

# *S-Plus Basics*

*for students in Biometry  
(ZOO 631)*

---

---

## About this guide

This guide summarizes the procedures to be used during the course (Biometry, ZOOL 631). It does not cover every way of doing a particular procedure, nor does it provide every detail about these procedures; use the extensive on-line help (which includes the full programming and statistical manuals) if you want to learn more.

These instructions use only the menu/dialog window interface; there also is a rich command-line programming environment.

This guide was written originally for Release 6.1, and has not yet been fully updated for Release 7, though most of it should apply to Release 7 (and indeed some of the examples were generated in Release 7). Much of it also applies to earlier releases.

The first three chapters describe general features of using S-Plus, managing data, and working with graphs. After that, methods are presented in the same order as in the course, except that:

- Power analysis and other methods relevant to designing studies are presented in the last chapter, along with miscellaneous methods relating to probability distributions; in the course these are encountered at various points throughout the semester.
- Bar charts for displaying categorical variables are presented in the **Describing distributions** and **Describing relationships** chapters but in the course these may be deferred to the final chapter.

# INTRODUCTION TO S-PLUS

---

## S, S-Plus and R

- “S”  
(now “New S”) is a statistical programming language or environment (not a software package) developed by AT&T’s Bell Labs.
- “S-Plus®”  
is a commercial implementation of the “S” language, developed and sold by Insightful Corporation. Its most important addition to the “New S” foundation is the MS Windows graphical user interface.
- “R®”  
“can be considered a different implementation of S” (quoting from “What is R?” at [www.r-project.org/](http://www.r-project.org/)). It is developed and distributed for free by The R Project for Statistical Computing. Its user interface is not as polished as that of S-Plus, so its learning curve is steeper at first, but obviously the price can’t be beat.

Since all these flavors of R/S are programming environments, they are extensible: users can write new procedures (analyses, graphs, or whatever). There are indeed large collections of procedures available as add-ons (called “libraries” in S-Plus; see below, and “packages” in R). R in particular has attracted ecologists and evolutionary biologists: among its packages are cutting-edge statistical methods in these areas, programmed by the researchers who developed the procedures (see for instance the “ouch” package by our own Marguerite Butler).

---

## S-Plus Libraries

### Installing and loading libraries

Various libraries of specialized procedures are available from the Insightful Corp. website, at <http://www.insightful.com/downloads/libraries/>. (You can get there from S-Plus using

**Help ⇒ Insightful on the Web ⇒ Download Libraries )**

To install a library:

1. under the paragraph describing the library, click on the link for your operating system to download the compressed (.zip) file
2. double-click on the .zip file to extract the files (it doesn’t matter where they are put at this step, but you might as well have them put in the folder in step 3)

3. put the extracted folder and all its subfolders and files in the **library** folder of your S-Plus installation (typically, **c:\Program Files\Insightful\splus70\library** )

Once a library is installed on your computer, you then can load it into S-Plus whenever you need it. This is done by

**File ⇒ Load Library...**

In the dialog window that opens, scroll through the list of available libraries, select the one you want, check the Attach at top of search list option, and click OK.

## IPSdata Library

This guide assumes the **S+Resample Library** is loaded. If it is not, everything in this handout should still be correct, except that the resampling methods will not be available and some of the menus, dialogs, and output may be slightly different.

You also will find it convenient to load the **IPSdata Library**, which gives easy access to all the datasets in the text, and also automatically loads the S+Resample library. This library can be downloaded from <http://www.insightful.com/Hesterberg/bootstrap/> (reachable from the **Resampling Software for IPS** link on the textbook website <http://bcs.whfreeman.com/ips5e/> ). Download, install, and load this just like any other library (conveniently, it will be at the top of the scroll list in the Load Library window). It seems that some of the resampling procedures look a little different if only the S+Resample Library is loaded, rather than the IPSdata Library, but there shouldn't be any meaningful differences for our use of these procedures in this course.

# DATA ORGANIZATION, MANIPULATION, AND IMPORT/EXPORT

---

## Projects

S-Plus uses a large variety of data structures, graphs, reports, etc. It supposedly is possible to organize these into “chapters” but I haven’t bothered to figure out how this works. I find it sufficient instead to have a separate Windows folder for each “project” I am working on, and to specify the desired folder when S-Plus asks, while starting up, “Open the S-Plus project in which folder?” S-Plus will create .Data and .Prefs sub-directories in the specified project folder. These hold all sorts of folders and files, none of which is directly useful but which together constitute the project.

**Warning: It is difficult to move an existing project to a different folder;** apparently some of the needed files are in other folders and/or the paths pointing to necessary parts of the project become invalid. It is possible, though, to save S-Plus data sets or graphs as files, which then can be re-opened in S-Plus even if they have been moved.



The “Object Explorer” window, opened by clicking the icon to the left, can be used to access, work with, and simply keep track of the various parts of a project, the main ones being data, graphs, and reports.

---

## Data sets

S-Plus shows the data in one or more worksheets which look much like a sheet in a spreadsheet program such as MS Excel. Unlike Excel, in S-Plus columns have names and types. Column names, types, etc., can be seen or changed by selecting the column, right-clicking, and selecting Properties. S-Plus is better than Minitab at keeping rows intact (as observations) but care still needs to be taken, for instance when sorting.

It is not necessary to save data sets, as this is done automatically. Data sets which have been closed can be re-opened using

**Data ⇒ Select Data...**

and then selecting the desired data set from the pull-down list by Name: under Existing Data. As noted above, data sets can be saved if you want to move them.

You may notice in the Object Explorer that data files can be of various “classes.” Most will be of class data.frame, but sometimes you may see data.set or other specialized classes. As far as I can tell none of this should matter for the basic uses of this course.

## Entering data

Data can be entered directly into the worksheet from the keyboard or by cut-and-paste from other applications. Data saved from various spreadsheet and database programs can be imported using the menu sequence

**File ⇒ Import Data ⇒ From File...**

Alternatively,

**Data ⇒ Select Data...**

then select Import File from the choices at the left of the Select Data dialog.

The Import From File dialog then will open. Typically all that needs to be done is to specify the file to be imported. S-Plus determines the file format from the filename extension, so if your file doesn't have the standard extension you will need to tell S-Plus what kind of file it is, using the pull-down list of formats. You may want to change the name to be given the S-Plus data set being created. The Options tab can be used for changing which rows and/or columns are imported, and/or where column or row names are.

## Exporting data

Data sets can be exported in a wide variety of formats, including text, Excel, Matlab, Minitab, and SAS.

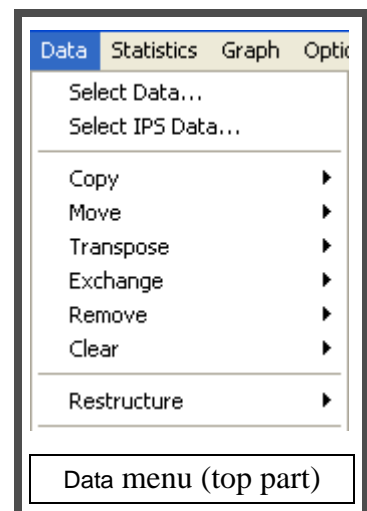
**File ⇒ Export Data ⇒ To File...**

then select the Data Frame to export, specify the File Name: to export it to, and select the format from the Files of Type: pull-down list.

## Data set manipulation

The Data menu (right) provides many ways of manipulating the data set. The most useful of these are:

- **Select Data...**  
This opens a pull-down list of available data sets, so you can re-open a closed data set. You also can use this to import data or create a new empty data set.
- **Select IPS Data...**  
This is only available **if** you have loaded the IPS library (File ⇒ Load Library...). It gives access to all the text data sets.
- **Copy, Move, Remove, and Clear**  
all act on columns, rows, or a block of cells, in the usual ways.

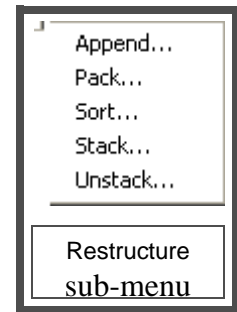


- Restructure

opens the secondary menu to the right, providing more complicated reorganizations of the data. The most useful of these are:

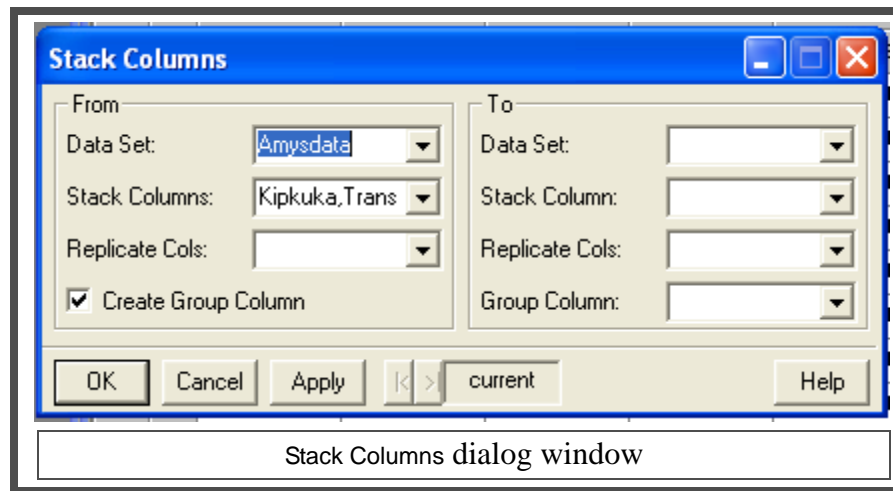
Sort...

does just that: sorts one or more columns, according to the values of one or more columns. The sorted columns can be put in the same or new columns. Just as in Minitab or Excel, **if you sort only some of the columns, associations between variables from the same observation may be destroyed.**



Stack...

stacks two or more columns into a single column. The columns to be stacked are specified in the From part of the dialog (see below), in the Stack Columns: box.



The column to create is specified in the To part of the dialog, in the Stack Column: box. If the Create Group Column option is checked, another column will be created, containing integers indicating which of the “From” columns an observation was from; the column to hold these subscripts is specified in the Group Column: box. Replicate Cols: will be duplicated in parallel with the stacked column, so that each row will contain a stacked value and the corresponding values of the replicate column.

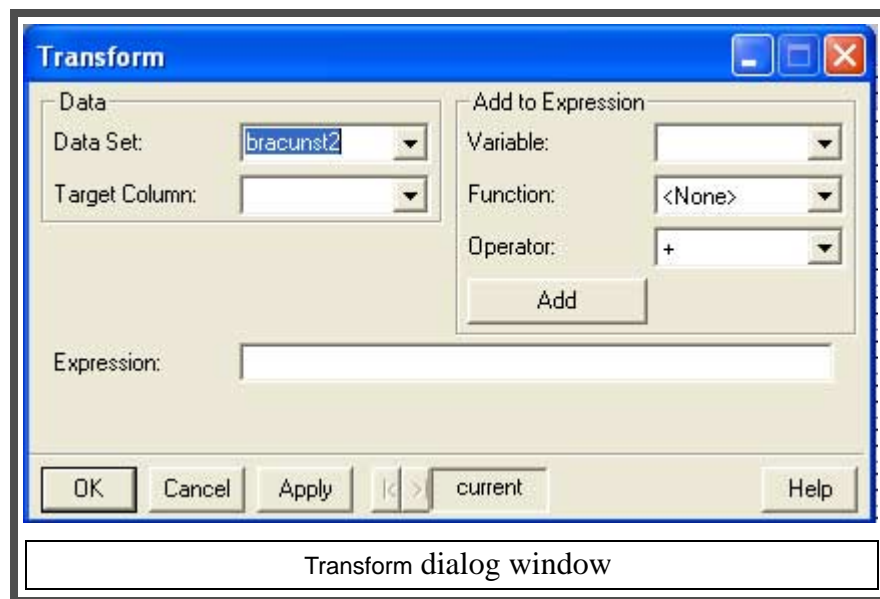
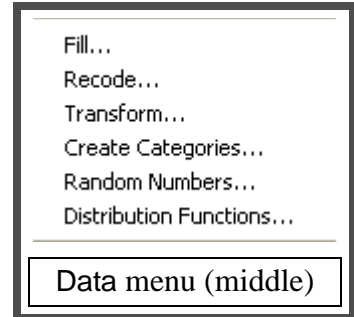
Unstack...

separates a column into several new columns, or a set of columns into several new sets, one new column (or set of columns) for each distinct value of the Group Column:.. The new columns are given names combining the value of the group column with the name of the original column.

- Transform...

The transformation dialog window (below) creates new columns as functions (mathematical and/or logical) of existing columns.

The column to be created is the Target Column:. The transformation can be typed directly into the Expression: box and/or built using the Add to Expression boxes. Unfortunately it does not appear that functions in the Function: box can be nested, as in the *arcsin-square root* transformation. I therefore mostly type the expressions directly, and sometimes use the Function: pull-down list to see what functions are available and how they are named.

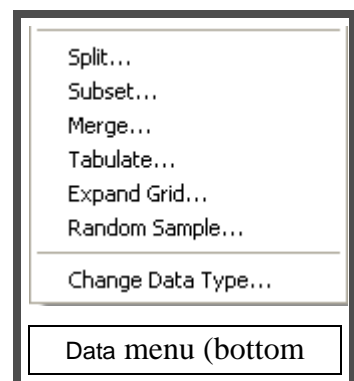


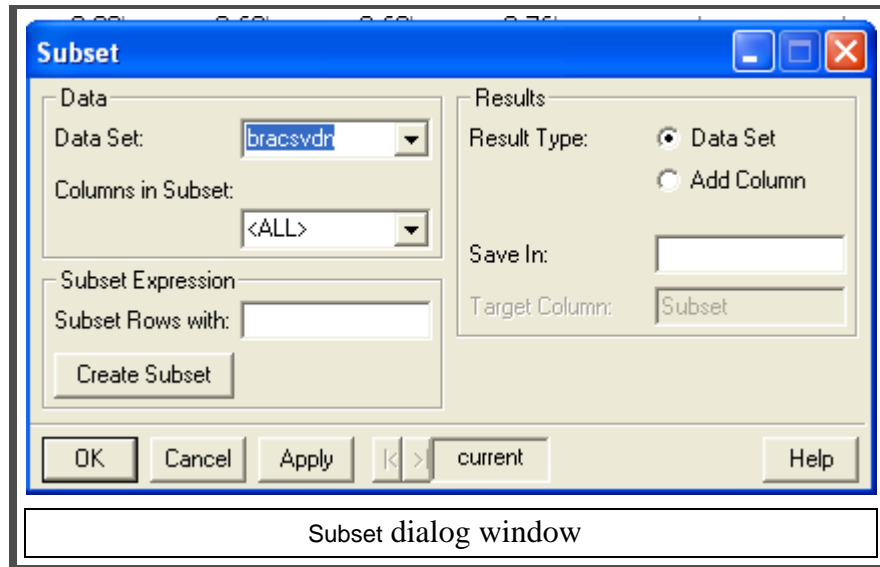
- Split...

separates a data set into several new data sets, according to the value of a Splitting Variable. If the splitting variable is quantitative, various ways of splitting it can be specified.

- Subset...

picks out the observations meeting a specified condition. The selected observations can become a new data set, if Result Type: is set to Data Set (the default; see dialog window below). Alternatively, a column can be added to the current data set containing True / False values depending on whether the given observation meets the condition or not; this is the Add Column option for Result Type:, and when it is selected the Target Column: box becomes available for naming this new column. This column can then be specified in the Subset Rows with: box when creating or

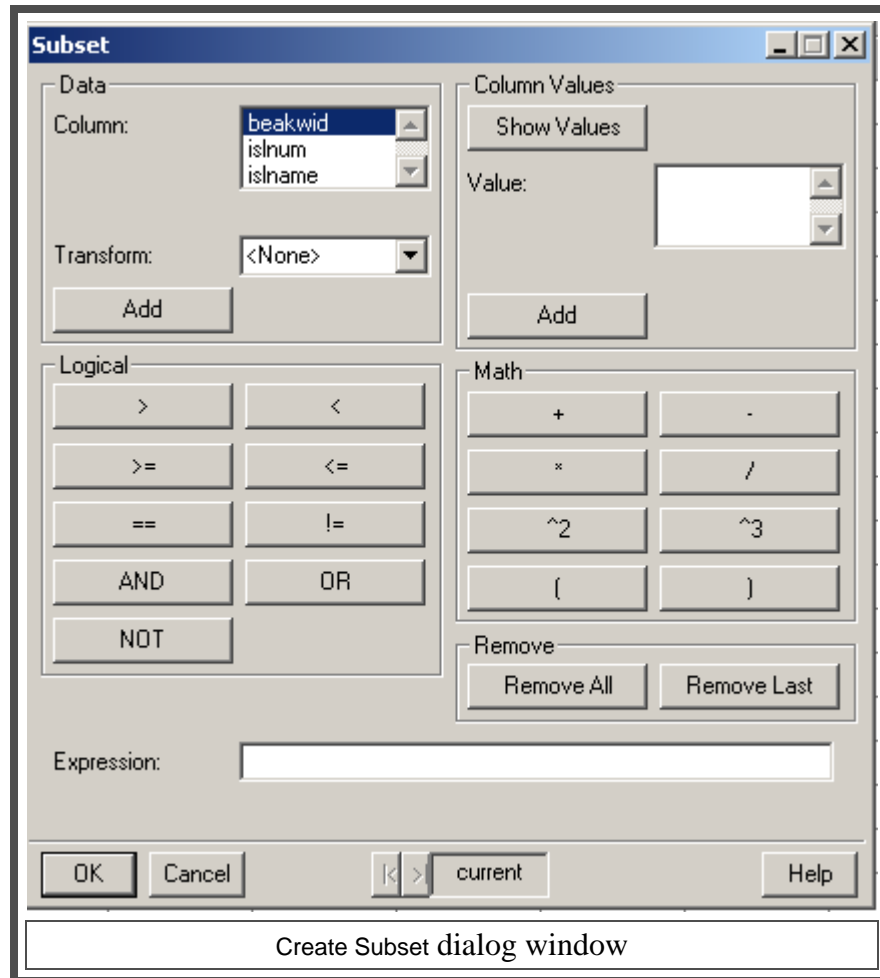




modifying graphs (as described below), causing only the True observations to be shown in the graph.

The condition defining the subset can be typed directly into the Subset Rows with: box, or clicking the Create Subset button opens a dialog window (next page) in which you can build the expression. The Transform: box in this dialog provides mathematical (e.g. square root) and statistical (e.g. median) functions. The Show Values button will cause all the values of the selected column to be listed in the Value: box; specific values can be selected from this list and incorporated into the expression.

- **Change data type**  
allows a column to be converted among the many types recognized by S-Plus. The most important of these are double (double-precision numeric), character, and factor (group identifiers, as for example the independent variable(s) in ANOVA).



## Exporting results

Text output from statistical procedures will be in Reports windows, which will be listed in the Object Explorer. All or part of a report can be copied and pasted into a word processing program, or can be printed (use the printer icon or the File menu).

Graphs can be exported in various ways, as explained in the next chapter.

## S-PLUS GRAPHING: GENERAL FEATURES

S-Plus has superb graphing capabilities. It combines a huge catalog of types of graphs with excellent WYSIWYG editing. It can easily make publication or presentation quality graphs, with direct export to PowerPoint. These capabilities come at the cost of more complexity than in programs such as Minitab, so in this section I will describe in detail the main ways of using S-Plus for graphing.

---

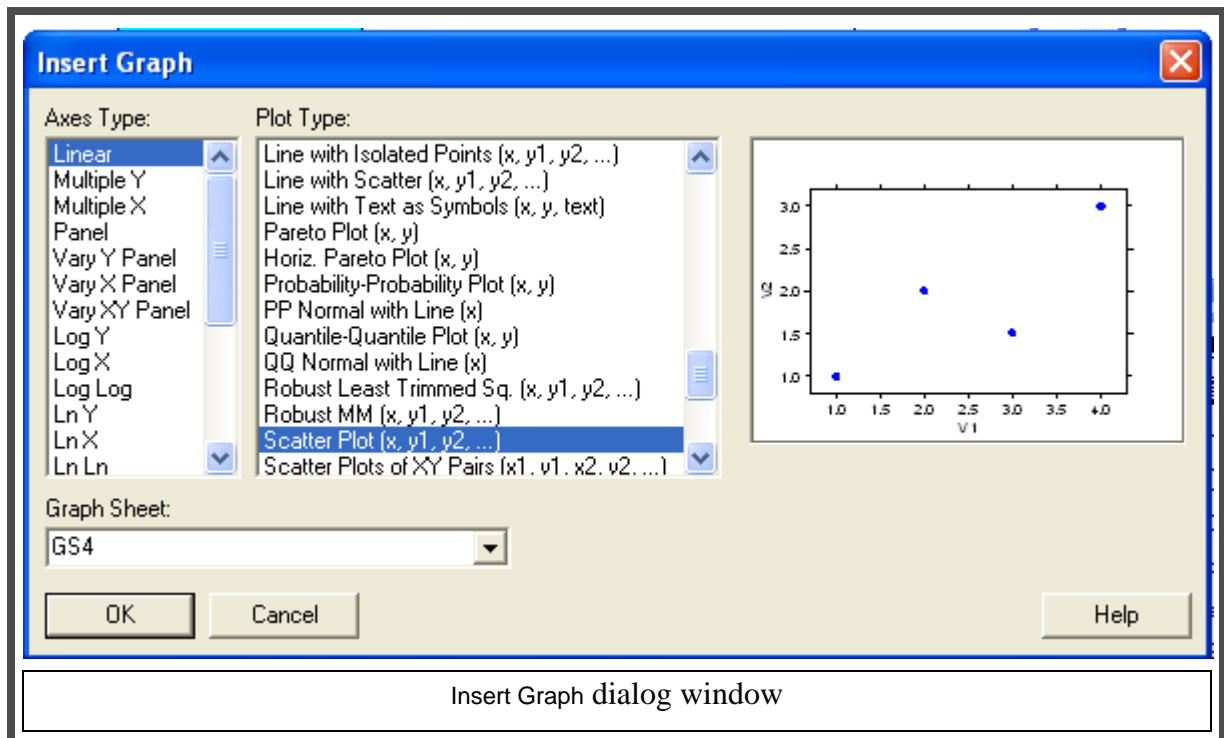
### Creating graphs

Graphs can be created in S-Plus by two routes: the Graph pull-down menu or a window showing a grid of buttons for various graph types.

#### Graph menu

**Graph** ⇒ **2-D Plot...**

opens the Insert Graph dialog window shown on the next page. The three input boxes of this window are:



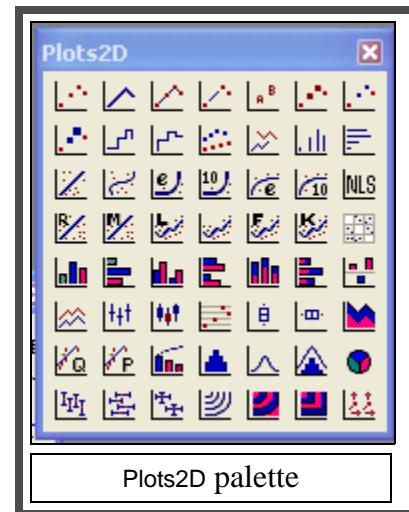
- **Plot Type**  
The middle box is where you specify the Plot Type you want; there is a vast variety of possible types.

- **Axes Type**  
The box at the left allows various rescalings of one or both axes, creation of multiple axes, etc.
- **Graph Sheet**  
A name for the graph can be entered in the Graph Sheet box; default names are GSn with sequential numbering.  
The box at the right of the dialog window simply shows an example of the type of plot being selected.

## Plot buttons



Alternatively, the toolbar icon to the left opens the selection palette to the right. The buttons on this palette represent the various graph types. What some of them represent is obvious, and for the others, if the cursor is held over a button for a few seconds a little box pops up identifying the graph type. The ones relevant to this course will be described below.



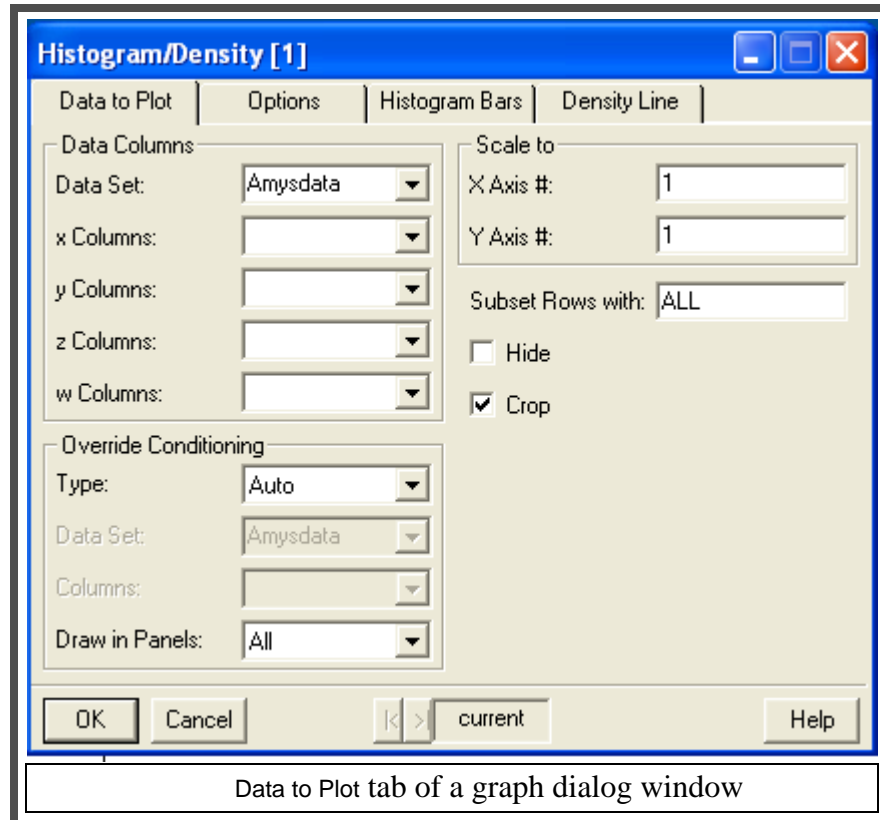
## Specifying the graph

If any columns are currently selected in the active data set, the plot will be created immediately using these selected columns, once you click OK in the Insert Graph dialog window or on one of the icons in the Plots2D selection palette.

