BIOMstat

for Windows

Statistical software for biologists

Version 3.3

User guide

F. James Rohlf
Dennis E. Slice

Department of Ecology and Evolution
State University of New York
Stony Brook, NY 11794

Exeter Software

47 Route 25A, Suite 2, Setauket, New York 11733

ISBN: 0-925031-29-1

Current printing: June 30, 1999.

Contents

† The methods are listed alphabetically within each group. Some methods belong to more that one group and thus are listed more than once. However, in the body of this manual all methods are simply listed alphabetically.

Introduction

BIOMstat performs many of the standard statistical computations needed in biological research. It includes modules for descriptive statistics, analysis of discrete and meristic data, analysis of variance and of covariance, regression and correlation analyses, and non-parametric analyses. BIOMstat is also designed to be easy to use and appropriate for use in the classroom. Analyses are selected from a hierarchical system of buttons within folders along the left side of the main window and the parameters for a particular analysis are specified by filling-in information and checking options in the form displayed to its right. The results of the computations appear in output windows whose contents can be examined, printed, copied and pasted to other Windows programs, or saved to a file. Icons are available on a toolbar for plots appropriate for the particular analysis just completed.

Version 3 of BIOMstat was developed to accompany the 3rd edition of the text *Biometry* (Sokal, R. R. and F. J. Rohlf, 1985. *Biometry*. W. H. Freeman and Co.: New York). An option can be selected so that the results of the computations are cross-referenced to specific pages in *Biometry* (relevant discussions in the text and Boxes showing the computational steps). However, the methods are general and this software can be used in conjunction with other text books.

Most of the statistical computations in this program are based on well-tested algorithms developed in the original FORTRAN versions of the program (BIOM, by FJR) first published as an appendix to the first edition of *Biometry* back in 1969. However, version 3 has major changes over the earlier versions. The previous suite of separate FORTRAN programs have been translated to Pascal and combined into a single program. A MS Windows-based user interface has been added which makes the program much easier to use than previous versions. In addition, new statistical methods have been added, taking into account some of the new methods added to the 3rd edition of *Biometry*. Version 3.2 added many types of plots appropriate for the analyses included in this program. There are, for example, histogram plots so one can see the general distribution of ones data, scatter plots to show relations between pairs of variables, and special plots appropriate for particular analyses. The help file has examples of all the types of plots. Version 3.3 added further refinements and it takes advantage of additional features that Windows 95/98/NT makes possible.

This version was written using Delphi 4.02 by F. James Rohlf and Dennis E. Slice. Dean Adams helped by testing the program and doing his best to find ways in which the program could be made to crash.

Once BIOMstat has been installed, users are encouraged to make use of the help file as it will be more up to date than the printed manual and because it includes additional information and examples not included in this user guide.

Installation

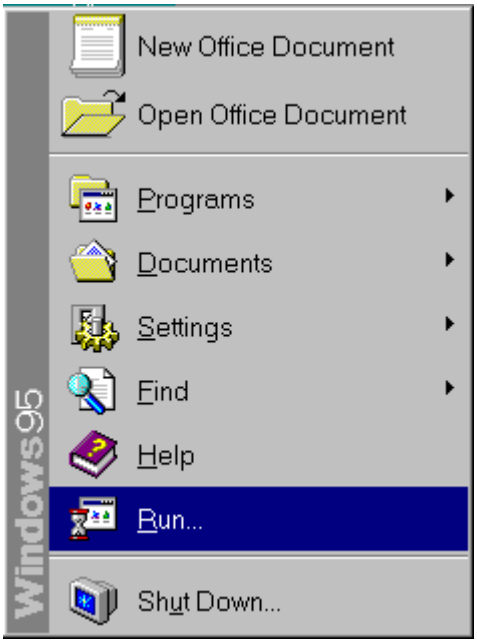
System requirements

1. A 486 or higher PC-compatible computer with MS Windows 95, 98, NT, or Windows 2000 installed.
2. A hard drive with at least 3 MB of available disk space.

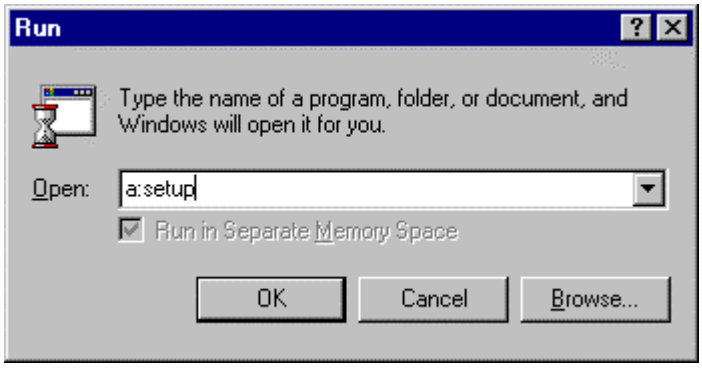
Installing BIOMstat

The following is a set of step by step instructions for installing the software. Minimal familiarity with using MS Windows is assumed.

1. Start Windows.
2. Insert the BIOMstat Install disk in drive `A:` or `B:` as appropriate.
3. Click on the "Start" menu and then chose "Run" from the pop-up menu as shown below.



In the Run window, The Command line box and then enter `a:setup` or `b:setup` as appropriate. Select OK to start the Setup program.



4. The Install program will display the user license information and the "readme" file that may have last-minute information and any corrections to this user guide. Click on the Next button to continue the installation process or the Back button to return to a previous screen.

5. The Install program will ask for the drive and directory where you would like to install BIOMstat. You may accept the suggestion or enter a new location. A new directory will be created if the directory does not exist already. Note: you can reuse the same directory as earlier copies of version 3 but it should not be installed in the same directory as version 2.1 or earlier versions. Click Next to the continue installation.

6. When the setup program finishes you should have a new folder on the Start menu called BIOM (unless you specified a different name) with icons for the BIOMstat and BIOMedit programs, their help files, and an Uninstall program (see below). The following files will be placed on your hard disk: `biomstat.exe`, `biomstat.ovr`, `biomstat.hlp`, `biomedit.exe`, `INETWH32.DLL`, `ROBO32.DLL`, `readme.txt`, and `Deisl1.isu`. Sample data files will be placed in the `BIOM\DATA` directory. In addition, the `uninst32.exe` will be placed in the windows directory, and the `bwcc.dll` file will be placed in the windows\system directory unless you already have a copy.

7. If you purchased the database version of BIOMstat then you need to install the Borland Database Engine, BDE, which will be included. That version of BIOMstat will not run unless the BDE is installed. The BDE Configuration Utility will be installed. It will allow you to define alias to facilitate access to your database files. The help file provides detailed information on how to do this.

When you run the BIOMstat program the first time you will be prompted to enter your name, institution, and the serial/registration number.

De-installation procedure

Should you have to remove this software from your computer you only need to double-click on the Uninstall icon in the BIOM folder on the startup menu. This will delete the files and the directory for BIOMstat. You do not need to delete BIOMstat in order to install a new version.

If you installed the Borland Database Engine you may also wish to delete it using the uninstall program.

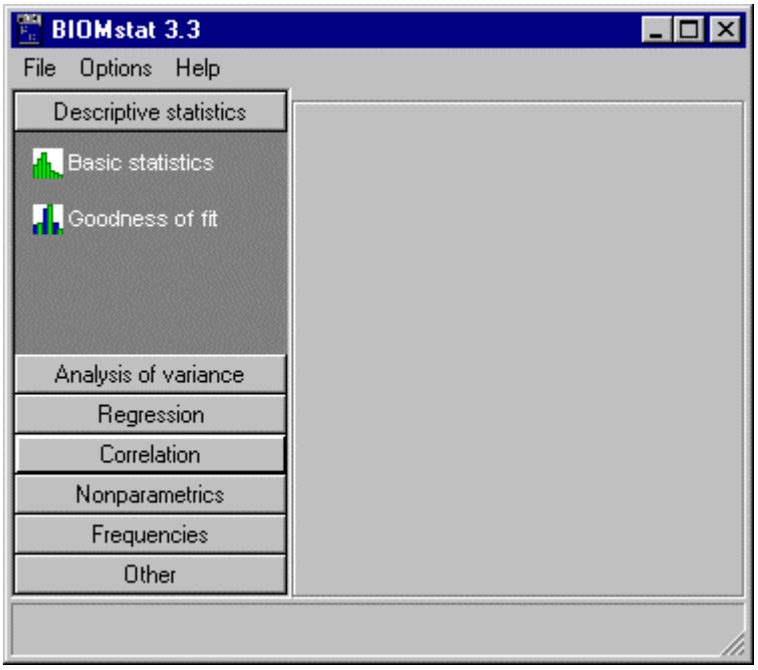
How to use BIOMstat

After installation, there should be a BIOMstat icon in the BIOM folder. Double-click on that icon to start the program. Before you can make use of the program to analyze your own data you will have to prepare data files in a format compatible with BIOMstat. Files can be either ASCII files or Excel files. With the database version of BIOMstat, dBase, Paradox, FoxPro, Access, and other standard database files can also be used. See the chapter on data file formats and the help file for more information.

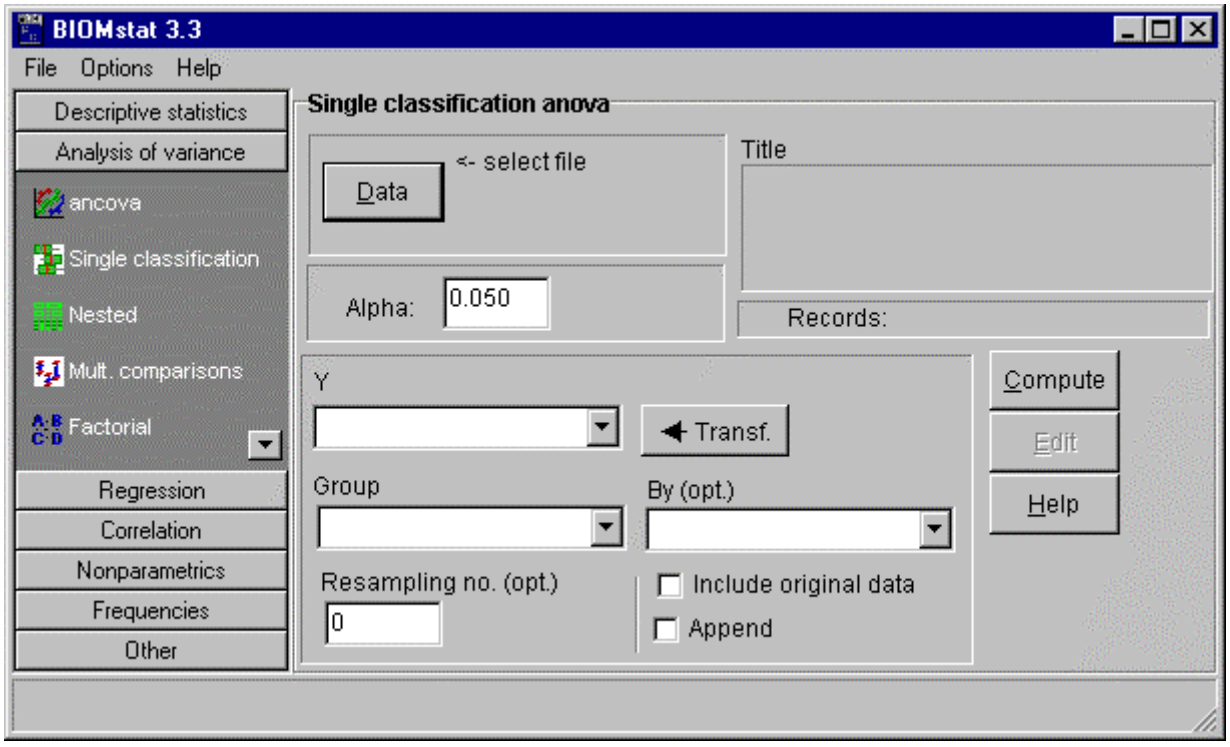
The next section gives a tutorial on how to use BIOMstat to carry out a single-classification analysis of variance using the sample data `singl.dta` (note: this `Courier font` will be used in this documentation for names of files, for data contained within them, and for information the user has to enter). Sample data files are placed in the `TESTDATA` subdirectory during installation and are also available as part of the help system. The Ariel font will be used to refer to information displayed on the screen by the program (such as names of fields where the user has to enter information).

A sample session

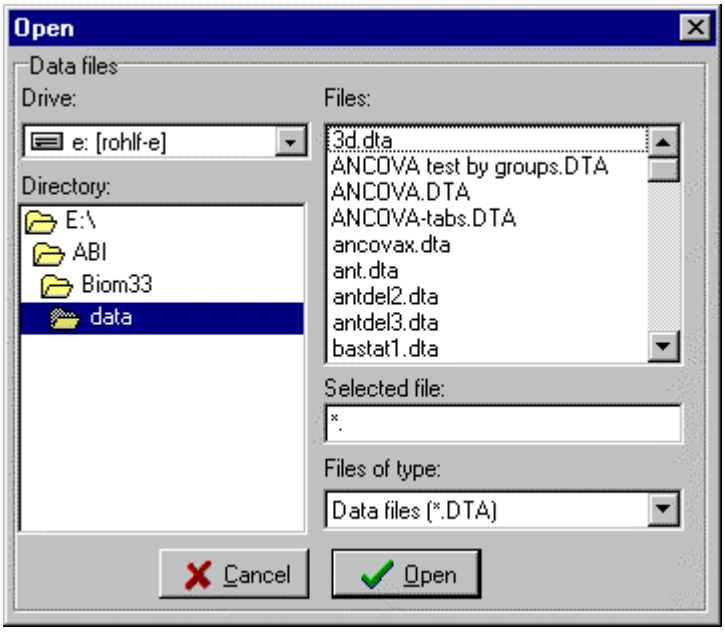
The initial screen consists of a vertical series of buttons grouped in folders along the left side of the window (see below). The first step is to find the folder corresponding to the desired type of method. For example, the "Single classification" button is located within the "Analysis of variance" folder. In the example shown below small icons are shown. They can be changed to larger icons in the Main Options dialog that can be opened from the Options menu.



