISER, is the independent or explanatory variable.

As there were 3 fertiliser treatments coded as 1, 2 or 3, FERTILISER is a categorical variable. By contrast, YIELD is a continuous variable.

To run the analysis select GLM and choose Yield as the dependent variable; the program will recognise that it is a continuous variable. Then select Fertiliser as an effects variable. As there is only one independent variable there are no interaction terms. When the GLM is run it produces the following analysis of variance table:

Source	DF	SS	MS	F	Prob
Full Model	2	10.823	5.411	5.702	0.009
Fertiliser	2	10.823	5.411	5.702	0.009
Residual	27	25.622	0.949		
Total	29	36.445			

This table shows that fertiliser does affect yield, with a probability of  $p = 0.009$ . This table is the same as that produced by a [one-way analysis of variance](#) [141].

The fertiliser data from Grafen and Hails

FERTIL	YIELD
1	6.27
1	5.36
1	6.39
1	4.85
1	5.99
1	7.14
1	5.08
1	4.07
1	4.35
1	4.95
2	3.07
2	3.29
2	4.04
2	4.19
2	3.41
2	3.75
2	4.87
2	3.94
2	6.28
2	3.15
3	4.04
3	3.79
3	4.56
3	4.55
3	4.55
3	4.53
3	3.53
3	3.71
3	7.00
3	4.61

### 9.3 A simple linear regression using a GLM

We use as an example the tree data in [Grafen and Hails](#) [18] page 29. The [demo data set](#) [21] is included with QED, filename: **tree.csv** and is listed at the bottom of this page.

The data set comprises the height and timber volume of 31 felled trees. The foresters wish to establish a relationship between height, which they can easily measure in the field, and the volume of timber in the tree. We can express this relationship using the word equation:

VOLUME = HEIGHT

The variable on the right-hand column in the data set below, VOLUME, is the dependent (response) variable we wish to predict. The variable on the left, HEIGHT, is the independent variable used to predict the volume.

To run the analysis, select GLM and choose Volume as the dependent variable; the program will recognise that it is a continuous variable. Then select height as a continuous effects variable ([covariate](#)<sup>[164]</sup>). As there is only one independent variable, there are no interaction terms. When the GLM is run it produces the following analysis of variance table.

Source	DF	SS	MS	F	Prob
Height	1	2901.189	2901.189	16.164	0.000
Error	29	5204.895	179.479		
Total	30	8106.084			

This shows that height is a good predictor of wood volume.

The regression coefficients were as follows:

Variable	Coef	SECoef	t	Prob.>t
Height	1.543	0.384	4.021	0.000
Constant =	-87.124			

so that the equation is:

$$\text{Volume} = -87.123 + 1.543 \times \text{Height}$$

Tree data from Grafen and Hails, page 29.

HEIGHT	VOLUME
70	10.3
65	10.3
63	10.2
72	16.4
81	18.8
83	19.7
66	15.6
75	18.2
80	22.6
75	19.9
79	24.2
76	21.0
76	21.4
69	21.3
75	19.1
74	22.2
85	33.8
86	27.4
71	25.7
64	24.9
78	34.5
80	31.7
74	36.3
72	38.3
77	42.6
81	55.4
82	55.7
80	58.3
80	51.5
80	51.0
87	77.0

## 9.4 Using more than 1 explanatory variable in a GLM

We use as an example the maths ability data in Grafen and Hails page 56. The [demo data set](#)<sup>[21]</sup> is included with QED, filename: **school maths.csv** and is listed at the bottom of this page.

The data set comprises a random sample of 32 children who sat a maths test (AMA - average maths ability). Their ages and heights were also measured. We will test the hypothesis that taller children are better at maths. Our first word equation is therefore

AMA = HEIGHT

Height is a continuous variable and is therefore a [covariate](#)<sup>[164]</sup>.

Running a GLM, with AMA as a continuous dependent variable and Height as the single continuous effects variable, gives the following result.

Source	DF	SS	MS	F	Prob
Height	1	412.774	412.774	726.866	0.000
Residual	30	17.036	0.568		
Total	31	429.811			

From this result we might conclude that mathematical ability was indeed related to height. However this is wrong, as we have not considered the effect of age. Rerunning the analysis, with Years also added as a second effects continuous variable, gives the following result.

Source	DF	SS	MS	F	Prob
Years	1	422.604	422.604	1702.428	0.000
Height	1	0.008	0.008	0.032	0.860
Residual	29	7.199	0.248		
Total	31	429.811			

From which it can be concluded that age rather than height is the key predictor of mathematical ability, and height is unimportant.

Here is the school maths data used in the example above.

AMA	Years	Height
10.31	5	129.44
10.77	5.2	129.46
10.16	5.4	131.81
11.73	5.6	130.54
12.20	5.8	130.73
11.40	6	134.31
11.80	6.2	134.07
13.39	6.4	135.27
12.30	6.6	135.24
14.49	6.8	137.21
14.41	7	137.66
13.83	7.2	137.88
14.71	7.4	141.74
15.11	7.6	142.28
14.90	7.8	141.71
15.93	8	144.73
16.24	8.2	145.88
16.47	8.4	145.5
17.73	8.6	144.3
17.89	8.8	145.23
17.81	9	145.66
18.62	9.2	146.06
18.66	9.4	148.19
18.99	9.6	151.27
19.06	9.8	149.4
19.38	10	150.62

20.72	10.2	153.26
20.67	10.4	156.84
21.72	10.6	155.97
20.78	10.8	155
21.85	11	157.13
22.62	11.2	158.04

## 9.5 Using sequential and adjusted sums of squares

Using a simple example we show how an examination of the adjusted and sequential sums of squares both within a single ANOVA table and between tables can help you to identify the best model. We use as an example the urban foxes data in [Grafen and Hails](#)<sup>[18]</sup> page 65. The [demo data set](#)<sup>[21]</sup> is included with QED, filename: foxes.csv and is listed at the bottom of this page.

The example data set used here derives from a study of fox survival over winter ([Grafen & Hales, 2002](#)<sup>[18]</sup>). Over a three year period 30 groups of foxes were followed and the following variables recorded:

Group size	- the number of individuals in a group
Weight	- the mean weight of adults in a group
Food	- an estimate of food availability in the group territory
Area	- the territorial area.

Neither of the following single independent variable models was significant:

WEIGHT = FOOD.  
WEIGHT = GROUP SIZE.

Both independent variables were then combined to give the model:

WEIGHT = FOOD + GROUP SIZE.

As is shown below, the result was significant for both variables.

Source	DF1	DF2	Seq SS	Adj. SS	MS	F	Prob
Food	1	27	0.063136	4.70392	4.70392	16.5873	0.000365
Group size	1	27	5.43795	5.43795	5.43795	19.1757	0.000161

An examination of the sequential (0.063) and adjusted (4.7) sums of squares for the Food variable shows that the addition of the Group size variable greatly improves the ability of food to explain the weight of the foxes. As Grafen and Hales (2002) explain this is because it is the available food per fox in the territory, not the total food present, that determines average weight.