If no columns are selected when you request a plot by one of the two routes above, a dialog such as below opens up for you to specify the plot you want. Some of the parts of this dialog are specific to the plot type, and will be described in the sections below for the various plots.

The Data to Plot tab, however, is the same for most if not all plot types, and will be described here.

- **Data set**  
The pull-down list for Data Set: allows any data set in the “project” (i.e. the current folder) to be selected.
- **Columns**  
The x Columns: and y Columns: selections are just what they say they are. The **z** and **w** columns have various uses; for instance, the z Columns: can be used to



specify a categorical variable defining groups to be shown by different symbols in a scatterplot.

- **Conditioning**

**Conditioning** is discussed below; the options in this box are to override the conditioning specified in the Multipanel dialog, and it is unlikely you will want to do this.

- **Scales**

The **Scale to** boxes allow multiple axes to be created, for instance two Y axes with different scaling, to overlay scatterplots of two variables, can be created by changing one to use **Y Axis # 2**.

- **Subsetting**

To show only some of the observations in the graph, create a column indicating the rows to be used, as described under **Subset: in Data set manipulations** above, and specify that column in **Subset Rows with:**.

---

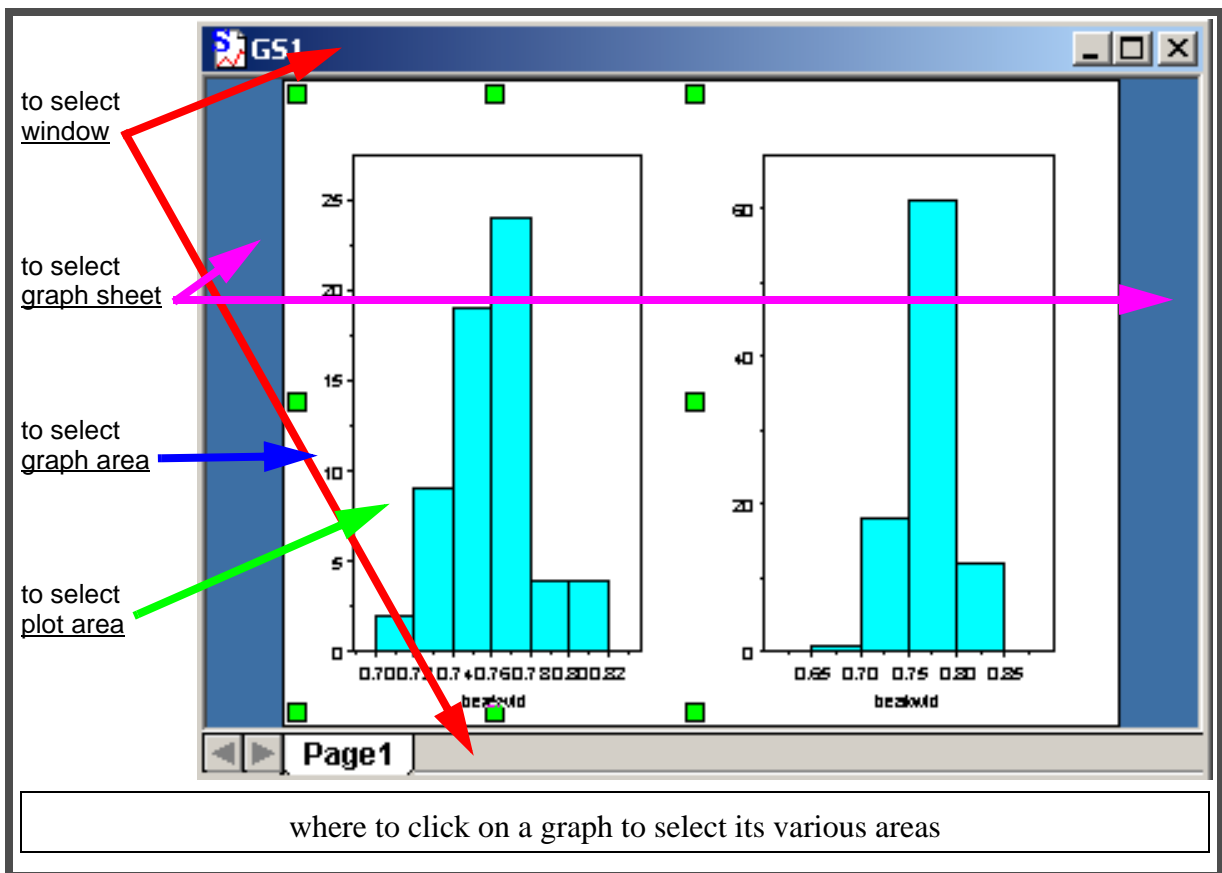
## Multiple graphs

Two or more graphs can be combined in one graph window, overlaid or in separate panels. Before explaining how this is done, the parts of a graph window and how to select among them need to be described.

## Window areas

Four kinds of area are defined in a graph window, in a nested hierarchy. Starting with the innermost (smallest) component, they are:

- plot area  
A “plot area” is the area within the axes of a plot. A plot area is selected by clicking within it; this will be indicated by green re-sizing knobs on the plot borders (similar to those in the example below, but on the corners of the axes.)
- graph area  
The “graph area” is the portion of the white page containing a plot with its axes, labels, etc. A graph area is selected by clicking within it but outside the plot area; this is shown by the green re-sizing knobs around the margin of the area, as in the example below.
- graph sheet  
A “graph sheet” is everything on one page. To select a sheet, click on the colored area of the page (but inside the frame of the window) No re-sizing knobs will appear.
- Window  
The graph window includes everything, and may have multiple pages. To select it, click on the bars across the top and bottom of the window.



## Overlaid plots

Overlaying graphs can be useful for showing two or more compatible variables or relationships. For instance, you may want to simultaneously plot two response variables against the same explanatory variable, or to the relationship between two variables across different groups.

To overlay plots:

1. Create one of them.
2. Select its “graph area” or “plot area” as described above.
3. either:
  - hold down the **shift** key while creating the second plot, e.g. while clicking a button on the Plots2D selection matrix
 or
  - use Insert ⇒ Plot... to open an Insert Plot dialog window exactly like the Insert Graph dialog described above under **Creating Graphs**.

In some cases, including scatterplots, multiple Y variables can be selected, prior to requesting the plot, which will cause all of them to be plotted in one plot.

## Multiple panels

It can be useful to have two or more graphs in a single window, perhaps with the same axis scaling, in order to compare them without the confusion overlaid plots can create. It also may be useful simply to group related graphs in one window to help keep track of them. Such multi-part graph windows can be created in two ways: by simply adding a new graph to an existing window, or by “conditioning” a graph on some other variable.

### Adding graphs

Adding one or more graphs to a graph window puts them in separate “graph areas.” To do this, select the “graph sheet” (not graph or plot area). Then either

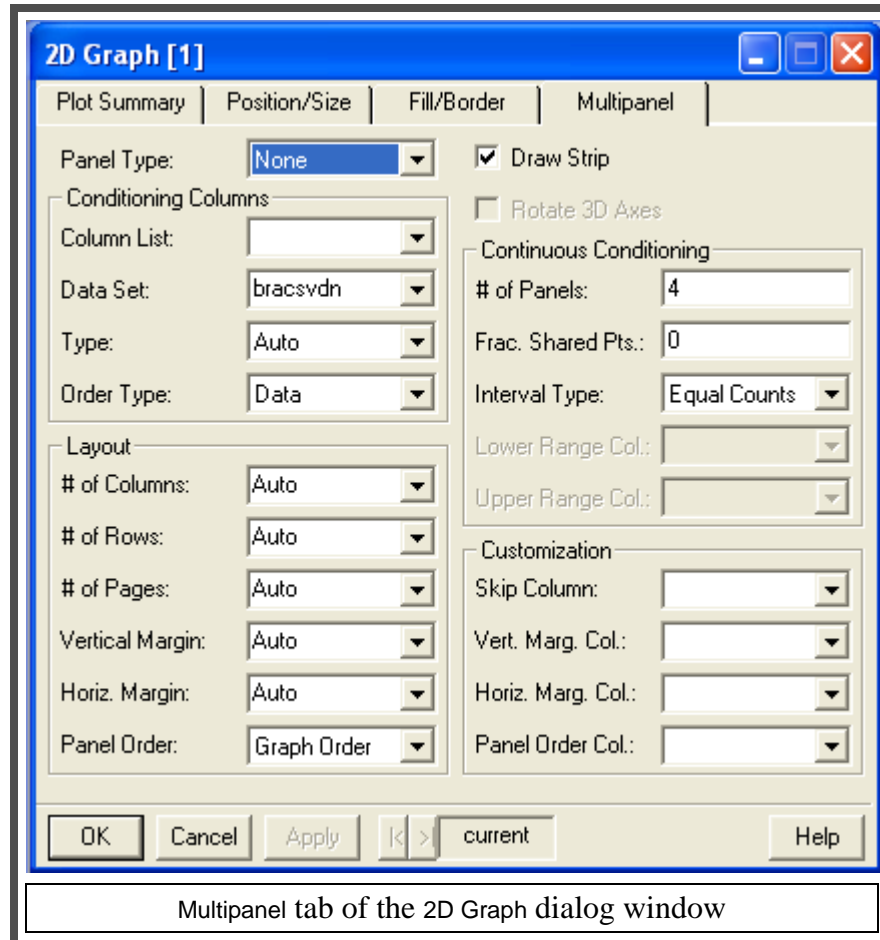
- hold down the **shift** key while creating the second plot
- or
- use Insert ⇒ Graph... to open the Insert Graph dialog.

### Conditioning

“Conditioning” a graph will split it into multiple panels based on values of a specified column (not one of the ones directly plotted). The panels will be within the same “graph area” (apart from perhaps spilling onto more than one page). The “**conditioning**” variable can be categorical, or quantitative; the range of a quantitative variable is divided into intervals defining the panels, and you can control how this is done.

To create a conditioned graph, first create the basic graph, without any conditioning. Then either

- select the conditioning variable in the data sheet and drag it onto the upper part of the graph
- or
- right-click on the graph area or plot area and select Multipanel... This will open the dialog window below. Change Panel Type: to Condition, then select the variable(s)



to condition on in the Column List: box (and if necessary, change the Data Set:). Various other parts of this dialog allow you to alter the number and/or layout of the panels. By default quantitative conditioning variables are divided into intervals having the same number of observations, but this can be changed in the Interval Type: box; the other options are to use equal ranges or to use another column to define the intervals.

Multipaneling can be modified, or eliminated, by selecting the graph, right-clicking, going to the Multipanel tab, and altering settings as desired.

---

## Editing graphs

Right clicking or double-left-clicking on a graph opens the same dialog window as used to create the graph, in which any desired changes can be made. (The menu that pops up lists each of the tabs of the dialog separately, e.g. Data to Plot..., Options..., Histogram Bars... and Density Line.... Selecting any one of these, however, gives access to all of them, via the tabs on the dialog window.)

### Changing axes

An axis can be modified by right clicking (or double clicking) on the axis. This brings up a dialog window with tabs. The range as well as the positions of the major ticks and labels are specified on the Range tab. The Display/Scale tab allows a choice between linear, log, ln, or probability scaling, without having to transform the variable. Other parts of the Display / Scale tab, and the Grids/Ticks tab, allow various changes in the appearance of the axis.

### Changing text

Any of the default text, including the axis tick labels, can be modified by right- or double-clicking on it. The actual text of graph titles and axis labels, as well as the font, can be modified. The format and font of tick labels can also be modified in this way; their locations, however, are determined by the axis specifications.

### Annotation

S-Plus has extensive tools for adding annotation to graphs, e.g. text, arrows, shapes, and various symbols. The best way to access these to activate the Graph, Annotation, and/or Graph Tools toolbars.

To change which tool bars are shown, use

**View ⇒ Toolbars...**

### Graph options

S-Plus assigns symbol style and color, line style and color, and other such plot characteristics, in designated sequences. These can be modified from the defaults using

**Options ⇒ Graph Styles ⇒ Color... (or Black and White...)**

This opens the Color Style dialog window, with tabs for the various characteristics that can be changed. When multiple plots are overlaid, or a plot distinguishes different groups, the styles, line types, colors, etc., are used in the sequences specified on the pertinent tabs.

## Saving and exporting graphs

### Saving graphs

Unlike data sets, graphs are not automatically saved, but when you close a graph window or exit S-Plus, you will be asked whether you want to save the graph(s). To save the active graph, use

**File ⇒ Save...**

(The filename extension for S-Plus graphs is **.sgr**.) Graphs can be saved to any directory, and opened later by

**File ⇒ Open...**

### Exporting graphs

Graphs saved as above can only be opened in S-Plus. To export a graph in a format usable by some other program, use

**File ⇒ Export Graph...**

and select the desired format.

For some applications, direct cut-and-(special) paste will work:

**Edit ⇒ Copy Graph Sheet Page**

followed by, in the target application

**Edit ⇒ Paste Special...**

and select Embedded SPLUS GraphSheet Object

Or, export graph directly to PowerPoint, using


**File ⇒ Create PowerPoint Presentation...**

# DESCRIBING DISTRIBUTIONS

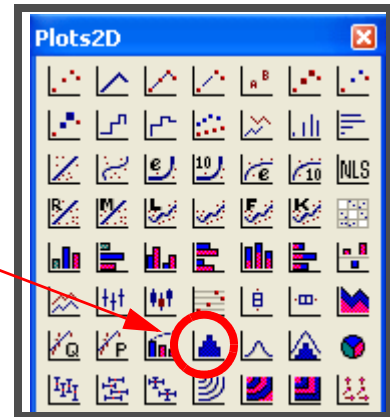
## Plots of distributions

Among the large array of graph types available in S-Plus are many which represent the distribution of a single variable. The most useful of these are the basics: histograms, boxplots, NQQ plots, and bar charts.

### Histograms

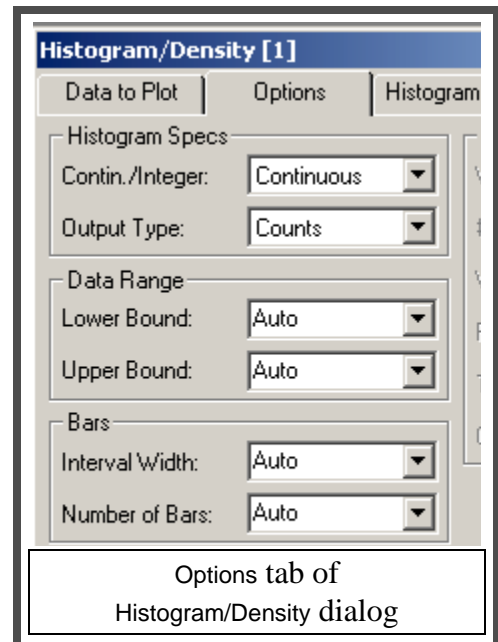
Select Histogram (X) as the plot type in the Insert Graph dialog, or click the  button in the 2-D Plot selection palette.

The variable whose distribution is plotted is considered the x Column. Either select the column before requesting the histogram, or specify it in the Data to Plot tab of the Histogram/Density dialog window.



To change the “binning” go to the Options tab (shown to the right) of the Histogram / Density dialog, either when first creating the graph or by right-clicking on the graph. It seems to work best to specify all four pertinent values: Lower Bound, Upper Bound, Interval Width, and Number of Bars.

Note that the horizontal axis tick marks and labels by default are put at the cut-points of the bins. The axis tick marks, however, can be edited independently of the binning of the histogram, so you can specify major ticks corresponding to bin mid-points. (Editing axes was described above in the **Editing graphs** section of **S-PLUS GRAPHING: GENERAL FEATURES**.)




Output Type lets you choose the scaling of the vertical axis.

## Comparing histograms

Histograms for two or more samples can be either overlaid on each other or shown as panels in a single graph window. This is done by creating them as separate plots then overlaying or panelling them as described above (**Multiple graphs in S-PLUS GRAPHING: GENERAL FEATURES**). To do this, the samples must either be in separate columns or be identified by separate “subset” columns (and in the latter case the graphs will have to specify the different columns in the Subset rows with: box).

To make overlaid histograms visible, it usually is necessary to edit the histogram bars to set the Fill Color to Transparent. Readability can be further enhanced by changing the Color and Weight of the histogram bars, and perhaps adding different Fill Patterns.

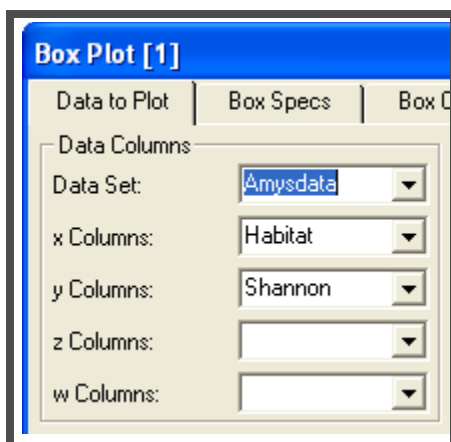
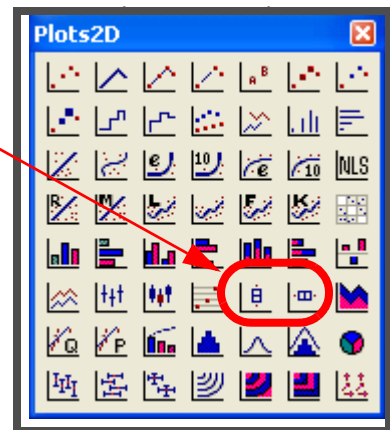
## Boxplots

Select Box Plot (x, grouping optional) as the plot type in the Insert Graph dialog, or click one of the  buttons in the 2-D Plot selection palette (the left button is for a vertical boxplot, the other for a horizontal boxplot).

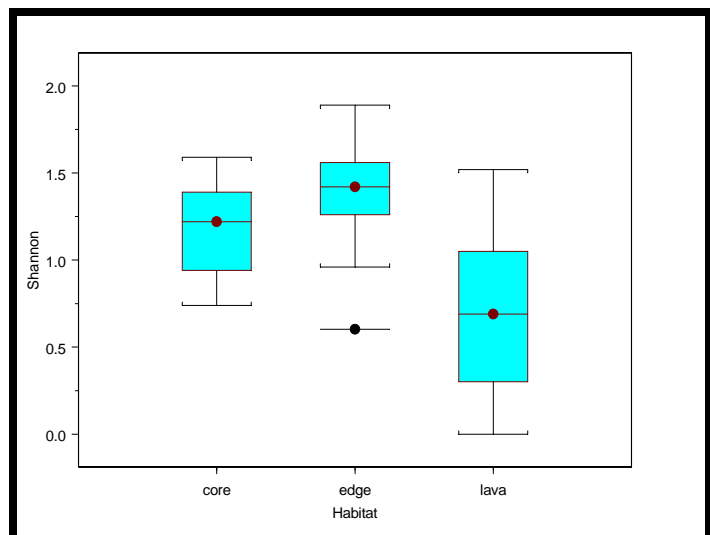
For a simple boxplot of one distribution, simply specify the variable as the y column (or as the x column if making a horizontal boxplot).

## Comparing boxplots

It is best to have the data in stacked layout (i.e. all observations of the variable of interest in one column, and group identifiers in another). As shown in the dialog window below, simply specify the grouping variable as the x column and the focal variable as the y column (or vice versa for a horizontal




part of Box Plot dialog window, and boxplots with grouping (x) variable

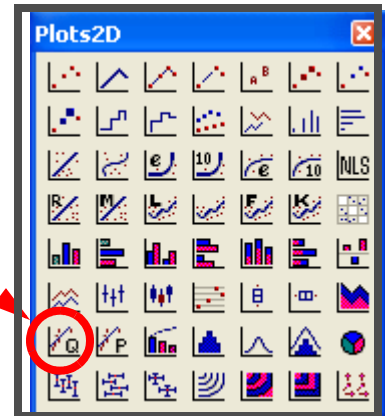


plot). The result will be as shown here:

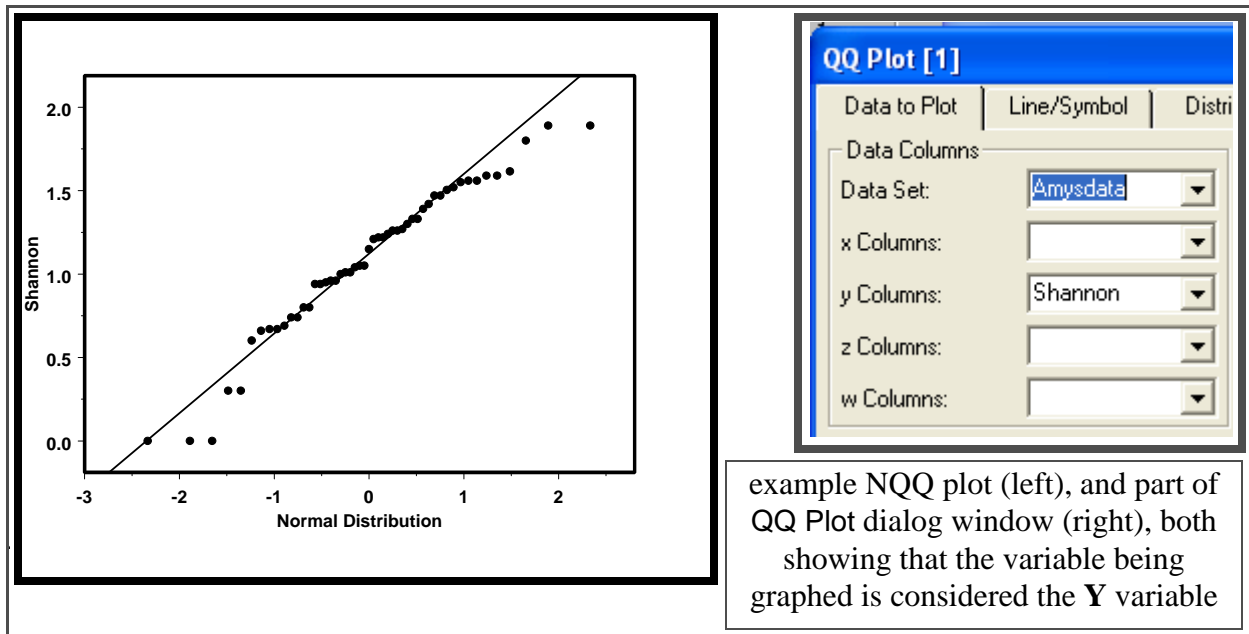
## NQQ plots

Select QQ Normal with Line (x) as the plot type in the Insert Graph dialog, or click the  button in the 2-D Plot selection palette.

In S-Plus, QQ (quantile-quantile) plots (see example below) put the observed values on the Y axis and the standard-normal scores on the X axis, the reverse of most NQQ plots. The variable therefore needs to be specified as the y column, with the x column blank, as shown in the dialog window below.



These plots implicitly standardize the variable: the reference line has a slope of one standard deviation and Y-intercept equal to the sample mean.





## Bar charts

To make a bar chart for a categorical variable, you specify a column containing the levels of the variable and another column containing their frequencies. You therefore may need to tabulate the data first, as follows:

**Data ⇒ Tabulate...**

and specify a name for a data set to put the results in. This data set is what you'll use to make the bar chart.

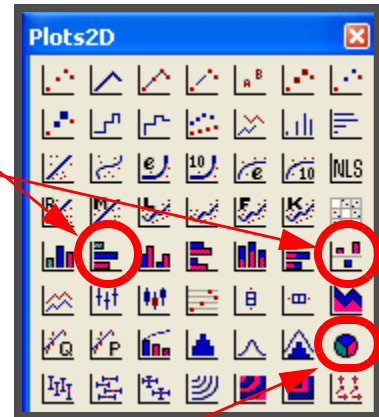
To make the chart, select either Bar with Base at Zero (x,y) or Bar, Horiz. (x,y) as the plot type in the Insert Graph dialog, or click either the  button or the  button in the 2-D Plot selection palette.

- Bar with Base at Zero 

creates a chart with vertical bars, based at 0 as is appropriate for frequencies. Specify the column containing the categories as the **x** column and the column containing the frequencies as the **y** column.


- Bar, Horiz. (x,y) 

creates a chart with horizontal bars. In this case, specify the column containing the categories as the **y** column and the column containing the frequencies as the **x** column.



## Pie charts

Pie charts are alternatives to bar charts, for showing relative frequencies of a categorical variable. Like bar charts, they also require the frequencies to be tabulated already.

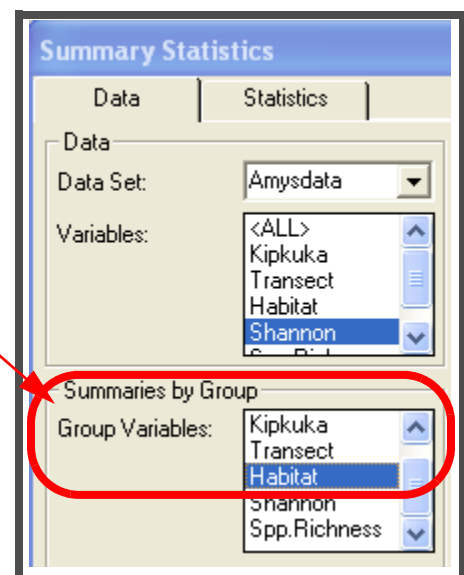
Select the  button in the 2-D Plot selection palette (interestingly, bar charts are not on the list of plot types in the Insert Graph dialog). Specify the columns as for a vertical bar chart.

## Descriptive statistics

### Statistics ⇒ Data Summaries ⇒ Summary Statistics...

then specify the variable(s) for which you want the statistics. To obtain descriptive statistics for subsets of the data, enter the variable(s) defining the groups (subsets) for which you want the statistics calculated as the Group Variables: in the Summaries by Group part of the dialog.

