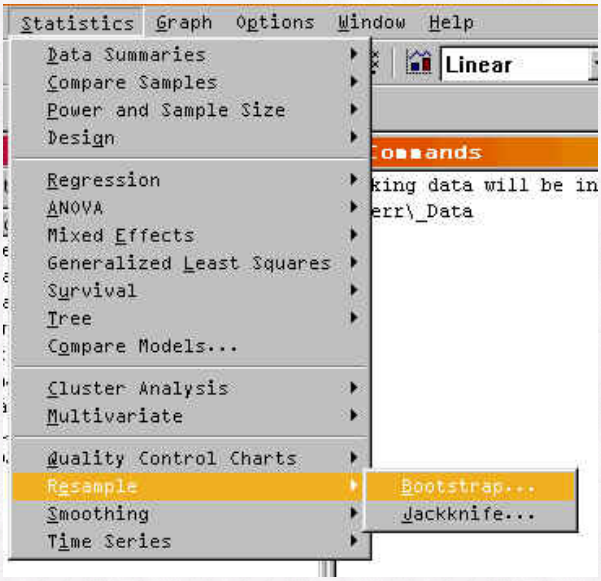
RSS Matters

Resampling Based Statistics in S-Plus

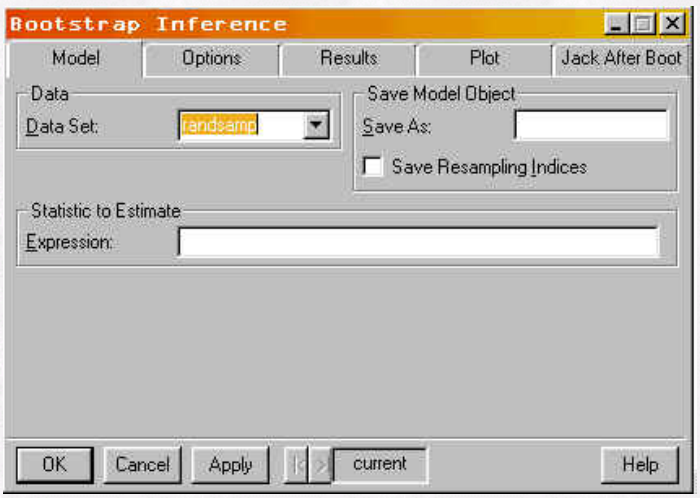
By Rich Herrington, Research and Statistical Support Services

This month we take a look at the bootstrap resampling capabilities of S-Plus. S-Plus has general bootstrapping functionality available so that nearly all statistical functions and expressions can be bootstrapped. S-Plus provides both parametric and nonparametric bootstrap confidence intervals.

From the main menu bar, we access the resampling menu from: **Statistics - Resample - Bootstrap**.



