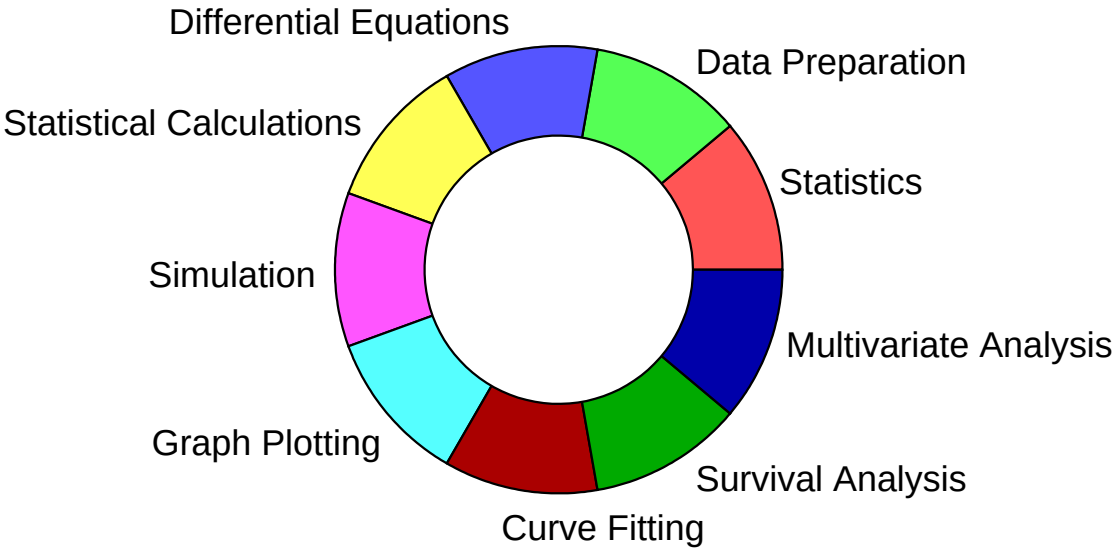
SIMFIT

SIMULATION, FITTING, STATISTICS, AND PLOTTING.

Tutorials
Test Files
and
Worked Examples

bill.bardsley@simfit.org.uk

http://www.simfit.org.uk



1 Summary

The SimFIT package is designed so that first time users can easily learn how to analyze their own data by following this sequence.

1. Decide which procedure is appropriate.
2. Open the correct SimFIT program.
3. Note that a default data set is provided to help you get started.
4. Examine the test data provided to appreciate the format required.
5. Analyze the file provided to observe what happens with correctly formatted data.
6. Prepare the experimental data in the correct format.
7. Submit the data for analysis.
8. Archive any results for retrospective use.

All SimFIT procedures are explained in the reference manual, but this was written with an emphasis on mathematical and statistical details that many users find daunting. For this reason there is a set of tutorials available as individual pdf documents (or in collected form as `w_examples.pdf`) to demonstrate how to use SimFIT as follows.

- Each tutorial presumes users have read this document and are aware how to select a procedure and use the test file provided.
- For every procedure the description avoids technical details as far as possible and merely presents the examples illustrated in the reference manual, but in a more user-friendly form.
- Users wanting more information would then be expected to consult the document `ms_office.pdf` and the SimFIT reference manual `w_manual.pdf`.

2 Additional documentation

These documents are available from the SimFIT website.

- **w_manual.pdf**
  The SimFIT reference manual.
- **w_examples.pdf**
  Collected tutorials with hyperlinks and an index.
- **ms_office.pdf**
  Explains how to transfer data from an Excel spreadsheet into SimFIT.

- **install.pdf**
  Describes how to to install SimFIT.
- **configure.pdf**
  Discusses the configuration options.
- **speedup.pdf**
  Shows advanced users how to switch off the numerous first-time user advisory messages which rapidly can become tedious.
- **pscodes.pdf**
  Informs users who have the Ghostscript package how to access additional graphical procedures such as placing one graph inside another or making collages.
- **source.pdf**
  Summarizes how to compile the SimFIT package from source code.