Now the third variable Area is added, giving the model:

WEIGHT = FOOD + GROUP SIZE + AREA.

The results below show that area is not significant and food has reduced in significance.

Source	DF1	DF2	Inc. SS	Adj. SS	MS	F	Prob
Food	1	26	0.063136	1.49379	1.49379	5.57004	0.026055
Group size	1	26	5.43795	5.8434	5.8434	21.7889	8.08E-05
Area	1	26	0.684071	0.684071	0.684071	2.55076	0.122325

A comparison of the adjusted sums of squares for Food between the two factor and three factor models shows a decline from 4.7 to 1.49. This indicates that food is less informative of fox weight when groups size and area are known. Since Area is not significant the above analysis suggest that the following model is most suitable:

WEIGHT = FOOD + GROUP SIZE.

The urban foxes data from Grafen and Hails.

Food	Group size	Weight	Area
0.3698	2	3.933	1.094
0.5314	2	5.702	2.05
0.4944	2	4.548	2.121
0.4513	2	4.309	1.294
0.7437	3	5.851	3.784
0.572	3	4.793	2.238
0.7387	3	5.372	2.752
0.4199	3	4.454	1.883
0.6815	3	5.614	3.769
0.6507	3	4.947	1.734
0.5114	3	4.565	2.209
0.9817	7	4.083	3.841
0.6024	4	3.557	3.018
0.7675	4	4.996	3.43
0.7289	4	4.558	2.651
0.7221	4	4.162	3.542
0.6612	4	4.964	2.446
1.2146	8	4.091	5.069
0.6832	4	4.685	2.59
0.7751	4	3.816	3.126
0.783	4	4.539	3.557
0.7982	4	5.83	3.348
0.6851	4	4.399	3.158
0.7136	4	4.98	3.042
1.0311	5	4.49	3.661
0.7794	5	4.29	3.917
0.7939	5	3.371	2.886
0.912	6	5.045	4.541
0.6716	4	4.391	2.754
0.4123	3	3.271	1.907

## 9.6 Combining continuous and categorical variables in a GLM

A general linear model can include both categorical and continuous variables within the same model. The resulting model therefore has features of both an analysis of variance and linear regression. We use as an example the body fat data in Grafen and Hails page 65. The [demo data set](#)<sup>[21]</sup> is included with QED, filename: **fat.csv** and is listed at the bottom of this page.

The objective is to determine if total fat can be predicted using weight and sex. Weight is a continuous variable and sex a categorical variable.

The model is:

FAT = WEIGHT + SEX.

In this example WEIGHT is a continuous ([covariate](#)<sup>[164]</sup>) variable and SEX is a categorical ([fixed](#)<sup>[163]</sup>) variable with the values 1 or 2. [Effect coding](#)<sup>[166]</sup> was used for the categorical (fixed) variable.

First, the Analysis of Variance table when sex is excluded indicates that there is no significant relationship between fat and weight.

FAT = WEIGHT

Source	DF1	Inc. SS	Adj. SS	MS	F	Prob
--------	-----	---------	---------	----	---	------

Weight 0.751001	1	1.32824	1.32824	1.32824	0.104011
--------------------	---	---------	---------	---------	----------

When both sex and weight are included in the model the Analysis of Variance table clearly shows that both weight and sex are significant predictors of fat (see Adjusted SS below).

Source	DF1	Inc. SS	Adj. SS	MS	F	Prob
Weight	1	1.32824	87.1049	87.1049	33.9962	0.0
Sex	1	176.098	176.098	176.098	68.7293	0.0

The output from the GLM also gives the coefficients for the regression model linking sex and weight to fat.

Variable	Coefficient	Std.Error	t	Prob.>t
Weight	0.217	0.037	5.831	0.000
Sex 1	3.952	0.477	8.290	0.000
Constant =	13.010			

So that the regression equation for Sex 1 (Females) is:

$$\text{Fat} = 13.010 + 0.217 \times \text{Weight} + 3.952.$$

The equation for the males (Sex 2) can be easily calculated, as the sum of the Sex coefficients = 0. So the equation is:

$$\text{Fat} = 13.010 + 0.217 \times \text{Weight} - 3.952.$$

Total fat data from Grafen & Hails (2002). Males = 2 & Females = 1.

WEIGHT	FAT	SEX
89	28	2
88	27	2
66	24	2
59	23	2
93	29	2
73	25	2
82	29	2
77	25	2
100	30	2
67	23	2
57	29	1
68	32	1
69	35	1
59	31	1
62	29	1
59	26	1
56	28	1
66	33	1
72	33	1

## 9.7 Studying interactions in a GLM

As an example we use the results from a factorial experiment to determine the optimal conditions for growing tulips, described on page 120 of [Grafen and Hails \(2002\)](#)<sup>[18]</sup>. The [demo data set](#)<sup>[21]</sup> is included with QED, filename: **tulip.csv**, and is listed at the bottom of this page.

The word equation for this study is:

$$\text{BLOOMS} = \text{BED} + \text{WATER} + \text{SHADE} + \text{WATER} * \text{SHADE}$$

BED, WATER and SHADE are all categorical variables and are therefore [fixed](#)<sup>[163]</sup> variables. There were 3 levels of shade and water regimes, giving a total of 9 combinations. The experiment was carried out on 3 beds, each divided into 9 plots which each received one of the 9 treatment combinations. [Effect coding](#)<sup>[166]</sup> was used for the categorical (fixed) variables.

The Analysis of Variance table for the above model using adjusted sums of squares for the tests was:

Source	DF1	DF2	Inc. SS	Adj. SS	MS	F	Prob
Bed	2.000	16.000	13811.350	13811.350	6905.675	3.880	0.042
Water	2.000	16.000	103625.781	103625.781	51812.891	29.112	0.000
Shade	2.000	16.000	36375.938	36375.938	18187.969	10.219	0.001
Water*Shade	4.000	16.000	41058.141	41058.141	10264.535	5.767	0.005
Error	16.000		28476.836		1779.802		
Total	26.000		223348.047				

We conclude that the number of blooms is sensitive to water and shading and the response to the watering depends on the level of shade.

The regression coefficients were as follows (those in italics were calculated as from the others to give a sum of zero)

Variable	Coef	SECoef	t	Prob.>t
Bed1	-31.870	11.482	-2.776	0.014
Bed2	13.589	11.482	1.183	0.254
<i>Bed3</i>	<i>18.28</i>			
Water1	-77.725	11.482	-6.769	0.000
Water2	3.846	11.482	0.335	0.742
<i>Water3</i>	<i>73.87</i>			
Shade1	51.436	11.482	4.480	0.000
Shade2	-19.667	11.482	-1.713	0.106
<i>Shade3</i>	<i>-31.77</i>			
Water1*Shade1	-72.665	16.238	-4.475	0.000
Water1*Shade2	12.945	16.238	0.797	0.437
<i>Water1*Shade3</i>	<i>59.73</i>			
Water2*Shade1	29.924	16.238	1.843	0.084
Water2*Shade2	-6.483	16.238	-0.399	0.695
<i>Water2*Shade3</i>	<i>-23.44</i>			
<i>Water3*Shade1</i>	<i>42.75</i>			
<i>Water3*Shade2</i>	<i>-6.46</i>			
<i>Water3*Shade3</i>	<i>-36.29</i>			

Constant = 128.994

So the equation for bed 1 with watering regime 1 and shade level 1 is:

Blooms = 128.994 - 31.87 - 77.72 + 51.44 - 72.67.