The default output will be the mean, standard deviation, median, upper and lower quartiles, minimum, maximum, the number of observations, and the number of observations with missing values, for each column specified. The selection of statistics can be changed on the Statistics tab of the Summary Statistics dialog.




## DESCRIBING RELATIONSHIPS

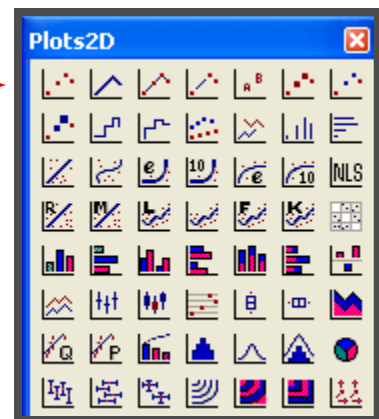
This section focuses on relationships between **two quantitative variables**. Describing the relationship between **one quantitative** variable and one or more **categorical** variables amounts simply to comparing (between levels of the categorical variable) descriptions of the distribution of the quantitative variable, using methods presented in the “Comparing distributions” parts of the preceding section. Describing relationships between **two categorical variables** is covered at the end of this section.

---

### Scatterplots

Select Scatter Plot (x, y1, y2, ...) as the plot type in the Insert Graph dialog, or click the  button in the 2-D Plot selection palette. Specify the X and Y columns, and optionally use the other tabs in the dialog to make changes in the appearance of the plot.

If two or more columns are already selected in the data set when the plot is requested, the first one selected is used as the X variable, and all the others as Y variables.



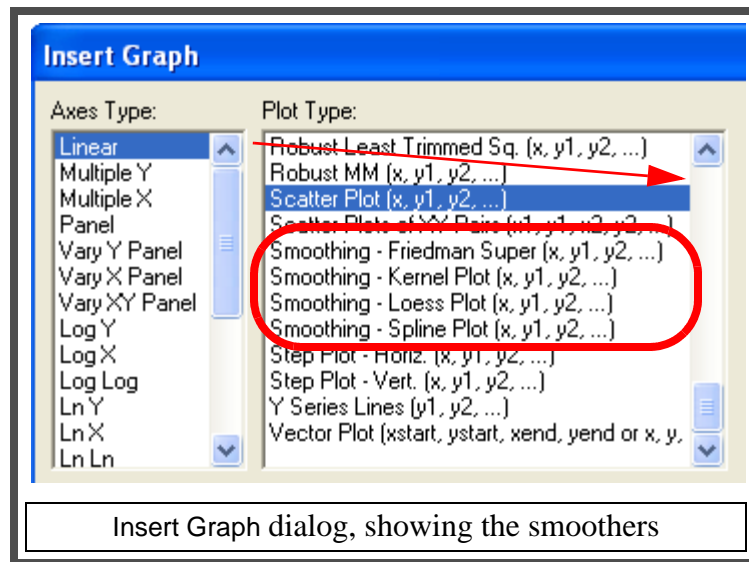
### Smoothers

S-Plus has four scatterplot smoothers: Friedman’s super-smoother, kernel, Loess (same as Minitab’s LOWESS) and spline. I don’t know of any particular advantages of any of these over any others.

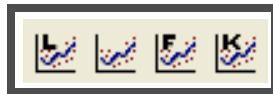
To create a scatterplot with a smoother, either

- select Smoothing - ... (x, y1, y2, ...) as the plot type (with one of the smoother types substituted for the ...) in the Insert Graph dialog, shown below,

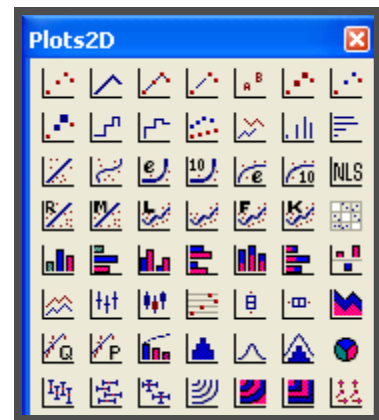
Or



- click one of the buttons below in the 2-D Plot selection palette (left to right: Loess, spline, Friedman, Kernel).



Specify the X and Y columns as for a scatterplot. Parameters for the smoothers, determining how much smoothing they do, can be adjusted on the Smooth/Sort... tab.



A smoother also can be added to an existing scatterplot by selecting (left clicking in) the plot, right clicking, and going to the Smooth/Sort... tab, similarly to the method described below for adding a fitted regression line to a plot. On this tab you select which type of smoother you want, and if desired, alter its default smoothing parameters. You also must go to the Line tab of the Line/Scatter Plot dialog and specify a line type other than None; otherwise, the smoother will not be drawn.

## Showing groups

How to create a scatterplot in which different groups are distinguished by different symbols (and with separate smoothers or regression fits, if these are desired) depends on how the data are organized.

## Unstacked

- One X

If the groups are unstacked with a single column for the X variable and separate columns for the Y variable for the groups, simply creating a scatterplot selecting all the columns will produce overlaid plots with different symbols.

- XY pairs


If the groups are unstacked with separate X and Y columns for each group, either

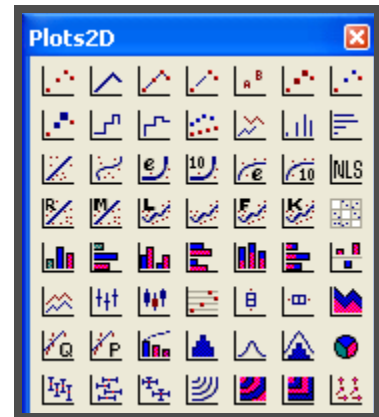
- create the plot for one group then overlay the plots for the other groups, as described above (**Multiple graphs in S-PLUS GRAPHING: GENERAL FEATURES**)

or

- select

Scatter Plots of XY Pairs ( $x_1, y_1, x_2, y_2, \dots$ ) as the plot type in the Insert Graph dialog, or click




the  button in the 2-D Plot selection palette. (If selecting columns before clicking the selection button, select them in the order  $X_1, Y_1, X_2, Y_2$ , etc.)

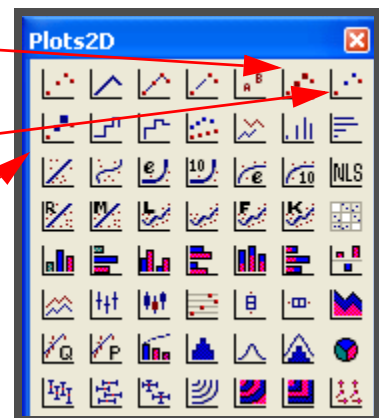


After making the plot by either of these methods, the symbols for any of the response variables can be modified by selecting its points, right clicking, and going to the Symbols tab.

## Stacked

If the data are stacked with the explanatory variable in one column, the response variable (all groups) in another column, and one or more columns with group identifiers, the groups can be identified by different sizes and/or colors of symbols, as well as by different types of symbols.

- Bubble Plot ( $x, y, \text{size}$ )  varies the size of the points according to the value of a third variable
- Color Plot ( $x, y, \text{color}$ )  varies the color of the points according to the value of a third variable, and
- Bubble Color Plot ( $x, y_1, \text{size}, \text{color}$ )  varies the size of the points according to the value of a third variable and the color of the points according to the value of a fourth variable; the same variable can be specified for both size and color if desired.



These plot types can be selected in the Insert Graph dialog, or by clicking on the appropriate buttons in the 2-D Plot selection palette as shown above. When selected in

**Line/Scatter Plot [1]**

Data to Plot | Line | Syn

Data Columns

Data Set: classdata

x Columns: ht

y Columns: age

z Columns: sex

w Columns: college

**Line/Scatter Plot [1]**

Data to Plot | Line | Symbol | Vary Symbols | Smooth/Sort

Vary Size By: z Column

Vary Color By: w Column

Vary Style By: None

Vary Symbol Size

Minimum Height: 0.08

Maximum Height: 0.25

Vary Symbol Color

Colors To Use: Range

Number of Colors: 16

Minimum Color: Blue

Maximum Color: Lt Cyan

Column:

**Bubble Color Plot example.**

Above left: Data to Plot tab of Line/Scatter Plot dialog window showing selection of **x**, **y**, **z**, and **w** variables.

Above right: Vary Symbols tab of Line/Scatter Plot dialog window showing the use of the **z** and **w** variables to determine symbol size and color.

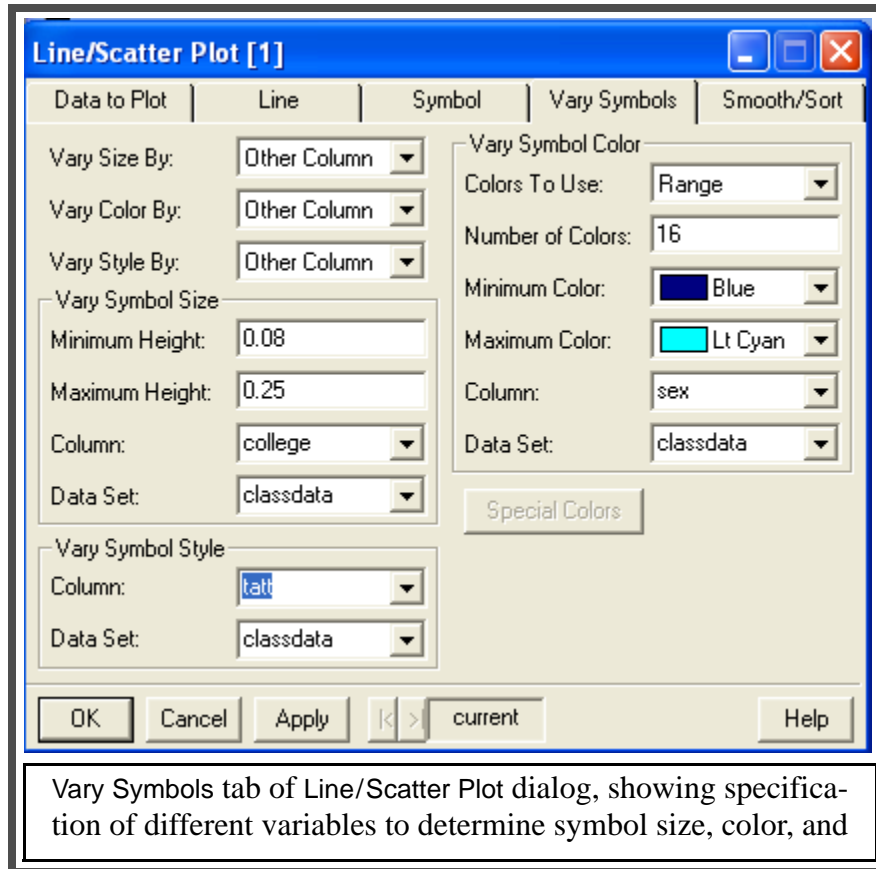
Right: Example of plot produced by the dialogs above (except with symbol type changed to solid circle to enhance visibility in this document).

this way the variable which determines symbol size or color is considered the **z** variable (or **z** and **w** for bubble-color plots). This is shown in the dialog windows below. (Note that these variables can be quantitative. In this case the plot shows the relationship among the three or four variables, as a two-dimensional alternative to three-dimensional scatterplots.)

An example of a bubble-color plot is shown below the dialog examples on the next page (with the symbol changed to a solid circle to enhance visibility in this document).

Alternatively, grouping variables can be added to an existing scatterplot by selecting it, right clicking, and going to the Vary Symbols... tab. On this tab variables can be specified to determine the symbol size, color, and/or style. These can be the **z** or **w** columns specified on the Data to Plot tab, or can be specified by selecting Other Column

as the Vary ... By: column, and then designating the variable in the Column: box in the appropriate section of the dialog window, as shown on the next page.



## Styles

The range of symbol sizes and/or the number and range of colors used to show the grouping variables can be altered on the Vary Symbols... tab, when creating the plot or later, as shown as part of the bubble-color-style specification example below.

A tip: When the grouping variable determining the color is considered by S-Plus to be numeric (rather than a “factor”), it is helpful to set the Number of Colors equal to the number of groups, and select a color range (Minimum Color and Maximum Color) to give a clear difference between groups. For better control of the colors, have the grouping variable be of type “factor.”

The sequences of symbol types and colors can be modified as describe earlier (**Graph options** in **S-PLUS GRAPHING: GENERAL FEATURES**).

## Numerical summaries

### Correlation coefficient


**Statistics** ⇒ **Data Summaries** ⇒ **Correlations...**

then specify the two (or more) variables.

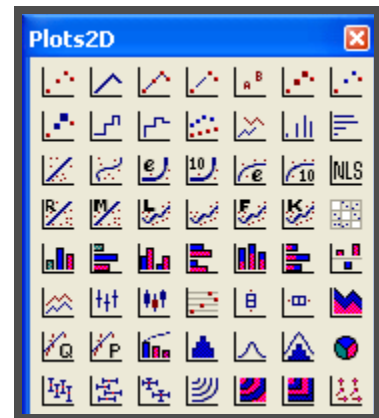
### Regression

As a descriptive tool, a regression can be obtained as a fitted line on a scatterplot or as a full regression analysis.

#### Fitted-line plot

Select Fit - Linear Least Squares Plot (x, y1, y2, ...) as the plot type in the Insert Graph dialog, or click the  button in the 2-D Plot selection palette. Specify the X and Y columns, and optionally use the other tabs in the dialog to make changes in the appearance of the plot.

A fitted regression line also can be added to an existing scatterplot by selecting (left clicking in) the plot, right clicking, going to the Smooth/Sort... tab, and choosing Least Squares as the Smoothing Type. You also must go to the Line tab of the Line/Scatter Plot dialog and specify a line type other than None; otherwise, the regression line will not be drawn.

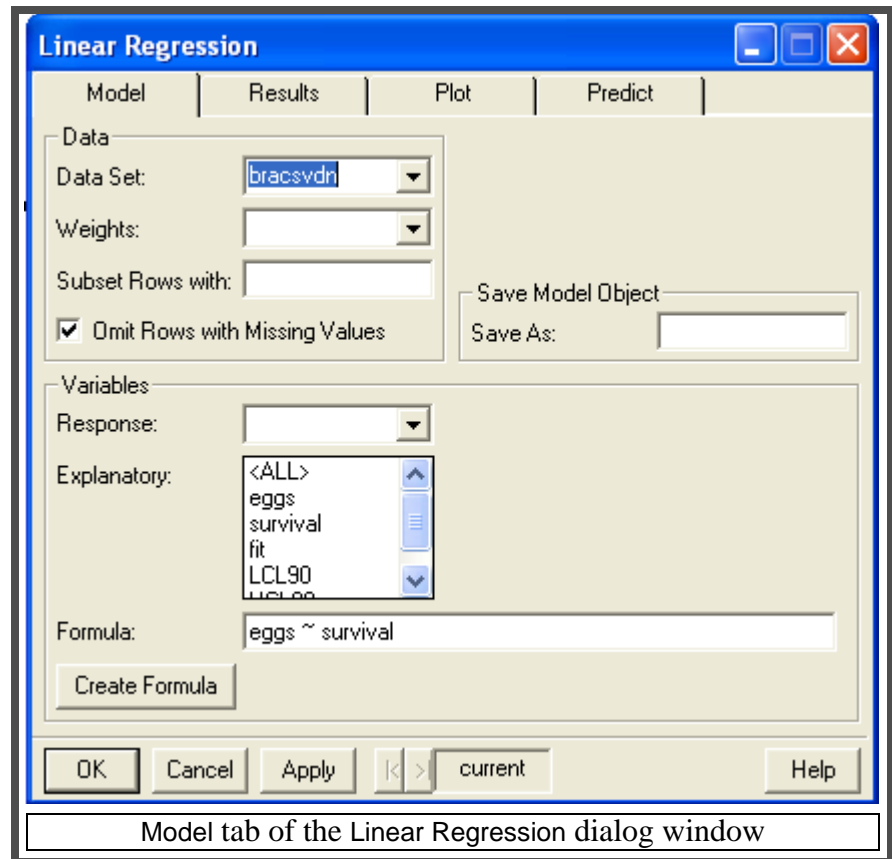


#### Regression model

**Statistics** ⇒ **Regression** ⇒ **Linear...**

opens the dialog window below.

For simple linear regression (i.e. only one explanatory variable), all that needs to be specified in this dialog is the Response and the single Explanatory in the Variables section of the window; selecting them will automatically produce the Formula.



### Residual plots

In the Linear Regression dialog window above, the Plot tab provides a variety of diagnostic plots, including a scatterplot of Residuals vs Fit and a Residuals Normal QQ. A smoother can be put on the residuals vs fits plot, and the most extreme residuals can be individually identified (the defaults are to include the smoother and to flag the 3 most extreme residuals).

The Results tab of the Linear Regression dialog window allows you to store the residuals and fits in a data set. These can then be used for additional diagnostic plots or other analyses.

---

## Categorical variables

As with simple bar charts for a single categorical variable, to display the relationship between two categorical variables by a stacked or grouped bar charts requires that the frequencies first be tabulated.

### Tabulating

**Data ⇒ Tabulate...**

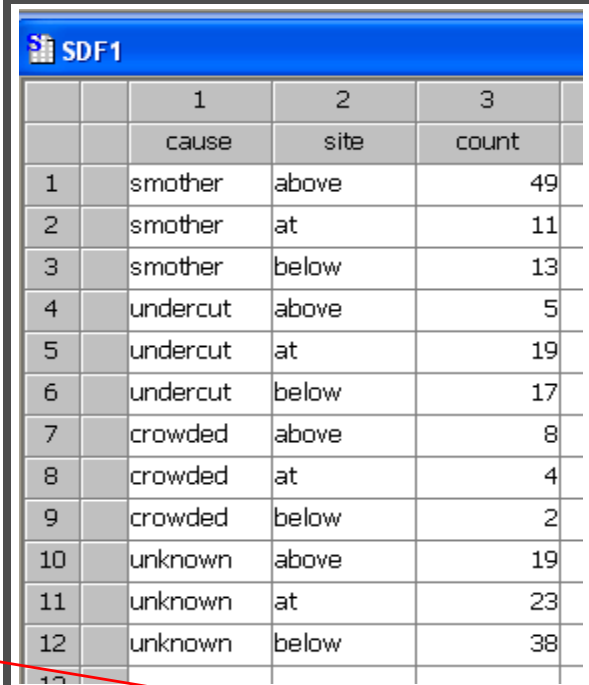
Select the two categorical variables in the pull-down list of Variables:. Also give a name for the data set to contain the cross-tabulation.

The tabulation will create a data set in stacked format, as shown in the example to the right: there will be columns for the two categorical variables and a column for the frequencies of the combinations of the categorical variables.

Unfortunately, to create stacked or grouped bar charts, the data must be laid out like a contingency table (i.e. unstacked), with one column containing one of the categorical variables, and columns of frequencies for each of the levels of the other variable. This layout can be produced from the stacked tabulation either by

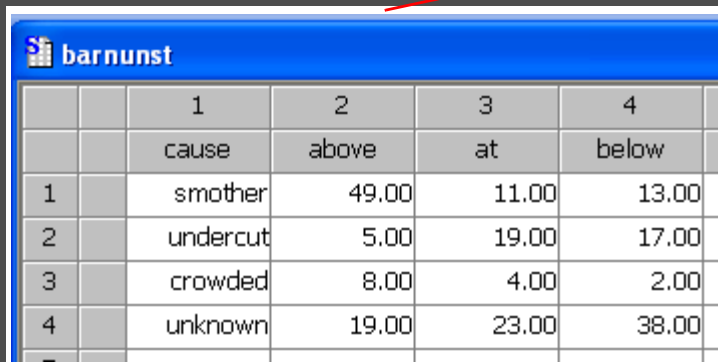
**Data ⇒ Restructure ⇒  
Unstack...**

or in most cases more simply by cut-and-paste. The resulting data set needs to be as below.



		1	2	3
		cause	site	count
1		smother	above	49
2		smother	at	11
3		smother	below	13
4		undercut	above	5
5		undercut	at	19
6		undercut	below	17
7		crowded	above	8
8		crowded	at	4
9		crowded	below	2
10		unknown	above	19
11		unknown	at	23
12		unknown	below	38

example of datasheet produced by tabulation







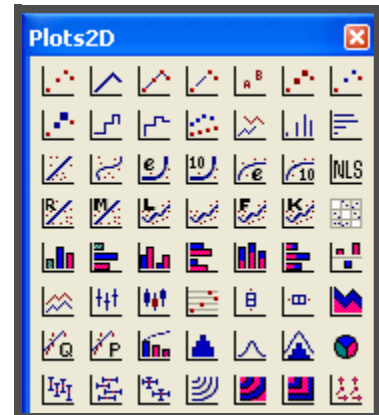
		1	2	3	4
		cause	above	at	below
1		smother	49.00	11.00	13.00
2		undercut	5.00	19.00	17.00
3		crowded	8.00	4.00	2.00
4		unknown	19.00	23.00	38.00

tabulated data in unstacked "contingency table" layout

## Stacked or grouped bar charts

With the data arranged as above, a stacked or grouped bar chart can be created by selecting one of the following in the Insert Graph dialog or clicking the corresponding button on the 2-D Plot selection palette:

- Bar - Grouped ( $x, y_1 \dots y_n$ ) 
- Bar - Stacked ( $x, y_1 \dots y_n$ ) 
- Bar, Horiz. - Grouped ( $x_1 \dots x_n, y$ ) 
- Bar, Horiz. - Stacked ( $x_1 \dots x_n, y$ ) 



For vertical bars, the  $x$  column will be the one containing the levels of one of the variables (cause, in the example above), and the two or more columns containing the counts (above, at and below) will all be  $y$  columns. In a stacked chart there will be a bar for each level of the  $x$  variable, each with segments for each of the  $y$  columns. In a grouped chart there will be a cluster of bars for each level of the  $x$  variable, each cluster containing bars for each of the  $y$  columns.

For horizontal plots, the  $x$  and  $y$  columns are reversed (i.e. the single  $y$  column will define the bars or groups, and the several  $x$  columns give the counts for the segments or within-group bars)

## ONE-SAMPLE PROCEDURES

---

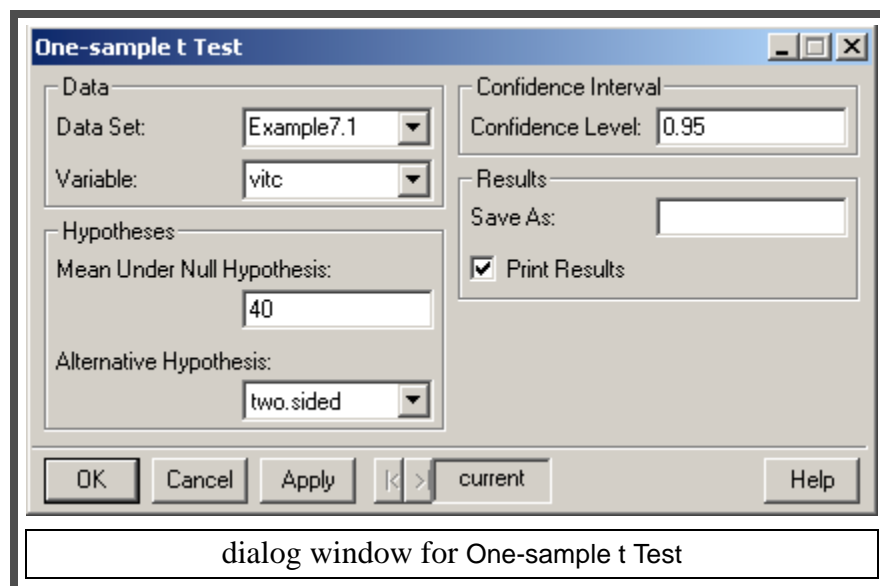
### *t* procedures

Statistics ⇒ Compare Samples ⇒ One Sample ⇒ t test...

#### Dialog

This procedure provides both a hypothesis test and a confidence interval. NOTE: If a one-sided test is requested, the confidence “interval” will actually be a confidence “bound,” i.e. it will have only one end.

Specifying the *t* procedure is straightforward:



- select the data set and variable
- set  $\mu_0$ , if not equal to 0
- request a one-sided test if desired
- set the confidence level, if not 0.95.
- optionally, name a data set to put the results in and/or turn off printing of the results.