3 Choosing a procedure

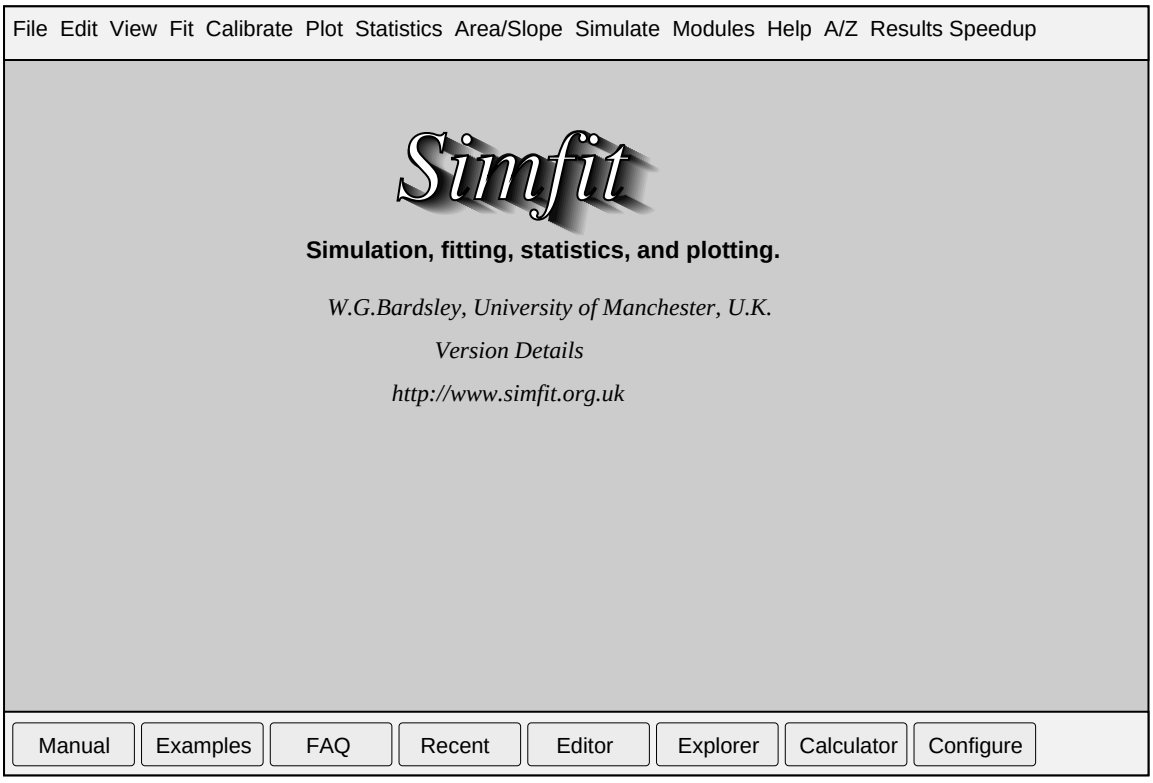


Figure 1: The main SimFIT menu

Figure 1 will be called the main SimFIT menu, and from it there are two distinct ways to choose a procedure, e.g. fitting exponential functions.

1. **Choosing the [Fit] options.**
   This provides options for various fitting procedures and you would see that program **exfit** is suggested for fitting simple exponentials, while program **qnfit** is available for more advanced curve fitting.

2. **Choosing the [A/Z] option.**
   This provides an alphabetical list of all the programs available in SIMFIT.

Once the [Fit] or any other menu item has been selected and the name for the appropriate program has become familiar, the [A/Z] list might subsequently be easier to use. There is one exception to this. Program **simstat** can only be selected in comprehensive mode from the [A/Z] menu, but the [Statistics] option enables subsections of **simstat** to be activated. If you only wanted just one statistical procedure, such as ANOVA, then it probably would be best to use the [ANOVA] option from the [Statistics] menu. However, to move backwards and forwards between the options within program **simstat**, it would be advantageous to use the [A/Z] technique.

4 Choosing test data

There are four possibilities.

1. **A data matrix.**
   A simple rectangular table of information is required. Here the only consideration is the meaning of the rows and columns in such a data matrix.

2. **A data matrix with extra information.**
   Sometimes it is advantageous to add additional data to the data, such as supplying parameter limits and starting estimates for constrained nonlinear regression.

3. **Several data matrices.**
   Often several tables must be supplied, such as sets of coordinates for fitting.

4. **Interactive input.**
   Occasionally users are required to input numerical values interactively, such as when requiring power as a function of sample size.

5 Alternative ways to input data

- **From a data file**
  This is by far the most versatile and useful technique because, once the data have been formatted this way, it is easy to perform repeat analysis. Data files can be generated and edited using the dedicated SIMFIT programs **makmat** and **editmt** for arbitrary data matrices, **makfil** and **editfl** for curve fitting data, or even a simple text editor like **notepad**, once you understand the format required.

Note that the SimFIT distribution also has a MS Excel macro called `simfit6.xls` for comprehensive data file preparation which will be found in the `\doc` folder of the SimFIT installation.

- **From the clipboard**
  This just requires that a rectangular matrix of data values be pasted in when data input is requested. The data must have no missing values, and every row must have the same number of columns. Note that, once such a data set has been copied to the clipboard, it can easily be written to a SimFIT data file using program **maksim**.

- **From a spread sheet**
  Data can be saved from spreadsheets in space separated, comma separated, HTML, or XML format, and SimFIT provides macros to do additional procedures such as adding column and row labels, or estimating missing values.

- **Typing in from the console**
  This can sometimes be the best way as, for instance, with simple calculations of power as function of sample size, or performing chi-square tests on small contingency tables. Note that this is only available when the speedup procedure concerned is switched on from the [Speedup] option on the main SimFIT menu.

6 Importing graphs into documents

The best way to archive SimFIT graphics for retrospective use such as including in documents depends on the quality and versatility users require. First note that, for complete future-proof archived graphics files, the industry standards are now .png for compressed bitmap files, and .svg for vector graphics files. For more details consult `ms_office.pdf`.

6.1 PostScript users

This category automatically includes Linux users, but also Mac or Windows users who are prepared to download and install Ghostscript and a PS viewer such as GSview. You should archive graphs using the [PS] graphics option followed by the [File] option, then save as a file with the .eps file extension. The advantages of doing this are these.