The tulip data from Grafen and Hailes (2002) is tabulated below:

BED	WATER	SHADE	BLOOMS
1	1	1	0
1	1	2	0
1	1	3	111.04
1	2	1	183.47

1	2	2	59.16
1	2	3	76.75
1	3	1	224.97
1	3	2	83.77
1	3	3	134.95
2	1	1	80.1
2	1	2	85.95
2	1	3	19.87
2	2	1	213.13
2	2	2	124.99
2	2	3	65.48
2	3	1	361.66
2	3	2	197.13
2	3	3	134.93
3	1	1	10.02
3	1	2	47.69
3	1	3	106.75
3	2	1	246
3	2	2	135.92
3	2	3	90.66
3	3	1	304.52
3	3	2	249.33
3	3	3	134.59

## 9.8 Fixed or categorical variables

A fixed variable is a categorical variable that can only have certain fixed levels, and must run from 1 to n. An example of a fixed variable is the sex of a subject, that can usually only be male and female, so this would be coded as 1 (male) and 2 (female). Another example would be a study using 4 levels of fertilizer application which could be coded as 1, 2, 3 or 4.

If you have fixed variables which do not run from 1, provided they are evenly spaced (for instance, 5, 10, 15, 20, 25), you can use the Divide by Constant function on the [Working Data](#)<sup>[89]</sup> tab; in this instance, choose Divide by Constant from the Transform box, select the column containing the variables from the drop-down menu (because you only wish to transform one column, not the whole data set), and enter 5 as the constant, then press Submit.

In comparison, a [covariate variable](#)<sup>[164]</sup> is a variable that can take a range of values that are not fixed by the experiment. For example, rainfall, pH of rivers or the minimum nighttime temperature.

As an example, consider the fat data in [Grafen and Hails](#)<sup>[18]</sup> page 99. The [demo data set](#)<sup>[21]</sup> is included with QED, filename: fat.csv and is listed below. The total body fat of 19 students was recorded, together with their sex. In this example sex is a fixed variable and weight is a covariate variable.

WEIGHT	FAT	SEX
89	28	2
88	27	2
66	24	2
59	23	2
93	29	2
73	25	2
82	29	2
77	25	2
100	30	2
67	23	2
57	29	1
68	32	1
69	35	1
59	31	1

62	29	1
59	26	1
56	28	1
66	33	1
72	33	1

## 9.9 Covariate variables

A covariate variable is a variable that can take a range of values that are not fixed by the experiment. For example, rainfall, pH of rivers or the minimum nighttime temperature.

In comparison, a variable is termed a [categorical variable](#)<sup>[163]</sup> if it can only have certain fixed levels. An example of a fixed variable is the sex of a subject that can usually only be male and female, so this would be coded as 1 (male) and 2 (female). Another example would be a study using 4 levels of fertilizer application which would be coded as 1, 2, 3 or 4.

As an example, consider the fat data in Grafen and Hails page 99. The data set is included with QED, filename: fat.csv and is listed below. The total body fat of 19 students was recorded together with their sex. In this example sex is a fixed variable and weight is a covariate variable.

WEIGHT	FAT	SEX
89	28	2
88	27	2
66	24	2
59	23	2
93	29	2
73	25	2
82	29	2
77	25	2
100	30	2
67	23	2
57	29	1
68	32	1
69	35	1
59	31	1
62	29	1
59	26	1
56	28	1
66	33	1
72	33	1

## 9.10 Coding categorical variables

To use [categorical](#)<sup>[163]</sup> or fixed predictor variables in a GLM or other forms of regression analysis, they must be recoded. For each categorical variable it is necessary to create a set of new variables to represent the various levels of the categorical variable.

To understand why they must be recoded, consider the following simple example. Fish are reared using 3 different diets, and weighted after 6 weeks:

Diet 1: mean weight = 45  
 Diet 2: mean weight = 62  
 Diet 3: mean weight = 88

If we plot weight against diet (1, 2, or 3), it would look like we actually had a linear relationship between these variables. However, if we had coded the diets as 2, 1, 3 (instead of 1, 2, 3) the

relationship would look quadratic instead. The parameter estimates from a regression of weight on diet using either of these two coding schemes would not be interpretable, due to the fact that the independent variable here is only a measure of group membership and is therefore arbitrary.

To overcome this problem, if there are  $k$  different categories (diets) we create  $k-1$  new variables to describe the membership. Pedhazur (1982) described the 3 common types of coding scheme.

[Dummy coding](#) <sup>165</sup>

[Effect coding](#) <sup>166</sup>

[Orthogonal coding](#) <sup>167</sup>

### 9.10.1 Dummy coding

For each categorical variable with  $k$  levels,  $k-1$  coded vectors consisting of 1s and 0s are created. In each vector, subjects in one of the groups are assigned 1s, and all the other groups are assigned 0s. Using this method, one of the groups will always be assigned all 0's.

The following example shows the creation of two vectors to code for groups A, B and C. As there are 3 groups,  $K=3$  and we therefore need  $k-1 = 2$  new variables. Note that membership to group A is coded as (1,0), group B as (0,1) and group C as (0,0).

Subject	Group	Measurement	Coded vector 1(V1)	Coded vector 2(V2)
1	A	y1	1	0
2	A	y2	1	0
3	B	y3	0	1
4	B	y4	0	1
5	C	y5	0	0
6	C	y6	0	0

The group that is assigned all zeros (group C above) is referred to as the comparison (or control) group. In the regression equation, the constant term is equal to the mean of the comparison group. Each slope coefficient is equal to the difference between the mean for the  $i$ 'th group (assigned 1s in the vector) and the mean of the comparison group. The test of significance for each of the  $b$ 's is a test of this difference (i.e., is it significantly different from 0).

The regression model is

$$Y = C + b_1CV1 + b_2CV2$$

In a GLM using the orthogonal coding scheme, the constant term is the "grand" mean for the  $Y$ s. The "grand" mean is unweighted, it is the average of the group averages, and will equal the overall average for  $Y$  only when the sample sizes are equal.

For group A the equation becomes:  $Y = \text{Const} + b_1$   
 For group B the equation becomes:  $Y = \text{Const} + b_2$   
 For group C the equation becomes:  $Y = \text{Const}$

Dummy coding is often used for its simplicity, even though there might not be a meaningful control or comparison group. It is indifferent to equal or unequal group sizes.

Dummy coding is useful when testing for differences of groups to a control group.

