Statistical Computing Tips

By **Rich Herrington**, ACS Statistical Consultant (**richherr@unt.edu**)

Sampling with Replacement in SPSS: Creating Bootstrap Confidence Intervals for the Correlation Coefficient

Resampling based approaches to estimating a statistic and determining the properties of the estimator are becoming increasingly identified with what is sometimes referred to as "modern methods" of data analysis (Fox and Long, 1990). Resampling, using subsamples of an original sample, allows one to incorporate all of the original characteristics of the observed empirical distribution into the significance estimates and confidence intervals for the statistic under consideration. Efron (1979) proposed a resampling plan, which he called the "bootstrap." The bootstrap technique has received considerable attention over the years since its popularization by Efron (an overview of these developments can be found in Efron and Tibshirani, 1993). In a bootstrap resampling scheme, the initial sample of observations is treated as if they constitute the population under study. By randomly resampling with replacement from this proxy population, the sampling error of the original population can be estimated and confidence intervals constructed for most statistics that are being evaluated.

For the present illustrative example, 11 pairs of observations constitute the sample/population. From these 11 pairs of data, sampling with replacement is used to create a new sample of size 11 (a bootstrap sample). In this new sample, the data pair (.18, .20) might appear only once or could appear multiple times in the sample of 11 data pairs. A Pearson's correlation coefficient is calculated for each of these bootstrap samples. The standard deviation of these bootstrap correlations is calculated thus giving an estimate of the standard error of the correlation coefficient. The calculation of the 2.5th and 97.5th percentiles of the distribution of the bootstrap correlations will give an estimate of the 95th confidence intervals for the correlation coefficient (percentile method). A bias adjustment for the percentile method (bias corrected percentile method) is discussed in Efron and Tibshirani (1993); here, we discuss only the unadjusted percentile method.

Thompson (1993) discusses using the bootstrap methodology in conjunction with traditional statistical significance testing to explore result replicability. Thompson's data set (page 370)is used as our example:

Data Set (From Thompson, 1993):

	Y	X
1.00	.18	.20
2.00	.54	1.88
3.00	-.49	-.76
4.00	..92	.42
5.00	.22	.32
6.00	.75	-.56
7.00	.66	1.55
8.00	-2.65	-1.21

9.00	-.51	-.66
10.00	.47	-.96
11.00	-.09	-.21

This data set produces the following statistics:

Normal Theory Significance and CI:

r = 0.56, p = .073, 95% CI = (-0.06, .868)

The SPSS syntax included here uses the SPSS INPUT PROGRAM to generate 1000 samples (n=11 per sample)of randomly sampled case id's (sampling with replacement). The MATCH FILES procedure is used to copy data from the original file (the x,y pairs) into the working data file.

```
**** Bootstrap Confidence Intervals for
**** the Correlation Coefficient
**** create 1000 bootstrap samples of size
**** n=11, use sampling with replacement
input program.
loop samp=1 to 1000.
+ loop #i=1 to 11.
+    compute id=trunc(uniform(11))+1.
+    end case.
+ end loop.
+ leave samp.
end loop.
end file.
end input program.
execute.
sort cases by id.
match files file=* /table='a:\thompson.sav' /by id.
sort cases by samp.
split file by samp.
execute.

**** calculate a correlation coefficient for each bootstrap sample
CORRELATIONS
/VARIABLES=y x
/PRINT=TWOTAIL SIG
/MISSING=PAIRWISE .
```

Once the correlation output has been saved to an output text file, one removes the sample ids, correlation values, and pvalues from the output file of the 1000 bootstrap samples:

```
SET WIDTH=80.
FILE TYPE NESTED FILE='a:\corr.out' RECORD=1-80 (A).
RECORD TYPE  SAMP:'.
DATA LIST / sample 9-16.
RECORD TYPE  X'.
DATA LIST RECORDS=3 / corr 13-18 // pvalue 16-19 .
END FILE TYPE.
FORMATS corr (F8.2) pvalue (F8.2)    sample (F8.2) .
execute.
```

Next, the lower 2.5th and upper 97.5th percentiles of the empirical distribution of correlation coefficients are calculated:

```
FREQUENCIES VARIABLES=corr
/FORMAT=NOTABLE
/PERCENTILES= 2.5 97.5
/STATISTICS=STDDEV MEAN.
CORR
Mean             .573        Std dev           .162

Percentile      Value        Percentile       Value
2.50            .179         97.50            .828

Valid cases     1000         Missing cases      0
```

Our bootstrap 95th percentiles are (.179, .828). Since these intervals do ot include 0, this is taken to

be a rejection of the null hypothesis, that the correlation coefficient is zero in the population Power estimation (Cohen, 1988) with the bootstrap is accomplished by counting the proportion of redrawn samples that lead to a statistically significant estimator (for a given alpha level):

```
**** Probability to reject an assumed false
**** null hypothesis (simulated power).
do if pvalue<<=.05).
compute count=1.
else if pvalue>>(.05).
compute count=0.
end if.
execute.
FREQUENCIES
 VARIABLES=count.
COUNT

                                      Valid     Cum
Value Label        Value  Frequency  Percent  Percent  Percent
                    .00        523     52.3     52.3     52.3
                   1.00        477     47.7     47.7    100.0
                              -        -        -
                   Total      1000    100.0    100.0
Valid cases 1000  Missing cases      0
```

Power estimate based on distributional assumptions (Using Cohen's power tables) = .460
Resampling based power estimate = .477

References:

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

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

Fox, J & Long, J.S. (1990). *Modern Methods of Data Analysis.* Sage Publications, Newbury, Park, CA.

Thompson, B. (1993). *The Use of Statistical Significance Tests in Research: Bootstrap and Other Alternatives.* Journal of Experimental Education, 61(4), 361-377.

If you have any problems or questions about this server, contact us as soon us as soon as possible. You can send mail to the following address: www@unt.edu