1. SimFIT .eps files are true vector graphics files with no embedded bitmaps, so they are very compact but can be printed at any magnification with no loss of resolution.

2. SimFIT .eps files have the unique feature that they can be edited in any text editor to change colors, plotting symbols, or line-types, and allow the addition of extra text, etc. This is explained in the SimFIT reference manual `w_manual.pdf` and also in the special document `pscodes.pdf`.

3. SimFIT .eps files can be edited and manipulated to generate arbitrary composite graphs or ordered collages by the program **editps**.

4. When it is required to import graphs into documents, the .eps file should be used to generate .png files using procedures available from the SIMFIT main menu or from the program **editps**, as .png files can be imported into all document preparation programs.

6.2 Non-PostScript users

For Windows users who are not prepared to install Ghostscript then graphs can be saved as enhanced Windows metafiles or .png files.

6.3 Internet graphics

Scalable vector graphics files (.svg) should be saved directly from the [Win] graphics option as these are true vector graphics files with no embedded bitmaps, so they are very compact but can be printed at any magnification with no loss of resolution. Such .svg files generated from .eps files using Ghostscript may not be true vector graphics files, as this will depend on the particular SVG generating engine distributed with your current Ghostscript installation.

7 Importing results files into documents

Every time SIMFIT performs an analysis, the results are written to a results log file for retrospective use. Tables can then be extracted from these for importing into documents in either tabbed-text, html, xml, or LaTeX when required.

7.1 Archiving results log files

SIMFIT maintains a current default archive called `f$result.txt` to store results from the program in use, plus up to 100 previous default archives named as follows.

```
f$result.txt
f$result.001
f$result.002
f$result.003
...
f$result.100
```

Each time a program is selected for analysis the list is rolled as follows. `f$result.100` is deleted, `f$result.099` is renamed as `f$result.100`, `f$result.098` is renamed as as `f$result.099`, and so until `f$result.txt` is renamed as `f$result.001` and a new `f$result.txt` is created.

7.2 The format of results log files

These are in standard ASCII text format and have several features.

1. They can be read and edited by any text editor, such as the program **notepad** which is present on most computers.

2. They will only appear as correctly formatted tables if they are viewed and incorporated into documents using a fixed font such as Courier, or in a format consistent for tables, such as html.

3. In order to always indicate the number of significant figures and to make the orders of magnitude immediately obvious, floating point numbers are always displayed in scientific notation. This will now be explained.

   For non-scientists unfamiliar with using powers of ten and standardized scientific notation it should be indicated that 172000, i.e., 1.72×10^6, is written as `1.72E+06`, or as `1.72e6`, or similar in computer listings. For instance.

```
1.234E+00 = 1.234
1.234E-01 = 0.1234
1.234E-02 = 0.01234
1.234E-03 = 0.001234
1.234E+01 = 12.34
1.234E+02 = 123.4
1.234E+03 = 1234.0
```

7.3 Importing tables into word processors

From the [Results] option on the SIMFIT main menu the following possibilities are available.

1. Open an archived results file for viewing.

2. Open an archived results file for printing.

3. Open an archived results file to Save As....

4. Open an archived results file for editing.

5. Open an archived results file to extract tables.

6. Open a selected results file from the SIMFIT results folder.

It is intended that users would save archived results to a named file in the SIMFIT results folder or elsewhere if long term archiving is required. However, if it is wished to extract a particular results table for including into documents then the option to extract tables would be used.

This can output tables in the format required for your word processor with exponential notation replaced by numbers in standard format, and with other options like replacing decimal points by commas as in continental usage, or substituting Greek characters for words (which is not done for tabbed-text). A full account of the techniques SIMFIT provides to do this will be found in the tutorial document called

`extracting_tables_from_simfit_results_files.pdf`.

Consider, for instance, this item copied from a SIMFIT results log file

```
 1-Way Analysis of Variance:  1 (Grand Mean  2.989E+01)

Transformation:-  x (untransformed data)
Source               SSQ       NDOF     MSQ          F           p
Between Groups    2.020E+02       4   5.051E+01   3.931E+00   0.0111
Residual          3.855E+02      30   1.285E+01
Total             5.875E+02      34

Kruskal-Wallis Nonparametric One Way Analysis of Variance

Test statistic   NDOF     p
    1.054E+01       4   0.0323
```

which, after the simple editing just discussed, looks like the following.

1-Way Analysis of Variance: 1 (Grand Mean 29.89)

Transformation: x (untransformed data)

Source	SSQ	$NDOF$	MSQ	F	p
Between Groups	202.0	4	50.51	3.931	0.0111
Residual	385.5	30	12.85		
Total	587.5	34			

Kruskal-Wallis Nonparametric One Way Analysis of Variance

Test statistic	$NDOF$	p
10.54	4	0.0323

A few simple worked examples will be given to make things clearer. However, you can also consult `ms_office.pdf` for more details for first time users on SIMFIT data file formats, and how to use SIMFIT in conjunction with word processor and spread sheet programs like those provided by MS Office, LibreOffice, and OpenOffice.