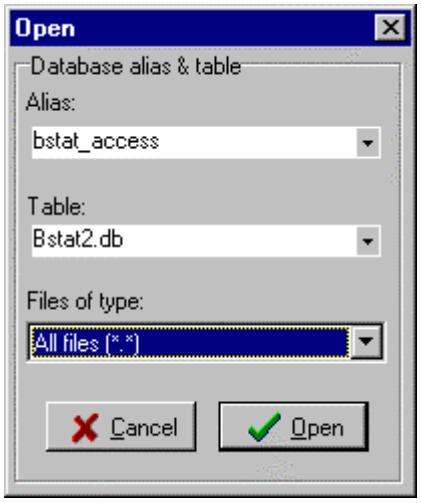
If you click on "Single classification" you would get the following display that lets you specify the data file and the particular options needed to carry out a single-classification analysis of variance.



If you have a data file ready for processing you should click on the "Data" button to load your data file by selecting it from the list in the file-open dialog box. An example of the file-open dialog box is shown below.

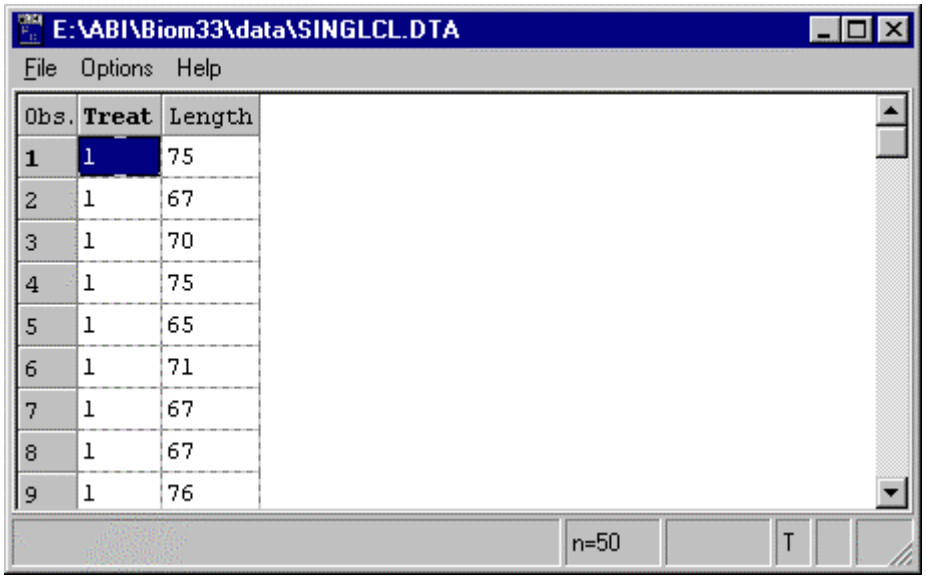


You can either type in the name of the file or use the controls shown above to click on a filename (perhaps after having changed to another directory). In the "Files of type" window you can also select "Excel files", "Database files", or "Database alias" (the latter two types are only available in the database version of BIOMstat). In the latter case the window is as follows:

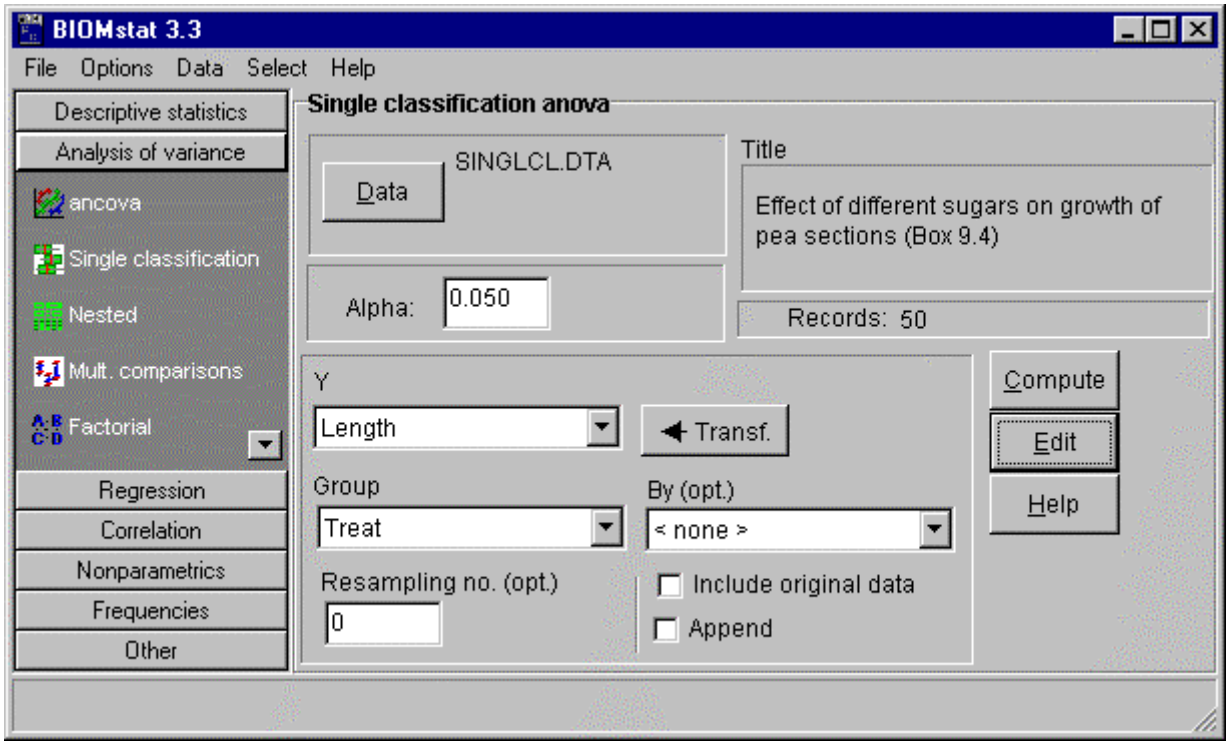


In this window you first select a previously defined database alias and then a data table within that database.

Alternatively, you can select "Edit" from the File menu to load the data editor that allows you to modify an existing file or to create a new one. Shown below is an edit session displaying the first part of the `singl.dta` file.



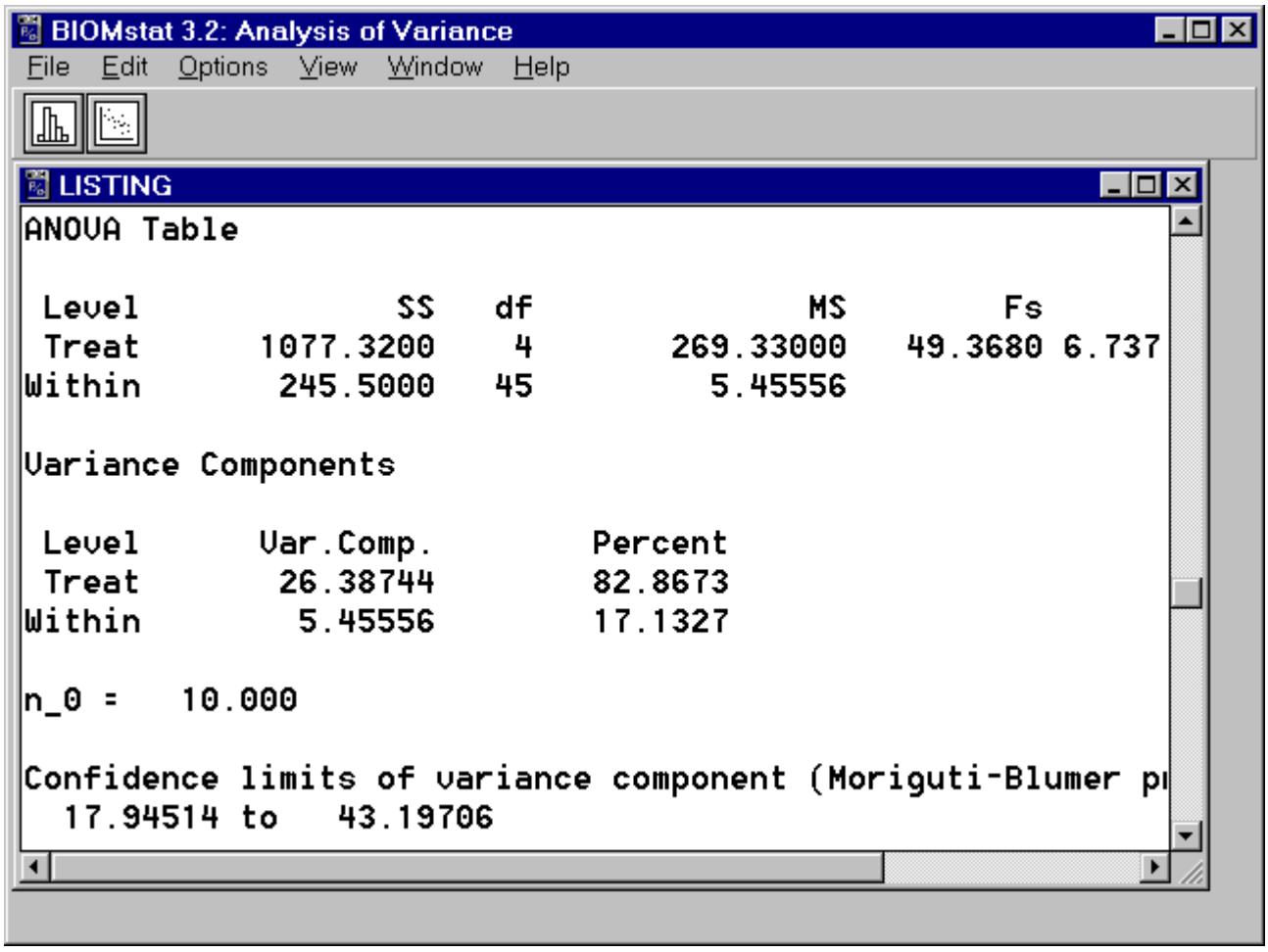
When the data file is complete, you can select the variable in your dataset that correspond to the dependent variable (Length in the `singl.dta` file) and the variable that groups the data into samples (Treat in this example). A completed dialog is show below. Note that the title and the number of data records in the dataset are displayed. Note that if you use the stand alone BIOMedit program then the file must first be saved to disk before it can be loaded into a BIOMstat module.



The "Include original data" checkbox is present in the dialog box for every analysis. If it is checked then the original input data will be included in the output listing (useful to make sure the program has interpreted your data

properly but may not be necessary if you are making many analyses with the same dataset). If the "Append" box is checked then the output of the next run will be appended to any previous output of this program. This can be useful if you are studying the effects of different options and you would like a single print-out of all the runs. The "By" field, present on most dialog boxes, is used to specify an input variable that divides the data into blocks that are to be analyzed separately. The output listings are not kept permanently so you should use the "File|Save as" item on the listing window's menu to save any you wish to keep (they will be saved as plain ASCII text files).

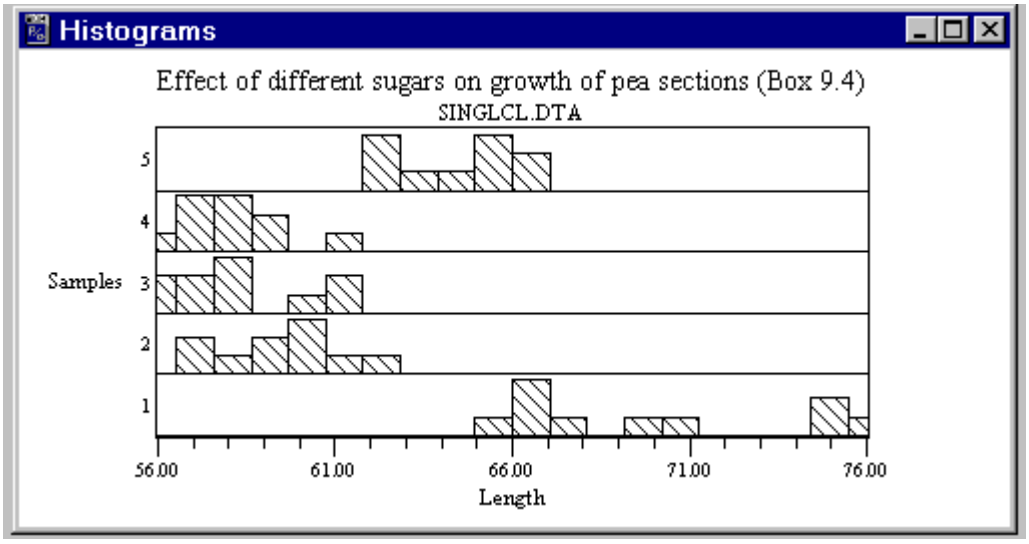
You can then select (e.g., click with the left mouse button) the "Compute" button to actually perform the computation. A new window will pop-up and display the listing file containing the results. An example is shown below in which part of the listing file is visible in the window.