### 9.10.2 Effect coding

For each categorical variable with  $k$  levels,  $k-1$  coded vectors consisting of 1s, -1s and 0s are created. Effect coding is similar to dummy coding, except that in effect coding the group previously assigned all 0s is now assigned all -1s.

The following effect coding example shows the creation of two vectors to code for groups A, B and C. As there are 3 groups  $K=3$  and we therefore need  $k-1=2$  new variables. Note that membership to group A is coded as (1,0), group B as (0,1) and group C as (-1,-1).

Subject	Group	Measurement	Coded vector 1(V1)	Coded vector 2(V2)
1	A	y1	1	0
2	A	y2	1	0
3	B	y3	0	1
4	B	y4	0	1
5	C	y5	-1	-1
6	C	y6	-1	-1

The result of this coding scheme is that the regression model corresponds now to the linear model commonly seen in ANOVA designs.

The regression model is

$$Y = C + b_1CV1 + b_2CV2$$

In a GLM using the orthogonal coding scheme, the constant term is the “grand” mean for the  $Y$ s. The “grand” mean is unweighted, it is the average of the group averages, and will equal the overall average for  $Y$  only when the sample sizes are equal.

For group A the equation becomes:  $Y = \text{Const} + b_1$

For group B the equation becomes:  $Y = \text{Const} + b_2$

For group C the equation becomes:  $Y = \text{Const} - b_1 - b_2$

In a GLM using the effect coding scheme, the  $b_i$  is the “grand” mean for the  $Y$ 's. The “grand” mean is unweighted, it is the average of the group averages, and will equal the overall average for  $Y$  only when the sample sizes are equal. Each  $b_j$  is equal to the difference between the mean of the  $i$ 'th group assigned 1s in the coding scheme, and the “grand” mean. This coding method is appropriate when you wish to generate a result similar to a conventional ANOVA and group sizes are the same.

Effect coding is useful when testing for differences of groups from the grand mean.

### 9.10.3 Orthogonal coding

This is also termed contrast coding.

For each categorical variable with  $k$  levels,  $k-1$  coded vectors are created. Each coded vector is designed to make a specific comparison between the means of the  $k$  groups.

The following orthogonal coding example shows the creation of two vectors to code for groups A, B and C. As there are 3 groups  $K=3$  and we therefore need  $k-1 = 2$  new variables. Note that membership to group A is coded as (1,-1), group B as (-1,-1) and group C as (0,2).

Subject	Group	Measurement	Coded vector 1	Coded vector 2
1	A	y1	1	-1
2	A	y2	1	-1
3	B	y3	-1	-1
4	B	y4	-1	-1
5	C	y5	0	2
6	C	y6	0	2

The regression model is

$$Y = C + b_1CV1 + b_2CV2$$

In a GLM using the orthogonal coding scheme, the constant term is the “grand” mean for the Y’s. The “grand” mean is unweighted, it is the average of the group averages, and will equal the overall average for Y only when the sample sizes are equal.

For group A the equation becomes:  $Y = \text{Const} + b_1 - b_2$

For group B the equation becomes:  $Y = \text{Const} - b_1 - b_2$

For group C the equation becomes:  $Y = \text{Const} + 2b_2$

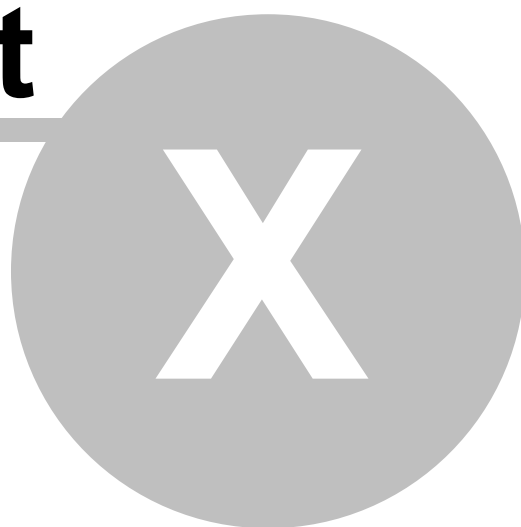
The predicted  $y$  values are the means of the respective groups.

With unequal sample sizes, an analysis using orthogonal coefficients no longer gives independent pieces of information.

Orthogonal coding is useful when testing hypotheses about patterns of group means with equal group sizes.

# Part

---



## 10 Printing and saving results

Output can be saved as a file, copied to the clipboard or printed. See topics below for further details.

[Printing text and grid output](#)

[Preparing charts for output](#)

[Printing charts](#)

[Exporting charts](#)

### 10.1 Exporting charts

The chart can be saved in a number of different file formats:

Enhanced Metafile (\*.emf),

Bitmap (\*.bmp),

JPEG (\*.jpg), or

Each file format has advantages and disadvantages.

- The advantage of Enhanced Metafile is that, if pasted into, for instance, a Word document, it can be resized by dragging, without losing resolution.
- Bitmaps are a lossless method of saving; the stored file will not lose any of the original's detail. Because of this, bitmaps tend to be much larger than compressed files such as Enhanced Metafiles or JPEGs.
- JPEGs are file formats which can be compressed to take up less space - useful if you wish to send one by email, put it on a website, or paste it in to a document. If they are compressed too heavily, they can lose resolution and detail, and spoil colours.

### 10.2 Printing charts

The graphs and other results can be printed using **File|Print**, or the Print button on the toolbar.

The Print Preview option will show the page layout and allow image size, margins, and paper orientation to be changed.

If you have Adobe Acrobat installed on your computer, you will be able to convert the chart directly to a .pdf file by selecting Acrobat Distiller or Acrobat PDF Writer from the list of available printers in the Print dialog box.

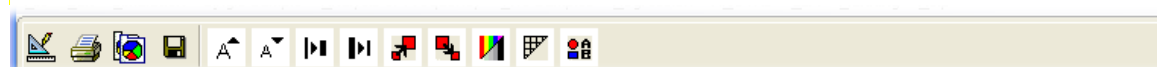
### 10.3 Printing and exporting text and grid output

With an output grid showing simply choose **File|Print** and a dialog box for printing options will be activated.

To save the results to file choose **File|Export** and follow instructions in the dialogue box.

### 10.4 Preparing charts for output

The graph option buttons on the Chart Toolbar are described in order from left to right below.



**Edit - (Set square and pencil)** This button will offer a wide range of options to change a wide range of aspects of the graph, add titles, and use various tools to customise the chart. It is also

used to export or copy your graph to file, and even to email it, using the 'Send' button. For more information on chart editing use the TeeChart help system available from the Help button on the chart edit box.