8 Example: Cochran Q test

This example describes how to use a test file where the data are in the form of a simple table. The sequence of steps is as follows.

1. Choose [Statistics] from the main SIMFIT menu.
2. Choose Standard statistical tests from the [Statistics] menu.
3. Choose Cochran Q test (on 0/1 integer matrix) from the Statistical test menu.

This will lead to Figure 2, the SIMFIT test file selection control.

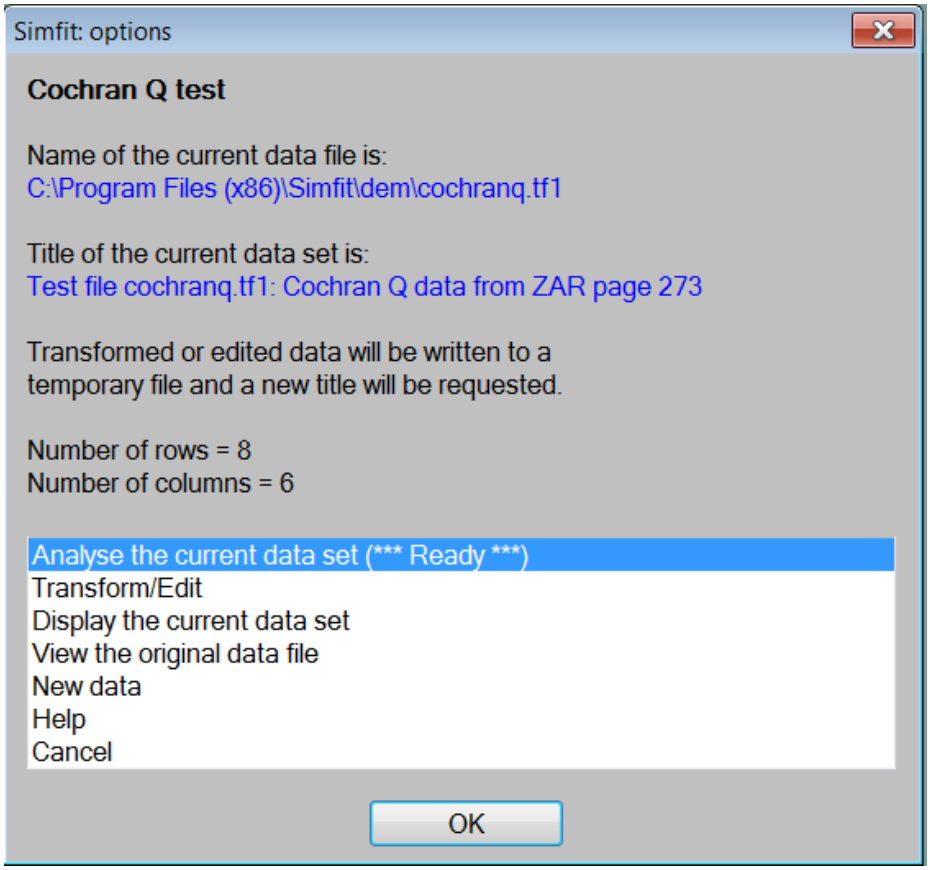


Figure 2: Example 1: Cochran Q test file selection

This gives the name of the test file, the title of the data set, the number of rows and the number of columns followed by a list box with these items.

1. **Analyze the current data set (***Ready***)**
   This informs you that a data set is ready to analyze

2. **Transform/Edit**
   This is for more advanced use when the effect of data transformation or editing is of interest

3. **Display the current data set**
   This displays the data as a simple numerical table

4. **View the original data file**
   This allows the test data file to be examined

5. **New data**
   This allows users to input their own data

6. **Help**
   This enlarges on the use of this test file selection control

7. **Cancel**
   This exits from the Cochran Q test procedure

The most useful option at this stage is to view the original data file which will display as follows.

```
Test file cochranq.tf1: Cochran Q test data
 8   6
 1   0   0   0   1   0
 2   1   1   1   1   1
 3   0   0   0   1   1
 4   1   1   0   1   0
 5   0   1   1   1   1
 6   0   1   0   0   1
 7   0   0   1   1   1
 8   0   0   1   1   0
10
Information about this data set.
Column 1 is just the observation number and can be omitted.
If Column 1 does contain observation numbers they must
be in increasing order as illustrated.
If column 1 represents data it must only contain the
values 0 or 1.
Columns 2 to 5: these are data values which must be 0 or 1
Any row of data consisting of all 0 or all 1 is not counted.
Therefore the above data set has 7 blocks of 5 observations.
cochranq.tf2 is the same data set without column 1 and row 2.
```

The file illustrated (`cochranq.tf1`) has the following sections

- **The header section**
  Line 1 is the data title, and line 2 records the number of rows and columns.

- **The data**
  There are 8 rows and 6 columns of data.

- **The trailer section**
  This consists an indicator noting that 10 rows of extra information have been appended, followed by the extra information.

At this stage it is well to point out that the only vital part of the above 8 by 6 table is actually the following 7 by 5 data matrix.