## Output

The first part of the output (the only part, if the Resampling Library is not loaded), also is straightforward:

```

Usual Students-t inferences (assume zero skewness):

      One-sample t-Test

data:  vitc in Example7.1
t = -6.883, df = 7, p-value = 0.0002
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 16.48795 28.51205
sample estimates:
 mean of x
      22.5
  
```

example output from one-sample *t* test

If the resampling library is loaded, the output also gives the *P*-value and CI calculated by a kind of resampling method called “saddlepoint” inferences, as in the example below. The parametric *t* procedures are somewhat sensitive to asymmetry, especially for one-sided tests; these saddlepoint results are robust to asymmetry. This is a very new method which I do not understand and will not teach, so be prudent in using and reporting it.

```

Saddlepoint inferences (allow non-zero skewness):
      method = saddlepoint
p-value = 0.0078
95 percent confidence interval:
 17.77263 26.50299
  
```

additional output from one-sample *t* test if the Resampling

## Resampling *t* procedures

**Statistics ⇒ Compare Samples ⇒ One Sample ⇒ t test/Resample...**

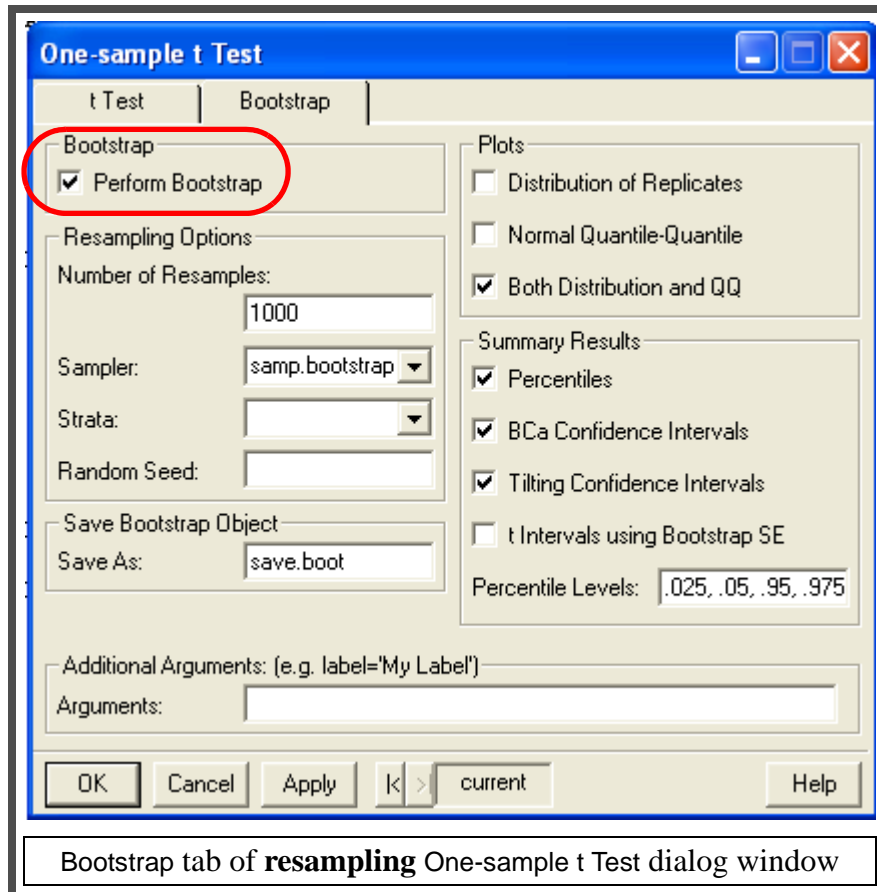
or

**Statistics ⇒ Resamples ⇒ One-sample t...**

## Dialogs

The first tab (t Test) of the dialog window is the same as for the parametric  $t$  procedures above, with the addition of a box in which you can specify a fraction (%) of the smallest and the largest observations to exclude for the trimmed mean.

The second tab (Bootstrap) is used to request and specify bootstrap estimation of the population mean (see example below). The only necessary part of this dialog is to



check the Perform Bootstrap checkbox. The two plots available (selected in the Plots section; the default is to do both) are a histogram and a NQQ plot of the distribution of means for the replicate re-samples. The Summary Results section allows selection of various ways of calculating the CI; the “Tilting” method is supposed to be the most efficient but is very new and certainly not well known by biologists. You also can choose what confidence levels you want.

## Output

The output (see example below) first gives the details of how the procedure was specified, and then gives the bootstrap estimate of bias of the mean. This bias estimate comes from comparing the observed mean to the mean of the resample means.

```

*** Bootstrap Results ***
Call:
bootstrap(data = Example7.1$vitc, statistic =
Number of Replications: 1000

Summary Statistics:
      Observed Mean      Bias      SE
mean      22.5      22.5      -0.004625      2.443
Percentiles:
      2.5%      5%      95%      97.5%
mean 17.50625 18.25 26.625 27.24688
BCa Confidence Intervals:
      2.5% 5% 95% 97.5%
mean 17.39111 18 26.5 27.125
Tilting Confidence Intervals:
      2.5%      5%      95%      97.5%
mean 17.33422 18.16515 26.24398 26.89661

```

example output of bootstrap estimation  
in resampling one-sample  $t$  test

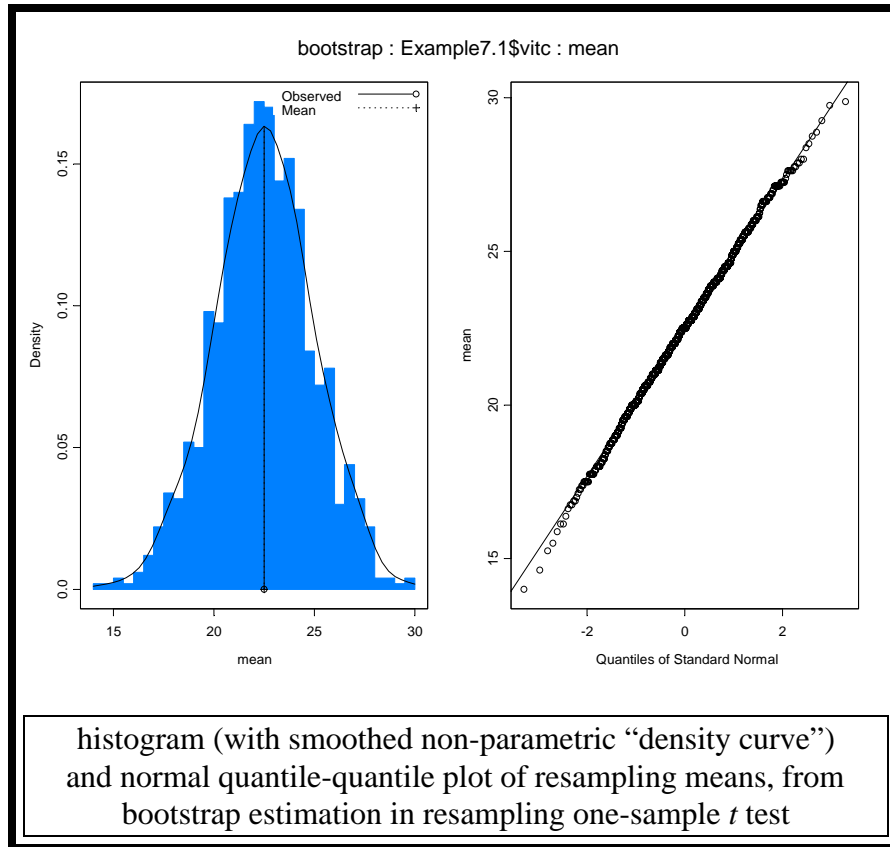
Then the several types of CIs (at the several confidence levels) are listed. In this example, the 95% CIs would be:

- (17.51, 27.25) by the percentile method
- (17.39, 27.125) by the BCa (accelerated bias corrected) method
- (17.33, 26.90) by the “tilting” method.

For comparison, the parametric  $t$  95% CI was (16.49, 28.51).

## Graphs

The default output includes the two plots of the distribution of re-sampled means shown below. For this example, the distribution is very close to normal, suggesting that the parametric  $t$  procedures are valid. In addition, the mean of the resample distribution is very close to the observed mean, indicating little bias (as already shown in the quantitative measure of bias in the output above).



## Signed-rank

**Statistics  $\Rightarrow$  Compare Samples  $\Rightarrow$  One Sample  $\Rightarrow$  Wilcoxon Signed Rank Test...**

Note that the signed-rank procedure in S-Plus only conducts hypothesis tests. It does not produce confidence intervals.

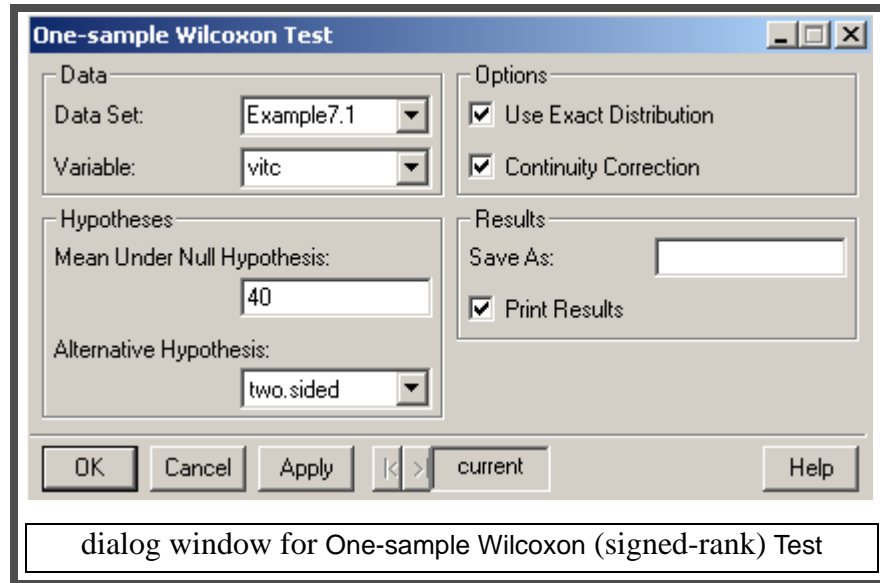
### Dialog

Specifying the signed-rank test is very similar to specifying a  $t$  test. One difference is that since no CI is produced, no confidence level is needed.

More substantial differences are that you can request two Options (the default being to use them both)

- Use Exact Distribution

This requests that the exact distribution of the test statistic (under  $H_0$ ), rather than the large-sample normal approximation, be used to determine the  $P$ -value. The exact calculation is not possible if there are ties in the data, and can be computer-intensive for large data sets.



- Continuity Correction

This option adjusts the  $z$  value used to get the  $P$ -value by the normal approximation, to account for using a continuous distribution to approximate a discrete one.

## Output

The output depends on whether the exact test was used. If it was, the test statistic is reported, along with the  $P$ -value.

```
Exact Wilcoxon signed-rank test
data: vitc in Example7.1
signed-rank statistic V = 12, n = 8, p-value = 0.4609
alternative hypothesis: true mu is not equal to 25
```

output of Wilcoxon signed-rank test when exact null distribution is used

If the exact test was not requested, or was not possible due to ties, the output reports the  $Z$  score (continuity-corrected if this was requested) and  $P$ -value. In either case the output also states  $H_a$ .

```
Wilcoxon signed-rank test
data: vitc in Example7.1
signed-rank normal statistic with correction Z = -2.4565, p-value = 0.014
alternative hypothesis: true mu is not equal to 40
```

example output of Wilcoxon signed-rank test when normal approximation is used

## Sign test

S-Plus does not explicitly provide the sign test, but it nonetheless can be done fairly easily.

You first need to know the count of observations greater than the hypothesized median, and the number of observations to use in the test (if you are excluding 0s). For small data sets these can simply be counted; it helps to sort the data first. For large data sets it is safer to have the software do this tabulation for you.

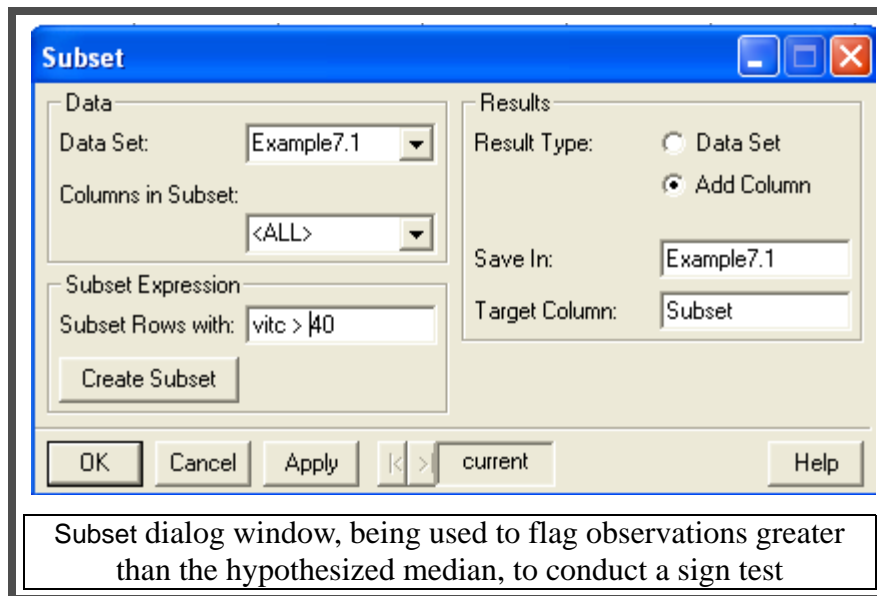
One approach is to use

**Data ⇒ Subset...**

to create a variable flagging whether the observation is or is not greater than the hypothesized median. For example, starting with the datasheet to the right, the Subset dialog below will produce an additional column in the datasheet, as shown to the right of the dialog.

	vitc
5	11.00
7	14.00
4	21.00
6	22.00
3	23.00
1	26.00
2	30.00
8	31.00

example data for sign test



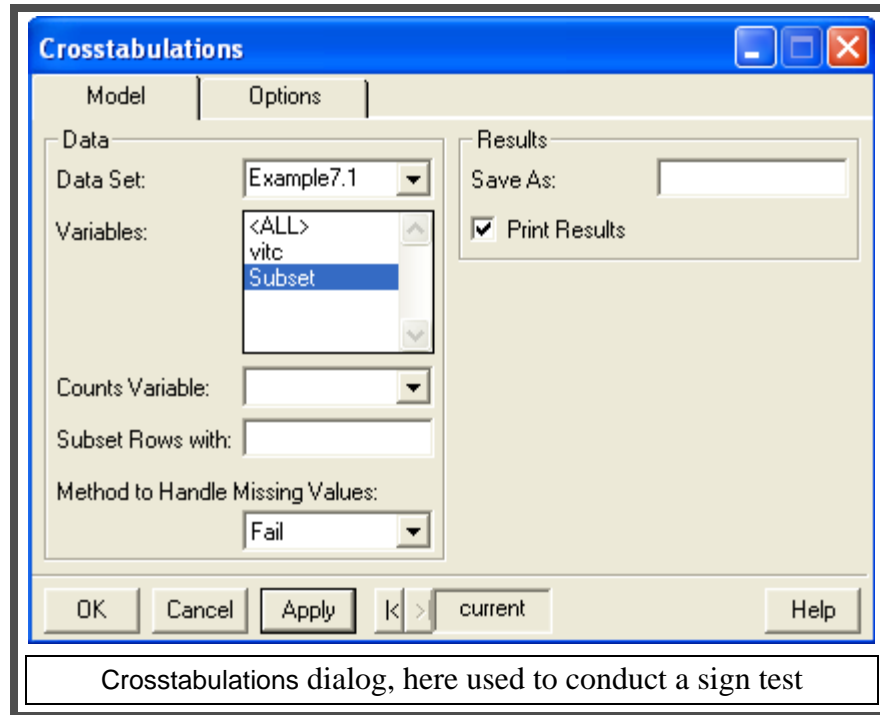
	vitc	Subset
5	11.00	F
7	14.00	F
4	21.00	F
6	22.00	F
3	23.00	F
1	26.00	F
2	30.00	F
8	31.00	F

example data for sign test, with Subset column flagging values greater than the hypothesized median

Then use

**Statistics ⇒ Data Summaries ⇒ Crosstabulations...**

to count the observations with the subsetting variable equal to True, as in the dialog window shown below. (The Options tab of this dialog window can be used to turn off most of the standard output, such as the chi-square test, which is not needed for the sign test.)



The cross tabulation produces simple results in the Report window, as in the example to the right. (In most cases the table will also have a row for TRUE, i.e. the count of observations with values greater than the hypothesized median used to define the Subset variable.)

Once you know the necessary counts, use

**Statistics** ⇒ **Compare Samples** ⇒  
**Counts and Proportions** ⇒  
**Binomial Test...**

In the resulting dialog box (next page), enter the count of values greater than the hypothesized median as the No. of Successes, and the number of observations to be used in the test as the No. of Trials. If a one-sided test is desired, select it in the Alternative Hypothesis box.

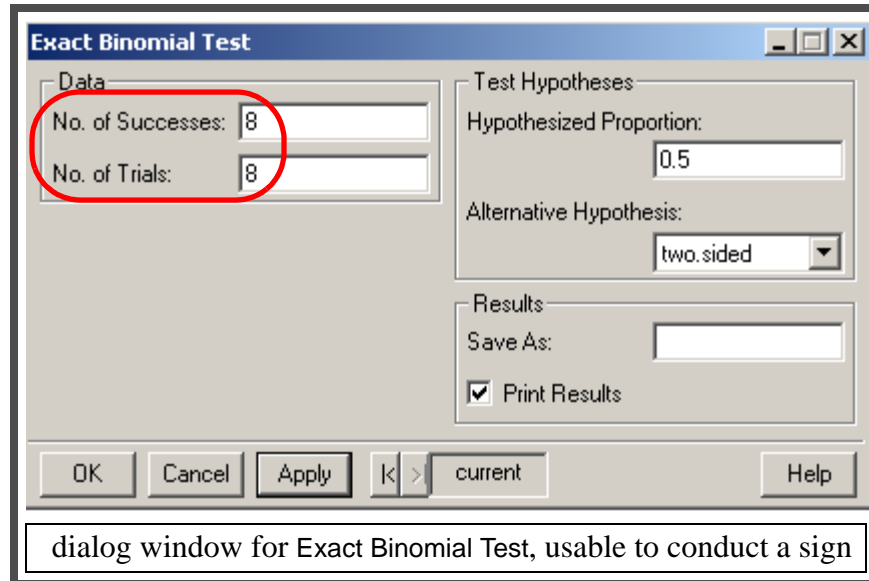
```

*** Crosstabulations ***
Call:
crosstabs(formula = ~ Subset,
           = na.fail, drop.unused.1)
8 cases in table
+-----+
|N      |
+-----+
Subset |
-----+-----+
FALSE |8      |8      |
      |      |11     |
-----+-----+
ColTot|8      |8      |
      |1      |      |
-----+-----+

```

example output from Crosstabula-

The only output (other than the data you gave) is the  $P$ -value, as shown on the next page, below the dialog window.



Exact binomial test

data: 8 out of 8

number of successes = 8, n = 8, p-value = 0.0078

alternative hypothesis: true p is not equal to

example output from Exact Binomial test,  
usable to conduct a sign test

## PAIRED-SAMPLE PROCEDURES

---

### Analyzing differences

Any of the paired-sample procedures can be implemented by first calculating within-pair differences, using the Transform dialog on the Data menu, and then applying the appropriate one-sample procedure (as described in the previous chapter) to the differences.

Note: While these tests usually are applied to within-pair differences, there could be circumstances in which some other within-pair comparison, e.g. a ratio, might be more appropriate. In this case, simply compute the desired within-pair measure and apply the one-sample procedures as usual.

---

### Paired-sample tests

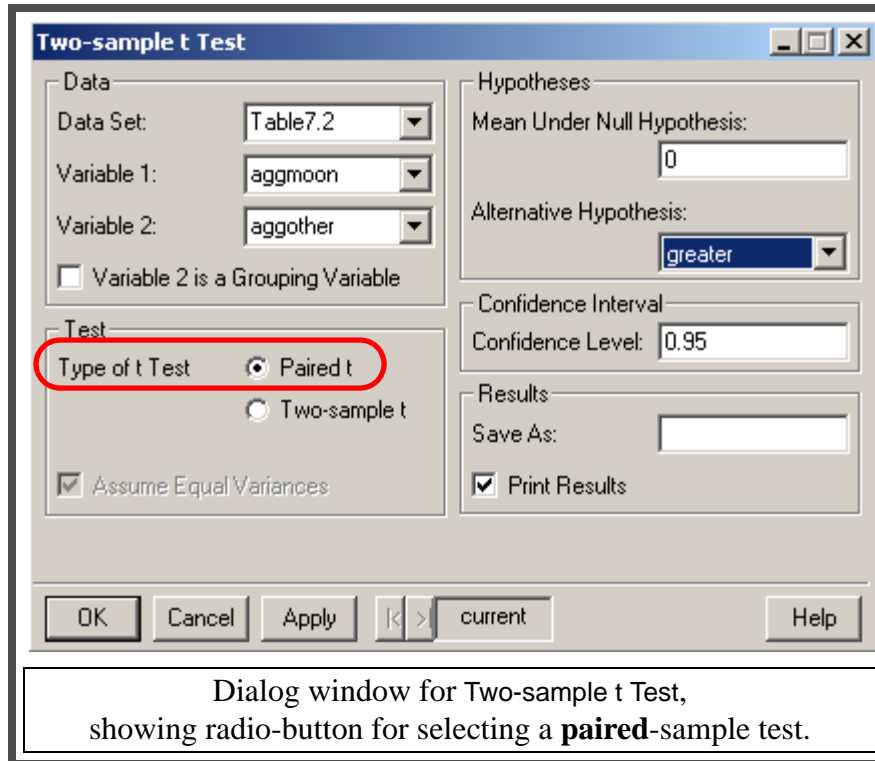
The paired-sample tests that can be applied without first computing within-pair differences all are found on the menus as two-sample procedures, for which there are toggles giving a choice between paired- and independent-sample analyses. These two-sample procedures are described in detail in the next chapter; the descriptions in this chapter therefore are sketchy, emphasizing the differences from the two-independent-sample use of these procedures.

#### *t* test

**Statistics ⇒ Compare Samples ⇒ Two Samples ⇒ t test...**

#### Dialog

To obtain the paired procedure, select the Paired t radio button for Type of t Test, as shown in the example dialog window below. Specify the two columns containing the paired observations, and (optionally) modify the null and alternative hypotheses and confidence level.



## Resampling $t$ test

**Statistics** ⇒ **Compare Samples** ⇒ **Two Samples** ⇒ **t test/Resample...**

or

**Statistics** ⇒ **Resample** ⇒ **Two Sample t...**

### $t$ Test

The top tab of the dialog,  $t$  Test, is very similar to that shown above for the standard  $t$  test. The only difference is the presence of a Trimmed Mean section (below the Test section) in which you can request that the extreme observations be omitted, by specifying a percent of the data set to omit.

### Bootstrap

There is a tab on which you can request Bootstrap estimates for the population mean within-pair difference. The dialog and output are like those described above for the single-sample resampling  $t$  test.

### Permutation test

There also a tab for a Permutation test. This test maintains the paired structure of the data, randomly varying the sign of each difference. The test statistic calculated for each permutation is the mean difference; significance is determined from the permutation distribution of essentially this statistic.

The dialog for the permutation test is simple: on the Permutation tab, just click the checkbox to request the test.

The output gives the observed mean difference, the mean and standard deviation of this statistic for the permutations,  $H_a$ , and the  $P$ -value. Graphs (histogram and NQQ plot) of the mean differences for the permuted samples also are produced; these are similar to those shown in the next chapter for the two-sample case.

```

*** Permutation Test Results ***
Call:
permutationTestMeans(data = Table7.2$aggmoon, data2 =
Number of Replications: 999
Summary Statistics:
      Observed      Mean      SE alternative p.value
Var      2.433 0.02341 0.7285      two.sided 0.002
  
```

example of output from paired-sample permutation  $t$  test

## Signed-rank

**Statistics** ⇒ **Compare Samples** ⇒ **Two Samples** ⇒  
**Wilcoxon Rank Test...**

### Dialog

The dialog (see next page) is the same as for the two-sample Wilcoxon (rank-sum) test described in the next chapter. To obtain the paired procedure, select the Paired  $t$  radio button for Type of  $t$  Test. As for the paired-sample  $t$  procedures described in the previous section, you specify the two columns containing the paired variables, and optionally modify the null and alternative hypotheses. In addition you can choose whether to have the  $P$ -value determined from the exact distribution of the test statistic (rather than from a normal approximation) and/or to have the continuity correction applied to the normal approximation.

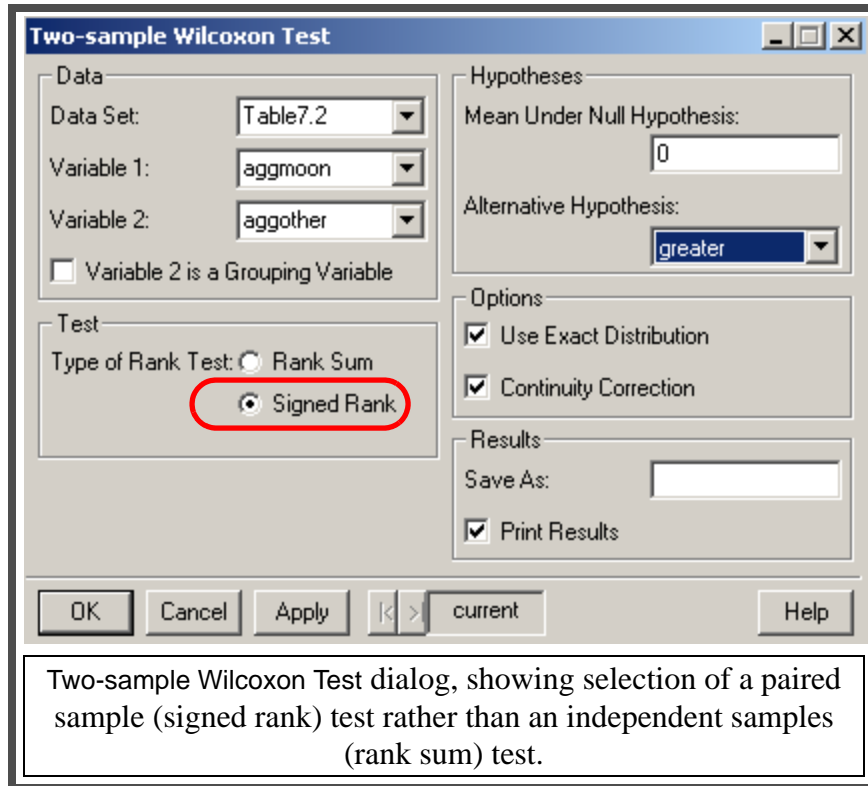
### Output

The signed-rank test output is simple. If the exact test is requested, the output is labeled accordingly, and the actual test statistic is given, along with the exact  $P$ -value.

```

Exact Wilcoxon signed-rank test
data: x: aggmoon in Table7.2, and y: aggoother in Table7.2
signed-rank statistic V = 119, n = 15, p-value = 0.0001
alternative hypothesis: true mu is greater than 0
  
```

example output from signed-rank test, when exact probabilities are used



If the exact test is not requested, the output headline does not call it an exact test, and rather than the test statistic, its normal approximation Z score is given (after applying the continuity correction if requested).

```

Wilcoxon signed-rank test

data:  x: aggmoon in Table7.2 , and y: aggothor in Table7.2
signed-rank normal statistic with correction Z = 3.3226, p-value = 0.0004
alternative hypothesis: true mu is greater than 0

```

example output from signed-rank test, when normal approximation is used

## Sign

To perform the sign test, the within-pair differences must be computed and then analyzed as described in the previous chapter for the one-sample situation.