Scroll the window up and down to study the results, save it to a file, print it, or copy part of it and paste it into some other Windows program -- such as a wordprocessor. To close the output window, you can either double-click on the upper left corner of the window, press the Alt-F4 keys or you can also click on

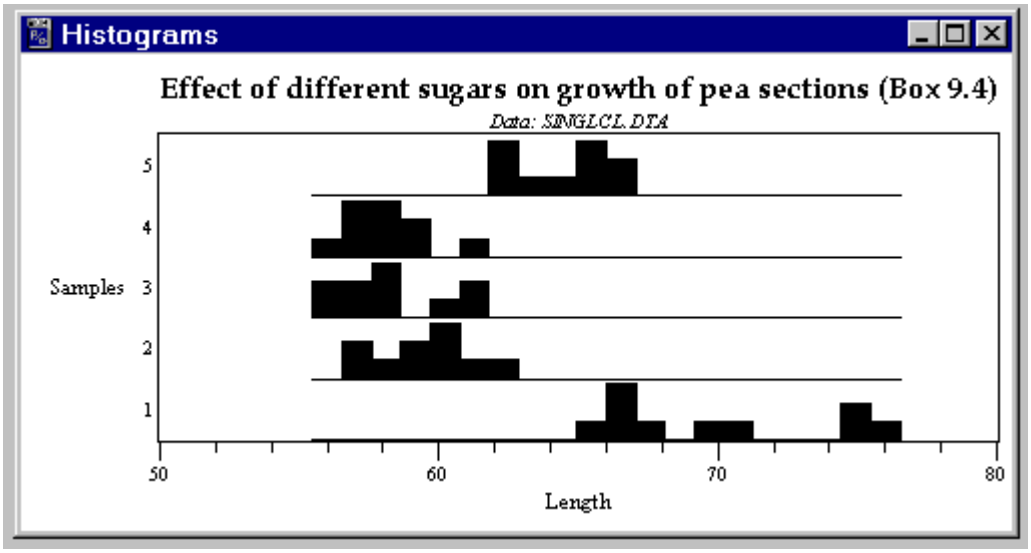
the ☒ in the upper right corner of the window). Selecting "File|Close" on the menu minimizes the listing window rather than closing it.

Finally, one can click on the graphics icons on the toolbar just below the menu. Clicking the histogram icon results in the following plot.



The Plot Options window for this plot can be opened from the Options menu of the output window or by right clicking on the plot. This window allows you to change the class marks, class interval, scales for the abscissa, and the properties (color, line thickness, and pattern) of the histogram bars. Examples of the various graphics options dialog boxes are shown in the Graphics section below (page 12).

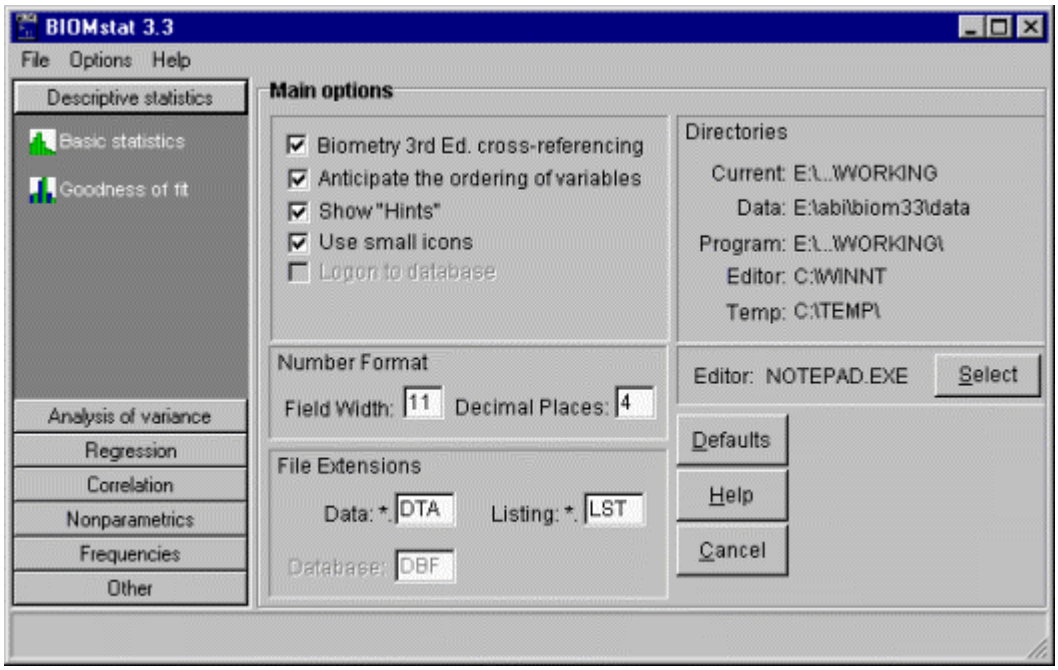
An example of such an "enhanced plot" is shown below.



Main Options

If you select "Options|Main Options" on the menu of the main window, a dialog (see below) will be displayed that lets you to set a number of options such as:

- Whether the output listing cite references to Biometry.
- Whether small or large icons will be used in the main window.
- Whether the dialog boxes should try to guess which variables go in which fields? They assume that grouping variables (if present) correspond to the initial variables and the dependent variable is listed last.
- Whether "fly-by" help hints are to be shown when the mouse cursor passes over a control in one of the dialog boxes.
- What 3 character extension is used for data, listing, and database files.
- The default field width and decimal places for listing data. The use of this field can force more precision in the output listings.
- The editor can be changed from the default Windows notepad editor. The "Select" button to change this to an editor of your choice. If you select a wordprocessor then make sure you save data files as plain ASCII text files.



Graphics

Most of the computational modules provide plots that help one visualize the results of an analysis or to detect departures from its assumptions. The set of available plots is indicated by the icons on the toolbar in the Results Window that contains the Listing Window. The right-most icons are always (except for

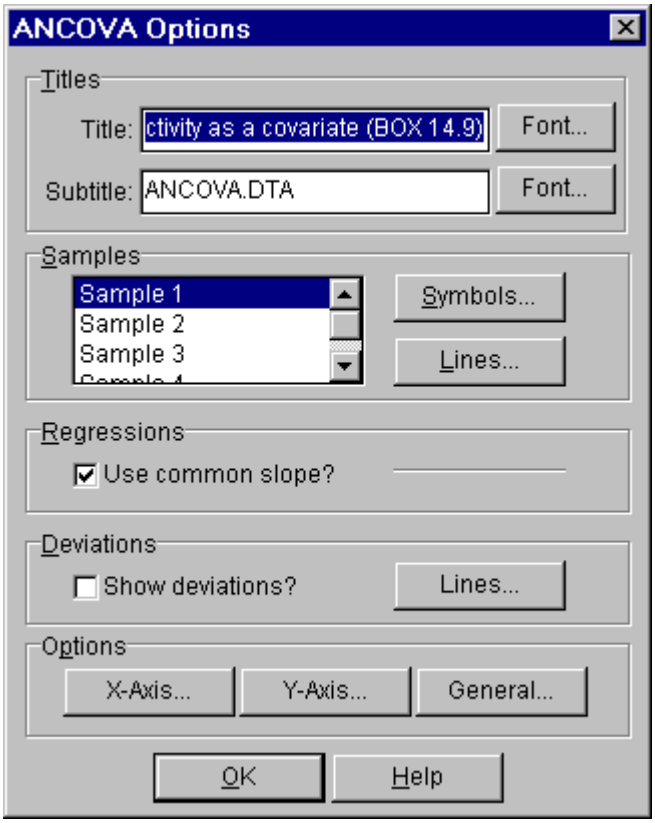
the Mantel module) for a simple 2D x,y scatterplot if the current data file contains at least two variables and a 3D x,y,z scatterplot if the current data file contains at least three variables. You can use these modules to plot any variable against any other. This is sometimes useful for detecting problems in the data. Place the mouse cursor over an icon to receive a "hint" about the type of plot it produces. Click on the desired icon to produce a plot. You can also select a plot from the View menu item.

The size and aspect ratio of a plot can be changed by dragging a corner of the plot or maximize its window for better viewing. You can also minimize the Listing Window to get it out of the way.

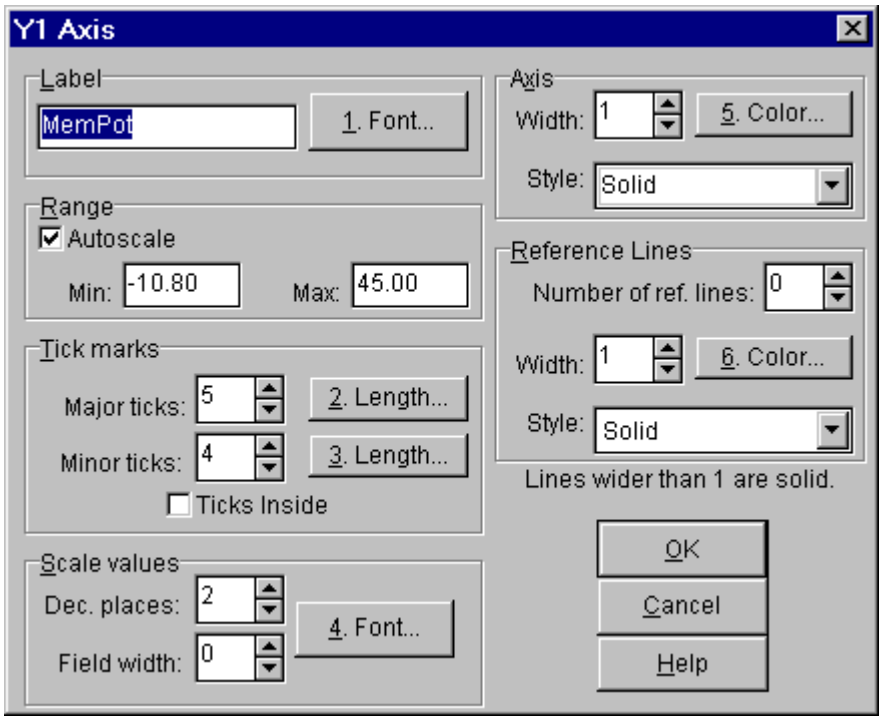
The File menu options allow you to preview or print a plot and setup your printer.

Select the "Options|Plot Options" menu item to display a dialog that allows you to set options specific to the current plot and also to set general plot options. The options specific to each plot allow you to customize the color and other attributes for the titles, subtitles, and any symbols, lines, and bars used in a plot. The axes buttons allow you to adjust the scale and appearance of each axis. Depending on the module, there may be check boxes that allow you to include or suppress the plotting of certain items. The General button brings up the general plotting options (see below). Press the "OK" button to accept the changes and close this window.

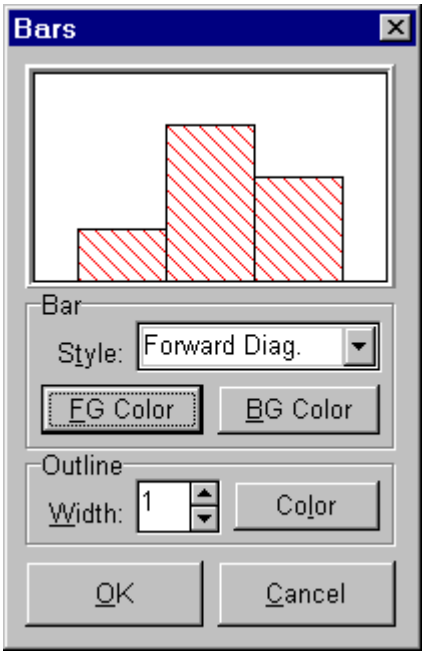
An example plot-options dialog is shown below for the ANCOVA module. In addition to the standard options, it allows one to change the attributes for the points and regression lines used for each sample, show the fit to a common slope or to separate slopes, and to show residuals as vertical deviations from points to the regression line for their sample.



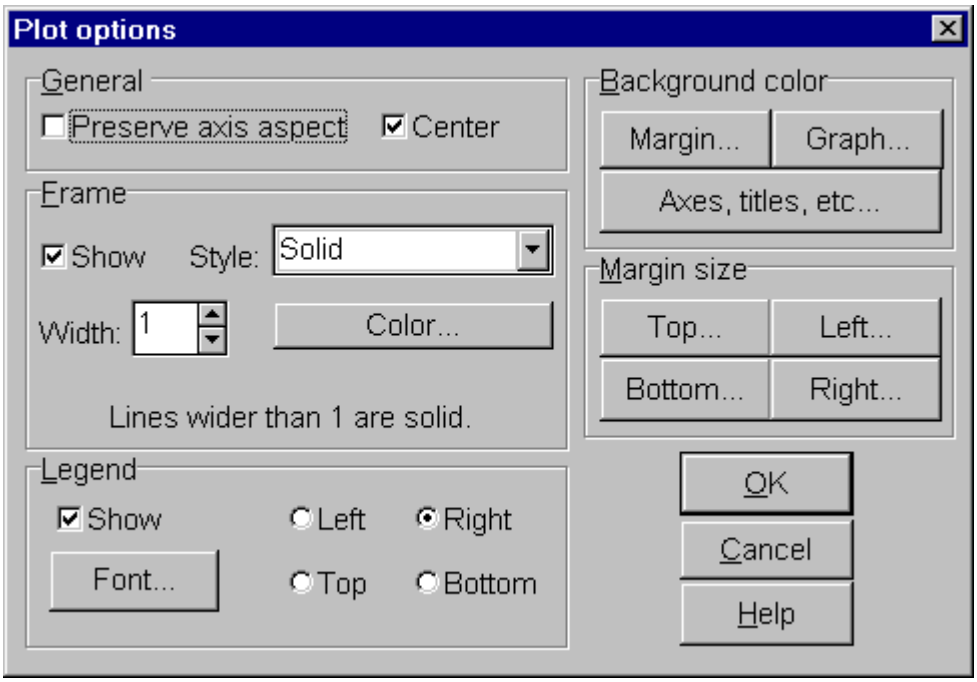
Pressing the Y-axis button results in the following dialog box that allows you to change many aspects of the Y-axis in the current plot.