0	0	0	1	0
0	0	0	1	1
1	1	0	1	0
0	1	1	1	1
0	1	0	0	1
0	0	1	1	1
0	0	1	1	0

That is because the header and trailer sections are only advisory, column 1 (in this case) simply numbers the cases, and rows consisting of all 1 or all 0 are ignored for the analysis. So, to analyze your own data you would simply copy such a simple table from your spreadsheet to the clipboard and paste into SimFIT as described in detail in `ms_office.pdf`.

Each time SimFIT performs a data analysis procedure a copy of the results is written to an output results log file. SimFIT archives the ten most recent results log files which can be browsed, saved, or printed using the [Results] option from the main SimFIT menu. So it only remains to list the output of analysis written to the output log file, which would appear like this.

Results for Cochran Q test

Number of blocks (rows)	7	Rows suppressed: 1 (all 0 or all 1)
Number of groups (columns)	5	Columns suppressed: 1 (not data)
Cochran Q value	6.947	
$P(\chi^2 \geq Q)$	0.1387	
95% chi-square point	9.488	
99% chi-square point	13.28	

Note that, instead of merely listing a p value, SimFIT explicitly lists the appropriate upper, lower, or two-tail probability for the test statistic in question, along with corresponding percentage points, given the degrees of freedom, etc. Since in this case the significance level (i.e. $p = 0.1387$) exceeds 0.05 (i.e. the 5% point), there is no evidence to reject the null hypothesis that the binary response is the same for all subjects.

9 Example: K-means clustering

Copying data from your spreadsheet to the clipboard for pasting into SIMFIT is very limited as sometimes an analysis procedure, such as K means clustering, requires additional information to be appended to the data.

In such instances it is best to create a SIMFIT data file from your spreadsheet, then edit it in a text editor such as **notepad** to append the additional information using the commands

begin{extra data} ... end{extra data}

that will now be explained.

To illustrate this, consider a K means cluster analysis using the following steps.

1. Choose [Statistics] from the main SIMFIT menu
2. Choose [Multivariate statistics] from the Statistics menu
3. Choose [Clusters: K means] from the multivariate statistics menu
4. Choose [View the original data file] from the test file selection control

The test file `kmeans.tf1` displayed in full on the next page will be seen to have header, data, and trailer sections.

You should study this data file and note that, in addition to the data values, the trailer contains advice about the data set and the following important sections.

- begin{values} . . . end{values}
  This section lists the starting cluster centroids.
  There are 3 starting clusters (i.e. $K = 3$) each with 5 coordinates.

- begin{indicators} . . . end{indicators}
  This section indicates with a 1 which variables to include and with a 0 which variables to suppress.
  There are 5 values of 1 indicating that all variables are to be included.

- begin{labels} . . . end{labels}
  This section lists the row labels (cases) followed by the column labels (variables).
  There are 20 case labels and 5 variable labels.

Data for 5 variables on 20 soils (G03EFF, Kendall and Stuart)
```
 20       5
77.3     3.0     9.7     1.5     6.4
82.5     0.0     7.5     1.5     6.5
66.9     0.6    12.5     2.3     7.0
47.2     3.8    19.0     2.8     5.8
65.3     0.5    14.2     1.9     6.9
83.3     0.0     6.7     2.2     7.0
81.6     2.7     5.7     2.9     6.7
47.8     6.5    15.7     2.3     7.2
48.6     7.1    14.3     2.1     7.2
61.6     5.5    12.9     1.9     7.3
58.6     6.5    14.9     2.4     6.7
69.3     2.3     8.4     4.0     7.0
61.8     0.8     7.4     2.7     6.4
67.7     5.3     7.0     4.8     7.3
57.2     1.2    11.6     2.4     6.5
67.2     2.7    10.1     3.3     6.2
59.2     1.2     9.6     2.4     6.0
80.2     3.2     6.6     2.0     5.8
82.2     1.1     6.7     2.2     7.2
69.7     0.7     9.6     3.1     5.9
...
...
```
The next line defines the starting clusters for K = 3
```
begin{values} < −− token to flag start of appended values
82.5    10.0     7.5     1.5     6.5
47.8    36.5    15.7     2.3     7.2
67.2    22.7    10.1     3.3     6.2
end{values}
```

The next line defines the variables as 1 = include, 0 = suppress
```
begin{indicators} < −− token to flag start of indicators
  1       1       1       1       1
end{indicators}
```

The next line defines the row labels for plotting
```
begin{labels} < −− token to flag start of row and column labels
A
B
C
...
...
V4
V5
end{labels}
```

Proceeding with the analysis of data in test file `kmeans.tf1` from the starting clusters appended to the data file and using all of the variables we can assign the cases to clusters, then inspect various tables and plots of the assignments.

In particular, Figure 3 clearly illustrates in a plot of the principal component scores how the cases, indicated by the row labels appended to the data file, have been assigned to 3 clusters.