## TWO-SAMPLE PROCEDURES

---

### A Warning: Which sample is which?

Be careful when interpreting results of the two-sample analyses or when specifying one-sided tests: **be sure you know which sample is being subtracted from which**. This of course is particularly important for one-sided tests, since it determines the direction of the inequality in the alternative hypothesis.

When the data are in the unstacked arrangement this should not be a problem: you explicitly identify variable **1** and variable **2**, and the difference is **1 – 2**. For stacked data, however, some of the procedures determine which group is 1 and which is 2 based on their order of appearance in the data set, and others do it based on the alphabetical order of the values of the grouping variable. In particular, although the dialogs for the standard *t* procedure and the resampling *t* procedure are almost identical, they define the difference (i.e. which group is subtracted from which) differently.

---

### *t* procedures

Statistics ⇒ Compare Samples ⇒ Two Samples ⇒ *t* test...

#### Dialog

How the variables are specified depends on the data arrangement: whether the samples are stacked in one column with another column defining groups, or are unstacked, each sample in a column.

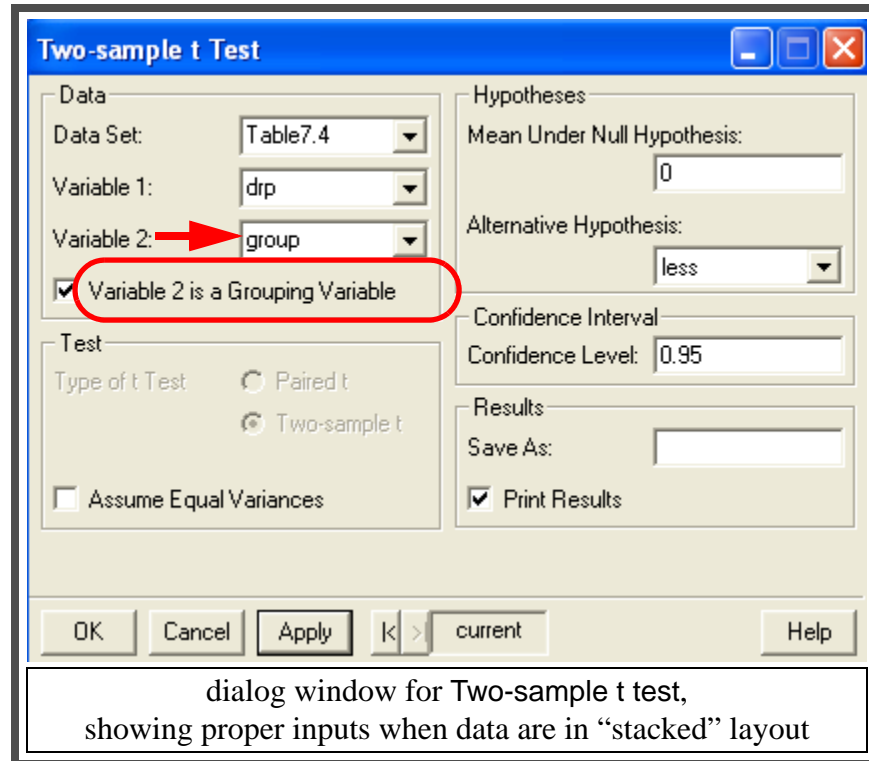
#### Stacked

As shown in the dialog window below, when the data are “stacked” the column with the variable of interest goes in the Variable 1 box, the grouping variable goes in the Variable 2 box, and the checkbox for Variable 2 is a Grouping Variable is checked. (The latter turns off the choice between paired and two-sample tests)

In this procedure the two groups defined by the grouping variable are ordered by the alphabetical order of the group labels. The analysis is of the difference between the group with the alphabetically prior label (here, I’ll call it group **A**) and the other group (**B**): the hypotheses are in terms of the quantity  $(\mu_A - \mu_B)$ .

#### Unstacked

When the data are “unstacked” the checkbox for Variable 2 is a Grouping Variable must be left unchecked; be sure also that the Type of *t* Test is set to Two-sample *t*. The two columns are named in the two Variable *n* boxes; which matters only for inter-



preparing the direction of the difference in means. As noted above, the analysis is of the difference between sample 1 and sample 2: the hypotheses are in terms of the quantity  $(\mu_1 - \mu_2)$ . Samples 1 and 2 correspond to variables 1 and 2 as you identify them in the dialog.

## Output

The output from this procedure includes both the  $t$  test result and a CI for the difference. The difference, as stated in the second line of the output, is presented as the “ $x$ ” variable minus the “ $y$ ” variable, corresponding to Variable 1 minus Variable 2 for unstacked data, or the alphabetically first group minus the other for stacked data. In the example below, the difference is the (perhaps counterintuitive) Control – Treat. This would result from either two columns with those names being specified in that order as Variable 1 and Variable 2, or with stacked data with the values of the grouping variable being Control and Treat.

If the test is one-sided, so is the CI (so it is a confidence bound, really). This is indicated by one of the limits being given as NA. In the example output below, the “CI” is interpreted as “we are 95% confident the true difference [control – treatment] is - 2.691293 or less (further negative).”

The output heading “Welch Modified Two-Sample t-test” refers to the version of the two-sample test which does not assume equal variances; this is the default.

```

Welch Modified Two-Sample t-Test
data:  x: drp with group = Control , and y: drp with group = Treat
t = -2.3109, df = 37.855, p-value = 0.0132
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  NA -2.691293
sample estimates:
mean of x mean of y
 41.52174  51.47619

```

example output from two-sample  $t$  test

## Resampling $t$ procedures

**Statistics  $\Rightarrow$  Compare Samples  $\Rightarrow$  Two Samples  $\Rightarrow$   $t$  test/Resample...**

or

**Statistics  $\Rightarrow$  Resample  $\Rightarrow$  Two Sample  $t$ ...**

### Dialog

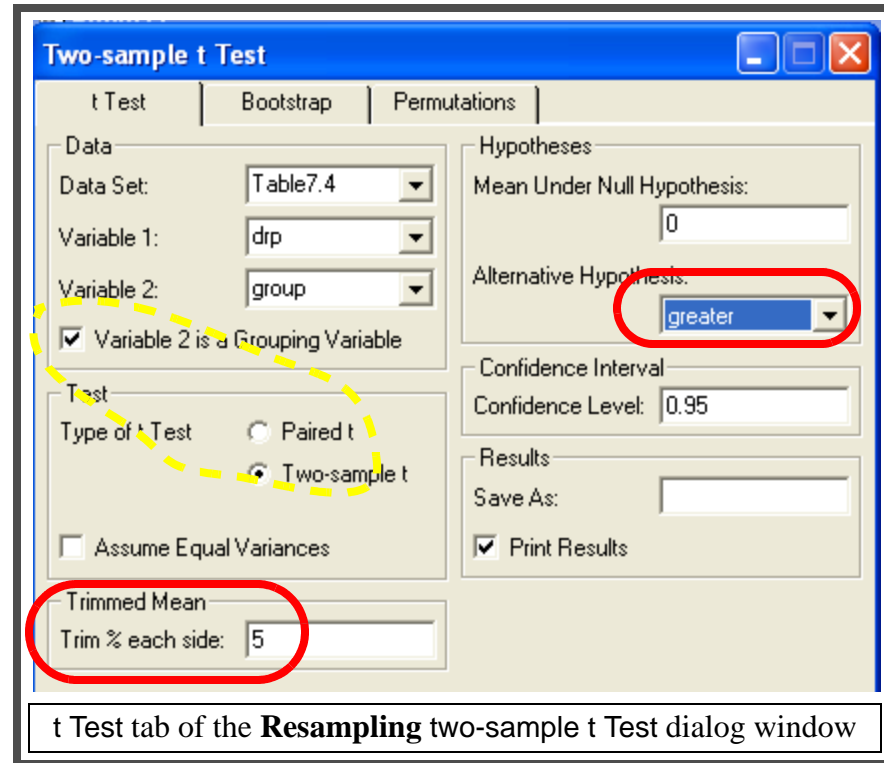
As with the one- and paired-sample resampling  $t$  procedures, this dialog has three tabs.

#### **t Test tab**

The first tab,  $t$  Test (shown on the next page), is nearly identical to the standard two-sample  $t$  described above. (One unfortunate difference is that specifying Variable 2 is a Grouping Variable does not turn of the Paired  $t$  option for type of test.)

#### Trimmed means

The only useful difference from the standard  $t$  test dialog is that you can specify a fraction of extreme observations to trim when computing means. Note that if trimmed means are requested (by entering any value other than 0 in this box), no simple  $t$  test results are produced, and trimmed means are used in the bootstrap estimation and permutation test.



### One-sided $H_a$

In contrast to the standard  $t$  test (invoked by  $\Rightarrow$  Two Samples  $\Rightarrow$   $t$  test), when the data are stacked so that Variable 2 is a grouping variable, the ordering of the two groups — i.e. which mean is subtracted from which — appears to be determined by the order of first occurrence in the data set. Thus for the data used in the  $t$  test example above, since a `Treat` observation is first in the data set, the analysis is of `Treat` minus `Control`, so the direction of the Alternative Hypothesis is reversed from that used in the previous section. Be careful.

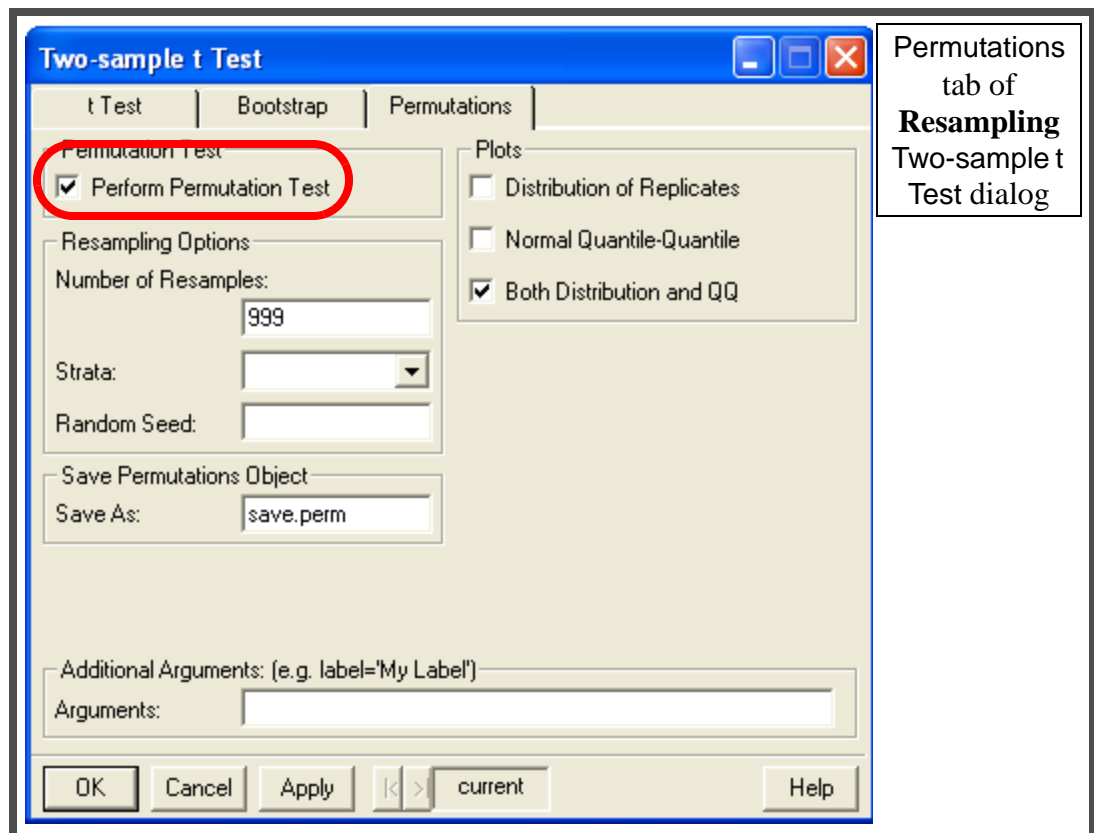
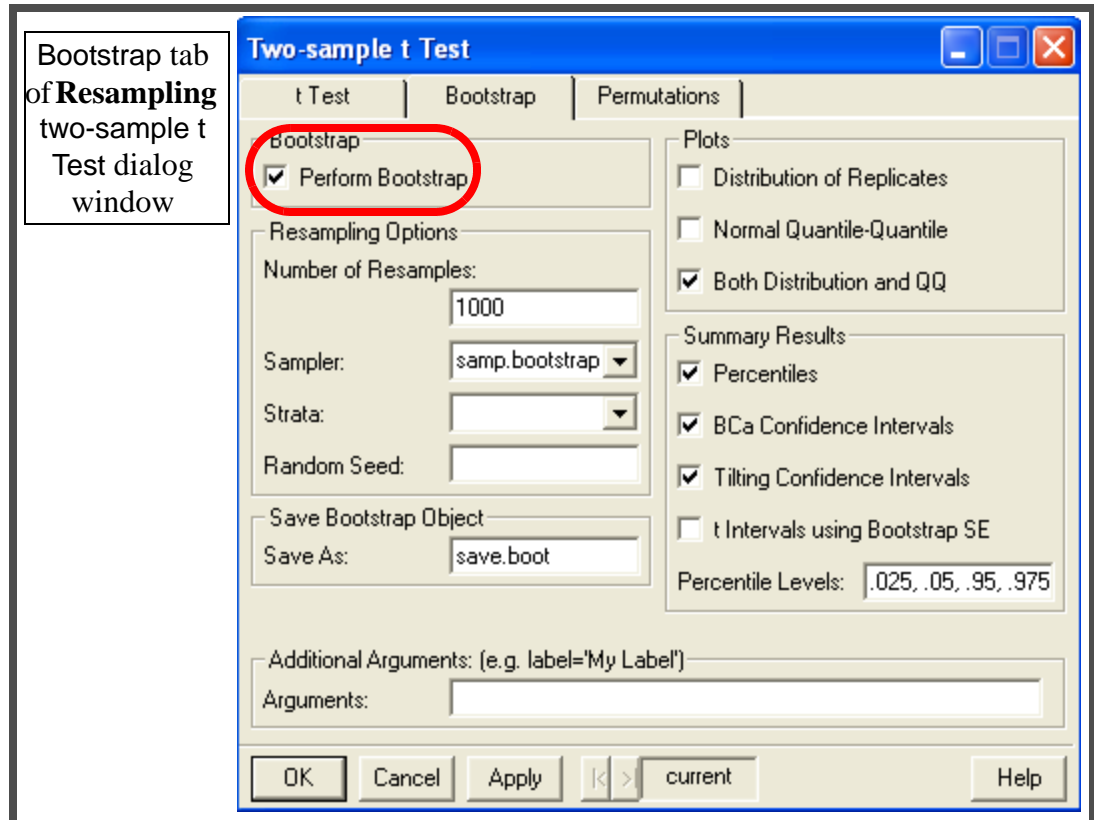
### Bootstrap tab

The second tab (shown on the next page) is for Bootstrap estimation of the difference in means. The only necessary part of this dialog is the check box in which you request it be done. Options include which methods to use, how many resamples to do, and what plots of the resample distribution to produce.

Since this estimation does not assume the populations are identical, the resampling maintains the independence of the two samples. For each iteration, a resample is taken from each original sample and the difference in means is calculated. The distribution of these differences is used to obtain the CI.

### Permutation test tab

The third tab (shown on the next page below the Bootstrap tab) is for hypothesis testing by Permutations. Again the only necessary part of this dialog is the check box



in which you request it be done; options include which how many iterations to do and what plots of the resample distribution to produce.

The randomization for this test is done assuming the null hypothesis is true: there is no difference between the two population distributions. This implies that any of the observations could have appeared in either sample. Each iteration therefore randomly divides the observations into two samples (with the same sample sizes as in the data) and calculates the difference in their means. The  $P$ -value is determined by comparing the observed difference to the distribution of differences from the permutations.

## Output

What output is produced depends on which of the resampling or robust methods are invoked. Notice that all results are in terms of `Treat - Control`, which is the opposite of the results for the same data from the standard  $t$  test shown above This is because the first observation in the data set is a `Treat` observation.

### Trimmed means

If the trimmed mean is requested — by specifying, in the `t Test` tab of the dialog, a percent of extreme observations to trim — the standard  $t$  test results are not given and instead simply the difference in trimmed means is reported.

```
Difference in 5% trimmed means: 10.94236
```

example output from Resampling  $t$  test when  
trimmed means are requested

If the trimmed mean is not requested, the standard  $t$  test results are given.

### Bootstrap estimates

If bootstrap estimation is requested, it produces output (shown on the next page) as in the one-sample bootstrap, including plots as shown two pages down. Note that the bootstrapping used trimmed means.

```

*** Bootstrap Results ***
Label: bootstrap 5% trimmed mean: Table7.4$drp
Call:
bootstrap2(data = Table7.4$drp, statistic = mean, ...
  trace = F, args.stat = list(trim = 0.05), ...
  "bootstrap 5% trimmed mean: Table7.4$drp",
  save.indices = T)
Number of Replications: 1000
Summary Statistics:
      Observed Mean      Bias      SE
mean    10.94 10.64 -0.2991 4.075
Percentiles:
      2.5%      5%      95%      97.5%
mean  2.760025 3.73396 17.38083 18.67914
BCa Confidence Intervals:
      2.5%      5%      95%      97.5%
mean  3.205514 4.021816 17.89195 19.04683
Tilting Confidence Intervals:
      2.5%      5%      95%      97.5%
mean  1.710333 3.439095 18.34711 19.77674

```

example output  
from Bootstrap  
analysis in  
Resampling  
Two-sample t Test

### Permutation test

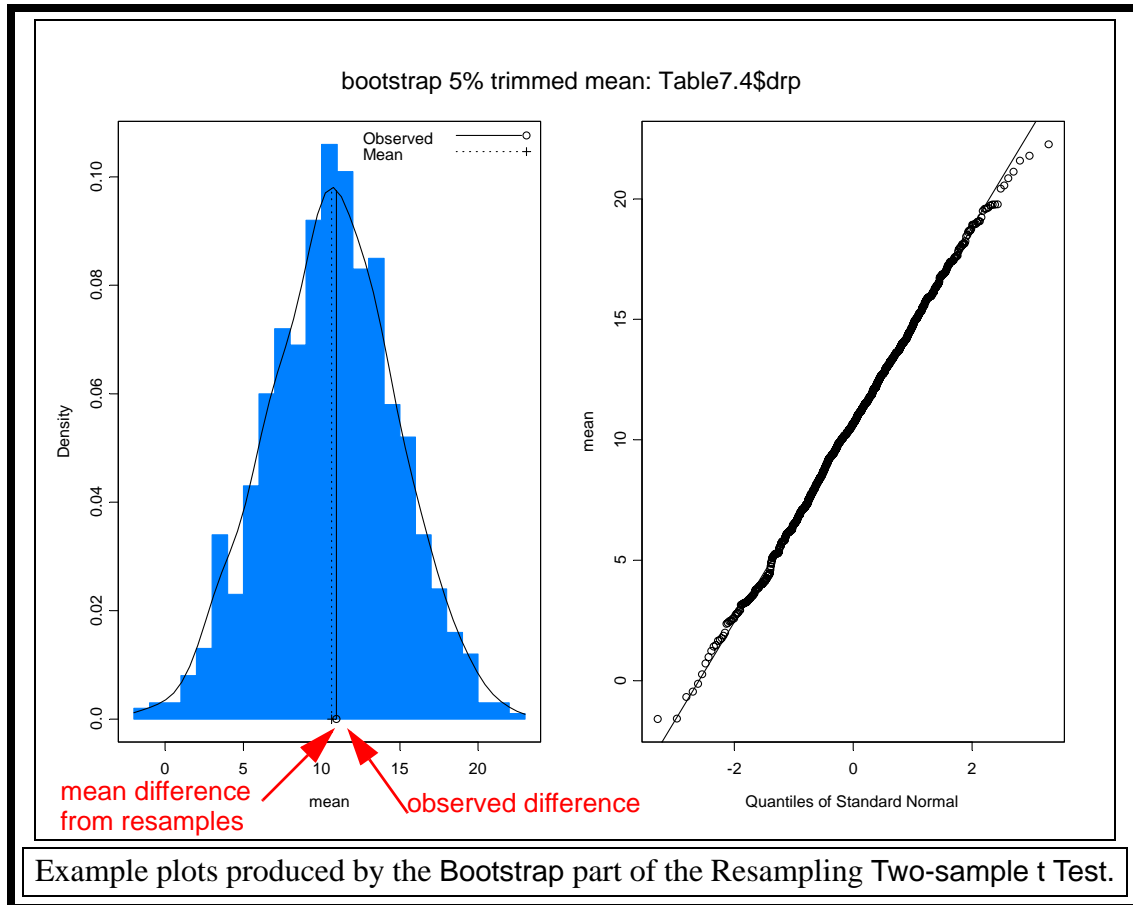
If the permutation test is requested, the output gives the test results (in the final line: the observed difference in means, the mean and standard deviations of the distribution of differences from the permutations, a statement of the alternative, and the  $P$ -value.

```

*** Permutation Test Results ***
Label: permutation 5% trimmed mean: Table7.4$drp-Table7.4$group
Call:
permutationTest2(data = Table7.4$drp, statistic = mean, treatment =
Number of Replications: 999
Summary Statistics:
      Observed      Mean      SE alternative p.value
Var    9.954      0.219 4.726      greater 0.018

```

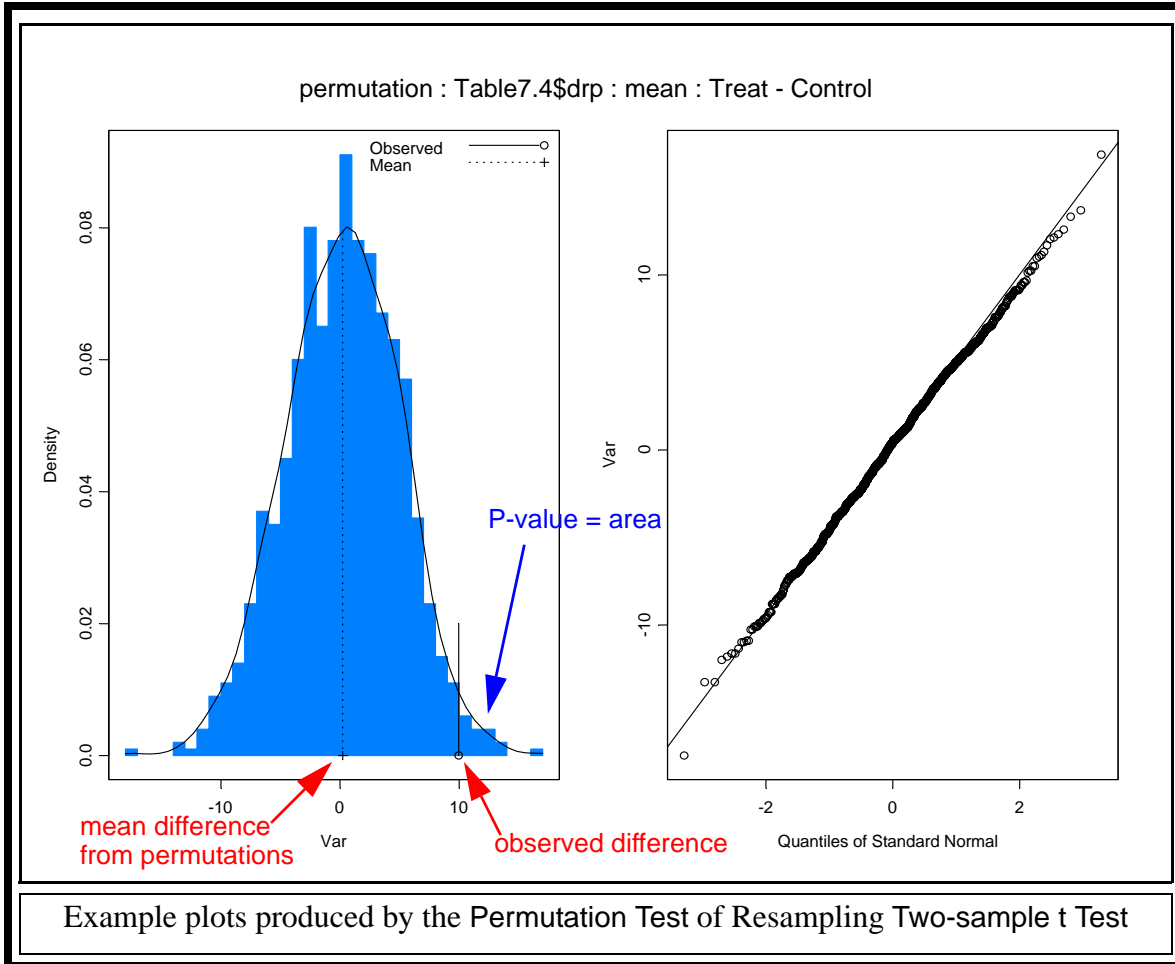
example output from Permutation Test in Resampling Two-sample t Test



## Graphs

Graphs are produced for the bootstrap estimation (above) and the permutation test (next page), if they are requested. If both methods are performed, the graphs will be on two pages of a single graph window.

The bootstrap and permutation graphs both show the distributions of the difference between the group means, obtained in the resamples. Because the randomization is done separately for each method, these distributions will not be identical, but they will be very similar in shape. Both histograms indicate the observed difference and the mean of the resample differences. Since the bootstrap distribution is from resamples maintaining the distinction between the two samples, it is centered very near the observed difference in means (about 10 in the example above), while the permutation distribution is centered very near the difference according to the null hypothesis, i.e. 0.