**Print - (Printer icon)** Use this button to print the graph.

**Copy - (pie chart icon)** Use this option to copy the graph to the clipboard.

**Save - (Floppy disc icon)** Save the file in a variety of different formats.

**Increase font size - (large A icon)**

**Decrease font size - (small A icon)**

**Increase line thickness**

**Decrease line thickness**

**Increase point size**

**Decrease point size**

**Change from colour to black and white**

**Add or remove grid from graph**

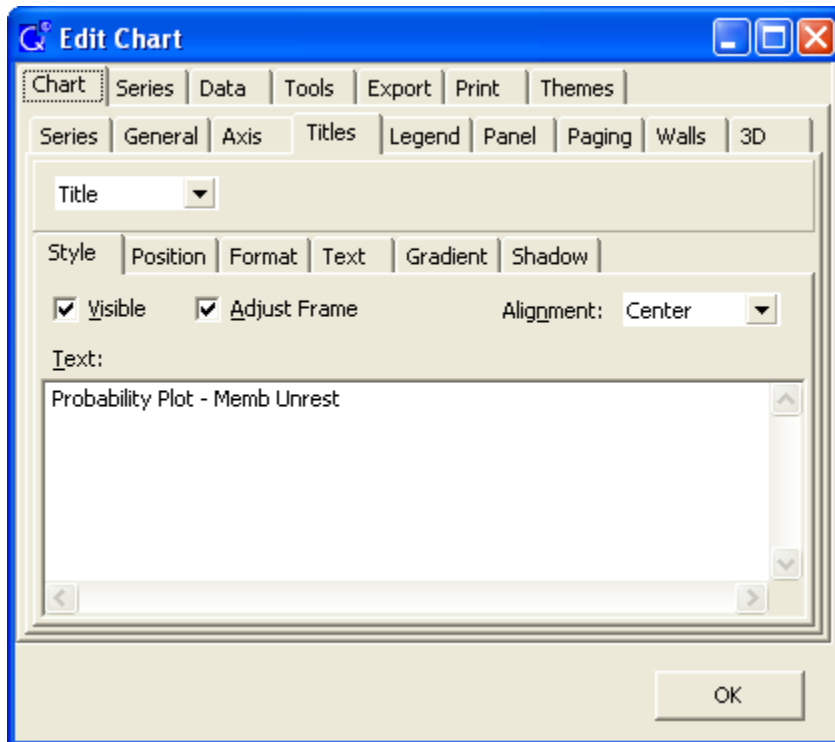
**Add or remove legend**

To edit elements of the graph, such as titles, points, line colours, etc. hover over the element you wish to edit, press the Shift key on your keyboard, and the cursor will change to a hand pointer:

[Probability Plot - Memb Unrest](#)



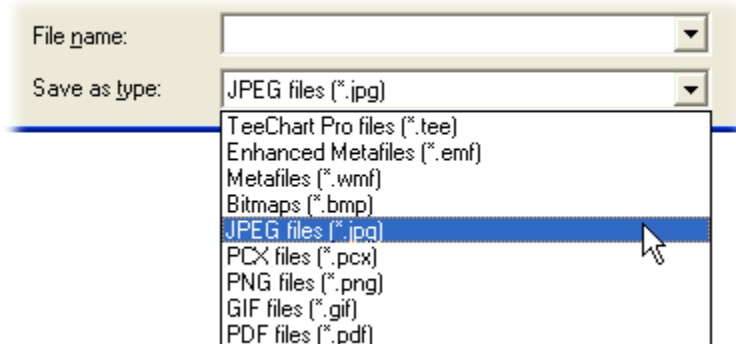
Click on the element, and the Edit Chart dialog will appear at the appropriate page for you to make the changes you require:



Alternatively, to alter the chart title, with the chart showing, click on the set-square icon to show the editing screen, select the Chart tab from the top row of tabs, then the Titles tab from the second row.

### Saving charts as files.

The chart can be saved in a number of different file formats:



Each file format has advantages and disadvantages.

- The advantage of Enhanced Metafile is that, if pasted into, for instance, a Word document, it can be resized by dragging, without losing resolution.
- Bitmaps are a lossless method of saving; the stored file will not lose any of the original's detail. Because of this, bitmaps tend to be much larger than compressed files such as Enhanced Metafiles or JPEGs.
- JPEGs are file formats which can be compressed to take up less space - useful if you wish to send one by email, put it on a website, or paste it in to a document. If they are compressed too heavily, they can lose resolution and detail, and spoil colours.

# Index

## - % -

% Zeros 97

## - . -

.bmp 169

.jpeg 169

.jpg 169

.pdf 169

.wmf 169

## - + -

+ve infinity 15

## - 1 -

1/1 line 63, 64, 65, 66

1-tailed 129

1-way 78

1-way ANOVA - results 73, 76

1-way anova - setup 72, 74

## - 2 -

2 x 2 contingency table - results 45

2 x 2 contingency table - setup 44

2-sample Chi-squared 60

2-sample Chi-squared - results 61

2-tailed 129

2-way ANOVA - results 77

2-way ANOVA - setup 76

## - A -

acknowledgement 17

add constant 91

add or remove grid from graph 169

add or remove legend 169

adding variables 138

Adj. SS 82

adjust data 89

adjusted R squared 66

adjusted sums of squares 82, 159

Adobe Acrobat 169

alert 15

alerts 15

alternative hypothesis 86

always load last-used data file at startup 27

analysis of covariance 154

analysis of frequency - methods of 117

analysis of frequency drop-down menu 42

analysis of variance 66, 71, 72, 73, 74, 76, 77, 78, 141, 143, 145, 147, 154

analysis of variance table 156, 159

ANCOVA 154

ANOVA 66, 71, 72, 73, 74, 76, 77, 78, 79, 141, 143, 144, 145, 147, 154, 155

ANOVA (1-way) - setup 72, 74

ANOVA (2-way) - setup 76

ANOVA example - GLM 155

ANOVA table 159

ANOVAs drop-down menu 71

arcsin 91

arcsin square root 91

ASCII 25, 94

association 119

association between variables 117, 132, 133, 134

assume equally likely 60

average 30, 100

## - B -

backward stepwise 68, 69

backward stepwise linear regression 138

base10 91

best-fit 136

between 73, 77

between columns 77

between rows 77

bin 35

binary 92

binary using mean 92

binary using median 92

binned data 33

binned frequency 107

binning 35

binomial distribution 111

bins 33, 107

bitmap 169

blank cells 90

bmp 169

Bonferroni adjustment 145

Bonferroni test 145

borders 26  
 box and whisker 29  
 box and whisker plot 28, 33, 106  
 box plot 33, 106  
 breakdown of calculations 96  
 browser 98  
 by maximum value 92  
 by mean 92  
 by StdD 92