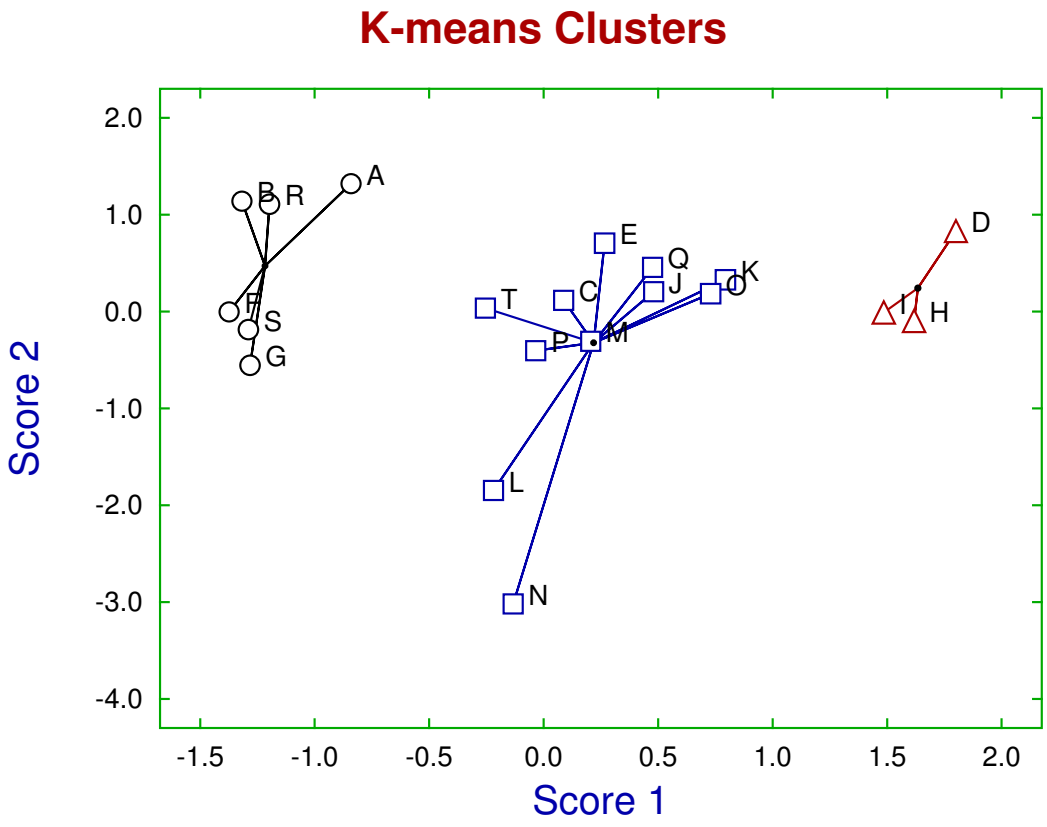


Figure 3: K means clusters

10 Example: 1-way ANOVA

It is not always possible to contain the whole of a 1-way ANOVA data set in a rectangular table, e.g. when there are several samples of different sizes. For instance, ANOVA can be performed following the next sequence.

1. Choose the [Statistics] option from the SIMFIT main menu.
2. Choose the [Analysis of variance] option from the Statistics menu.
3. Choose 1-way and Kruskal-Wallis nonparametric from the ANOVA menu.

This will lead to the ANOVA file selection control shown next.

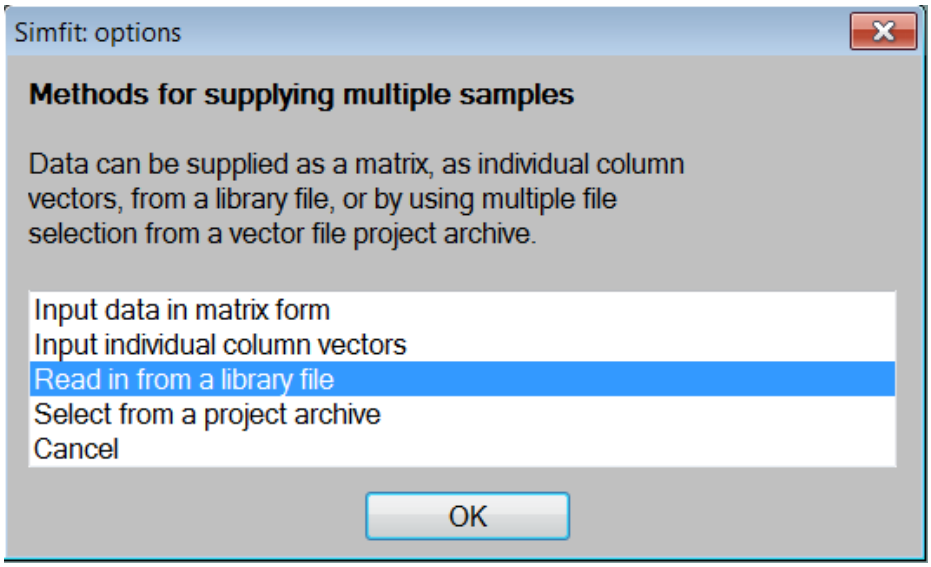


Figure 4: ANOVA file selection

These options can be used as follows.

- **Input data in matrix form**
  This is only possible if the sample sizes are all identical, so that the whole data set can be input as a matrix with rows as observations and columns as subjects.
- **Individual column vectors**
  This is only possible if the columns of observations are held in individual vector files, but this procedure is likely to prove inconvenient for routine use.
- **Read in from a library file**
  This has the great advantage that the filenames for all of the vector files for individual samples can be supplied as a single library file.