For various plots, you can also change fonts, colors, plotting symbols, and histogram bars. These each have their own options window. For example, the dialog box for changing the attributes of the bars in the Single classification anova histograms is shown below.



Clicking on the General button on a plot options dialog brings up a dialog box where you can control the general layout of the plot. You can change the color and other attributes of the plot background, the frame around the plot, the margin outside a plot, and the location of the legend (if appropriate). An example is shown below.



Data file formats

To use your own data with the program you first have to prepare data files. Data are in the form of a table or matrix with columns corresponding to variables and rows to the observations. They may be stored as ASCII text files with special keywords (as described below), as Excel files (see below), or as database files (*e. g.*, dBase, Paradox, or Access table files). The variables consist of both the measurements being analyzed and also numerical codes indicating the grouping of observations into samples.

ASCII files

There are four kinds of input records. Each of them starts with a keyword and ends with a semicolon.

`title` Title or comment about the data file. It is used to label the output listing. This record is optional.

`vars` List of the names of the variables (columns) in the data table. This record is required and must come before the data since it defines the number of columns that the program should process. Qualitative variables (those using alphanumeric codes) are indicated by the presence of the "$" character in the variable name. This character can be changed in the options dialog window. No spaces are allowed within a variable name (but you can use the underscore "_" character).

`data` The actual data follow this keyword. The number of columns must match the number of variable names provided in the vars record. The number of rows corresponds to the total number of observations. The entries for each row must be separated by at least one blank, comma, or tab character. Note: the Mantel module is unique in that it uses the "`symmatrix`" keyword instead. It requires the data in the form of a lower ½ matrix including the diagonal. There is no predefined limit to the number of records in any one dataset.

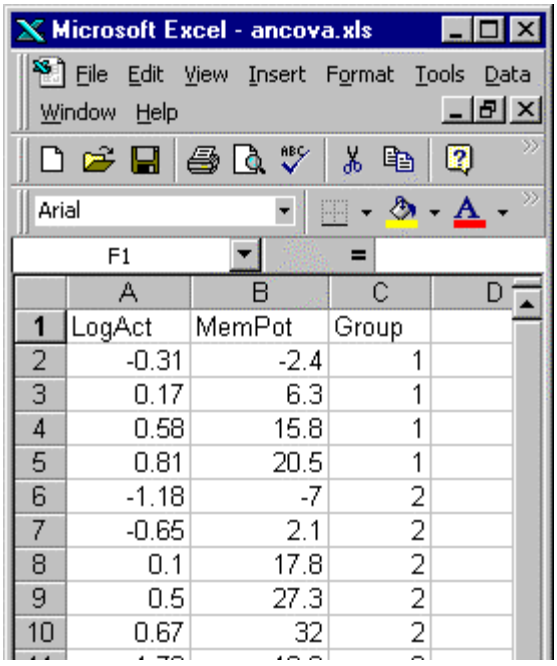
`end` This indicates the last line of the file to be processed. Information following this record will be ignored.

Excel files

The format for an Excel data file is quite simple – each variable corresponds to a column in the spreadsheet with the first value in each column taken as the name of the variable. However, the variable labels must not contain any blanks. As with the BIOMstat data files, qualitative variables are indicated by the presence of a "$" character within the variable name. There can be only one dataset in a spreadsheet.

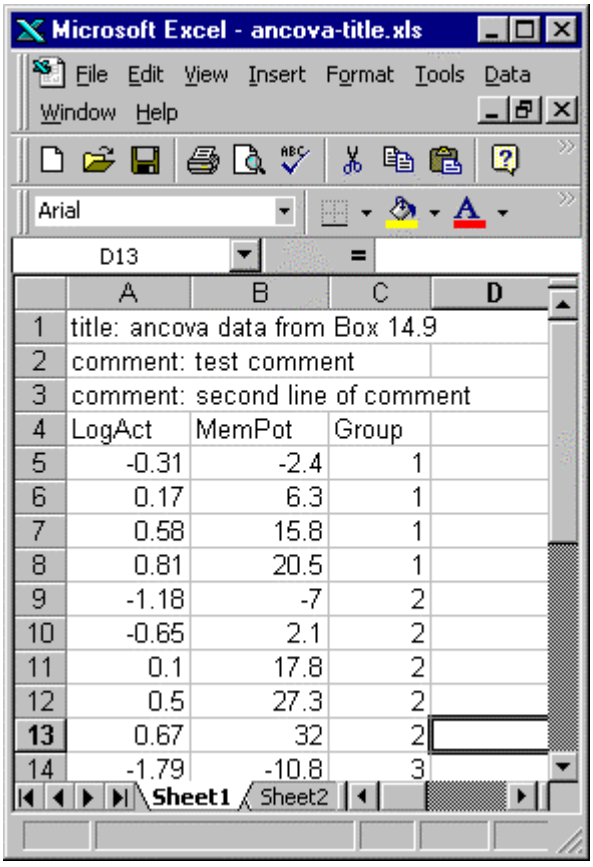
Optionally, the variables can be preceded by title and comment lines. The title cell (there can be only one) must start with 'title:'. For example, cell A1 could contain the following string: '`title: data from nutrition experiment`'. One or more comment cells can be provided. Each must contain 'comment:'. For example, cells A2 and A3 could contain the following: '`comment: variable 1 was measured in mm`' and '`comment: variable 2 is in degrees Celsius`'.

Note: BIOMstat can only read Excel files (values changed in BIOMedit but they *cannot* be saved back to an Excel file). The changed file can be saved as an ASCII data file. An example of a data file in Excel is shown below (there are no title or comment cells in this example).



	A	B	C	D
1	LogAct	MemPot	Group	
2	-0.31	-2.4	1	
3	0.17	6.3	1	
4	0.58	15.8	1	
5	0.81	20.5	1	
6	-1.18	-7	2	
7	-0.65	2.1	2	
8	0.1	17.8	2	
9	0.5	27.3	2	
10	0.67	32	2	

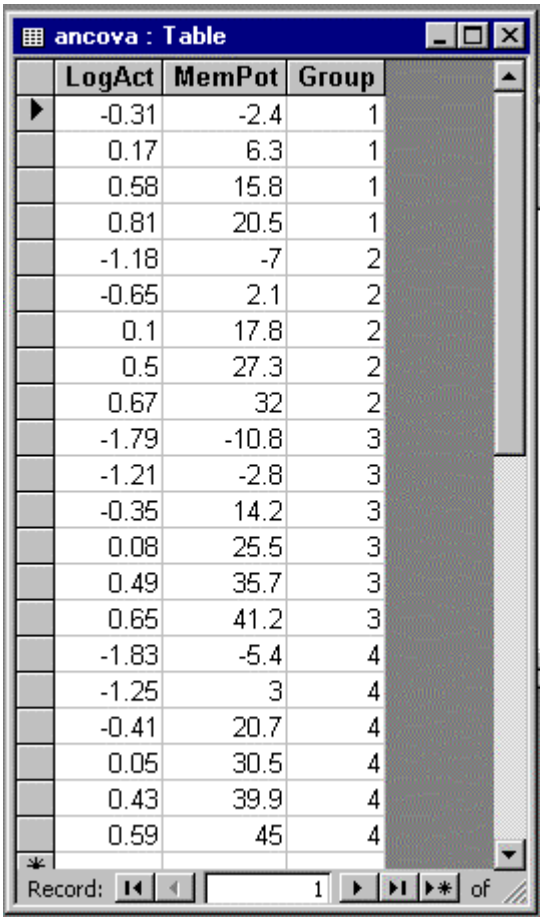
An example with both a title cell (A1) and two comment cells (A2 and A3) is shown below.



Database files

With the database version of BIOMstat, you can use supported database files such as dBase, Paradox, or Access table files. The database version of BIOMstat includes the "Borland Database Engine" which provides a common database interface to BIOMstat for all compatible database files. You must specify in the Options dialog box (see page 12) what file extension will be used to identify database files or else use the BDE Administrator program to define an alias for each database you plan to use.

In a database table file the information is already in the format needed by BIOMstat. The fields correspond to BIOMstat variables and the records correspond to observations. Variable names in database files must also contain a "$" for qualitative variables. An example of a database file is shown below. It is a MS Access table called "ancova" within a MS Access database file called "`BIOMstat.mdb`".



ancova : Table

	LogAct	MemPot	Group
▶	-0.31	-2.4	1
	0.17	6.3	1
	0.58	15.8	1
	0.81	20.5	1
	-1.18	-7	2
	-0.65	2.1	2
	0.1	17.8	2
	0.5	27.3	2
	0.67	32	2
	-1.79	-10.8	3
	-1.21	-2.8	3
	-0.35	14.2	3
	0.08	25.5	3
	0.49	35.7	3
	0.65	41.2	3
	-1.83	-5.4	4
	-1.25	3	4
	-0.41	20.7	4
	0.05	30.5	4
	0.43	39.9	4
	0.59	45	4
*			

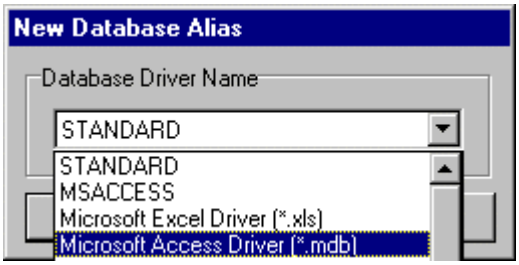
Record: 1 of

An error message will be displayed if entries are found that cannot be interpreted as valid values. The database files can be prepared using either database programs or other software (such as MS Excel) that can save files in one of the standard database formats.

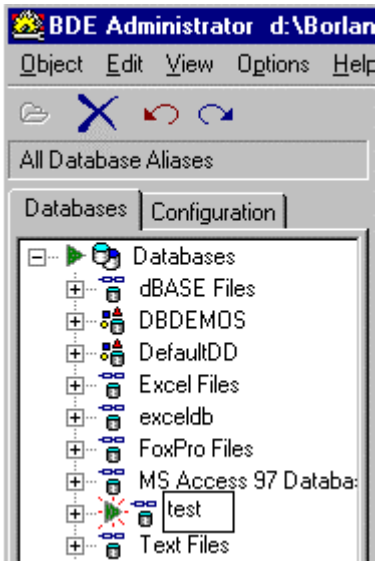
Creating an alias to access a database

A database alias is a shortcut name for a database. Once setup you no longer will have to specify the drive, directory path, and file name for a database. Its use is optional for database files such as dBase where a dbf file consists of a single database table. Its use is required for databases such as MS Access where more than one table can be stored in a single mde file.

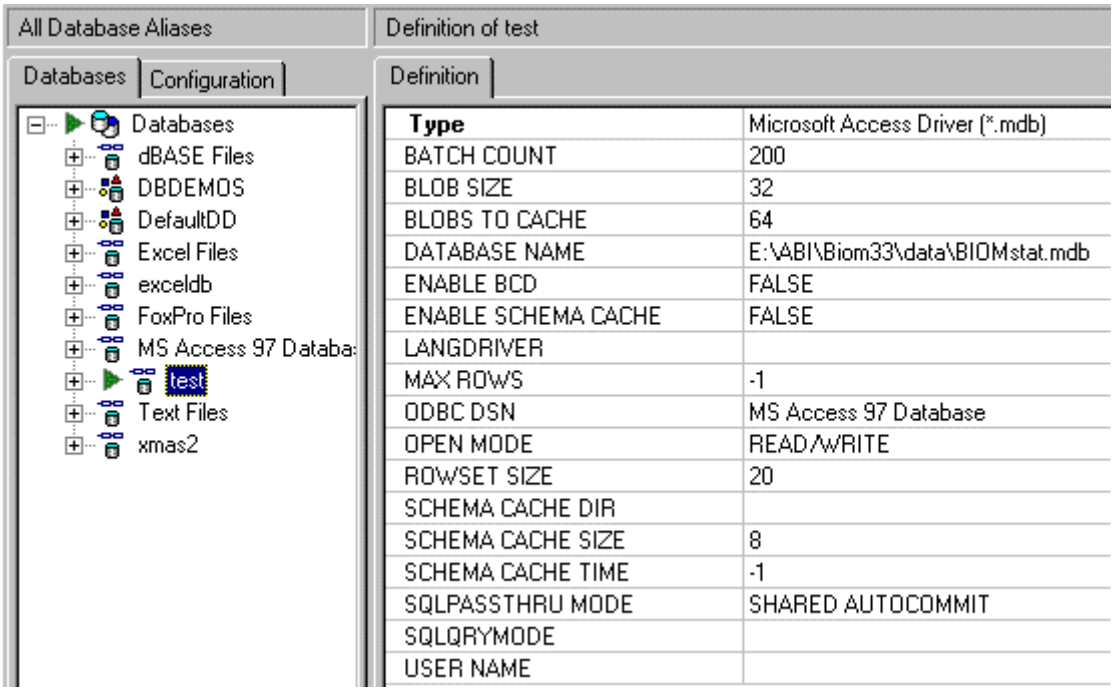
First, open the BDE Administrator program included with the database version of BIOMstat. Select the Databases tab in the left pane and then right click anywhere in the left pane. In the pop-up menu select “New” and the following window will open. Select the desired driver. For example, select “Microsoft Access Driver” to connect to a MS Access database.