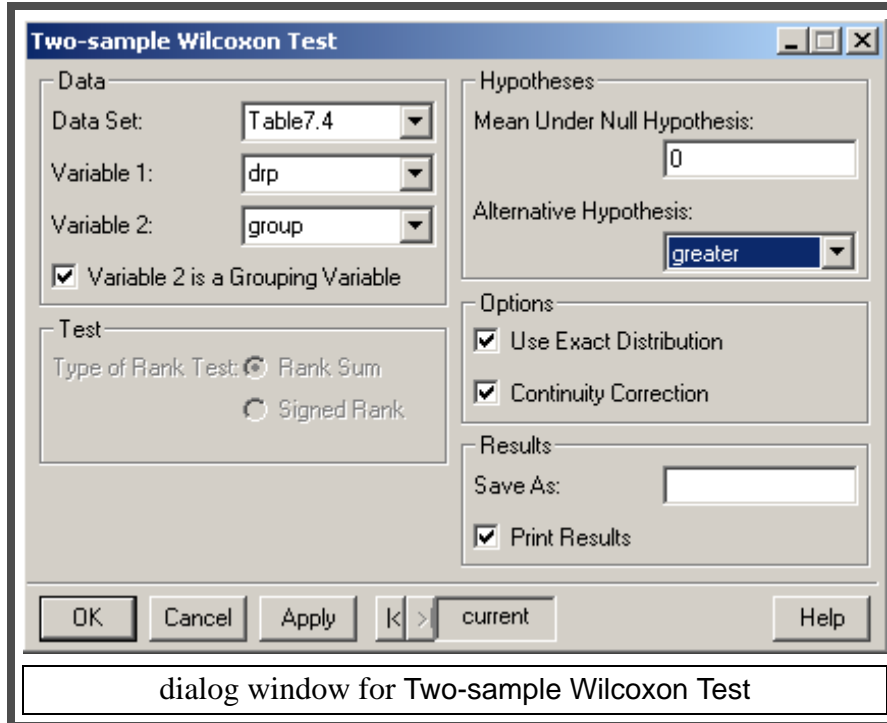


## Rank-sum test

Statistics  $\Rightarrow$  Compare Samples  $\Rightarrow$  Two Samples  $\Rightarrow$   
 Wilcoxon Rank Test...

### Dialog

As with the two-sample  $t$  procedures described above, the data for the rank-sum test can be either stacked or unstacked. If they are stacked, select the Variable 2 is a Grouping Variable option (which will make the Type of Rank Test choice unavailable, since paired data cannot be in the stacked arrangement). If the data are unstacked make sure the Variable 2 is a Grouping Variable option is not checked, and that Rank Sum is selected as the Type of Rank Test. Specification of the variables, and the options concerning the hypotheses and the exact or normal-approximation tests, are the same as for one- and paired-sample procedures described above. As before, the exact test cannot be performed if there are ties in the data.



## Output

If the exact test is requested (and possible), the output is labeled accordingly, and the actual test statistic is given, along with the exact  $P$ -value.

```
Exact Wilcoxon rank-sum test
data:  x: drp.1 with group = Control , and y: drp.1 with group = Treat
rank-sum statistic W = 407, n = 23, m = 21, p-value = 0.0044
```

example output from Two-sample Wilcoxon Test when the exact test is performed

If the exact test is not requested or is not possible due to ties, the output does not call it an exact test, and rather than the test statistic, its normal approximation  $Z$  score is given (after applying the continuity correction if requested).

```
Wilcoxon rank-sum test
data:  x: drp.1 with group = Control , and y: drp.1 with group = Treat
rank-sum normal statistic without correction Z = -2.5965, p-value = 0.0047
alternative hypothesis: true mu is less than 0
```

example output from Two-sample Wilcoxon Test when the exact test is not performed

## (Mood's) Median test

This test is not directly available in S-Plus. It can be implemented as follows:

1. determine the pooled median: stack the samples and use Statistics  $\Rightarrow$  Data Summaries  $\Rightarrow$  Summary Statistics...
2. create a column describing whether an observation is above the pooled median or not: use Data  $\Rightarrow$  Subset..., as described above for the one-sample sign test
3. Do a Fisher's exact or chi-square test on the two variables: the grouping variable and the True/False variable created in step 2. (These tests are described in the section on inference for contingency tables, near the end of this guide.)

## SEVERAL-SAMPLE PROCEDURES

---

### ANOVA

The data must be in the stacked arrangement: the response variable in one column and a grouping variable in another column. The grouping variable must be of type “factor” rather than numeric (even if the level labels are numbers); this can be determined using the column properties (right-click on the column).

A basic one-way ANOVA is accessed by

**Statistics** ⇒ **Compare Samples** ⇒ **k Samples** ⇒ **One-way ANOVA...**

This procedure produces only the ANOVA table; graphs, contrasts, multiple comparisons, storage of residuals, etc., are not available. It therefore generally will be preferable to use the procedure describe in the following.

### Dialog

**Statistics** ⇒ **ANOVA** ⇒ **Fixed Effects...**

#### Model

The first tab is the Model tab shown to the right, on which the data set and the Response and Explanatory variables are specified.

As with most S-Plus procedures, the ANOVA can use a subset of the data, by specifying a subset column or subsetting criterion in the Subset Rows with: box.

For one-way ANOVA the Formula parts of the dialog can be ignored.

This tab also gives you the option (on the right side, not shown here) of saving the analysis — the calculations, not just the output — as an S-Plus data object; this can be useful if you want to extract more detailed output, e.g. for contrasts.

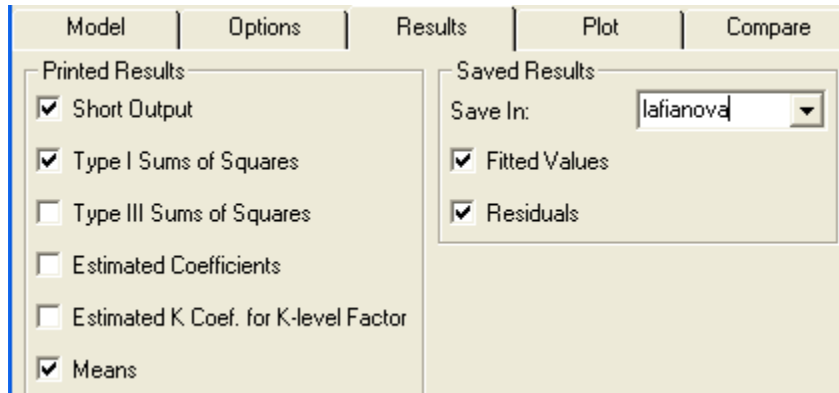
#### Options

The second tab, labelled Options, allows contrasts among the groups to be selected. Given the importance and complexity of contrasts, they are discussed in detail below.

The screenshot shows the 'Model' tab of the ANOVA dialog box. It has three tabs: 'Model', 'Options', and 'Results'. The 'Data' section includes a 'Data Set' dropdown set to 'laf187fe', a 'Weights' dropdown, and a 'Subset Rows with:' text box. A checkbox labeled 'Omit Rows with Missing Values' is checked. The 'Variables' section has a 'Response:' dropdown set to 'beakwid' and an 'Explanatory:' list box containing '<ALL>', 'beakwid', 'islnum', and 'island', with 'island' selected. The 'Formula:' text box contains 'beakwid~island'. A 'Create Formula' button is at the bottom.

## Results

The third tab (see below) lets you choose what Results are produced. The default choices of Short Output and Type I Sums of Squares usually are appropriate for one-way ANOVA. You may also want to have the group Means listed, and/or have the Fitted Values and Residuals stored.



## Plot

The fourth tab lets you choose from a long list of plots. These include the familiar Residuals vs Fit and Residuals Normal QQ. Two other plots — Partial Residuals and



Cook's Distance— are more advanced methods for assessing influence of individual observations, and are covered in Advanced Biometry. The other choices are not covered in either Biometry course but seem like they could be useful.

The Options to the right of the Plot tab apply to the various plots selected. Include Smooth adds a Loess smoother to the plots, and the chosen Number of Extreme Points are identified on plots by observation number; both these are generally useful. A Rugplot, however, seems more appropriate for regression situations.

## Compare

The final tab provides for (unplanned) multiple comparisons. These are discussed below.

## Output

The default output produces a very basic ANOVA table with rows for the factor and “error” (= Residuals) terms (but not the total SS) and the Residual standard error (the square-root of MSE).

```
*** Analysis of Variance Model ***
Short Output:
Call:
  aov(formula = beakwid ~ island, data = lafi87fe, qr = T,
Terms:
              island  Residuals
Sum of Squares 0.01507044 0.05512248
Deg. of Freedom      2          89
Residual standard error: 0.02488682
Estimated effects may be unbalanced

      Df Sum of Sq   Mean Sq F Value   Pr(F)
island  2 0.01507044 0.007535218 12.16626 0.00002133428
Residuals 89 0.05512248 0.000619354
```

If means are requested on the Results tab, the output includes the total mean, and the group means and sample sizes.

```
Tables of means
Grand mean
0.76811
island
  Laysan  North SouthEast
0.760    0.800    0.777
rep 62.000 10.000 20.000
```

## Contrasts

As noted above, contrasts among group means can be requested on the Options tab of the ANOVA dialog. There are several pre-defined contrasts or you can define your own. Standard output will be only the estimates of the contrasts; to get tests of their significance requires a bit of command-line programming.

## Order of groups

A contrast is defined by a sequence of coefficients for the groups (levels of the factor). To make sense of this you need to understand the ordering of the groups. To find out the ordering of levels of a factor column you can sort the data set according to that column, or you can enter the command `levels(factor)` in the Commands Window (enter the name of the factor variable in place of *factor*).

As best as I can figure out, the ordering is based on the numerical/alphabetical ordering at the time the factor column is created. To control the ordering yourself, you could use numeric coding, in the desired order, when importing or creating the variable, then after changing its type to “factor” you could re-code the values to meaningful labels.

### Predefined contrasts

There are four predefined sets of contrasts:

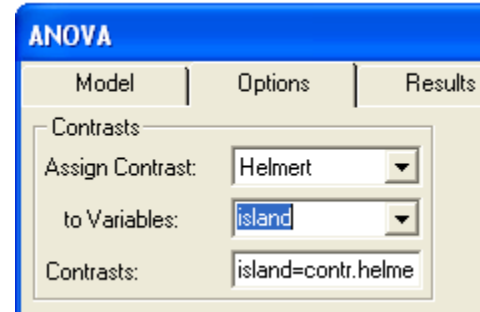
- **Helmert**

contrasts compare each level to the combination of all the preceding levels. For example, with four groups the sets of coefficients would be

$$\begin{aligned} C_1: & 1 \quad -1 \quad 0 \quad 0 \\ C_2: & 1 \quad 1 \quad -2 \quad 0 \\ C_3: & 1 \quad 1 \quad 1 \quad -3 \end{aligned}$$

(Note that rather than use fractional constants to average two or more levels, these contrasts sum the levels to be combined, and multiply the other level accordingly.)

Helmert contrasts are the default for unordered factors. If the levels are properly ordered these contrasts often will be sufficient. For instance, in Data Set #1, the contrasts compared Laysan to (North + Southeast), and then compared North to Southeast. If Laysan is the last level, these are the Helmert contrasts.



- **Orthogonal Polynomial**

contrasts can be used when the levels of the factor are ordered, and are best when the levels are evenly spaced. In that case they represent linear, quadratic, cubic, etc., trends with respect to the levels. With four levels, the sets of coefficients are (with some rounding):

$$\begin{aligned} C_L: & -0.6708 \quad -0.2236 \quad 0.2236 \quad 0.6708 \\ C_Q: & 0.5 \quad -0.5 \quad -0.5 \quad 0.5 \\ C_C: & -0.2236 \quad 0.6708 \quad -0.6708 \quad 0.2236 \end{aligned}$$

(The non-integer constants are used to make the contrasts independent of each other, i.e. “orthogonal” in statistical jargon.)

- **Sum**

contrasts compare each level to the last level. There are the allowable  $k-1$  contrasts, but they are not independent. With four groups the coefficients are:

$$\begin{aligned} C_1: & 1 \quad 0 \quad 0 \quad -1 \\ C_2: & 0 \quad 1 \quad 0 \quad -1 \\ C_3: & 0 \quad 0 \quad 1 \quad -1 \end{aligned}$$

(These contrasts are similar to Dunnett’s multiple comparisons, if the last level is the “control” level, except that no adjustment is made for multiple testing.)

- **Treatment**

“contrasts” are actually just dummy variables for all the groups but the first one, and so are not true contrasts.

$$C_1: 0 \quad 1 \quad 0 \quad 0$$

$$C_2: \begin{matrix} 0 & 0 & 1 & 0 \\ C_3: & 0 & 0 & 0 & 1 \end{matrix}$$

### Custom contrasts

To define your own contrasts, you must do so in the Commands window. In the following I use the contrasts described in the text (Ch. 12, Sect. 2). These are

$$\begin{aligned} \Psi_1 &= (\mu_{sk} + \mu_{un}) - 2\mu_{su} \\ \Psi_2 &= \mu_{sk} - \mu_{un} \end{aligned}$$

where the subscript *sk* is for skilled workers, *un* is for unskilled workers, and *su* is for supervisors. The first contrast compares supervisors to workers, and the second compares skilled to unskilled workers.

To define these contrasts, and have them used when performing the ANOVA, do the following:

1. Open the Commands Window by

**Window** ⇒ **Commands Window**

2. Make a local copy of the ANOVA independent variable, e.g.

```
> jc <- jobcat
```

where `jobcat` is the name of the independent variable, in a data set that is already open (the `>` is the prompt given by S-Plus).

3. Check the ordering of the levels of the independent variable, e.g.

```
> levels(jc)
[1] "skilled" "superv" "unskill"
```

4. Create a matrix of your contrasts. For example,

```
> contrast.mat <- matrix(c(1,-2,1, 1, 0, -1),ncol=2)
```

Here `contrast.mat` is the name of the matrix being created, all the coefficients are given in one string, and the `ncol` value shapes the matrix to have that many columns. Note that because the levels of the factor `jc` are in alphabetical order, the order of the coefficients for the two contrasts is different than in the equations defining the contrasts.

5. Set the contrasts for your factor equal to the matrix just created:

```
> contrasts(jc) <- contrast.mat
```

6. Check that the contrasts are as you wanted:

```
> contrasts(jc)
```

which should give the output:

```
      [,1] [,2]
skilled    1    1
superv     -2    0
unskill    1   -1
```

7. Run the ANOVA:

```
> sciaov <-aov(SCI ~ jc)
```

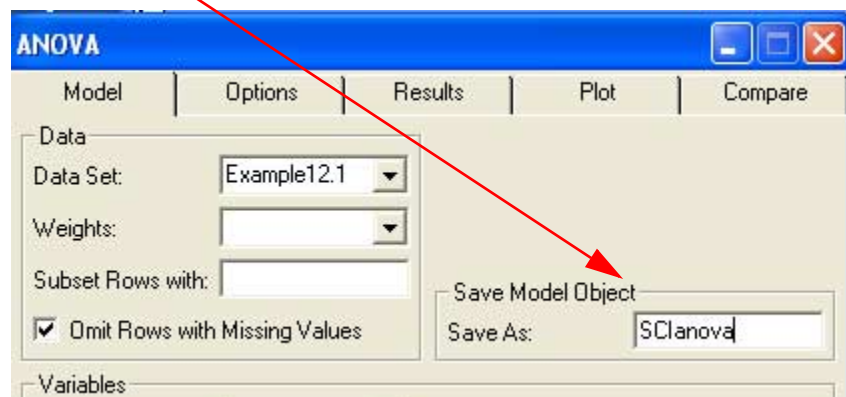
Here `sciaov` is the name under which the “Model Object” (i.e. the results of the ANOVA) will be stored. `aov` is the procedure name. `SCI ~ jc` specifies the model, in the format *response ~ explanatory*).

The preceding will not produce any output. To get the output, see the following section (“Testing contrasts”).

### Testing contrasts

To get the results of testing contrasts, you must save the “Model Object” and then enter a command in the Commands window.

1. On the Model tab of the ANOVA dialog window, enter some name in the Save Model Object; Save As: box.



2. Open the Commands Window by  
**Window ⇒ Commands Window**
3. If you need to check what contrasts were used, enter

**`contrasts (factor)`**

where the name of the independent variable (the ANOVA factor) is entered in place of `factor`. This will produce output like:

```
> contrasts (jobcat)
      [,1] [,2]
skilled  -1  -1
superv   1  -1
unskill   0   2
```

These, the default Helmert contrasts for an unordered factor, are

$$\Psi_1 = (-\mu_{sk}) + \mu_{su}$$

$$\Psi_2 = -(\mu_{sk} + \mu_{su}) + 2\mu_{un}$$

4. Enter

```
> summary.aov(myanova, split = list(factor =
+ list(c1=1, c2=2 ...)))
```

In this command:

- do not enter the `>` and `+` prompts; they are given by the program.

- in place of **myanova** give the name of the “Model Object” saved when the ANOVA was run.
- in place of **factor** give the name of the explanatory variable in the ANOVA.
- in place of **c1**, **c2**, etc., give names for the successive contrasts (see “Order of Groups” and “Predefined Contrasts” above to determine what the contrasts are, if you did not define them yourself).

The output will be put in the Commands Window, for example:

```
> summary.aov(sciaov2, split = list(jobcat = list (f = 1, s = 2)))
              Df Sum of Sq  Mean Sq  F Value    Pr(F)
  jobcat      2    4662.2  2331.116  7.136969 0.00086604
jobcat: f     1    1672.6  1672.634  5.120952 0.02400315
jobcat: s     1    2989.6  2989.599  9.152986 0.00259179
Residuals 587   191729.2   326.626
```

In this ANOVA table the first line is for the factor as a whole, just as given in the ANOVA table when the analysis was first run. The next lines partition the *df* and SS for this term into those corresponding to the contrasts. The *F Value* will be the square of the *t* statistic for the contrast.

## Multiple comparisons

Multiple comparisons are requested on the Compare tab of the ANOVA dialog.

### Dialog

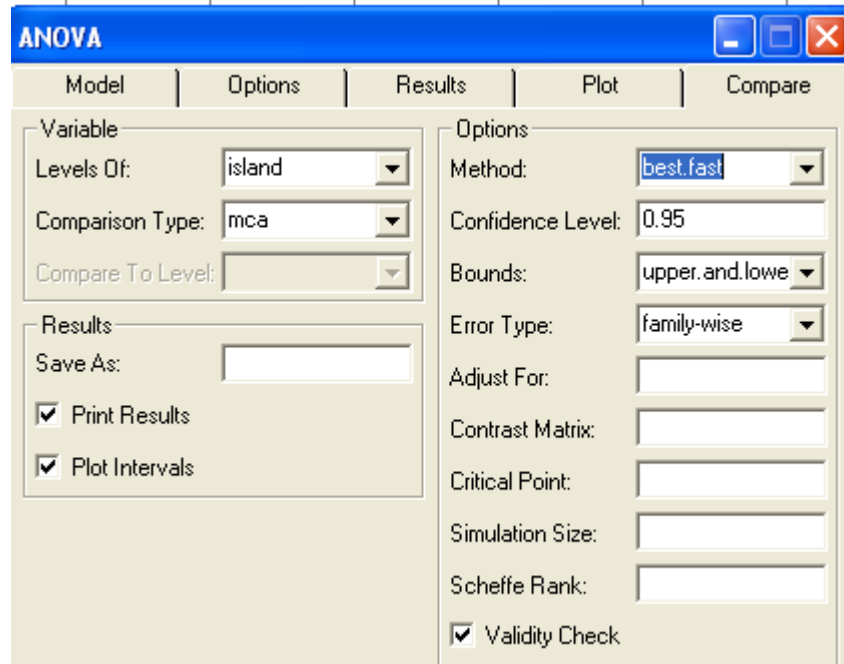
The only item in this dialog (see below) that must be specified is the Variable - Levels Of:, which must be the factor in the ANOVA model. The default options generally are appropriate.

The principal parts of this dialog that you may want to change are:

- Comparison Type

This gives a choice among:

- mca: mean comparisons among **all** groups (i.e. all pairwise comparisons; this is the default);
- mcc: mean comparisons with **control**, with the “control” level specified in the Compare To Level box; and
- none: CIs for the level means rather than pairwise differences.



- Method

This gives a choice among:

**best:** uses whichever of the following methods gives the greatest power (smallest CI).

**best.fast:** also uses the most powerful of the following procedures, excluding the Simulation method; this choice is the default and generally will use the Tukey method for two-ended CIs and the Sidak method if only upper or only lower bounds are requested (as described below).

**Bonferroni:** the usual.

**Dunnett:** can only be used if mca is selected as the comparison type (and only Dunnett can be used in that case)

**Fisher.lsd:** simply does pairwise  $t$  tests with no adjustment for the multiple comparisons, so can only be used if Error Type is set to comparison-wise (in which case only Fisher.lsd can be used).

**Scheffe:** for unplanned contrasts more complex than pairwise differences; very conservative and you probably will use it rarely if ever.

**Sidak:** similar to Bonferroni but somewhat more powerful.

**Simulation:** This uses Monte Carlo simulation. For nonstandard families of comparisons or unbalanced designs, this method will often be substantially more efficient than other valid methods. The simulation size is set by default to provide a critical point whose actual error rate is within 10% of the nominal  $\alpha$  (with 99% confidence). This method can take a lot of time with large data sets, but for the Laysan Finch example used here it needed only a few seconds.

**Tukey:** the usual.



- **Confidence Level**  
if you want something other than the default 95%.
- **Bounds**  
if you want one-ended “intervals,” i.e. only upper bounds or only lower bounds, rather than the default two-ended intervals; this could be appropriate if  $H_a$  specified an ordering of the  $\mu_i$  rather than simply that there is a difference somewhere.

## Output

The output (see below) is in the form of the difference, its SE, and the resulting CI, for each of the pairs of differences. Significant comparisons are flagged. If the Plot Intervals option is checked, a graph showing all the pairwise CIs is produced, as shown below after the text output.

```

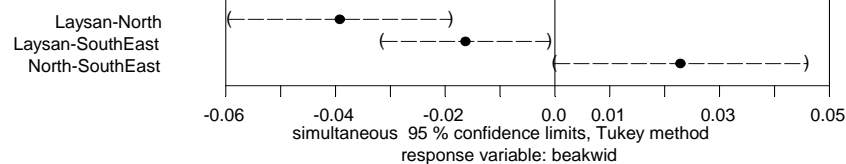
95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.3836
response variable: beakwid

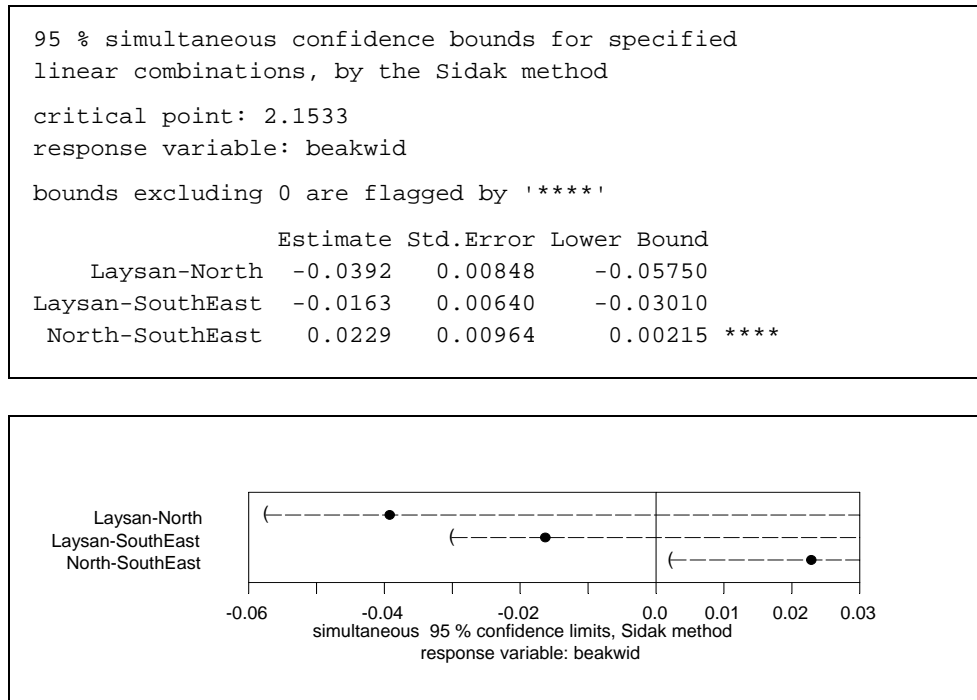
intervals excluding 0 are flagged by '****'

      Estimate Std.Error Lower Bound Upper Bound
Laysan-North  -0.0392   0.00848  -0.0594000  -0.01900 ****
Laysan-SouthEast -0.0163   0.00640  -0.0315000  -0.00104 ****
North-SouthEast  0.0229   0.00964  -0.0000741   0.04590

```



If only lower bounds (or only upper bounds) are requested, the output is similar except that of course only the requested bound is shown, and the plots are similarly one-ended.



## Test for Equal Variances

S-Plus does not appear to include any of the tests for equality of variances. Levene's test could be approximated by storing the residuals, computing their squares or absolute values, and then performing ANOVA on these. The simple  $f_{\max}$  test can be done by getting the variances for all the groups, calculating the ratio of the largest to the smallest, and comparing this to  $F$  critical values gotten using `Data ⇒ Distribution Functions...`

## Resampling ANOVA

There does not appear to be a resampling / randomization method specifically for comparing several samples. Such an analysis presumably could be conducted using the resampling version of linear regression (as described below), with dummy variables to represent the levels of the ANOVA factor.

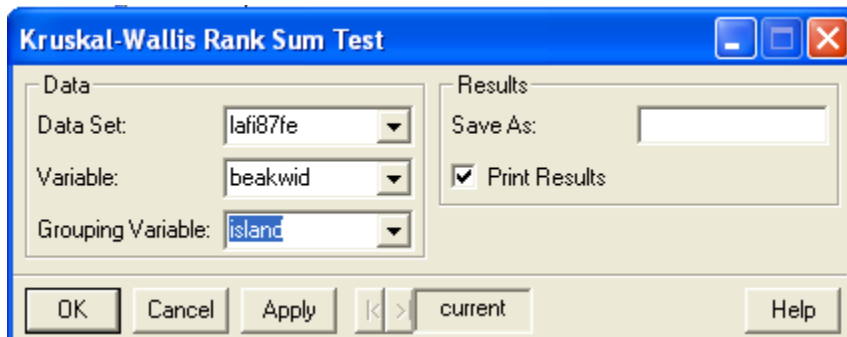
## Kruskal-Wallis test

The data must be in the stacked arrangement: the response variable in one column and a grouping variable in another column. The grouping variable must be of type “factor” rather than numeric (even if the level labels are numbers); this can be determined using the column properties (right-click on the column).