- **Select from a project archive**
  This technique is for advanced users who are familiar with the SIMFIT project archive technique.

- **Cancel**
  This option returns users to the ANOVA menu.

If the library file option is chosen followed by the [Demo] option to select `anova1.tfl` and this file is viewed it will be seen to be as below.

```
library file for 1-way ANOVA with column1.tf?
column1.tf1
column1.tf2
column1.tf3
column1.tf4
column1.tf5
```

Information about SIMFIT library files

1. Line 1 is an arbitrary title for the library file.
2. Lines 2 to $n + 1$ are names of n files grouped together by the library file for plotting, statistics, curve-fitting, etc.
3. The first blank line is taken to be the end of the library file and everything after the first blank line is ignored.
4. Library file are usually only valid if all n files specified do exist
5. However, library files analyzed by some SIMFIT programs (e.g. **qnfit** and **deqsol**) can have % to indicate a missing data set.
6. This is a SIMFIT test file so local names are given for the SIMFIT test files grouped together for analysis. Your own library files must have fully qualified file names i.e. path plus filename as in:

   ```
   C:\You\Documents\Simfit\usr\mydata.one
   C:\You\Documents\Simfit\usr\mydata.two
   C:\You\Documents\Simfit\usr\mydata.three
   ```

   and not just local unqualified filenames as in:

   ```
   mydata.one
   mydata.two
   mydata.three
   ```

Note that, given filenames for the individual vector files, the SIMFIT program **maklib** can be used to create a library file.

The data files referenced by the library file and results from the analysis are listed below.

```
C:\Program Files (x86)\Simfit\dem\column1.tf1
C:\Program Files (x86)\Simfit\dem\column1.tf2
C:\Program Files (x86)\Simfit\dem\column1.tf3
C:\Program Files (x86)\Simfit\dem\column1.tf4
C:\Program Files (x86)\Simfit\dem\column1.tf5
```

Results for 1-Way Analysis of Variance: Grand Mean 29.89

Transformation: x (untransformed data)

Source	SSQ	$NDOF$	MSQ	F	p
Between Groups	202.0	4	50.51	3.931	0.0111
Residual	385.5	30	12.85		
Total	587.5	34			

Kruskal-Wallis Nonparametric One Way Analysis of Variance

Test statistic	$NDOF$	p
10.54	4	0.0323

Tukey Q-test with 5 means and 10 comparisons

5% point = 4.142, 1% point = 5.034

Columns	Q	p	5%	1%	N_B	N_A	
3, 1	5.538	0.0042	*	*	6	5	
3, 5	2.922	0.2609	NS	NS	6	8	$N_B < N_A$
3, 4	[[2.557	0.3880]]	No-Test	No-Test	6	8	$N_B < N_A$
3, 2	[[2.191	0.5398]]	No-Test	No-Test	6	8	$N_B < N_A$
2, 1	3.806	0.0792	NS	NS	8	5	$N_B > N_A$
2, 5	[[0.7890	0.9801]]	No-Test	No-Test	8	8	
2, 4	[[0.3945	0.9987]]	No-Test	No-Test	8	8	
4, 1	[[3.460	0.1307]]	No-Test	No-Test	8	5	$N_B > N_A$
4, 5	[[0.3945	0.9987]]	No-Test	No-Test	8	8	
5, 1	[[3.114	0.2066]]	No-Test	No-Test	8	5	$N_B > N_A$

[5%] and/or [[1%]] No-Test results given for reference only

Note the following about this results log file.

- It lists the fully qualified path-filenames for the individual files specified by the library file.
- The Bonferroni correction should be considered if the p values from both parameteric and nonparametric ANOVA are to be consulted.
- It is clear from the Tukey post-ANOVA test that the reason for rejecting the null hypothesis of equal column means is the significant difference between columns 1 and 3.

11 Example: Power and sample size

There are several places in SIMFIT where it can be convenient to type in parameters directly to perform interactive analysis.

For instance, Figure 5 is the menu for calculations of power as a function of sample size obtained by the following sequence.

1. Choose [Statistics] from the main SIMFIT menu
2. Choose [Statistical calculations] from the statistics menu
3. Choose [Statistical power and sample size] from the statistical calculations menu