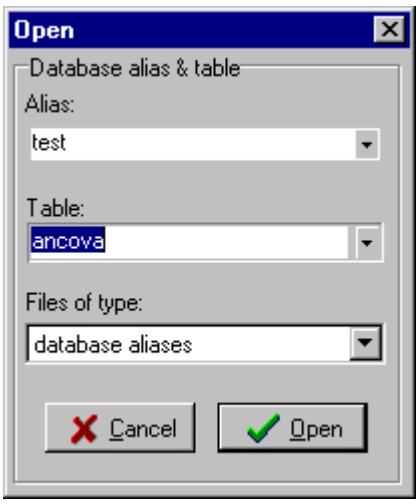
When you click the OK button there will be a new database entry. In the example below we have renamed it "test". This is the new alias name.



Next you must complete the definition of this new alias. Highlight the new entry and then enter the Database name (the full path to the database. The ODBC DSN field is set to reflect that we are using a MS Access database. A user name could be provided if the database on a fileserver. A completed example is given below.



To use this database within BIOMstat click the Data button and change the Files of type to "database aliases". In the Alias combo box select "test" and in the Table combo box select the "ancova" table and then click the Open button. A filled-in example is shown below. If a Database logon dialog is displayed by Windows, just click the OK button unless the database actually requires a userid and password.



Examples of ASCII data files

- Data file with a single variable (such as might be used with the Basic Statistics module). The data elements cannot be entered with several on the same line since each line corresponds to an observation.

```
title Aphid stem mother femur length data (BOX 2.1);
vars FemLen;
data
3.3
3.5
3.6
3.6
3.6
3.6
3.8
3.8
3.8
3.8
3.9
3.9
3.9
4.1
4.1
```

```
4.2
4.3
4.3
4.3
4.3
4.4
4.4
4.5
4.7
4.4
;
end;
```

- Data file for a 2×2 contingency table (such as used with the R×C or Fisher modules). The first two variables indicate the rows and columns and the third variable gives the cell frequencies.

```
title Ant invasion data (Box 17.7);
vars Species Invaded? f;
data
 1   1     2
 1   2    13
 2   1    10
 2   2     3
;
end;
```

- Example of a SYMMATRIX data file as used by the Mantel module. The matrix is a 10x10 lower ½ matrix (including diagonal).

```
title Genetic distances (Box 18.5);
vars 3G 8L-11P 8F 8ABC 11YZ 11D 11X 3RS 8O 15L;
SYMMATRIX
0
.02040  0
.04143  .03944  0
.01213  .02549  .05285  0
.01246  .01362  .04473  .01682  0
.02868  .03908  .04109  .02195  .04362  0
.03393  .03347  .03313  .04774  .04928  .04567  0
.03023  .02816  .08473  .02804  .02605  .05730  .05661  0
.03784  .03229  .03276  .02445  .02767  .02986  .04330  .04805  0
.05698  .06308  .07491  .03722  .03377  .06550  .10242  .05331  .03661  0
;
end;
```

There are many examples of ASCII data files in the help file. There are also sample listings and plots (with annotations) from every computational module.

Descriptions of methods

Introduction

Furnished below (listed alphabetically) are descriptions of the modules included in the BIOMstat program. Each section lists the purpose of the statistical method, its assumptions, how to setup the dialog box, and a description of the output listing. These sections are not intended to replace a good statistics text (such as *Biometry*) which would give much more detail about each method and how the results should be interpreted.

There are two standard check-boxes on all of the dialog boxes statistical analyses (i.e., except for the estimation of *n* and the statistical calculator).

- If "Include original data" is checked then a listing of the input data will be included in the output listing window. This is useful to make sure that BIOMstat has correctly interpreted your data file.
- If "Append" is checked then the output listing will be appended to the end of any previous listing for the current computational module.

There are also four standard buttons on the dialog boxes for each statistical analysis.

- Data, loads a data file into the program.
- Compute, the entries in the dialog box are checked and the corresponding module executed.
- Edit, the built-in edit program is executed using the currently selected data file. You may choose which edit program is called if an error is encountered in reading the file (by default the Windows Notepad program is used) in the Options dialog box (page 12).
- Help, the help file is loaded and set to display help for the current dialog box. You can also just push the F1 key. Note: the help is extensive. It includes examples of data sets for every statistical method.

Transformations

The following data transformations are available for use in most of the computational modules:

$\|Y\|$	absolute value of Y
$\sqrt{Y}$	square root of Y
$\sqrt{(Y+.5)}$	square root of (Y+½)
Y^2	square of Y
ln Y	natural log of Y
log Y	log to base 10 of Y
log (Y+1)	log to base 10 of (Y+1)
$\sin^{-1}(\sqrt{Y})$	arc sine of square root of Y (assumed to be a proportion)
$\sin^{-1}(\sqrt{\%})$	arc sine of square root of $Y/100$ (Y assumed to be a percentage)
$1/Y$	reciprocal of Y
Y^p	Y to the power p, p furnished in the power field of the transformation dialog box.
$(Y^p-1)/p$	The Box-Cox transformation. The optimum value of p is estimated and then the power transformation used. Only the Basic Statistics (page 26) and the Homogeneity of Variances (page 33) modules allow the use of the Box-Cox transformation.

Analysis of covariance

Purpose

The analysis of covariance, ancova, is used to test whether sample means differ after taking into account covariation with another variable (called a covariate). Plots are available that show the dependent variable as a function of the independent variable within each group (using either common or separate slopes) and that show the residuals from fitting the regression lines.

The linear model for ancova is:

$$Y_{ij} = \mu + \alpha_i + \beta_{within}(X_{ij} - \overline{X}_i) + \varepsilon_{ij}$$

Where $\overline{X}_i$ is the mean of the independent variable for the ith sample, X_{ij} the jth observation in the ith sample for the independent variable, Y_{ij} the jth observation in the ith sample for the dependent variable, μ the population mean for the dependent variable, α_i the y-intercept for the ith sample, β_{within} the slope of the regression line (assumed the same in all samples), and ε_{ij} the random error for the jth observation in the ith sample

Assumptions

In addition to the usual anova assumption that the residual error is independent and normally distributed with the same variance in all samples, it is also assumed that the slope of the regression line is the same in all samples. This module provides a test for homogeneity of slopes. The plots can be used to verify the reasonableness of these assumptions for a set of data.

How to use

First use the Data button to specify a data file. Since this module accepts only raw data (not summary statistics such as the mean and variance), use the Variable and Covariate fields to select the corresponding variables in the data file. Use the Samples field to select the variable in the data file that indicates the sample to which each observation belongs. Set the Alpha field to the desired probability level to be used in setting (1-α)100% confidence limits. If a value greater than 0 is entered in the "Resampling no." field then bootstrap estimates of the common slope will be computed. Optionally, one can also select a By variable to split the data file into a series of separate analyses. Note: while the observations within an analysis may be in any order, those in different By-groups must be contiguous in the input file.

Once the computations are performed one should view the plot of the dependent variable as a function of the covariate and also the plot showing the residuals from each group. This is important in order to detect many types of potential problems in the data.

Output listing

A table of the group means, adjusted means, and the standard error of the adjusted means is listed. Then, there is a table giving sums of squares and cross products and regression equations for each group separately as well as for within-groups, among-groups, and total variation. Next, there is an "`Analysis of covariance`" table giving the sums of squares explained, "`SSYhat`," and unexplained, "`SSY.X`," by regression both within and among-groups. Finally, there are "`Summary ancova tables`" which include tests for differences among the adjusted means and for heterogeneity in the slopes for the various groups (the results of this test should be checked first).

If the resampling number is greater than 0, a bootstrap estimate of the common slope and its standard error will be displayed.

See the help file for an example of an output listing (with annotations and plots).

Basic statistics

Purpose

Provides descriptive statistics for continuous data. Computes mean, minimum, maximum, range, standard deviation, variance, median, g_1 (skewness), g_2 (kurtosis), and their confidence limits. Also performs a Kolmogrov-Smirnov test for goodness of fit to a normal distribution using the estimated mean and variance. The Box-Cox transformation can be used to find the power transformation that results in the best fit to the Normal distribution.

Assumptions

These methods are based on the assumption that the variables are continuous, independent, and normally distributed (or can be transformed to approximate normality).

How to use

First use the Data button to specify a data file. The module accepts either raw data in the form of a single column of values or a frequency distribution with the Variable field indicating the class marks and the Frequencies field indicating the frequencies for each class. Set the Alpha field to the probability level to be used in setting (1-α)100% confidence limits. If the Resampling no. field contains a value greater than 0 then bootstrap estimates of the mean and coefficient of variation will be computed. Plots are available giving a histogram of the data and a normal quantile plot (deviations from a straight line imply a lack of normality).

Output listing

The output consists of a table of the various statistics computed—mean, median, variance, standard deviation, coefficient of variation, g_1, and g_2. These are followed by their standard errors and confidence intervals where applicable. The results of the Komolgorov-Smirnov test for goodness of fit to the normal distribution using the estimated mean and variance is given at the end. Both D_{max} and Khamis' δ-corrected statistics are given (see Section 17.2 of *Biometry*). Histogram and normal quantile plots are available.

If the resampling number is greater than 0, bootstrap estimates and their standard errors will be given for the mean and the coefficient of variation.

See the help file for an example of an output listing (with annotations and plots).

Correlation

Purpose

This module computes the Pearson product-moment correlation coefficient, *r*, for a pair of variables. In addition, the module computes the means, variances, standard deviations, standard errors, and covariance for the variables, and confidence limits to *r*.

The equation for the principal and minor axes of an equal-frequency ellipse are given. The confidence limits for the slope of the principal axis are computed and coordinates of points are given for hand plotting of confidence ellipses for the bivariate mean. The slope of the principal axis corresponds to the slope of the major axis regression line (a technique for Model II regression). The slope and its confidence limits are also given for the reduced major axis regression line. X,y-scatterplots of the data with a confidence ellipsis and principal axes can be viewed. Optionally, a random permutation test can be carried out for the correlation coefficient and for the slope of the major axis.

Assumptions

The computations are based on the assumption that the pair of variables follow a bivariate normal distribution (or can at least be transformed to approximate this distribution). The random permutation test does not require the assumption of a bivariate normal distribution.

How to use

First use the Data button to specify a data file. Then the Y1 and Y2 columns correspond to the two variables to be correlated. If the variables actually correspond to class marks in a 2-way frequency distribution, then also enter the variable giving the frequencies for each pair. Set the Alpha field to the probability level to be used in setting (1-α)100% confidence limits. To perform a random permutation (randomization) test, enter the number of replication (the observed data are considered to be the first replication). Be sure to view a plot of the data to make sure the apparent correlation is linear and not due to the presence of outliers.

Output listing

The means, variances, standard deviations, and standard errors for each of the variables, and the covariance between the two variables are displayed. Next, the correlation and its 100(1-α)% confidence limits are given along with the equations of the principal and minor axes and 100(1-α)% confidence limits to the slope of the principal axis. The coordinates for eight pairs of points are given to be used for plotting a confidence ellipse for the bivariate mean. These points correspond to points *a* through *h* shown in Box 15.5 of *Biometry*. The slope of the reduced major axis regression line and its confidence limits are given next.

Finally, the results of the optional random permutation tests are given. Note the complication in the test for the major axis -- slopes of zero and infinity correspond to no association and a slope of ±1 corresponds to the maximal association. Thus reciprocals are computed if a slope exceeds ±1.

See the help file for an example of an output listing (with annotations and plots).

Estimate sample size in anova

Purpose

This module is used to estimate the sample size needed in anova in order to achieve a desired level of statistical power. This module implements the method given in Box 9.14 of *Biometry*.

Assumptions

The computations assume comparisons between means in a single classification anova with *a* groups and an equal number of replicates in each group. The dependent variable is assumed to be normally distributed with the same variance in each group.

How to use

Fill out the values in the input parameter edit boxes. No. of groups is the number of groups, *a*, in the planned single classification anova. Type I error The desired type I error rate, α. Desired power is the power, $P = 1 - \beta$, that one would like to achieve in detecting a difference as small as δ at the α level of

significance. Note that only the ratio of the standard deviation to the smallest expected difference (the reciprocal of the effect size) is important—not their absolute values.

Output listing

The output consists of simply the estimated value for *n* and it is displayed in the result edit box.

See Section 9.8 of Biometry for a discussion of sample size and power in anova.

Factorial anova

Purpose

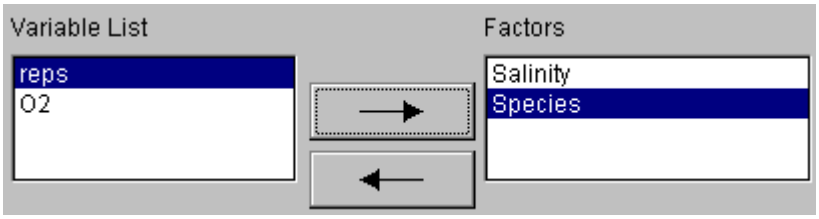
This module can be used for up to a nine-way analysis of variance with no replications or an eight-way analysis with equal replications. The module produces the standard anova table and, optionally, provides a table of deviations for all possible one-way, two-way, three-way, *etc.* tables. Note that there is a separate module for two-way analyses of variance (see page 56).

Assumptions

This analysis is based on the standard anova model which assumes homogeneous, independent, and normally distributed error within each cell of the data table. This module only handles completely balanced designs with equal numbers of replicates in each cell.