### Dialog

**Statistics** ⇒ **Compare Samples** ⇒ **k Samples** ⇒  
**Kruskal-Wallis Rank Test...**

You specify the Data Set, response Variable, and Grouping Variable. Optionally you can turn off printed output and/or have the results saved as a data object.



There is no capability for multiple comparisons, contrasts, or any plots of the results, of residuals, etc.

### Output

The output is similarly simple: the test statistic (expressed as a chi-square statistic), the *P* value, and a statement of the alternative hypothesis.

```
Kruskal-Wallis rank sum test
data:  beakwid and island from data set lafi87fe
Kruskal-Wallis chi-square = 20.0268, df = 2, p-value = 0
alternative hypothesis: two.sided
```

## (Mood's) Median test

This test is not directly available in S-Plus. It can be implemented as described above in the two-sample situation.

# REGRESSION

---

## Linear least-squares regression

Obtaining a simple linear regression was described earlier, in the **Describing Relationships** section. Here I repeat the key points and add a description of how to obtaining confidence intervals for the mean response.

### Dialog

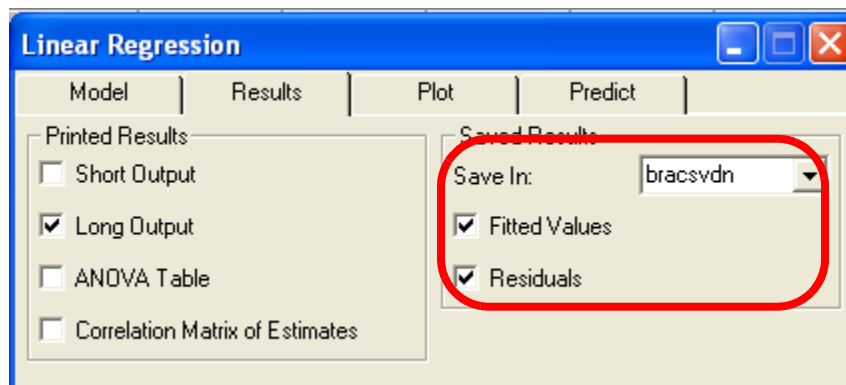
**Statistics** ⇒ **Regression** ⇒ **Linear...**

### Model

On the Model tab you specify the Data Set (in the Data section of the window), and the Response and Explanatory variables (in the Variables section).

### Results

On the Results tab the principal option of interest is to save the Fitted Values and the Residuals. If no data set name is given in the Save In box, a new data set will be created which contain **only** the fits and residuals; to have them stored with the original data, specify the original data set in this box.



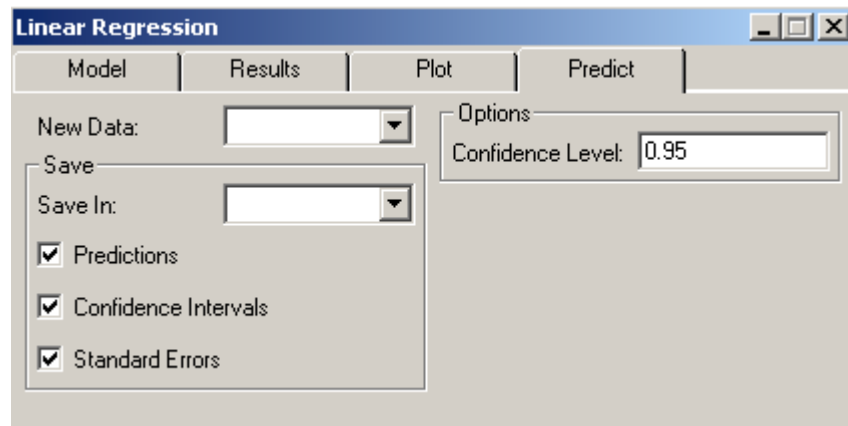
### Plot

The Plot tab provides a variety of diagnostic plots, including a scatterplot of Residuals vs Fit and a Residuals Normal QQ. A smoother can be put on the residuals vs fits plot, and the most extreme residuals can be individually identified (the defaults are to include the smoother and to flag the 3 most extreme residuals).

### Predict

The final tab is labelled Predict. All that needs to be done on this tab is to check the boxes to select which of the three sorts of “predictions” is desired. If New Data is left

blank, predictions will be produced for the observations in the data set used in the regression analysis.,



To get predictions for new values of the explanatory variable these values can be put in another data set, which would be specified in the New Data box; alternatively, the new values of the explanatory variable can be entered in the original data set with missing values for the response variable, so that the regression will not use them but will produce predictions for them along with the rest of the data set.

If no data set is specified in the Save In box, a new data set will be created containing **only** the predictions output. To have the predictions put in the data set being used for the regression, or the data set specified for New Data, simply specify the desired destination in the Save In box.

## Output

The default output (despite being called “Long”) is minimal. Most parts of it are

```

Residuals:
  Min       1Q   Median       3Q      Max
-0.5976 -0.08508 0.009312 0.1167 0.4024

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  0.7046   0.0438   16.0707  0.0000
eggs        -0.0089   0.0018   -4.8654  0.0000

Residual standard error: 0.1852 on 83 degrees of freedom
Multiple R-Squared: 0.2219
F-statistic: 23.67 on 1 and 83 degrees of freedom, the p-value is 5.36e-006

```

familiar, the important ones being the  $t$  test and  $R^2$ . The  $F$  test is exactly equivalent to the  $t$  test when there is only one explanatory variable. The Residual standard error is the standard deviation of the residuals, i.e.  $s$ . The five-number summary of

the residuals could be used to look for skew or outliers, but it would be more effective to store the residuals and examine them graphically

### Storage

If storage of residuals and/or fits is requested on the Results tab, they are added to the data set specified on that tab; if the original data set is specified, the result will be like:

		1	2	3	4
		eggs	survival	fit	residuals
34		16.00	0.38	0.56	-0.19
35		16.00	0.75	0.56	0.19
36		16.00	0.31	0.56	-0.25

Similarly, if predictions are requested (and specified to go into the original data set), the result will be as below. `se.fit` is the standard error of the fit (i.e. the estimated mean  $Y$ ), and `LCL95` and `UCL95` are the lower and upper confidence limits, respectively. (Also note that the column label `fit` has been moved from column 3, in which the fits were stored above, to column 5.)

		1	2	3	4	5	6	7	8
		eggs	survival		residuals	fit	se.fit	LCL95	UCL95
4		16.00	0.38	0.56	-0.19	0.56	0.02	0.52	0.61
5		16.00	0.75	0.56	0.19	0.56	0.02	0.52	0.61
5		16.00	0.31	0.56	-0.25	0.56	0.02	0.52	0.61
7		16.00	0.50	0.56	0.00	0.56	0.02	0.52	0.61

## Regression - resampling

The resampling library includes a procedure to produce bootstrap confidence intervals for the coefficients of a linear regression (simple or multiple).

This can be used to indirectly test hypotheses about the coefficients, i.e. by determining whether the hypothesized value is within the CI. Alternatively, the statistical significance of the relationship can be tested by a randomization test, using the resampling correlation procedure described below.

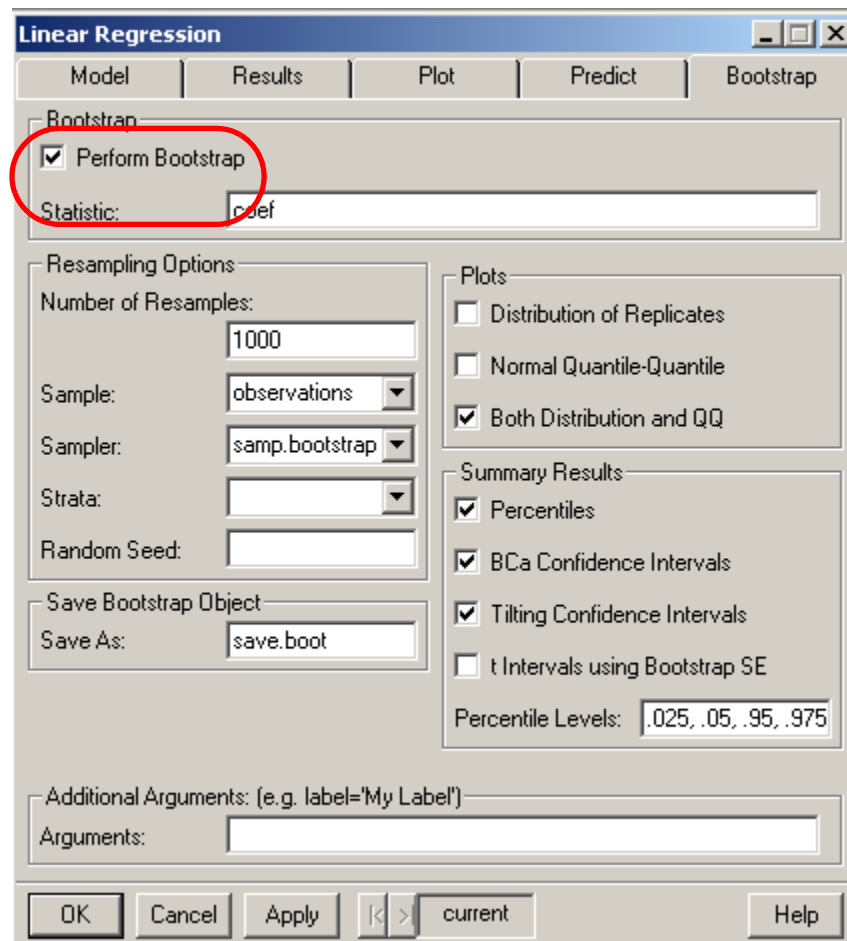
## Dialog

**Statistics** ⇒ **Regression** ⇒ **Linear/Resample...**

or

**Statistics** ⇒ **Resample** ⇒ **Linear Regression**

The dialog window has five tabs. The first four — Model, Results, Plot, and Predict — are exactly as for the standard linear regression described in the preceding section. The fifth tab is labelled **Bootstrap** (see next page). The only necessary input on this tab is to check the box to have the procedure performed. The selection of kinds of plots and confidence intervals to be produced can be modified also.



## Method

The bootstrap resampling (with replacement) is of entire observations, i.e. values of the response and explanatory variable(s) for a given observation are kept together in the resampling.

## Output

The output consists of bootstrap confidence intervals for each of the coefficients in the model, produced by each of the methods selected in the dialog window. In the example shown here the most useful output (circled) gives the BCa 95% CI for the slope.

Summary Statistics:				
	Observed	Mean	Bias	SE
(Intercept)	0.704579	0.706380	0.0018010	0.045063
eggs	-0.008914	-0.009038	-0.0001239	0.001488
Percentiles:				
	2.5%	5%	95%	97.5%
(Intercept)	0.61021603	0.62900610	0.777467951	0.793145841
eggs	-0.01185643	-0.01153092	-0.006498775	-0.005988335
BCa Confidence Intervals:				
	2.5%	5%	95%	97.5%
(Intercept)	0.6068326	0.61743755	0.775389002	0.787171309
eggs	-0.0116381	-0.01114356	-0.006164669	-0.005499124

The default output also includes graphs (histogram and NQQ plot) of the bootstrap distribution for each coefficient.

## Robust regression

One of the strengths of the S language, on which S-Plus is based, has always been its diverse set of methods for fitting regression models and smoothing scatterplots. Two of the most useful of these are methods for fitting linear models using criteria which are more robust than least squares: Robust MM Regression and Least Trimmed Squares Regression. Their primary use, within the scope of this course, is for comparison with standard least-squares regression results, as a way of assessing the influence of unusual observations.

There are several other methods on the Regression menu which are for fitting linear models with categorical response variables, and therefore also use criteria other than least squares; these specialized analyses are beyond the scope of this course and so will not be covered in this handout.

### Robust MM regression

**Statistics** ⇒ **Regression** ⇒ **Robust MM...**

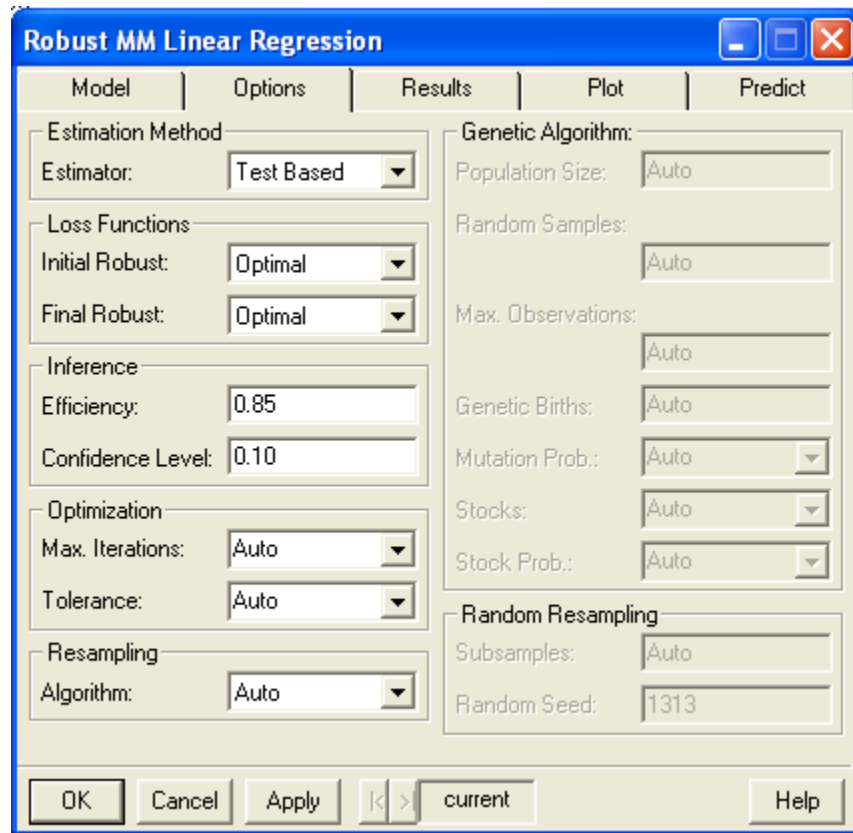
This method minimizes the sum, over all the observations, of a function of the residuals, as does least squares. This “loss function” in the “robust MM” method, however, does not give as much weight to very large residuals as does least squares. As a result

it is much less sensitive to outliers in either the response or the explanatory variable(s) than is least-squares regression.

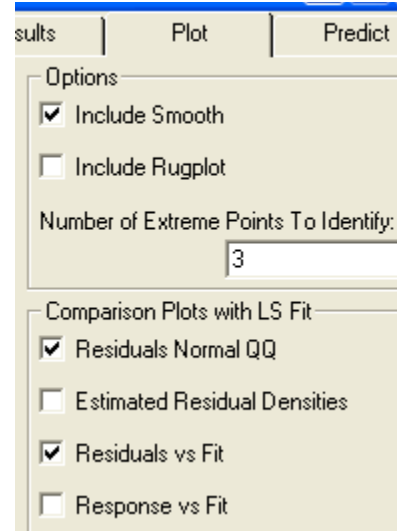
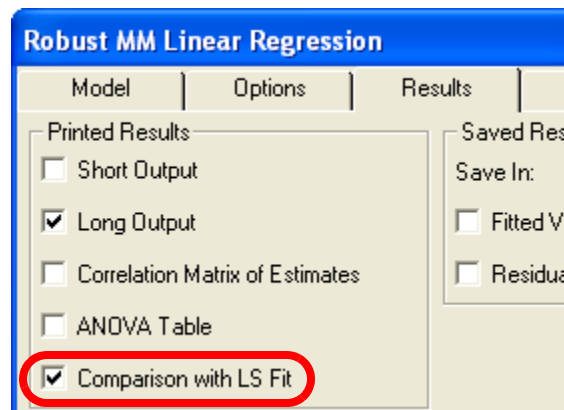
The Robust MM regression procedure provides hypothesis tests and confidence intervals for coefficients, as well as the usual array of residual plots. **To plot the robust fit you will need to have the “fits” stored (select this on the Results tab) and then construct the plot.**

## Dialog

Most of the dialog for specifying the robust MM regression is the same as for ordinary least-squares regression as described above. The principal difference is the addition of an Options tab (see next page), on which you can modify parameters relating to the fitting criterion. Unless you understand the details of the model-fitting criterion, leave these options alone.



Another other useful feature is that on the Results tab you can request a comparison between the robust MM and least-squares results, and on the Plot tab you can request that various residual plots be compared between robust MM and least-squares models (see example dialogs below).



## Output

The basic output from Robust MM regression is similar to that from LS regression. The most important parts again are the test for the slope, and the measure of the fraction of the variation in the response explained by the regression (analogous to  $R^2$ ). The output also includes tests for bias of the robust-MM and the least-squares models.

```

Residuals:
  Min       1Q   Median       3Q      Max
-0.6453 -0.1068 -0.006805  0.07244  0.3547

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  0.7781    0.0601   12.9456  0.0000
          eggs -0.0111    0.0024   -4.5230  0.0000

Residual scale estimate: 0.1605 on 83 degrees of freedom

Proportion of variation in response explained by model: 0.2488

Test for Bias
              Statistics P-value
M-estimate      0.647  0.7238
LS-estimate      7.235  0.0269

The seed parameter is : 1313

```

If the comparison of robust-MM and least-squares results is requested, most of the preceding output is repeated, but accompanied by the corresponding LS results. The Residual Scale Estimates are the standard error of the LS residuals and its robust analog from the MM regression. The Correlations at the bottom of this out-

put are correlations among the coefficients (i.e. the estimates of the regression parameters).

```

Residual Statistics:
      Min      1Q      Median      3Q      Max
LS.Fit -0.5976 -0.08508  0.009312  0.11674  0.4024
Robust.Fit -0.6453 -0.10680 -0.006805  0.07244  0.3547

Coefficients:
              Value Std. Error t value  Pr(>|t|)
LS.Fit_(Intercept)  0.704579   0.043842  16.071  0.00000000
Robust.Fit_(Intercept)  0.778141   0.060108  12.946  0.00000000
LS.Fit_eggs  0.008914   0.001832  -4.865  0.00000536
Robust.Fit_eggs -0.011067   0.002447  -4.523  0.00002011

```

```

Residual Scale Estimates:
  LS.Fit : 0.1852 on 83 degrees of freedom
Robust.Fit : 0.1605 on 83 degrees of freedom

Proportion of variation in response(s) explained by model(s):
  LS.Fit : 0.2219
Robust.Fit : 0.2488

Correlations:
  LS.Fit
      eggs
(Intercept) -0.8888928

Robust.Fit
      eggs
(Intercept) -0.8921922

```

## Robust LTS

### Statistics ⇒ Regression ⇒ Robust LTS...

This method minimizes a sum of squared residuals, as in least-squares regression, but in LTS regression only some of the observations are included in this sum: those with the smallest residuals. Typically just over half the observations are used. This procedure therefore produces a model that fits at least half the data points well but totally ignores the remainder of the data. This fit therefore is very robust to unusual observations and indeed can be very useful for detecting clusters of unusual observations.

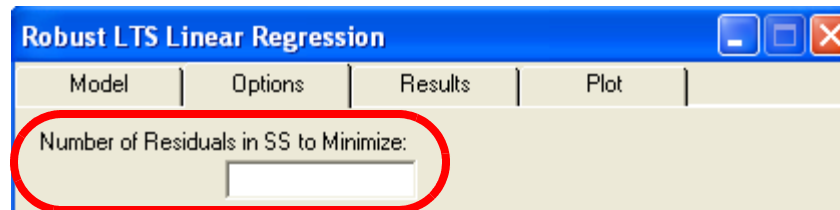
The Robust LS regression procedure does not provide any inference; it simply fits the model. It does provide some residual plots, but again **to plot the robust fit you will**

**need to have the “fits” stored** (select this on the Results tab) **and then construct the plot.**

### Dialog

Most of the dialog for LTS regression is similar to those for LS or MM regression, though there are fewer options available on the Results and Plot tabs.

There is only one item on the Options tab: how many residuals to include in the sum that is to be minimized. The default value is  $(n+2)/2$  (rounded down if  $n$  is odd); increasing this number causes the procedure to “ignore” less of the data, and therefore be less robust.



### Output

As already noted, LTS regression does not perform inference. The output therefore does not include standard errors or tests for the coefficients. It simply gives the estimates and some statistics describing the residuals and the explained variation.

```

Coefficients:
  Intercept    eggs
    0.7684   -0.0108

Scale estimate of residuals: 0.1685

Robust Multiple R-Squared: 0.3666

Total number of observations: 85

Number of observations that determine the LTS estimate: 76

Residuals:
  Min. 1st Qu.  Median 3rd Qu.  Max.
-0.6389 -0.1025 -0.0025  0.0703  0.3611

Weights:
 0  1
 2 83
  
```

# "CORRELATION"

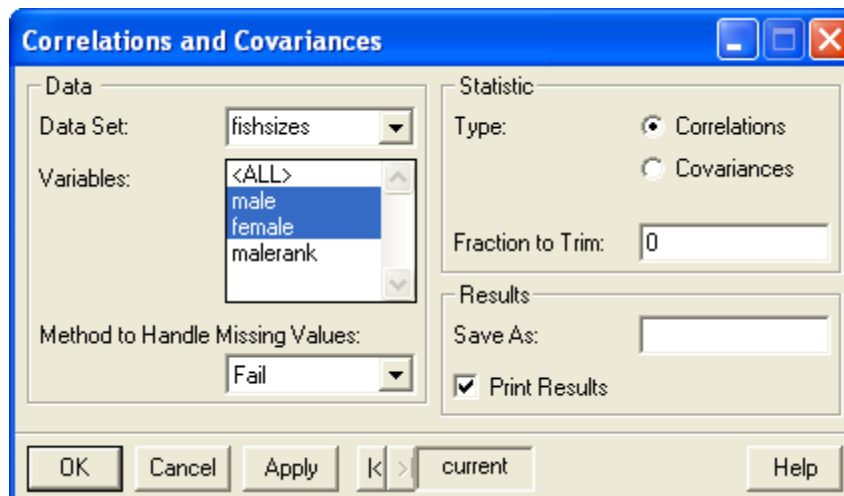
---

## Correlation

### Dialog

**Statistics ⇒ Data Summaries ⇒ Correlations...**

then specify the two (or more) variables. (Also be sure the Print Results box is checked.) If the Resampling Library is loaded, the option to have the most extreme observations trimmed from the data will be part of this dialog. I am not sure how "extreme" is measured in this bivariate setting.



### Output

**This only produces the correlation coefficients; no inference is performed.** To test the hypothesis of no correlation, use the regression procedure above.

```
          male   female
male 1.0000000 0.8879311
female 0.8879311 1.0000000
```

---

## Correlation - resampling

**Statistics ⇒ Data Summaries ⇒ Correlations...**

or

**Statistics ⇒ Resample ⇒ Correlations**

## Dialog

Select the variables for which correlations are desired, as shown above for the standard correlation analysis. On the **Bootstrap** and/or **Permutations** tabs check to have these procedures done, and optionally request different output.

## Methods

The bootstrap procedure, which produces confidence intervals for the population correlation, uses resampling (with replacement) of paired observations, i.e. keeping values of  $X$  and  $Y$  from a given observation together. This is the same method as in the bootstrap procedure for regression.

The permutation test is of the null hypothesis of no correlation:  $H_0: \rho = 0$ . It randomizes (shuffles) the observations of one variable in relation to the observations of the other variable. This is analogous to Kendall's test.

## Output

The bootstrap output is similar to that of other bootstrap procedures described already, the useful part being the list of the various types of CIs.

The output from the permutation test is simply the observed  $r$ , the mean and standard deviation of the distribution of  $r$ s from the randomization, a statement of  $H_a$  (despite there being no apparent way to request a one-sided test), and finally the  $P$ -value.

Summary Statistics:					
	Observed	Mean	SE	alternative	p-value
cor(male, female)	0.8879	-0.004179	0.2565	two.sided	0.002

Both the bootstrap and the permutation test also produce the usual histogram and NQQ plot of their respective distributions.

---

## Spearman's rank correlation

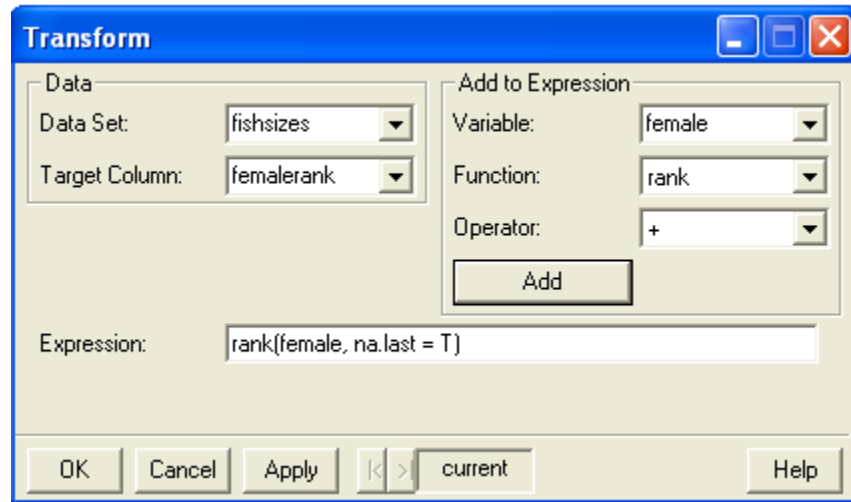
S-Plus does not explicitly provide Spearman's rank correlation. It can, however, be conducted fairly easily.