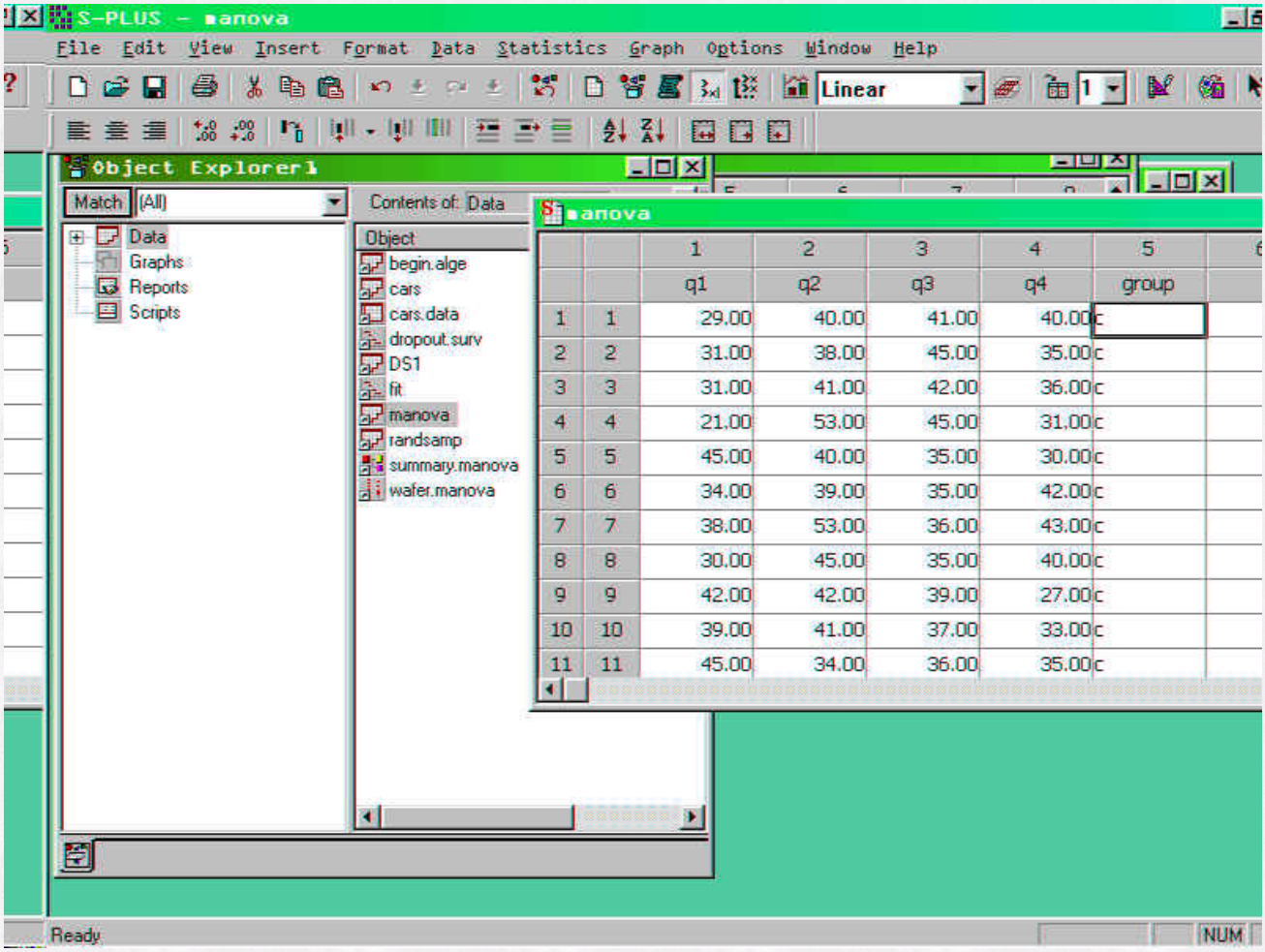
The menu for the Bootstrap facilities has five entry areas for initializing the Bootstrap analysis: **Model, Options, Results, Plot,** and **Jack After Boot**. Each of these option tabs are initialized with default values. However, the critical entry field which does not have a default entry is the **Expression** entry field. Entering an expression to bootstrap can be tricky as this assumes that the user has some knowledge of the syntax of the S-Plus language.



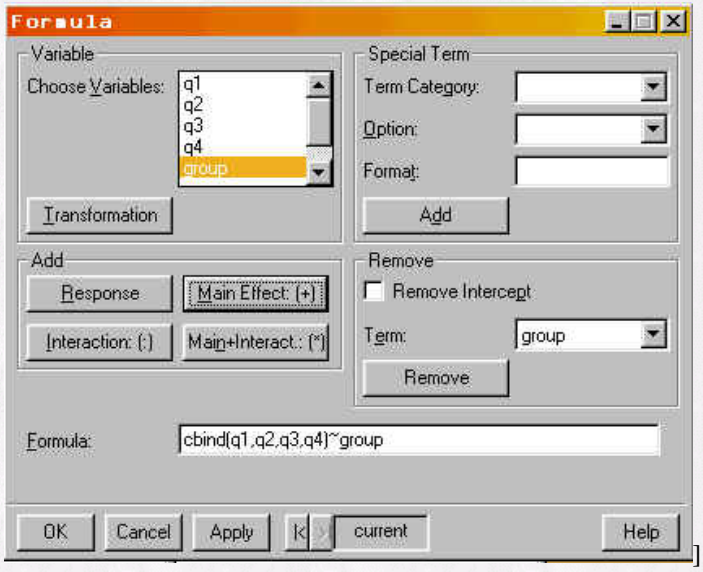
One way of avoiding having detailed knowledge of the syntax used to generate a particular analysis, is to generate the analysis before hand from the drop down menu system. Once this analysis has been run, the syntax used to generate the analysis is displayed. Essentially, the drop down menu system generates the syntax as entry fields are filled in. After an analysis is run from the menu system, this syntax can be saved, cut and pasted back into the **Expression** entry field. In the following example we will perform a four-group MANOVA with four dependent measures.