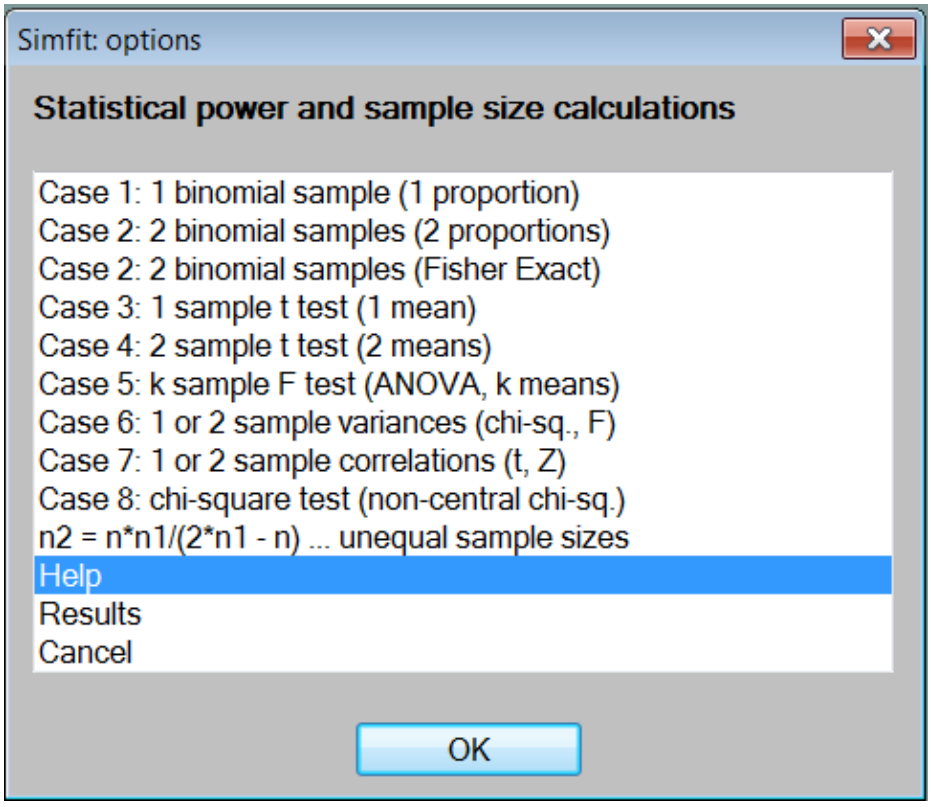


Figure 5: Power and sample size

A frequent use of this procedure is to examine the power, i.e. $100(1-\beta)\%$, as a function of sample size n, where two samples are assumed to come from normal distributions with the same variance but possibly different means, so that an unpaired t test is justified.

This corresponds to Case 4 (i.e. option 5) in Figure 5 which leads to the control illustrated in Figure 6.

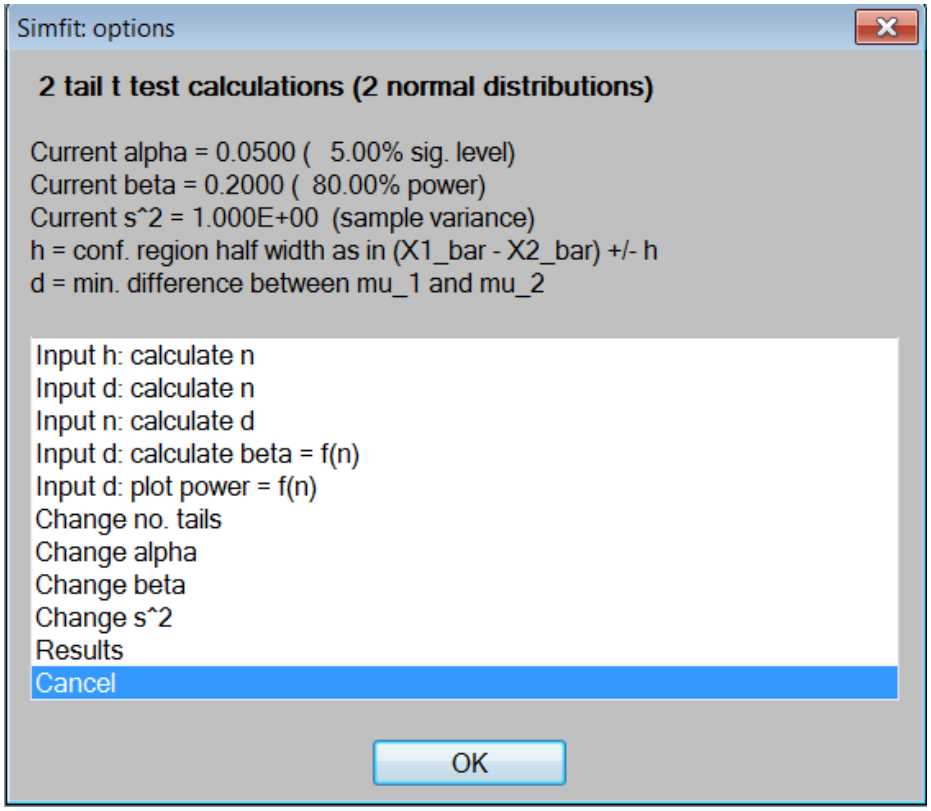


Figure 6: t test power

Selecting items one to four sequentially from the options displayed in Figure 6 then leads to the results below.

Results from power analysis for 2 normals (t test)

$h = 0.5$	$\alpha = 0.05$	$s^2 = 1$	$n = 32$	
$d = 0.5$	$\alpha = 0.05$	$\beta = 0.2$	$s^2 = 1$	$n = 64$
$n = 64$	$\alpha = 0.05$	$\beta = 0.2$	$s^2 = 1$	$d = 0.4991$
$n = 64$	$d = 0.5$	$\alpha = 0.05$	$s^2 = 1$	$\beta = 0.1986$