## - C -

calculations 95  
 candlestick chart 106  
 categorical 155  
 categorical data 117  
 categorical variables 154, 160, 161, 163, 164, 165, 166, 167  
 CD 2  
 cdf 108  
 cell does not contain a float value 12, 15  
 change data 13, 89, 90  
 change from colour to black and white 169  
 chart 169  
 chart export 169  
 chart format 169  
 chart toolbar 169  
 charts 169  
 check distribution 107  
 checklist 17  
 Chi 34, 35, 45, 46, 47, 48, 49, 60  
 Chi-squared 12, 34, 35, 45, 46, 47, 48, 49, 60, 110, 127  
 Chi-squared based methods 119  
 Chi-squared test 117  
 Chi-squared test - contingency table 118  
 Chi-squared test for normality 107  
 Chi-squared two sample test 127  
 choose number of samples 83, 84, 85, 86  
 choose sample 38  
 choose samples 66  
 choose variable type 80  
 choosing best method 98  
 citation 17  
 cite QED 17  
 class intervals 110  
 classification variables 155  
 clipboard 169  
 close 24  
 coded vectors 165, 166, 167  
 coding 80, 164, 165, 166, 167  
 coefficient of determination 66, 69  
 coin toss 111  
 column delete 6  
 column effect type 72, 76, 78  
 column headers 6  
 column insert 6  
 column labels 6  
 column statistics 97  
 column titles 6  
 columns 92, 93  
 combine continuous and categorical variables 160  
 comma-delimited text 10  
 comma-separated values 10  
 compare distributions 127  
 compare means 123, 141, 143  
 compare medians 128  
 compare two samples 50  
 comparing frequencies 127  
 comparing frequency distributions 60  
 comparing means 51, 52, 54, 58, 123  
 comparing means - equal size and variance 124  
 comparing means - unequal size, equal variance 125  
 comparing means - unequal variances 126  
 comparing medians 59, 129  
 comparing variances 57  
 comparison 145  
 comparisons 60, 144, 145  
 computer 2  
 concordant 64  
 constant 66, 69, 91  
 contact pisces 14  
 contingency coefficient 45, 47, 49, 119  
 contingency table 42, 48, 117, 121  
 contingency table (2 x 2) - results 45  
 contingency table (2 x 2) setup 44  
 contingency table (G-Test) - results 49  
 contingency table (G-Test) setup 48  
 contingency table (R x C) - results 47  
 contingency table (R x C) setup 46  
 contingency table Chi-squared test 118  
 contingency table data 12  
 contingency table G-Test 120  
 contingency table tests 119, 120  
 contingency tables 42, 120  
 continuous 155  
 continuous effects variable 156  
 continuous variable 158  
 continuous variables 154, 160

copy 3, 27, 95, 169  
 copying 169  
 corrected 134  
 correlation 62, 64, 65, 86  
 correlation coefficient 63, 66, 69  
 correlation coefficients 132, 133, 134  
 correlation methods 132  
 correlation results 63, 64, 65  
 covariate variable 80, 158  
 covariate variables 163, 164  
 Cramer's V 45, 47, 49, 119  
 create contingency table 12  
 create data set 87  
 create data set - contingency table 12  
 create new data set 6, 10  
 csv 10, 25, 94  
 cumulative distribution function 108  
 cumulative normal probability 32

## - D -

data 3, 6  
 data entry wizard 6, 10  
 data file 10, 13  
 data problems 17  
 data set 10  
 data set size 2, 13  
 data set statistics 97  
 data set structure 10  
 data set templates 10  
 data sets 21  
 data structure 4  
 data structure - contingency tables 12  
 data transformation 90  
 data transformations 13  
 decimal places 27  
 decrease font size 169  
 decrease line thickness 169  
 decrease point size 169  
 degrees of freedom 38, 39, 40, 45, 47, 49, 58, 61, 63, 73, 76, 77, 79, 82, 127, 129, 148, 150  
 delete 93  
 delete columns and rows 6, 87  
 demo data 21  
 demonstration data sets 21  
 demonstrations 98  
 demos 21  
 dependent variable 66, 68, 69, 80, 135, 156  
 dependent variable variability 148  
 dependent variables 138

dependent variance 148  
 deselect column 93  
 deselect data 93  
 deviation 101  
 DF 38, 39, 40, 45, 47, 49, 51, 53, 55, 56, 58, 61, 63, 73, 76, 77, 79, 148, 150  
 DF1 82  
 DF2 82  
 difference between variances 127  
 difference squared 65  
 direct 6  
 directly 6  
 directly related variables 139  
 discordant 64  
 distribution 33, 101, 102, 103, 106, 107, 111, 112  
 distribution - exponential 113  
 distribution test 105  
 distribution testing 100, 101  
 distributions 28  
 divide by constant 91, 163  
 drag 35  
 drop-down menus 24  
 dummy coding 80, 164, 165, 166, 167

## - E -

edit 169  
 edit charts 169  
 edit data 13, 87  
 edit data set 6  
 edit menu 27  
 effect coding 80, 160, 161, 164, 166, 167  
 effects 147, 155  
 effects variable 80  
 eliminating variables 138  
 enhanced metafile 169  
 enter data 12, 87  
 entering 3  
 entering data 6  
 equal 124  
 equal size and variance 124  
 equal variance 51, 52  
 equal variances 86  
 equality of medians 147  
 equation 95, 127  
 error 15  
 error messages 12, 15  
 error term 155  
 errors 15  
 errors in data file 10

Exact 12  
Exact test 42, 117  
Exact test (Fisher) 43  
example data sets 21  
examples 98  
Excel 3, 4, 25, 94  
Excel files 10  
existing data 13  
exit 24  
expand tab 95  
expected frequencies 121  
expected values 44, 46, 48  
explanatory variables 154, 158  
explore 28  
explore distributions 28  
explore tab 96  
exponential distribution 113  
export 13, 24, 95  
export data 25, 89, 94  
exporting 25, 169  
extreme configurations 117  
extremity 117

## - F -

F 57, 73, 76, 77, 82, 148, 150  
F statistic 82  
F Test 57, 127, 145, 154  
factorial designs 145  
FAQ 14  
file 24  
file menu 24  
first quartile 106  
Fisher 42, 43, 84  
Fisher's 42  
Fisher's Exact 12, 84, 117  
Fisher's Exact test 42, 117  
Fisher's Exact test - results 43  
fixed 147  
fixed effects 76, 155  
fixed effects model 148  
fixed variable 80  
fixed variables 163, 164  
float value 15  
floating point division by zero 15  
floating point error 15  
follow steps in calculations 96  
font 26  
font size 169  
footer 26

format 25, 26, 27  
formatted 27  
forward stepwise 68, 69  
forward stepwise linear regression 138  
frequency 42, 107, 110, 111, 118, 121  
frequency data 33  
frequency distribution 35  
F-test 57  
F-test results 57

## - G -

G 120  
general 92  
general linear model 80, 132, 141, 154, 155, 156, 163, 164, 165, 166, 167  
general linear model example 158, 160, 161  
general linear model results 82  
generate data 28  
generate distributions 28  
generate simulated data 28  
GLM 80, 82, 141, 154, 163, 164, 165, 166, 167  
GLM drop-down menu 80  
GLM example 155, 156, 158, 159, 160, 161  
GLM preferences 80  
GLM results 82  
GLM settings 80  
GLM setup dialog 80  
goodness of fit 110  
grafen 18  
grand total 121  
graph export 169  
graphics 169  
graphs 169  
G-Test 48, 117, 120  
G-Test contingency table - results 49  
G-Test contingency table - setup 48  
guaging normality 107  
guides 98

## - H -

h 79  
Hails 18  
hard disk space 2  
header 26  
headers 6  
help 14, 98  
henderson 2  
histogram plot 33, 107

homogeneity of variances 143  
 honestly 144  
 honestly significant difference test 144  
 how many samples 83  
 HTML 25, 94