How to use

First use the Data button to specify a data file. Next The Y field corresponding to the dependent variable. Then in the Factors field select *k* variables corresponding to the *k* factors in a *k*-way anova. Up to nine variables may be selected. For example, for a 2-way anova select a variable containing the row number of each observation and another variable giving the column number for each observation. Click on the right arrow to move a variable from the left to the list of factors. Click on the left arrow to return a variable at the right to the list of available variables.



Then click the Compute button to perform the computations. The rows on the input data file need not be in any particular order. If more than one observation is given for a cell then it is considered to be a replicate (but note that this program *requires* an equal number of replicates in each cell). If the Residuals field is checked then a table of residuals will be added to the output listing.

Output listing

If requested (i.e., the print residuals box is checked), a table of the deviations of the means are included in the output. There are two types of identifications. First, the level numbers for each of the factors are given. This is followed by a column labeled "`ID NO.`" which correspond to various sources of variation in the anova table presented below. All the deviations with the same identifying number are used to compute the sums of squares for a particular source of variation with the same identifying number in the anova table. This table can be inspected for evidence of heterogeneity within the interaction terms or to indicate which means are most deviant.

In the anova table the various sources of variation, their sums of squares, degrees of freedom, and mean squares are given. *F*-tests are not provided since the appropriate tests depend upon which combinations of factors correspond to fixed treatment effects (Model I) versus random effects (Model II).

See Section 12.5 of *Biometry* for a discussion of how the computations for a variety of completely balanced anova designs can be based on a complete factorial analysis.

See the help file for an example of an output listing (with annotations).

Fisher's exact test

Purpose

This module performs Fisher's exact test for independence in a 2×2 contingency table.

Assumptions

Since the hypergeometric distribution is used, this method is "exact" only for Model III data (*i.e.*, for the rather unusual situation in which both the row and column totals are fixed by the design of the experiment.

How to use

First use the Data button to specify a data file. Next select the Row and Column variables for the rows and the columns of the 2-way table. Then the Frequency variable indicating the cell frequencies must be entered and the Compute button clicked to perform the computations.

Output listing

The probability of the observed frequencies is computed. If the determinate of the 2×2 table is zero then the message "`Exact fit, determinant = 0.0`" will be displayed and no computations performed since the sample data show perfect independence and *all* other combinations of frequencies would be "worse" cases. In other cases, tables of successively worse cases (combinations of frequencies with lower probability assuming the two factors are independent) are computed and their probabilities summed. The probabilities are summarized at the end of the listing.

See the help file for an example of an output listing (with annotations).

Goodness of fit

Purpose

This module performs a variety of goodness of fit tests for frequency data. For example, it can fit an observed frequency distribution to a binomial or Poisson distribution (using specified parameters or parameters estimated from the data) or to an arbitrary set of expected frequencies provided as a variable in the dataset. A *G*-test for goodness of fit is carried out. Adjusted G-values using Williams' correction are also furnished. The dataset can contain a series of samples of frequency distributions. These will be individually tested for goodness of fit, a pooled distribution will be computed and tested, and a test for heterogeneity among the samples will be performed. The frequency distributions (including the expected frequencies computed by this program) can be plotted as bar graphs.

Assumptions

The *G*-test is based on the assumption that the observed frequencies follow the multinomial distribution with means given by the expected frequencies (from whatever distribution is used).

How to use

First click the Data button to specify a data file. Then the column in the input data that gives the frequency of each class must be selected. The distribution must be selected from the list. The predefined distributions (binomial and Poisson) are marked with an "*". Variables in the input dataset are also listed allowing them to be selected as a source of arbitrary expected frequencies or ratios. The other fields that must be selected depend upon the type of analysis to be performed. The Variable field does not need to be selected when arbitrary expected frequencies are used. It also does not have to be provided for the Binomial and Poisson distributions if the observed frequencies are given for all classes starting with $Y = 0$.

The Samples field should be filled-in for replicated goodness of fit tests (leave blank otherwise). The check mark should be cleared from the Estimate parameters box if parameters are to furnished. For the Binomial distribution enter p and k. For the Poisson distribution enter the mean μ.

The frequency distributions can be plotted as bar graphs. Plots are provided for each By group. The plotting options dialog box allows you to specify whether the pooled distribution is also plotted.

Output listing

The results of one or more *G*-tests (including Willams' correction) are displayed. If several samples of observed frequencies have been provided then a pooled observed frequency distribution is computed and the results of a *G*-test for heterogeneity shown. If there are two or more distributions they are plotted one above the other. If the pooled distribution is plotted it is given at the top of the plot as if it were an additional sample.

See the help file for an example of an output listing (with annotations and plots).

Homogeneity of variances

Purpose

This module performs the Bartlett, F_{max}, and the Scheffé-Box (log-anova) tests for homogeneity of variances. The Box-Cox transformation is provided to determine the power transformation that best transforms the data to normality and makes the variances as homogeneous as possible. A plot of the variances as a function of the means can be generated (useful since a common reason for heterogeneity is that sample from populations with larger means tend to vary more than those from populations with smaller means). Optionally, a randomization test can be performed using Levene's F-statistic.

Assumptions

An important assumption is that the data are normally distributed. Note that outliers can greatly increase the variance of the samples in which they are found and thus cause an apparent heteroscedasticity. The Scheffé-Box test assumes that the data values in each sample are ordered randomly since the program (arbitrarily) groups adjacent observations into each subsample. The randomization test assumes only that the variates were sampled at random from some population.

How to use

First use the Data button to specify a data file. If the data file consists of raw data then the variable that corresponds to the dependent variable and the variable that indicates to which sample each observation belongs must be entered. If the data file consists of summary statistics including variances and sample sizes then enter these variables in the fields for Variance and Sample size (leave them blank otherwise). If the Resampling no. field is greater than 0 then a randomization test using Levene's (1960) statistic will be performed and a bootstrap estimate of the common variance computed. Click the Compute button to perform the computations. If raw data are provided then a plot can be made of the variance as a function of the mean. There should not be any correlation. The presence of a positive relationship (samples with larger means have larger variances) implies that a log transformation might be appropriate.

Output listing

The sample sizes and variances (ordered by increasing magnitude of the variances) are listed followed by the average variance and its degrees of freedom. The F_{max}-test statistic is displayed next. For Bartlett's test the uncorrected X^2 statistic is displayed followed by the correction factor, *C*, the adjusted X^2 value, its degrees of freedom, and probability.

The exact results of the Scheffé-Box (log-anova) test depend upon how the variates in each group are grouped into subsamples. Therefore the module displays the sample sizes of each subgroup it forms. If, for example, the module displays "Sample 1: 3 3 2", then the first 3 variates have been grouped to form a subgroup, the next 3 form the next subgroup, and the last 2 form the third subgroup for group 1. This is followed by a table that gives an anova of the logs of the variances of the above subgroups.

The results of Levene's anova of the absolute values of the deviations of each observation from its sample mean is given next. If the number of resamples is greater than zero then a randomization test is performed using the F statistic from Levene's test. Each sample is a random permutation of the residuals of the original observations from their sample medians. In addition, a bootstrap estimate of the pooled within-groups variance is also given.

See the help file for an example of an output listing (with annotations and plots).

Isotonic regression

Purpose

This method allows one to relax the assumptions of linear regression and test simply for the presence of a consistent increase (or decrease) in the dependent variable, *Y*, as a function only of the rank order of the values of an independent variable, *X*.

Assumptions

The method assumes that the dependent variable is normally distributed with the same variance for all values of the independent variable.

How to use

First use the Data button to specify a data file. If the data file contains raw data then the variable that corresponds to the dependent variable and the variable

that gives the hypothesized ordering must be entered (this later variable also serves to group the values of the dependent variable). If the file contains summary statistics then the variables that contains the mean of the dependent variable, the sample size, the sample variance, and the hypothesized ordering must be entered. In either case, enter the number of iterations of random sampling that are to be used to generate a reference population for comparison with the E^2 statistic for the observed data. Click the Compute button to perform the computations. Plots can be made showing the relationship between the dependent variable and the rank ordering and of a histogram of the E^2 values obtained from the random sampling.

Output listing

Note: the computations may take a while if there are more than just a few groups and a large number of iterations is specified.

The means and sample sizes are listed in the specified order and then after merging so that they increase monotonically with the specified order. The value for the E^2 statistic (SS_{groups} divided by SS_{Total}) is then displayed followed by the results of a Monte Carlo sampling. The probability is estimated as $(c+1)/(nIter+1)$, where c is the number of random samples with an E^2 value equal to or larger than the observed and *nIter* is the number of random samples taken. Plots are available showing the relationship between the means and the specified order and of the distribution of E^2 values obtained from the random samples.

See the help file for an example of an output listing (with annotations and plots).

Kruskal-Wallis test

Purpose

This module performs the Kruskal-Wallis test, a non-parametric test for equality in the location statistics of two or more samples. This is a distribution-free analog of single-classification anova.

Assumptions

The observations are independent.

How to use

First click the Data button to specify a data file. Then the variable containing the dependent variable being analyzed and the variable indicating the sample to

which each observation belongs must be entered. Only raw data can be analyzed since individual variates need to be ranked. Then click the Compute button to perform the computations. A plot can be made showing histograms for each sample.

Output listing

The average rank for each sample is displayed followed by the Kruskal-Wallis statistic, *H*, (adjusted for ties if necessary).

See the help file for an example of an output listing (with annotations and plots).

Linear regression

Purpose

This module finds the least-squares best fitting regression line to predict the dependent variable, *Y*, as a linear function of the independent variable, *X*. There can be one or more Y-values corresponding to each distinct X-value. The following model is fit by the module:

$$Y_{ij} = \mu + \beta(X_i - \overline{X}) + D_i + \varepsilon_{ij}$$

where β is the slope of the regression line, X_i the *i*th distinct value of the independent variable, $\bar{x}$ the overall mean of the independent variable, D_i the deviation of the mean μ_i of the dependent variable from the mean of the dependent variable predicted by the regression line using X_i, and ε_{ij} is the error.

Optionally, a randomization test for *b* and a bootstrap estimate of *b* can be computed.

Assumptions

It is assumed that the errors (ε_{ij} values) of the dependent variable are independent and normally distributed with a mean of 0 and a variance of σ^2 for all values of the independent variable. The independent variable is assumed to be Model I.

How to use

First use the Data button to specify a data file. Next, enter the variables for the dependent and independent variables. If there are more than one Y values for each distinct value of X, then this variable will also break the data into groups. Enter in the Alpha field the probability level to be used in setting $(1-\alpha)100\%$

confidence limits. If the Analyze residuals box is checked then the residuals from the regression model will also be displayed in the output listing and residual plots will be available. If the value in the Resampling no. field is greater than 0 then a randomization test for *b* will be performed and a bootstrap estimate of *b* will also be computed. Click the Compute button to perform the computations. Plots should always be made showing the dependent variable as a function of the independent variable and of the residuals as a function of the independent variable.

Output listing

The means and variances for *X* and *Y* and the regression equation for predicting *Y* given *X* are listed first. The confidence limits of the slope are also shown. These are followed by the analysis of variance table. In the anova, the MS for linear regression is tested over the MS for deviation. The latter is tested over the error MS.

If the "Analyze residuals" box is checked then the anova table is followed by a table giving the *X* and *Y*-values ordered by the sequence of original input. The table also features predicted values (`Yhat`) and their confidence limits. This is followed by another table giving $d_{Y.X}$ (`Deviation`), leverage coefficients (`h_i`), and standardized residuals for each value of *X*.

If the resampling number is greater than 0, then a randomization test is performed for the regression coefficient. Counts are made of the number samples with a regression coefficient that is less than, equal to, or greater than the observed coefficient. A bootstrap estimate and its standard deviation are given for the regression coefficient.

See the help file for an example of an output listing (with annotations and plots).

Logistic regression

Purpose

Logistic regression is used to fit a linear relationship between the logit transformation of a proportion and one or more independent variables. This is in contrast to tests of independence (RxC tests for 2-way tables and log-linear tests for 3-way tables in BIOMstat) which do not take the ordering of the states of the independent variables into account. The logit of a proportion *p* is $\ln(p/(1-p))$. The model is similar to that used in linear regression analysis. For a single independent variable, the model being fit is

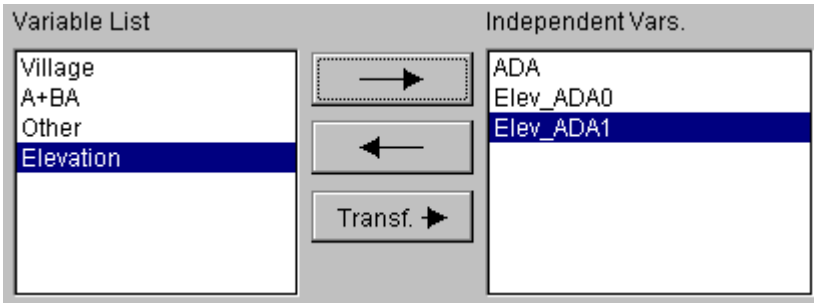
$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta X + \varepsilon$. However, in logistic regression the error, ε, is not assumed to be normally distributed and thus special methods must be used.

Assumptions

The proportions should be independent and follow the binomial distribution. The independent variables are assumed to be Model I (fixed effects known with little or no error).

How to use

First use the `Data` button to select a data file. Next the variables corresponding to the dependent variable (giving, for example, the frequency for which some attribute is present) and the "complementary variable" (giving, for example, the frequency for which the attribute is absent) in each of a series of samples. Raw counts, *not* percentages or proportions must be given. Then select one or more variables as the independent variables as one would do in multiple regression analysis. Each of the independent variables can be transformed if desired. Click on the right arrow to move a highlighted variable at the left to the list of independent variables. Click on the left arrow to return the highlighted variable on the right to the Variable List. An example is shown below.