1. First transform the data to ranks:

### Data $\Rightarrow$ Transform...

The Target Column is the column you want the ranks put into. In the Add to Expression section, select the Variable you want to be ranked, select rank on the Function pull-down list, then click the Add button. Once the window looks like the following, click OK or Apply (the latter is convenient if you want to rank several variables,

since the window stays open and you only need to change the parts that name the variables).



2. Calculate the correlation coefficient as above using the columns of ranks, and/or run a regression (the latter provides a hypothesis test).

## CATEGORICAL VARIABLES

---

### Single proportion

Inferences about a single proportion can be done in three ways: exact binomial hypothesis test, normal-approximation hypothesis test and confidence interval, and resampling confidence intervals.

### Binomial test

This procedure only performs a hypothesis test.

**Statistics** ⇒ **Compare Samples** ⇒ **Counts and Proportions** ⇒  
**Binomial Test...**

### Data

To use the binomial test you must already know the count of “successes” and the number of “trials.” (If you do not have these numbers already, you can use the Tabulate procedure, discussed below, and specify only the one variable of interest, and a column of counts if appropriate.)

### Dialog

The dialog then requires only that you enter the numbers of successes and of trials and (optionally) modify the null and alternative hypotheses.

**Exact Binomial Test**

Data

No. of Successes: 14

No. of Trials: 21

Test Hypotheses

Hypothesized Proportion: 0.5

Alternative Hypothesis: greater

Results

Save As:

Print Results

OK Cancel Apply < > current Help

## Output

The output is short, since there is little to this test.

```
data: 14 out of 21
number of successes = 14, n = 21, p-value = 0.0946
alternative hypothesis: true p is greater than 0.5
```

## Normal approximation

This procedure is less accurate than the binomial test, but can provide confidence intervals for population proportions, which the binomial test does not.

**Statistics** ⇒ **Compare Samples** ⇒ **Counts and Proportions** ⇒ **Proportions Parameters...**

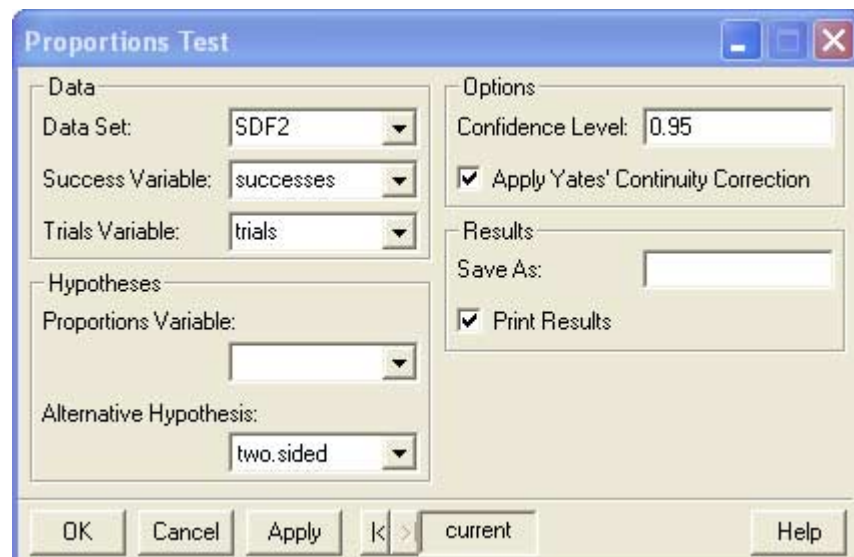
## Data

To use this procedure you must have the data in two columns, specifying the counts of successes and of trials. In the context of a single proportion, these columns would have only one row, as shown to the right.

	successes	trials
	14	21

## Dialog

All that needs be specified is the Success Variable and the Trials Variable. Optionally the confidence level can be changed and/or a one-sided test can be requested.



The default null hypothesis is that the population proportion (i.e. probability) equals 0.5. To tests a different null hypothesis, create a third column in the data set, with a single entry containing the hypothesized probability, and select it as the Proportions Variable.

## Output

The output includes the hypothesis test (expressed as a  $X^2$  test) and the confidence interval obtained using the normal approximation.

```
data: successes and trials from data set SDF2
X-square = 1.7143, df = 1, p-value = 0.1904
alternative hypothesis: true P(success) in Group 1 is not equal to 0.5
95 percent confidence interval:
 0.4310506 0.8451865
sample estimates:
prop'n in Group 1
```

## Resampling methods

These procedures provide the exact binomial hypothesis test and several versions of confidence intervals for the population proportion.

**Statistics ⇒ Compare Samples ⇒ Counts and Proportions ⇒  
Proportions Parameters/Resample...**

or

**Statistics ⇒ Resample ⇒ Proportions**

### Data

This procedure is an augmented version of the Proportions Parameters procedure just described, and thus requires the data to be in the same layout: in two columns specifying the counts of successes and of trials.

### Dialog

The first dialog tab, Proportions, is identical to that shown above for the Proportions Parameters procedure.

The other tabs, Bootstrap and Permutations, are just as in all the other resampling procedures. With only one proportion, however, the Permutations tab doesn't do anything. All that must be done on the Bootstrap tab is to check the boxes to have the bootstrapping done; optionally, the selection of what CIs are produced can be changed.

### Method

I'm not sure of this, but it appears that this procedure resamples (with replacement, as always), from a sample of  $x$  successes and  $n-x$  failures, where  $x$  is the number of successes in the actual data.

### Output

The output includes the standard normal-approximation results described above. It also includes, regardless of whether bootstrapping (or permutation testing) is

requested, the exact binomial test and a “Wilson” confidence interval (what the text calls a “plus four” confidence interval; see p. 539).

```
Z, Wilson estimates, and exact p-value for one proportion
z: 1.62
Wilson estimate: 0.64
95 percent Wilson confidence level: 0.4518 0.8282
```

If bootstrapping is requested the output also includes the various confidence intervals and the graphs of the bootstrap distribution, as described previously for other bootstrapping procedures.

## Contingency tables

The difference between two or more proportions — that is, the relationship between two categorical variables — can be tested using either a chi-square test or Fisher’s exact test.

### Data layouts

Contingency table data can be in any of three arrangements, but the various test procedures have limitations as to which arrangements they can handle.

- individual observations:

Two “factor” columns contain the values of the two categorical variables, and each observation is a separate row. The example to the right is in this layout.

Data in this layout can only be analyzed by any of the three procedures described below.

- stacked frequencies:

Two “factor” columns contain the values of the two categorical variables, and a third column contains the frequency of that row’s combination of values of the categorical values. The example below is in this layout.

This layout can be produced from a data set of individual observations using

**Data ⇒ Tabulate...**

.Select the two categorical variables in the pull-down list of Variables:. Also give a name for the data set to contain the cross-tabulation.

Data in this layout can only be analyzed by the Cross Tabulations procedure

	site	cause
	at	smother
	at	crowd
	at	unk
	at	smother
	at	smother
	at	crowd
	above	crowd
	above	crowd
	above	unk
	above	unk
	above	smother
	above	smother
	above	crowd
	below	crowd
	below	crowd

SDF1				
		1	2	3
		cause	site	count
1		smother	above	49
2		smother	at	11
3		smother	below	13
4		undercut	above	5
5		undercut	at	19
6		undercut	below	17
7		crowded	above	8
8		crowded	at	4
9		crowded	below	2
10		unknown	above	19
11		unknown	at	23

- contingency table (= unstacked frequencies):

Separate columns represent the different levels of one of the categorical variables, and entries in each row are counts for the levels of the other categorical variable. **The analyses use the entire data set as the contingency table, so it cannot contain any extra columns, such as a column of labels for the rows.** The example above is in this layout.

	1	2	3	4
above	49.00	11.00	13.00	
	5.00	19.00	17.00	
	8.00	4.00	2.00	
	19.00	23.00	38.00	

Data in this layout can be analyzed by the Chi-square Test or Fisher's Exact Test but not by the Cross Tabulations procedure.

## Cross Tabulations

**Statistics ⇒ Data Summaries ⇒ Cross Tabulations...**

This procedure can perform a chi-square test on two categorical variables, with any number of levels of each. It does not do the continuity correction.

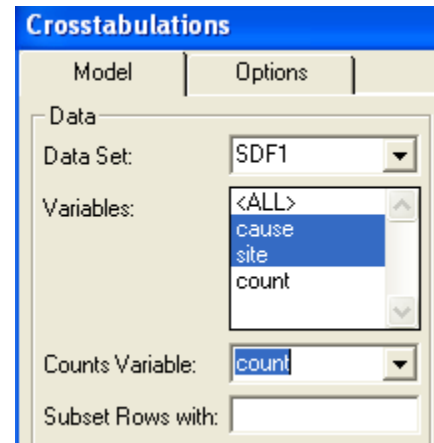
### Dialog

On the Model tab select the two categorical variables in the Variables box. If the data are as stacked frequencies, specify the Counts Variable.

On the Options tab be sure that the Run Chi-Square Test option is checked (this is the default).

### Output

The output (next page) includes the contingency table itself, by default with cell counts also expressed as fractions of the row, the column, and total. The table is followed by the chi-square statistic and test.



```

208 cases in table
+-----+
|N      |
|N/RowTotal|
|N/ColTotal|
|N/Total |
+-----+
cause  |site
      |above |at   |below|RowTotl|
+-----+-----+-----+-----+
crowded| 8     | 4   | 2   |14     |
      |0.57  |0.29 |0.14 |0.067  |
      |0.099 |0.07 |0.029|       |
      |0.038 |0.019|0.0096|      |
+-----+-----+-----+-----+
smother|49     |11   |13   |73     |
      |0.67  |0.15 |0.18 |0.35   |
      |0.6   |0.19 |0.19 |       |
      |0.24  |0.053|0.062|      |
+-----+-----+-----+-----+
underct| 5     |19   |17   |41     |
      |0.12  |0.46 |0.41 |0.2    |
      |0.062 |0.33 |0.24 |       |
      |0.024 |0.091|0.082|      |
+-----+-----+-----+-----+
unknown|19     |23   |38   |80     |
      |0.24  |0.29 |0.48 |0.38   |
      |0.23  |0.4  |0.54 |       |
      |0.091 |0.11 |0.18 |      |
+-----+-----+-----+-----+
ColTotl|81     |57   |70   |208    |
      |0.39  |0.27 |0.34 |       |
+-----+-----+-----+-----+
Test for independence of all factors
Chi^2 = 50.1379 d.f.= 6 (p=4.411115e-009)
Yates' correction not used
Some expected values are less than 5, don't trust stated p-value

```

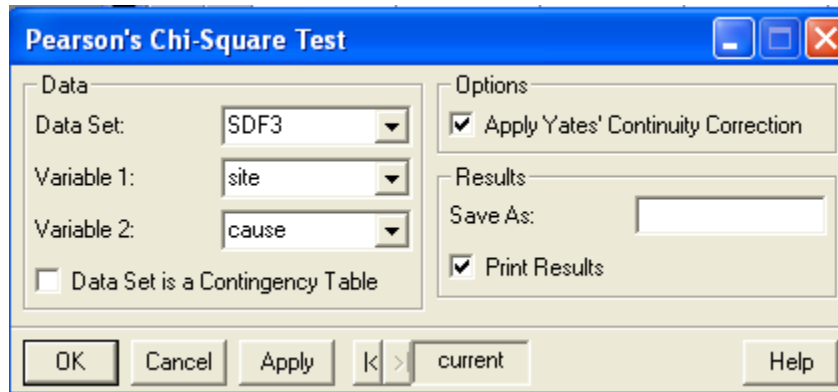
## Chi-square test

**Statistics ⇒ Compare Samples ⇒ Counts and Proportions ⇒  
Chi-square Test...**

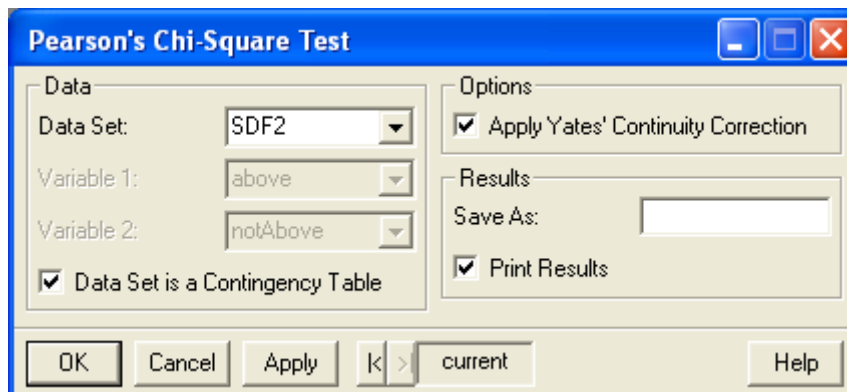
This procedure can perform a chi-square test on two categorical variables, with any number of levels of each. It supposedly can do the continuity correction, but when I prepared this example it would not do so.

## Dialog

Invoking this analysis is simple, with almost no options. If the data are as individual observations, simply select the two categorical variables as Variable 1 and Variable 2, as in the first example below.



If the data are in the contingency table layout, simply check the Data Set is a Contingency Table option, as in the second example below; Variable 1 and Variable 2 do not need to be specified. As noted above, the **entire** data set will be treated as being the contingency table, so there cannot be any additional columns.



The Apply Yate's Continuity Correction option is checked by default, but as noted above it is not always used even when checked.

## Output

In contrast to the Cross Tabulations procedure, the Chi-square procedure creates almost no output: simply the  $X^2$  statistic, the  $df$ , and the  $P$ -value.

```
Pearson's chi-square test without Yates' continuity correction
data: barnunst
X-square = 50.1379, df = 6, p-value = 0
```

---

## Fisher's exact test

Fisher's exact test is invoked exactly as for the chi-square test described immediately above (except that there is no option for the continuity correction). Because of the complex computations required, **this test will not run if there are more than 200 total observations in the table**. The output (above) again is minimal.

```
Fisher's exact test
data:  site and cause from data set SDF3
p-value = 0.7344
alternative hypothesis: two.sided
```

# POWER ANALYSIS AND MISCELLANEOUS PROBABILITY PROCEDURES

---

## Power/sample size analysis

Although S-Plus may have procedures for calculating power (and thus determining sufficient sample sizes) for more complex designs such as ANOVA, the only procedures I have been able to find using the menu system (GUI) are for one-, paired- and two-sample designs. There are slightly different versions for analysis of means (i.e.  $t$  tests) and analysis of proportions (i.e. binomial tests).

### Tests of means

**Statistics** ⇒ **Power and Sample Size** ▶ ⇒ **Normal mean...**

This procedure encompasses single-, paired-, and two-sample situations.

#### Dialog

In the **Select** section of the Model tab (shown below) are three radio buttons giving a choice as to which parameter — Sample Size, Power, or Min. Difference — will be computed (and thus implicitly which will have to be specified). Also in that section is a scroll box in which the Sample Type: is specified.

Which other parts of the dialog are available depends on the choices made as to sample type and which parameter to compute.

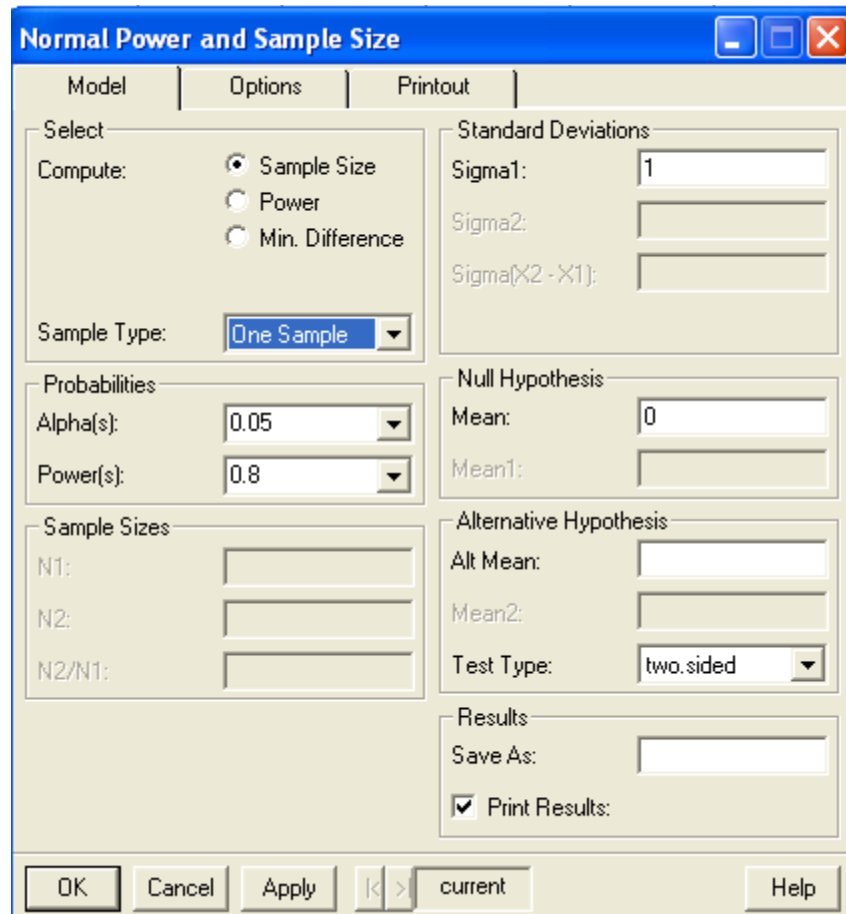
Note that for any of the quantities that are specified in the dialog, multiple values are allowed; they are to be separated by commas. This allows a range of possible scenarios to be explored quickly and easily.

#### Standard deviations

For any of the sample types and “parameters to compute,” the variability must be specified. For a single sample this is the standard deviation of that population, given as  $\text{Sigma1}$ . For paired samples the pertinent measure is the standard deviation of the differences:  $\text{Sigma}(X2 - X1)$ . For the two-sample case, the standard deviation of the first population,  $\text{Sigma1}$ , must be specified; the standard deviation for the other population,  $\text{Sigma2}$ , also can be specified, or it can be left as the default “ $\text{Sigma1}$ ,” meaning the two are assumed equal.

#### Probabilities

For all cases, the significance level of the test must be specified, in the Alpha(s): box. Unless the parameter to be computed is the power, the power also must be specified, in the Power(s): box. Both are entered as probabilities in decimal fraction format.



### Sample Sizes

For single- or paired-samples, there is only one sample size to consider, entered as N1. For two samples, both sample sizes need to be determined, by specifying any two of three quantities: the first sample size N1, the second sample size N2, and the ratio of the second sample size to the first, N2/N1.

### Null Hypothesis

For the single-sample situation, the Mean is specified; for paired- and two-sample situations, the mean of the first population is specified, as Mean1. The null hypothesis in the latter cases is assumed to be that there is no difference, i.e. that the mean of the second population also equals Mean1. The role of this parameter seems to be only that it together with the parameter specified for the Alternative Hypothesis actually define the magnitude of the difference under the alternative hypothesis.

### Alternative Hypothesis

For a single-sample test the population mean according to the alternative hypothesis is specified, called Alt. Mean; the “minimum difference” is the difference between this value and the Mean specified for the Null Hypothesis. For paired- and two-sample tests, the mean for the second population (or second condition, for paired data) is specified, called Mean2; the “minimum difference” is the difference between this value and Mean1 as specified in the Null Hypothesis section.

In all cases the test can be specified as two-sided, or as one-sided in either direction.

### Other tabs

There are tabs for Options and Printout (the output) but the defaults normally will be adequate.

### Output

The output from these procedures is in the form of a table such as shown below. There is a row for each combination of values of the various quantities specified in dialog. The columns will include two means, either `mean.null` and `mean.alt` for the single- and paired-sample cases, or `mean1` and `mean2` for two samples. It also will include one or two standard deviations, `sd1` and (if appropriate) `sd2`. Then will come a column labeled `delta`, which is the difference between the means according to the alternative hypothesis. Then are `alpha` and `power`, and one or two samples sizes, `n1` and (for two samples) `n2`.

An example of this output, for a single sample, with sample size being computed, and for two alternative hypothesis, is as shown here:

*** Power Table ***							
	<code>mean.null</code>	<code>sd1</code>	<code>mean.alt</code>	<code>delta</code>	<code>alpha</code>	<code>power</code>	<code>n1</code>
1	0	1	1	1	0.05	0.8	8
2	0	1	2	2	0.05	0.8	2

As this example shows, the output has no labeling as to what specifications went into producing it. In some respects this is not a problem. The output does not state which quantities were specified and which were computed, but it should be obvious since the computed parameter typically will not be an integer or neat fraction, and in any event the results are valid regardless of which parameters were specified and which were computed. The output also does not say which design the computations are for, but again inspecting what columns are in the output will indicate this.

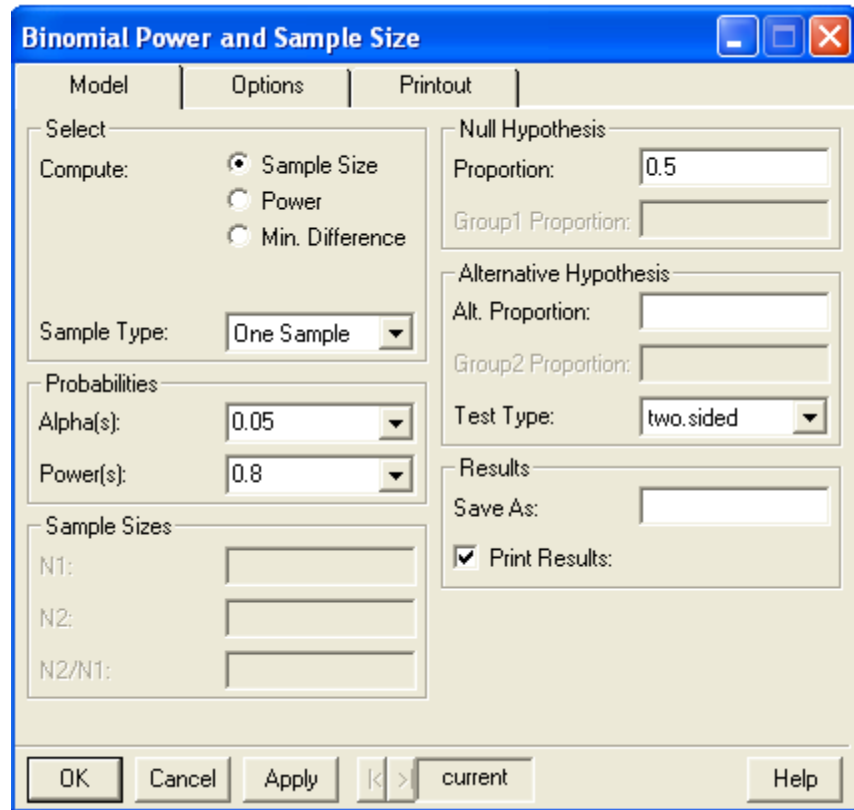
What is bothersome about the lack of labeling of the output is that the output gives no indication of whether it is for a two-sided or one-sided test, and there is no obvious way to deduce this from the results.

## Tests of proportions

### Statistics ⇒ Power and Sample Size ▶ ⇒ Binomial proportion...

This procedure is very similar to that described above for means, so it will be described only briefly. As shown by the dialog window below, this procedure can handle the same three sample situations as the procedure for means: a single sample, paired samples, and two samples. It also allows any of the three types of parameters — sample size, power, or minimum difference — to be computed. Specifying the probabilities and sample sizes is the same as in the procedure for means. Specifying the null

and alternative hypotheses is similar, except that the population parameters of interest are proportions rather than means. Finally, since the standard deviation of a binomial distribution is fully determined by the population proportion, there is no need to specify standard deviations.



## Random numbers

### Randomizing

S-Plus' random-number generator can be used to select a random sample or to randomize treatment over units. The most convenient way to do this is to:

1. create a data set containing a single column, in which are listed the labels of all the units in the population to be sampled or all the experimental units (which in many cases will be simply a sequence of integers).
2. Select

**Data ⇒ Random Sample...**

In the resulting dialog, name the data set containing the list of units and the desired sample size; if randomizing more than two treatments, set the sample size equal to the total of all samples, and simply divide the resulting randomized list into the desired number of groups. The other parts of the dialog can be ignored.

The result will be a new data set (you can give it a name in the preceding dialog if you wish) which will contain a column of the unit labels, in randomized order.

## Simulations

The random-number generator also can be used to simulate samples from a wide variety of distributions.

### **Data ⇒ Random Numbers...**

In the dialog for this procedure, name the Data Set and Target Column into which the random numbers are to go, and the Sample Size. Select a type of Distribution from the scroll list; there are 17 possibilities. Then specify the parameters for that distribution; of the long list of possible parameters, only the (typically) one or two relevant to the chosen distribution will be available for input.

---

## Probabilities

The procedures for probability distributions can be useful for determining critical values for confidence intervals or tests, or determining P-values for tests you have hand calculated. The menu selection is

### **Data ⇒ Distribution functions...**

You specify the Distribution and its parameters, and a Source Column containing the value(s) of interest; what these are depends on what sort of result is to be gotten, as described below. You can optionally change the name of the column in which the results will be stored.

This procedure, for any of the many distributions, has radio buttons to choose among Probability, Density, and Quantile.

- **Probability**  
This is the probability of all values of  $X$  less than or equal to the given value  $x$ , i.e. the area under the density curve to the left of the value, as in Table A in the text. For discrete values it does include the probability of the exact value.
- **Density**  
For discrete variables (e.g. binomial distribution) this is the probability of the given value  $x$ , as in Table C in the text. For continuous variables (e.g. normal distribution) it is the height of the density curve at the given value; this is not likely to be useful.
- **Quantile**  
This is the value corresponding to the specified quantile(s) of the distribution. In other words, this is a “critical value,” i.e. the value  $x$  which has cumulative proba-

bility equal to the quantile value you specify. This is as in Tables D, E, and F in the text, except that it is left-sided (i.e. equal to 1 minus the values in the text tables).

When the Source Column contains values of a random variable — for instance, a test statistic or column of test statistics — Probability and Density are the pertinent Result Type(s). Conversely, when the Source Column contains probability values — e.g. confidence or significance levels — the appropriate choice of result is Quantile.