## - I -

image format 169  
 import data from Excel file 4  
 importing 3  
 Inc. SS 82  
 increase font size 169  
 increase line thickness 169  
 increase point size 169  
 incremental sums of squares 82  
 independent variable 66, 68, 69, 80, 135, 148  
 independent variables 138, 154, 155  
 infinity 15  
 input 3, 6  
 input data 12, 87  
 insert columns and rows 6, 87  
 installation 2  
 instructions 2  
 integers and real numbers 6  
 interaction terms 154  
 interactions 80  
 interactions in GLM 161  
 internet explorer 98  
 interquartile range 106  
 intersection 135  
 introduction 2  
 introduction to qed statistics 2  
 invalid floating point 15  
 invalid floating point operation 15  
 issues 17

## - J -

jpeg 169  
 jpg 169

## - K -

Kendall 64  
 Kendall correlation 132, 133  
 Kendall correlation results 64  
 Kendall correlation setup 64  
 known mean 114, 115  
 Kolmogorov-Smirnov 108

Kramer 144, 145  
 Kruskal 78, 79, 147  
 Kruskal-Wallis 78, 79, 141, 147  
 Kruskal-Wallis - results 79  
 kurtosis 28, 29, 32, 38, 97, 103  
 kurtosis results 32

## - L -

labels 6  
 lack of symmetry 102  
 lambda 112, 113  
 last used file 27  
 layout 26  
 leptokurtic 103  
 levels for treatment 76, 78  
 likelihood 111  
 likelihood ratio 120  
 Lilliefors 34  
 Lilliefors test 38, 108  
 Lilliefors test for normality 107  
 Lilliefors test results 38  
 line thickness 169  
 linear regression 132, 135, 136, 154  
 linear regression example - GLM 156  
 linear regression results 66  
 linear regression setup 66  
 load data 10, 89  
 loading 2  
 loading from Excel 4  
 log 91  
 log 10 91  
 log e 91  
 lower quartile 106

## - M -

Macromedia Flash 98  
 main program window 24  
 make bins 35  
 MANCOVA 154  
 Mann Whitney test - results 56  
 Mann Whitney test - setup 55  
 Mann-Whitney 123  
 Mann-Whitney U test 128  
 MANOVA 154  
 marginal totals 121  
 margins 26  
 matrix 93  
 max 97

maximum 13  
 maximum likelihood 120  
 maximum size 13  
 mean 29, 30, 39, 41, 51, 53, 55, 58, 92, 97, 100, 111  
 mean results 30  
 mean squares 73, 76, 77, 82, 148, 150  
 mean, known 114, 115  
 means 51, 144, 145  
 median 29, 31, 92, 97, 100, 128  
 median results 31  
 medians 147  
 mesokurtic 103  
 metafile 169  
 Microsoft Excel 4  
 min 97  
 missing values 93  
 mixed model 154  
 mixed models 147  
 MLR 68, 69  
 model 80  
 model II 147  
 MS 73, 76, 77, 82, 148, 150  
 multicollinearity 138, 139  
 multiple 144, 145  
 multiple comparison tests 145  
 multiple comparisons 72, 76, 78  
 multiple comparisons tests 144, 145  
 multiple linear regression 68, 69, 132, 137, 138  
 multiple predictor variables 137  
 multiply by constant 91

## - N -

N 37, 39, 41, 51, 53, 55, 58, 63, 65  
 NAN 15  
 natural 91  
 negative infinity 15  
 new 24  
 new data 10  
 new data set 3, 6  
 Newman-Keuls test 144, 145  
 no. bins 33  
 non parametric 55, 56  
 nonparametric 55, 56, 59, 78, 79, 129, 147  
 non-parametric 59, 78, 79, 123, 128, 129, 147  
 non-zeros 97  
 normal 28, 34, 35  
 normal distribution 34, 35, 101, 105, 107, 108, 110, 111

normalisation - x 64  
 normalisation - y 64  
 normality 33, 38, 107, 108  
 normality testing methods 107  
 normality tests 34, 35  
 not a number 15  
 notation 27  
 null hypothesis 86  
 number 27  
 number of bins 33, 35  
 number of groups 145  
 number of observations 13  
 number of rows and columns 13  
 number of samples 83, 84, 85, 86  
 number of tails 39, 41, 51, 53, 55, 58  
 number of variables 13

## - O -

observed and expected 110, 120  
 observed and expected frequencies 118, 120, 121, 127  
 observed and expected values 60  
 observed values 44, 46, 48  
 Omega squared 73, 77, 148, 150  
 one or two tailed test 62, 64, 65, 84, 85, 86  
 one way anova - setup 72, 74  
 one-tail 129  
 one-tailed 129  
 one-way 72, 74, 78, 79, 147  
 one-way ANOVA 141, 143, 144  
 one-way ANOVA - example 148  
 one-way ANOVA - results 73, 76  
 open 24  
 open Excel file 4  
 open spreadsheet file 4  
 opening a data set 10  
 options 169  
 orientation 26  
 orthogonal coding 80, 164, 165, 166, 167  
 outliers 33, 136  
 outlying data 33, 107  
 output 25, 26, 56, 79, 95, 169  
 output of Fisher's Exact test 43  
 outputting 25

## - P -

page layout 169  
 pages 24

pair 123  
 paired 58, 59, 129  
 paired samples 59, 129  
 paired t 58  
 paired t-Test 123  
 paired t-Test results 58  
 paired t-test setup 58  
 paper 26  
 parametric 123  
 parametric test 123  
 paste 3, 27  
 PDF 169  
 peak 103  
 Pearson correlation 62, 132, 133  
 Pearson correlation - results 63  
 Pearson correlation setup 62  
 pick X sample 63, 64, 65, 66, 69  
 pick Y sample 63, 64, 65, 66, 69  
 planning 84, 85, 86  
 planning drop-down menu 83  
 platykurtic 103  
 plots 169  
 point size 169  
 Poisson distribution 112  
 pop ups 98  
 popups 98  
 positive infinity 15  
 power 83, 84, 85, 86, 91  
 precision 27  
 preferences 27  
 print 24, 95, 169  
 print layout 169  
 print settings 26  
 print setup 26, 169  
 printer 26  
 printer options 26  
 printer setup 24  
 printing 26, 169  
 printing grids 169  
 printing output 169  
 Prob 37, 39, 41, 51, 53, 55, 56, 58, 60, 61, 63, 64, 65, 82, 148, 150  
 probability 40, 105, 111, 112, 114, 115, 117  
 probability level 129  
 probability of events 84  
 probability plot 28, 29, 32, 105  
 probability plots 107  
 problem 15  
 problems 14  
 product moment 133