Enter in the Alpha field the probability level to be used in setting (1-α)100% confidence limits.

If the independent variables correspond to levels of two or more factors in an analysis of variance or analysis of covariance type design then the program will need to be run several times — each time with the variables corresponding to different factors left out. The difference in the *G*-values from the goodness of fit tests measures how important were the variables that were left out (the difference in the degrees of freedom give the degrees of freedom for testing the difference in *G*-values). Note: if the independent variables are too highly correlated then the iterative maximum likelihood estimation of the parameters may not converge to a solution. In such cases an error message will be displayed.

If Empirical logits is checked then empirical logits will be used in the computations (useful when sample sizes are small). If Residuals is checked then residuals from the regression line will be displayed in the output listing and residual plots will be available.

Output listing

A report is given first on the rate of convergence to the maximum likelihood estimate of the model. The parameters are then listed along with their standard errors, tests of significance, and confidence limits (both as logits and as odds ratios). This is followed by the covariance matrix among the parameter estimates. A *G*-test is given for the fit of the entire model.

Bivariate plots can be made of the dependent variable as a function of the other variables in the data. The regression line can be shown when the abscissa corresponds to an independent variable. If the "Residuals" box is checked then points corresponding to expected frequencies and deviation vectors connecting observed and expected values can be superimposed on the plot. Note: if there is more than one independent variable then the expected points will not fall along the regression line (they do fall on the regression plane but that cannot be shown on a bivariate plot).

If the "Residuals" box is checked then a list of residuals and their standard errors is given in the listing followed by their confidence limits in terms of logits and of proportions. The list of leverage values, h_i, for each observation is useful since larger values correspond to observations that are given higher weight for estimating the parameters of the model and thus would be more sensitive to the effects of outliers (if present). If residuals are computed then a plot can be made of the residuals as a function of the other variables in the data.

See the help file for an example of an output listing (with annotations and plots).

Log-linear analysis of 3-way tables

Purpose

This module fits a succession of log-linear models to frequency data in a three-way contingency table. The results of these tests can be used to determine the simplest model to fit the associations among the 3 variables.

Assumptions

The frequencies are assumed to follow a multinomial distribution. There are no constraints on the marginal totals based on the design of the experiment.

How to use

First use the Data button to specify a data file. Then the variables corresponding to the three factors and to the cell frequencies must be entered. Click the Compute button to perform the computations. If the 3-way interaction is large then one should split the data into subsets and analyze the 2-way tables using the RxC module. If List deviations is checked, then the residual deviations from each model will be displayed. The value in the Alpha field will be used as the probability level for setting (1-α)100% confidence limits.

Output listing

The program displays the results of fitting the following models: three-way interaction, independence of two factors given a third (conditional independence), complete independence of one factor from the other two, and the complete independence of all three factors. The G-statistic (with and without Williams' correction) is given for each test along with its degrees of freedom. If requested, each test will be followed by a three-way table giving the observed and expected frequencies and the Freeman-Tukey deviates for the model being fitted.

Next, one can find the simplest model that gives an adequate fit to the data. This can be accomplished using Figure 17.5 in *Biometry* as a guide to testing a hierarchical series of models. The increase in G-value due to a parameter being deleted from a model is computed by subtraction of the proper pairs of G-values.

See the help file for an example of an output listing (with annotations).

Mann-Whitney *U*-test

Purpose

This module provides a non-parametric test for differences in location statistics for two samples (see the Kruskal-Wallis module, page 35, for more than two samples). It is a distribution-free analog of the *t*-test for the difference between two means.

Assumptions

The observations are assumed to be independent.

How to use

First use the Data button to specify a data file. Then enter the variables corresponding to the dependent variable and the samples. Raw data, rather than summary statistics, must be provided since the data need to be ranked. Click the Compute button to perform the computations. A plot can be made showing the two samples as histograms.

Output listing

The average rank for each sample is displayed. This is followed by the Mann-Whitney test and its probability. A plot can be made showing the two samples as histograms.

See the help file for an example of an output listing (with annotations and plots).

Mantel test

Purpose

This procedure is used to test for association between two independent dissimilarity matrices describing the same set of variables or entities (see Page 16 for a description of the special data format required). This approach is very general since one of the matrices can be coded to represent a wide variety of statistical methods (*e. g.*, see Section 18.3 of *Biometry*).

Assumptions

The two matrices are assumed to have been obtained independently. One matrix cannot, for example, have been mathematically derived from the other.

How to use

Use the Matrix1 and the Matrix2 buttons to specify each of the two data files. Note that the BIOMedit program requires a special mode to edit a distance matrix. Then specify the number of random permutations to be used (the observed data are considered to be the first sample). If the Standardized Mantel Statistic box is checked then the Mantel statistic will be expressed in standardized form (a correlation coefficient). Plots are available showing the scatterplot of one matrix against the other and a histogram of results of the random permutations.

Output listing

The number of elements in the lower half matrix is displayed along with the mean and sums of squares of each lower half matrix. This is followed by the Mantel statistic, Z, and an asymptotic test. If the number of iterations is > 0 then the results of the sampling are reported as the number of samples which resulted in values $< Z$, $= Z$, or $> Z$ and the estimated probability of a random sample being $\geq Z$ (note: the observed value of Z is included in the counts and the computation of this probability). Plots are available showing the scatterplot of one matrix against the other and also a histogram of the results of the random permutations.

See the help file for an example of an output listing (with annotations and plots).

Multiple comparisons among means

Purpose

This module performs several different multiple comparison tests for differences among a set of means. These tests all allow for the fact that one is making multiple tests with the same set of data. The following methods are included:

- The T method

- T', Tukey-Kramer, and GT2 methods
- Games and Howell method (allows for unequal variances)
- Welsch step-up method
- The SS STP method

Plots are also provided of comparison limits for the T and GT2 methods.

Assumptions

The methods are all based on the usual anova assumptions of independent random samples from normally distributed populations. Except for the Games and Howell method, it is assumed that the variances are the same in all populations. The T and Welsch step-up methods require equal sample sizes.

How to use

First use the Data button to specify a data file. A variable must be selected that gives either the raw data values (which can be transformed) or the mean of each sample to be compared. If the file contains raw data then a variable must be selected that indicates the grouping into samples. If means were entered then variables must be selected that give the variances and sample sizes. Select the desired method from the list in the Method field. Click the Compute button to perform the computations.

Use the Alpha field to enter the probability level to be used in testing. Note that some methods permit testing at only a particular set of α values. T: 0.05, 0.01; T': 0.10, 0.05, 0.01; T-K: 0.05, 0.01; GT2: 0.10, 0.05, 0.01; and Welsch: only at 0.05. This is because these methods use special tables that are available only for those values of α.

Output listing

The form of the output depends upon the method selected. For all methods a list of the means sorted from low to high is displayed first followed by the average within-groups variance and its degrees of freedom.

T-method. A table is shown in which every mean is tested against every other mean. The differences between the two means is given in the row labeled "`Diff`" and the critical values for the T method are given in the following row. A difference between a pair of means is significant (and marked with an "*") if it is larger than this value. A plot of comparison limits is also available.

T',T-K, GT2 methods. A table is provided in which every mean is tested against every other mean. The differences between the two means is given in the row

labeled "Diff" and the critical values the T', Tukey-Kramer, and GT2 methods are given in the following three rows. A difference between a pair of means is significant (and marked with an "*") if it is larger than at least one of these values. A plot of comparison limits is also available for the GT2 method.

Games and Howell method. A table is shown in which every mean is tested against every other mean. The differences between the two means is given in the row labeled "Diff" and the critical values for the Games and Howell method is given in the following row. A difference between a pair of means is significant (and marked with an "*") if it is larger than this value.

Welsch step-up method. A table is given showing the Welsch statistics for maximal non-significant ranges of adjacent pairs of means, triplets of means, *etc.*

SS STP method. A list of the maximal non-significant sets is displayed. Each set consists of a list of the identification code numbers for those samples that are not significantly different, the mean for this set, and the sums of squares among these groups. Since it possible for there to be a very large number of such sets, computations will stop and the message "List of subsets incomplete" will be displayed if more than 100 sets are found (results are cut off at this point since it is very difficult to interpret large numbers of non-significant sets).

See the help file for an example of an output listing (with annotations and plots).

Multiple regression

Purpose

This module is used to predict a dependent variable given a suite of 1 or more independent variables using least-squares multiple regression analysis. The multiple regression model is:

$$Y_i = \alpha + \beta_{Y1} X_{i1} + \beta_{Y2} X_{i2} + \cdots + \beta_{Yk} X_{ik} + \varepsilon_i$$

Where Y_i is the ith observation of the dependent variable, α the Y-intercept, $\beta_{Y1.}$ the partial regression coefficient for Y on X_1 with all other independent variables held constant, X_{i1} the value of the first independent variable for the ith observation, $\beta_{Y2.}$ the partial regression coefficient for Y on X_2 with all other independent variables held constant, X_{i2} the value of the second independent variable for the ith observation, $\beta_{Yk.}$ the partial regression coefficient for Y on X_k with all other independent variables held constant, X_{ik} the value of the kth independent variable for the ith observation, and ε_i the random error term for the ith observation.

The program can also compute Kruskal's method of estimating the overall importance of each variable (the average correlation and partial correlation of the dependent variable with each independent variable for all possible combinations of the other independent variables being held constant).

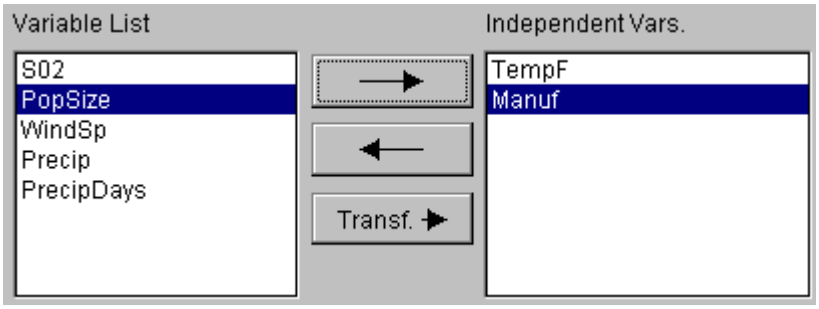
Optionally, a randomization test for R^2 and bootstrap estimates of the standardized partial regression coefficients can be computed.

Assumptions

The residual error is assumed to be independent, normally distributed, and with a homogeneous variance. The independent variables are assumed to be Model I (fixed effects with little or no error).

How to use

First use the Data button to specify a data file. Next select a variable as the dependent variable. Then select one or more independent variables by clicking on the "→" button (click on the "←" button to remove a variable from the list) as shown below.



Transformations may be applied to the independent variables by using the "Transf" button for the highlighted variable. Selecting the checkbox for Kruskal's method can greatly add to the computation time if there are more than just a few variables. Enter in the Alpha field the probability level to be used in setting (1-α)100% confidence limits.

If Analyze residuals is checked then the residuals will be displayed and tested. If Kruskal importance is checked then Kruskal's importance index will be computed for each independent variable. If a value greater than 0 is entered in the Resampling no. field then a randomization test for the multiple correlation coefficient will be performed. Bootstrap estimate of standardized partial regression coefficients will also be computed. Click the Compute button to perform the computations.

Both the plots of the dependent variable against the independent variables and the residual plot should be examined to make sure there are no influential outliers and that there is no systematic pattern to the residuals.

Note: if the independent variables are too highly correlated then the matrix inversion step may fail and estimation of the parameters will not be possible. In such cases an error message will be displayed. Even when the program runs to completion, the variance inflation factors should be checked to detect potential problems.

Output listing

The mean and standard deviation for each variable are listed followed by the variance-covariance and the correlation matrices for all k+1 variables. Given next is the inverse of the correlation matrix for the k independent variables. Variance inflation factors, VIFs, over 10 give cause for concern but the accuracy of the solution should still be satisfactory unless they are greater than 100.

The anova table gives F-values for the overall test of significance as well as tests for each variable (with all other variables held constant). Both the multiple correlation coefficient and its square are shown at the bottom of the anova table. Values of the partial regression coefficients and the standard partial regression coefficients are provided next along with their standard errors and confidence limits.

If an analysis of residuals is requested, then the predicted values, residuals, standardized residuals and the leverage coefficients (also called the diagonals of the "hat matrix"), h_i, are displayed for each observation.

If the resampling number is greater than 0, a randomization test is performed on the R^2 statistic. In addition, bootstrap estimates and their standard deviations are given for the standardized partial regression coefficients.

See the help file for an example of an output listing (with annotations and plots).

Nested anova

Purpose

This module performs a nested analysis of variance with up to nine levels. The module allows for unequal sample sizes at all levels and uses the Satterthwaite approximation in such cases to estimate variance components and to perform significance tests. If there is only a single level then the computations become those of a single-classification anova.

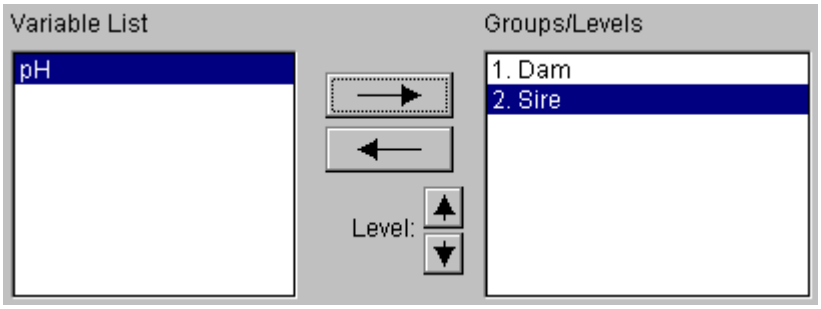
Assumptions

This analysis makes the usual anova assumptions of independent, normally distributed errors with homogeneous variances. While the highest level can be either Model I or Model II, the lower levels are assumed to be Model II factors.