Example

The data set we will use for our analysis will have four groups: a control group and three experimental groups (c1, e1, e2, e3). We see a screen capture of the object browser and the data worksheet:



From the main menu bar select: **Statistics - Multivariate - MANOVA**. Select the **Create Formula** tab. Fill out the create formula tab with the following specifics. First select q1 through q4 and click **Add Response**. Then select group and click **Add Main Effect**:



Select **OK** to return to the previous menu. Select **OK** once more to actually run the analysis. In the report window we see the following:

```
*** Multivariate Analysis of Variance Model ***

Short Output:
Call:
   manova(formula = cbind(q1, q2, q3, q4) ~
       group, data = manova, na.action =
       na.exclude)

Terms:
                group Residuals
Deg. of Freedom     3       123

Estimated effects may be unbalanced

Analysis of Variance Table:
            Df Pillai Trace approx. F
    group 3   0.0775       0.8089
Residuals 123
          num df    den df    P-value
    group     12       366    0.6413
Residuals
```

The calling function is listed under **Call**. Copy the **manova(formula.....)** and paste this into your **Commands window**. Use the **summary** function to summarize the call to the manova function. Assign this summary to an object, **man.out**, for example:

```
> man.out<-summary(manova(formula = cbind(q1, q2, q3, q4) ~
+       group, data = manova, na.action =
+       na.exclude))
> man.out
           Df Pillai Trace approx. F num df    den df    P-value
    group 3   0.0775       0.8089        12       366    0.6413
Residuals 123
> names(man.out)
[1] "row.names"    "SS"           "Df"           "Eigen.values" "Stats"        "test"
>
```

Typing **man.out** by itself displays the contents of this object.**names** displays the components of this list. We have six components to this list. To extract the fifth element **"Stats"**. We have to index the list in the following fashion:

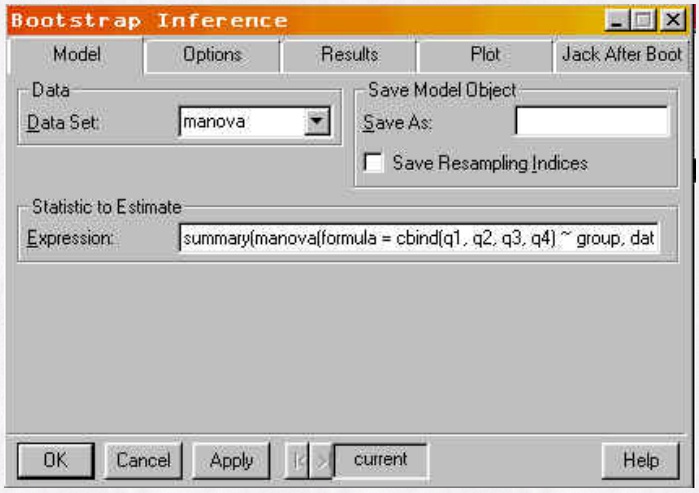
```
> man.out[[5]]

, , Statistic
          pillai wilks lambda hotelling-lawley roy largest
group 0.07751211    0.9240134       0.08059128  0.05130897

, , approx. F
         pillai wilks lambda hotelling-lawley roy largest
group 0.8089407    0.8029447        0.7969582    1.564923

, , num df
      pillai wilks lambda hotelling-lawley roy largest
group     12           12               12           4

, , den df
      pillai wilks lambda hotelling-lawley roy largest
group    366     317.7817              356         122

, , P-value
         pillai wilks lambda hotelling-lawley roy largest
group 0.6412999    0.6474587        0.6537033   0.1879375
> man.out[[5]][2]
[1] 0.9240134
> man.out[[5]][3]
[1] 0.08059128
> man.out[[5]][4]
[1] 0.05130897
> man.out[[5]][5]
[1] 0.8089407
> man.out[[5]][12]
[1] 4
>
```

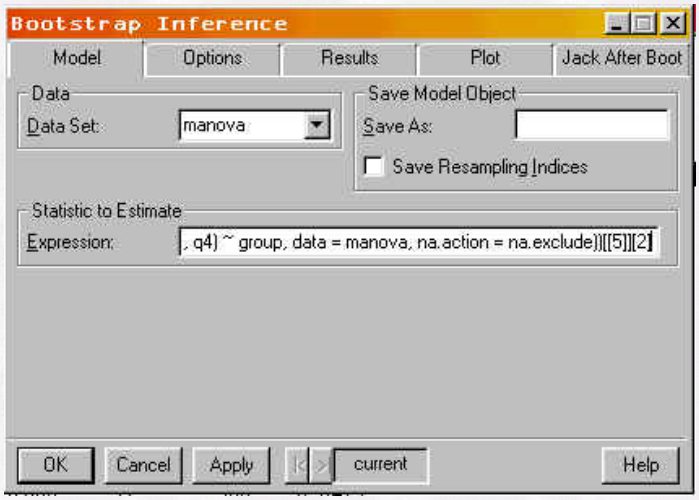
We see that wilks lambda (.9240) is the second index for the fifth element of the list, **man.out**. So the complete calling

function to the bootstrap function will be:

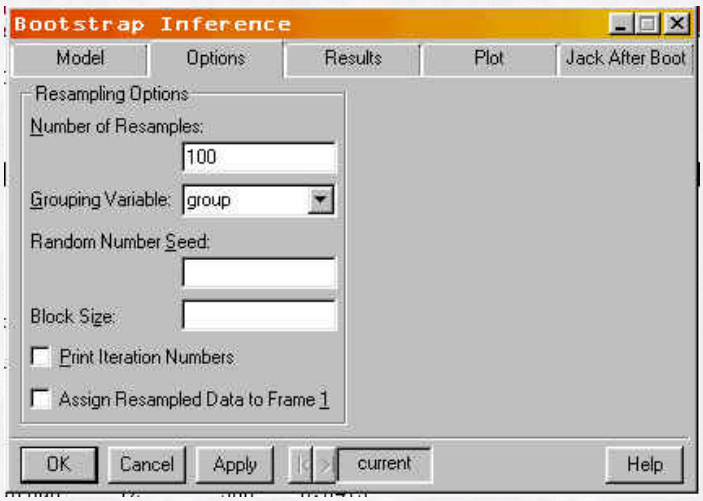
```
> summary(manova(formula = cbind(q1, q2, q3, q4) ~
+       group, data = manova, na.action =
+       na.exclude))[[5]][2]
[1] 0.9240134
>
```

This calling function returns a value of .9240 for wilks lambda for this particular data set. We need to copy this function call: **summary(manova.....))[[5]][2]**, into the **Expression** window on the bootstrap menu.

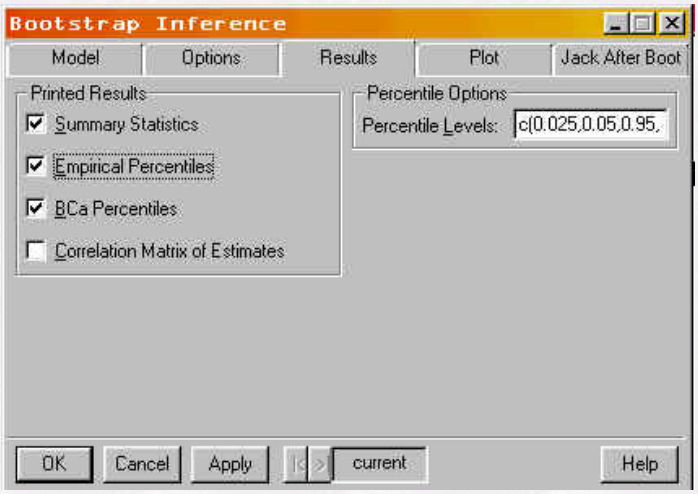




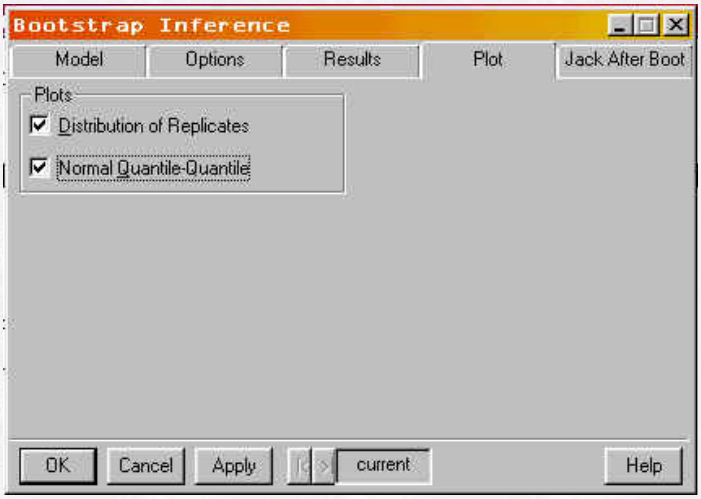
For the **Options** tab we need to select the grouping variable and how many bootstrap iterations we need:



For the **Results** tab we select **empirical percentiles**:



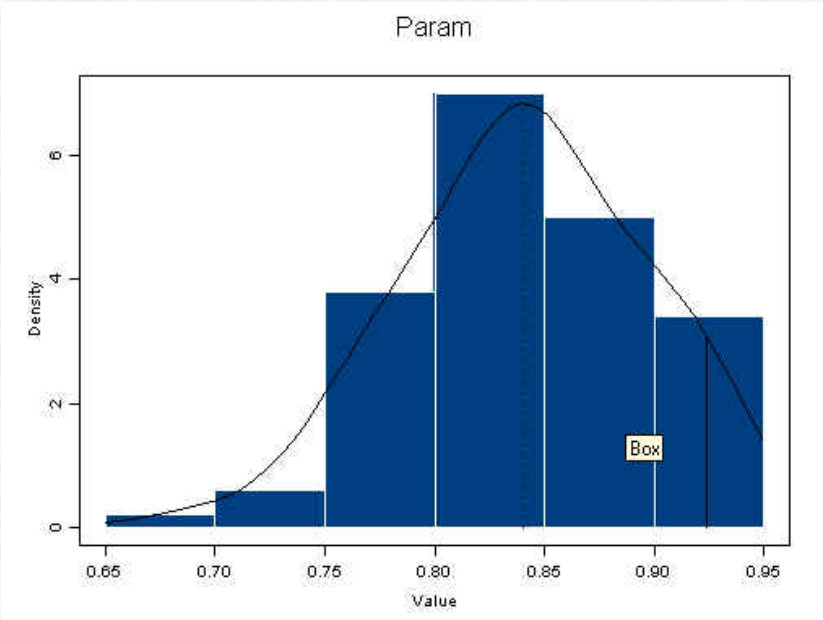
For the **Plot** tab we select **Normal Quantile-Quantile** to see how well the sampling distribution matches with "normal distribution" theory.

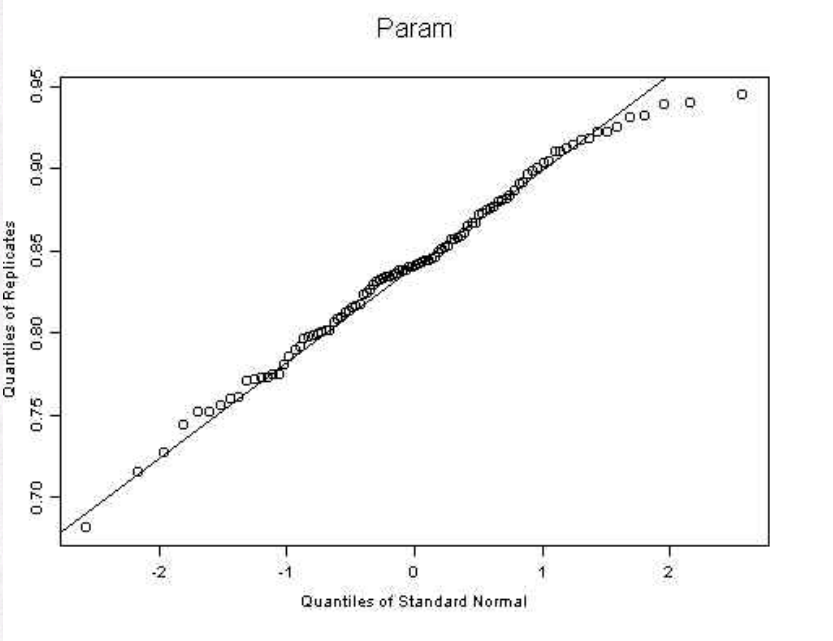


Selecting **OK** generates the following report window:

```
       *** Bootstrap Results ***
Call:
bootstrap(data = manova, statistic = summary(manova(formula = cbind(q1, q2, q3, q4) ~ group, data
        = manova, na.action = na.exclude))[[5]][2], B = 100, group = group, trace = F,
       assign.frame1 = F, save.indices = F)

Number of Replications: 100

Summary Statistics:
      Observed     Bias    Mean      SE
Param    0.924 -0.08357 0.8404 0.05491

Empirical Percentiles:
           2.5%        5%       95%     97.5%
Param 0.7348601 0.7518553 0.9250432 0.9352212

BCa Percentiles:
           2.5%        5%       95%     97.5%
Param 0.9100873 0.9213088 0.9448852 0.9448857
```

And the following plots:





We see that the empirically resampled sampling distribution for wilks lambda follows normal theory fairly closely except for the right tail region. We see that the upper and lower cut-offs for the 2.5/97.5th and 5/95th percentiles both contain the observed value of wilks lambda. We take this as a failure to reject the null hypothesis for wilks lambda. In general the BCa percentiles will be more accurate than the empirical percentiles.

Further Reading

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. San Francisco: Chapman & Hall.