program output 95  
 proportions 120  
 proving significant difference 83

## - Q -

QED Statistics website 14  
 quartiles 106

## - R -

r 66, 69, 133  
 r corrected 65  
 R squared 66, 69  
 r uncorrected 65  
 R x C contingency table - results 47  
 R x C contingency table - setup 46  
 RAM 2  
 random 147  
 random effects 76, 155  
 random variable 80  
 range 144, 145  
 range check error 15  
 rankit plot 105  
 rare events 112  
 raw data 6, 13, 87  
 real numbers and integers 6  
 recent 24, 27  
 recently used 24, 27  
 redo 27  
 references 18  
 regression 132, 135, 136, 137, 147, 155  
 regression coefficients 156  
 regression drop-down menu 62  
 regression example 156  
 regression line 63, 64, 65, 66  
 regression methods 62  
 regression results 66  
 related variables 138, 139  
 relativise data 90, 92  
 reload raw data 89, 91  
 removal of variables 82  
 remove rows 93  
 remove sparse data 93  
 removing variables 138  
 removing zeros 93  
 repeat 74, 76  
 repeated measures 74, 76, 143  
 replicate observations 145  
 reset 80

residual 66  
 results 56, 79  
 results tab 95  
 row delete 6  
 row headers 6  
 row insert 6  
 row labels 6  
 row titles 6  
 rows 92, 93  
 rows and columns 13  
 rows effect type 72, 76, 78

## - S -

sample data sets 21  
 sample statistics 29  
 sample variance 97  
 save 13, 24, 169  
 save as 13, 24, 25, 94  
 save data 94  
 save data changes 89  
 save new data file 13  
 save results 25  
 save working data 13  
 saving 169  
 saving data 13, 25  
 saving grids 169  
 saving output 169  
 saving working 13  
 scatter 136  
 scatter plot 66, 136  
 Scheffe 144, 145  
 scientific notation 27  
 sd 31  
 seaby 2  
 select interactions for GLM 80  
 select printer 26  
 select sample 38  
 select samples 51, 52, 54, 58, 62, 64, 65, 66  
 select variables 68  
 sequential sums of squares 159  
 set preferences 27  
 set up new data 10  
 Shapiro-Wilk test 34, 108  
 Shapiro-Wilk test for normality 107  
 Shapiro-Wilk test results 37  
 show calculations 95, 96  
 show equation 95  
 signed rank test 59, 129  
 significance 83  
 significance level 84, 85, 86  
 significant difference test 144  
 significant places 27  
 similar variance 141, 143  
 similar variances 143  
 simple linear regression 132  
 simulation 28  
 simulation drop-down menu 28  
 single sample drop-down menu 29  
 single sample test 100, 101, 102  
 single sample tests 100  
 single sample t-Test results 39  
 single sample t-Test setup 38  
 single sample z test 40  
 single sample z test results 41  
 single variable 29, 100  
 size of data set 2, 13  
 skew 28, 31, 38, 102  
 skewness 29, 31, 38, 97, 102  
 slope 66, 135  
 SNK 144, 145  
 somes 2  
 sparse data 93  
 Spearman rank correlation 134  
 Spearman rank correlation results 65  
 Spearman rank correlation setup 65  
 spread 101  
 spreadsheet 4  
 square root 91  
 squared deviation 101  
 SS 73, 76, 77, 148, 150  
 standard deviation 29, 101, 110, 111  
 standard deviation results 31  
 start data set 3  
 start new data set 6  
 StdD 31, 39, 41, 51, 53, 55, 56, 58, 101, 111  
 step size 35  
 stepped addition 82  
 steps in calculations 96  
 stepwise 68, 69  
 stepwise linear regression 138  
 structure of data set 4  
 student t 51, 52, 54  
 Student-Newman-Keuls 144, 145  
 style 27  
 submit data 91  
 subtract constant 91  
 subtracting variables 138  
 suggestions 98  
 sum 97

summary of data 97  
 summary statistics 97  
 sums of squares 73, 76, 77, 148, 150, 159  
 sumSqr 97  
 support 14  
 swap columns and rows 90, 93  
 symmetry 102  
 system requirements 2

## - T -

t 51, 53, 55, 58, 60, 63  
 t test 123, 124, 125, 126, 129  
 t value 38, 39  
 tabs 24  
 tailed 129  
 tails 38, 40, 129  
 Tau 64  
 templates 10  
 test 144, 145  
 test for differences between variances 123  
 test for normal 34  
 test for normality 35, 108  
 testing equality of medians 147  
 testing for normality 38, 107, 110  
 testing normality 29  
 testing single variable 100  
 third quartile 106  
 this file was not found 15, 27  
 titles 6  
 toolbar 24  
 tossing coins 111  
 total variance 97  
 transform data 90  
 transformations 89  
 transpose data 90, 93  
 transposing 93  
 t-Test 29, 38, 39, 52, 54, 58, 85, 86, 114, 123, 124, 125, 126, 129, 154  
 t-Test - comparing observations with a known mean 114  
 t-Test - equal variance balanced - results 51  
 t-Test - equal variance balanced - setup 51  
 t-Test - equal variance unbalanced - results 53  
 t-Test - equal variance unbalanced - setup 52  
 t-Test - unequal variances - results 55  
 t-Test - unequal variances - setup 54  
 Tukey 144, 145  
 Tukey-Kramer test 145  
 tutorials 98

two sample 127  
 two sample f test - results 57  
 two sample F test - setup 57  
 two sample tests 123  
 two way ANOVA - setup 76  
 two way table 46  
 two-sample Chi-squared - results 61  
 two-sample Chi-squared - setup 60  
 two-sample tests 50  
 two-tail 129  
 two-tailed 129  
 two-way 76  
 two-way ANOVA 141, 145  
 two-way ANOVA - example 150  
 two-way ANOVA - results 77  
 type of coding 80

## - U -

U 56, 128  
 U test 55, 56  
 unbalanced design 145  
 uncorrected 134  
 undo 27  
 unequal 125, 126  
 unequal size, equal variance 125  
 unequal variance 54  
 unequal variances 85, 126  
 unimodality 136  
 upper quartile 106  
 user preferences 27

## - V -

value to test 38, 40  
 variability 136  
 variable type 80  
 variables 155  
 variance 29, 31, 72, 74, 76, 101, 124, 125, 126, 127, 141, 143, 145  
 variance component models 147  
 variance ratio 57, 127  
 variance results 31  
 variances 123, 143

## ■ ■ ■

-ve infinity 15  
 vectors 165, 166, 167  
 Vista 2

## - W -

W 37, 108  
Wallis 78, 79, 147  
website 14  
Welch 126  
Welch's t-Test 126  
which method 98  
Wilcoxon 123, 129  
Wilcoxon matched pairs - results 60  
Wilcoxon matched pairs - setup 59  
Windows 2  
within 73, 77  
within groups / error 77  
within subjects 143  
wizard 6, 10  
wmf 169  
working data error 12  
working data grid 89

## - X -

xls 25, 94  
XP 2

## - Z -

z 41, 64  
z test 29  
z test - comparing observations with a known mean  
115  
z test results 41  
z test setup 40  
z value 40, 41  
zero 93  
zeros 90, 97

Endnotes 2... (after index)

Back Cover