This program has the constraint that the input data records must be ordered in a hierarchical manner to reflect the structure of the data.

How to use

First use the Data button to specify a data file. Next, select the variable to be analyzed. Then the variables corresponding to the levels in the anova must be entered in the Groups/Levels field. Note: the variables must be arranged in order with the highest level at the top of the list as shown below. The list of variables corresponding to the levels in the analysis using the special control illustrated below. Click on the "→" button to add a variable from the left to the list of Groups/Levels. Click on the + and -- keys to change the highlighted variable's position with the list. Click on the "←" button to return a variable at the right to the Variable List.



Correspondingly, the rows of the input data *must* be arranged so that lower level observations are nested within higher levels. The value in the Alpha field is used as the probability level for setting (1-α)100% confidence limits. Click the Compute button to perform the computations.

Output listing

A standard anova table is given showing the SS, df, MS, Fs, and P values for each level. If sample sizes are unequal then synthetic mean squares and their approximate degrees of freedom (using Satterthwaite's approximation) are given below each MS, df, and F-value. They are enclosed in parentheses to help distinguish them from the usual quantities. No pooling of mean squares is performed.

The anova table is followed by a list of the estimated variance components expressed both in the original units and as percentages, which is in turn followed by a table of the coefficients of the expected mean squares.

See the help file for an example of an output listing (with annotations).

Non-parametric tests of association

Purpose

This module computes Kendall's and Spearman's rank correlation coefficients to test for association between two dependent variables, Y_1 and Y_2.

Assumptions

The bivariate observations are assumed to be independent. The variables can be either Model I (fixed, measured without error) or Model II (random, measured with error).

How to use

First use the Data button to specify a data file. Then the variables Y_1 and Y_2 must be entered in their respective fields. Click the Compute button to perform the computations. Note: the computations can take a long time for moderate to large sample sizes. A scatterplot plot is available showing Y_1 vs. Y_2.

Output listing

The Kendall rank correlation and a test for whether it is different from zero are given first. These are followed by the Spearman rank correlation coefficient and the test for its significance. The closely related sum of squared difference in ranks is also given along with its test.

See the help file for an example of an output listing (with annotations and plots).

Non-parametric two-way anova methods

Purpose

This module performs several methods that are alternatives to standard two-way analysis of variance. The methods include Friedman's method for randomized blocks, Wilcoxon's signed-ranks test for two groups (paired comparisons), and the Scheirer-Ray-Hare two-way anova of ranks. Equal sample sizes are required in each cell.

Assumptions

This analysis assumes independent random sampling.

How to use

First select the Data button to select a data file. Next select the column in the data corresponding to the dependent variable. Then select the variables corresponding to factors A and B (use factor B for the blocks in a randomized blocks design). The rows on the input data file need not be in any particular order. If more than one observation is given for a cell then it is considered to be a replicate.

Output listing

The output depends on the analysis that is run – which in turn depends on the numbers of rows, columns, and sample size in each cell. The Scheirer-Ray-Hare method is used if the number of observations in each cell, n, is greater than 1. The Wilcoxon's signed-ranks test is used only if n=1 and factor A has just two levels.

See Section 13.12 of Biometry for a discussion of the computations for these methods. Box 13.10 describes Friedman's method, Box 13.11 describes Wilcoxon's signed-ranks test, and Box 13.12 describes the Scheirer-Ray-Hare procedure.

Polynomial regression

Purpose

This module carries out curvilinear regression to predict the dependent variable *Y* as a polynomial function of the independent variable *X*. While the resulting model may give good empirical predictions (at least within the range of the observed values for the independent variable), it is usually not used to suggest an underlying causal model for the dependent variable or to extrapolate beyond the observed values for X.

Assumptions

The residual error is assumed to be independent, normally distributed, and with homogeneous variance. The independent variable is assumed to be Model I (fixed, measured without error).

How to use

First use the Data button to specify a data file. Next The dependent variable. Then the independent variable and the degree, *k*, of the polynomial to be fit must be entered. The degree must be less than the number of observations (in practice the maximum power is often limited by computational accuracy when $k>4$). The value in the Alpha field is used as the probability level for setting $(1-\alpha)100\%$ confidence limits. If the Analyze residuals box is checked then an analysis of residuals will be performed. Click the Compute button to perform the computations. Plots are available of both Y as a function of X and of the residuals in Y as a function of X.

Output listing

The mean, variance, and standard deviation of the dependent variable, *Y*, are given followed by the mean and standard deviation of the independent variables (*X* raised to powers from 1 to *k*). Next, the correlation between each independent variable and *Y* and a matrix of the correlations among the independent variables is shown.

For each degree from 1 to *k* the following information is provided: partial regression coefficients, their standard errors, *t*-value, and variance inflation factors (VIFs). Note: it is not unusual for the VIF values to quite large (greater than 100 indicating very unreliable results for $k>4$). These are followed by the

residual SS, degrees of freedom, residual MS, squared multiple correlation coefficient, and the sample *F*-value and its probability. Note: these tests are for fit of the model, not the increase in the fit over the previous model. The latter is given in a summary table at the end of the listing.

If the "Analyze residuals" box is checked then a table is printed giving X, Y, predicted values (`Yhat`), residual, standardized residual, and the leverage (`h_i`) values for each observation. A plot of the residuals will also be available.

If $k>1$ then a summary table is given at the end of the listing file. It gives significance tests for the incremental improvement given by each increase in the degree of the polynomial being fitted.

See the help file for an example of an output listing (with annotations and plots).

Probability calculator

Purpose

This module is used to compute critical values and probabilities for samples from the Chi square, F, Normal, and t distributions.

How to use

First select the desired distribution from the list in the Distribution field. Fill in the degrees of freedom as appropriate. Then check the inverse box if you want to compute a probability rather than a critical value. If it is not checked then enter the desired type I error rate. Otherwise enter a value for the statistic. Finally, click the Compute button to do the computations and display the results.

Output listing

The result is simply a value displayed in the results edit box.

See Chapters 6 and 7 of Biometry for a discussion of these statistical distributions.

RxC test of independence

Purpose

This module performs a test of independence in an R×C contingency table by means of the *G*-test (with Willams' correction). Optionally, it carries out unplanned tests of all subsets of rows and columns in the R×C contingency table by Gabriel's simultaneous test procedure which finds all maximal non-significant sets of rows and columns.

Assumptions

The cell frequencies follow a multinomial distribution. There are no constraints on the row or column totals (Model I, only the total sample size is considered fixed by the design of the experiment).

How to use

First use the Data button to specify a data file. Next select variables to indicate the rows and columns of the contingency table. Then enter the variable giving the cell frequencies. Records in the input file can come in any order. Multiple entries for the same cell in the contingency table are summed. The value in the Alpha field is used in performing the STP tests. If List n.s. sets is checked then Gabriel's STP will be carried out and maximal non-significant sets will be listed.

Note: computations can take a long time for large datasets if maximal non-significant sets are requested.

Output listing

A G-value (both with and without Williams' correction) for the test of independence for the entire table is given first. If the input corresponds to a 2x2 table, then Yates' correction will also be applied.

If the "List n.s. sets" option is checked, then all combinations of rows and columns will be tested and the maximal non-significant sets will be computed. A maximal non-significant set is one that is non-significant but becomes significantly heterogeneous if any other row or column is added to the set. If all subsets are significant then the message "`Entire set significant`" will be displayed, otherwise the maximal sets will be listed. For each set the G-statistic will be given followed by a list of the rows and the columns in the given set. The number of sets can be quite large so the search is terminated with the message

"Too many non-significant subsets found, search truncated" if more than 100 sets are found. It can be very difficult to interpret the results in a simple way when there are many such sets.

See the help file for an example of an output listing (with annotations).

Robust line fit

Purpose

This module uses Kendall's robust line-fit method to estimate the slope of a regression line. It is computed as the median of slopes computed from all possible pairs of observations. It is less sensitive to the effects of outliers than is the usual linear regression estimate. Kendall's rank correlation is used to test whether the slope is different from zero.

Assumptions

The independent variable can be either Model I (fixed, measured without error) or Model II (random, measured with error).

How to use

First use the Data button to specify a data file. Then the dependent and the independent variables must be entered. Note: these computations can take a long time for larger sample sizes. Since slopes from all pairs of observations need to be computed and stored, a maximum of 129 observations can be used. If there are more than 129 observations then only a random sample (with a maximum size of 64K/8-1 pairs) will be used. Click on the Compute button to perform the computations. A progress bar is displayed to suggest the rate of completion of the computations. A plot is available showing Y as a function of X with the robust regression line superimposed.

Output listing

The listing displays the slope and *Y*-intercept based on Kendall's robust line-fit method. Kendall's rank correlation is also given to provide a test of the null hypothesis that the slope is zero.

See the help file for an example of an output listing (with annotations and plots).

Single-classification anova

Purpose

This module performs a single classification analysis of variance. Confidence limits are computed for the estimate of the variance component.

Optionally, a randomization test and a bootstrap estimate of the variance component can be computed.

Assumptions

This method is based on the usual anova assumptions of independent, normally distributed errors with homogeneous variances. The effect of groups can be either Model I or Model II.

How to use

First use the Data button to specify a data file. Next select the dependent variable, Y. Then the variable that indicates to which group each observation belongs must be entered. The value in the Alpha field is used to set (1-α)100% confidence limits. If the value in the Resampling no. field is greater than 0 then a randomization test for the *F* ratio will be performed. Bootstrap estimate of group mean differences (estimates of the α_i) will also be computed. Click the Compute button to perform the computations. A plot is available showing histograms for each of the samples.

Output listing

The number of levels (groups) is displayed. The usual anova table is then presented. This is followed by an estimate of the variance component, its confidence limits, and the percentage of variance explained by the variance component. The coefficient, n_0, for the variance component is also given.

See the help file for an example of an output listing (with annotations and plots).

Tukey's test for non-additivity

Purpose

This module performs Tukey's test for non-additivity. This method partitions the interaction sum of squares in a two-way anova. The presence of a large component for non-additivity may suggest the use of a log transformation of the dependent variable so as to reduce the size of the interaction.

Assumptions

The deviations from the model are assumed to be independent and follow the normal distribution with homogeneous errors. The method requires an equal number of replications in each cell of the two-way table.

How to use

First, use the Data button to specify a data file. Either raw data or means and sample sizes can be used as input to this module. Next select the variables that indicate the factors that define the rows and columns of the two-way anova table. If means are selected above, then the variable that indicates the sample sizes for each cell must be entered. Click the Compute button to perform the computations. A plot is available showing the cell means as a function of the product of the row and column deviations for each cell. A large SS for non-additivity implies that this plot will show a strong linear relationship between these two variables.

Output listing

A table is presented that furnishes a partitioning of the interaction sums of squares into a single degree of freedom component for non-additivity and a residual sums of squares.

See the help file for an example of an output listing (with annotations and plots).

Two-way analysis of variance

Purpose

This module is used for a two-way analysis of variance. The module produces the standard anova table. It can handle balanced (equal sample sizes in each cell) and unbalanced designs. One can indicate whether each factor is model I or model II so that the correct significance tests and estimation of variance components are performed.

Assumptions

This analysis is based on the standard anova model which assumes homogeneous, independent, and normally distributed error within each cell of the data table.

How to use

First select the Data button to select a data file. Next, select the column corresponding to the dependent variable. Then select the variables corresponding to factors A and B. For each factor specify whether it is Model I (fixed treatment effects) or Model II (random effects). The rows of the input data file need not be in any particular order. If more than one observation is given for a cell then it is considered to be a replicate. Click the Compute button to perform the computations.

Output listing

In the anova table the various sources of variation, their sums of squares, degrees of freedom, and mean squares are given. F-tests are provided that take the models (I, II, and mixed) into account. A plot of the cell means using on of the factors as the abscissa can be produced. A Plot can also be produced showing the cell means as a response surface in 3D. Error bars (+ and – one standard error) can be shown in both plots.

See Chapter 11 of Biometry for a discussion of how the computations for balanced two-way designs are performed. For unbalanced designs a general linear model is used to solve for the various parameters in the anova model. Separate SS are given for the main effects when taking into account or ignoring the presence of any interaction.

Other software from Applied Biostatistics Inc.

BIOMlab

A program to help you and your students develop a better understanding of statistical principles through the use of sampling experiments. You can draw random samples from various distributions (both single samples and groups of samples). Explore properties of the central limit theorem, confidence limits, type I and II errors, correlation, and regression.

NTSYS-pc

A data analysis system designed to help you discover patterns and structure in multivariate data. It includes computation of various similarity and dissimilarity coefficients (including association coefficients and genetic distances), cluster analyses (includes UPGMA and neighbor-joining), ordination analyses (includes PCA, Pcoord, CVA, and MDSCALE), Mantel test, partial least-squares, and canonical correlation analysis. It includes partial least-squares analysis and canonical correlation analysis for studying the relationships between two sets of variables.. It also includes Fourier analysis and the computation of partial-warp scores for both 2D and 3D coordinate data (useful for geometric morphometric studies).

For more information or to place an order contact Exeter Software:

47 Route 25A, Suite 2
Setauket, NY 11733
516-689-7838, *FAX:* 516-689-0103
e-mail: `sales@ExeterSoftware.com`
www URL: `http://www.ExeterSoftware.com`