

Laboratory #11. Bootstrap Estimates

Randomization methods so far have been used to compute p-values for hypothesis testing. Randomization methods can also be used to place confidence limits around estimates. Randomization methods can be used to place a confidence limit around any statistic, not just those that have known statistical distributions. As can be seen by referring to the key for choosing a frequency distribution (see Handouts), randomization methods using the "bootstrap" method are especially useful when errors around an estimate are not normally distributed, or when the statistical distribution for a statistic is unknown.

The bootstrap is a technique for estimating a statistic and its distribution. It uses resampling. The resampling for randomization tests (Lab 4) was carried out WITHOUT replacement. The values of the response variable (in Minitab column c1, for example) were sampled without replacement into a new arrangement (in Minitab column c2, for example). Each value in the first arrangement (c1) occurs only once in the second (c2). The advantage of repeated sampling and calculation, of course, was that a statistical analysis could be undertaken without making assumptions about the distribution of outcomes.

Bootstrap methods use sampling WITH replacement. That is, any individual observation in one column can be sampled repeatedly. Any single observation in column c1 can appear repeatedly in c2. Or it may not appear at all in c2.

The bootstrap was invented by Bradley Efron (*Annals of Statistics* 7:1-26). It is described in Manly (1991: *Randomization and Monte Carlo Methods in Biology* Chapman Hall).

The purpose of this lab is:

- to describe the bootstrap
- to show how to compute a bootstrap estimate using Minitab
- to show how to devise an accurate computational procedure in Minitab.

To complete this lab, you will need a package that returns a random sample of values from a column of data. Most statistical packages will do this. The basic version of commercially available spreadsheets do not have functions to do this directly.

You will also need a statistical package or spreadsheet that will execute a command repeatedly in order to accumulate the values of a statistic from repeated runs into a single column. This can be done readily with line code commands. It is hard to execute a batch file from a graphics interface.

The Bootstrap.

The idea behind the bootstrap is simple. A sample of n values of a variable quantity $Q = Q_1, Q_2, \dots, Q_n$ is taken from a population and used to estimate some parameter, such as the mean of the population, or the skewness of the population. The true value of the parameter can never be known exactly unless the entire population is sampled, but of course the parameter can be estimated from the n observations. The sampling variation in this estimate is assessed by taking random samples of size n . This collection of random samples of size n can be as large as we like. We are going to regard this observed distribution as the best approximation of the true distribution of all possible samples from the population.

The samples that make up this approximate distribution of Q are called bootstrap samples. Each sample is used to make a bootstrap estimate of the true parameter (mean, slope, variance, diversity index, etc). The distribution of these bootstrap estimates, which we require for hypothesis testing and stating confidence limits, can be obtained by repeated sampling and then constructing an observed frequency distribution.

Example: Mandible lengths of male golden jackals (Manly 1991)

$$Q = [120 \quad 107 \quad 110 \quad 116 \quad 114 \quad 111 \quad 113 \quad 117 \quad 114 \quad 112] \cdot \text{mm}$$

The observed mean value is $\text{mean}(Q) = 114.3$ mm. This is an estimate of a parameter, the true mean.

To demonstrate how bootstrap estimates work we will to obtain a bootstrap estimate of the mean. This will be done by taking 10 samples (WITH replacement) from the collection of 10 observations, computing the mean from each sample, and then collecting 500 of these estimates.

To do this, we will use Minitab to sample WITH replacement.

The next page shows a series of steps for determining how to accomplish this in Minitab.

Name _____

Table 11.1. Generic recipe for devising computational procedure in any statistical package.

1. State the computation, in words.
2. Find a set of commands in your package to execute the procedure
3. Check the commands with a sample set of made-up data.
4. If it is not correct, look for another set of commands.
5. Repeat until a correct procedure is found.

Step 1. State the computation. sample with replacement from c1 into c2

Step 2. Use the help file in your package to find the commands.

```
MTB > Help Commands
```

You may need to choose a collection of commands:

```
MTB > Help commands 15
```

Which command are you going to use to sample with replacement from c1 into c2?

Now display the help file for this command.

Step 3. Next, see if this command works on a small data set.
Put some data into a column for a test run.

```
MTB > set into c1
DATA> 3 4 5 6 7 8
DATA> end
```

Laboratory #11. Bootstrap Estimates

If you sample from c1 into c2 with replacement, what do expect to see in c2? (this will be marked as present absent, not on whether it is correct)

Now execute the command to sample from c1 into c2. Then display c1 c2.

You should see some of the numbers in c1 several times in c2. You should also see that some numbers from c1 are missing from c2. If this is not happening, the computational procedure is probably not correct.

Step 4. If it is not correct, look for another set of commands.

Step 5. Repeat until an accurate procedure is found.

* * *

Bootstrap estimates. Mean mandible lengths of golden jackals.

Now that we have a computational procedure for sampling with replacement, we can move on to making bootstrap estimates of a statistic.

The first statistic will be the mean for mandible lengths of male golden jackals.

Example: Mandible lengths of male golden jackals (Manly 1991)

$Q = [120 \ 107 \ 110 \ 116 \ 114 \ 111 \ 113 \ 117 \ 114 \ 112] \cdot \text{mm}$

Set these 10 observations into column 1 of your statistical spreadsheet

*Define Data
from keyboard*

Next, place a random sample of 10 observations from c1 into c2.

Name _____

Next, calculate the mean value of the random sample in c2 and store this in k1.

You may need to use the help facility to do this.

The mean in c1 and c2 should differ slightly.

Next, store your bootstrap estimate of the difference in two means at the top of column 3

You may need the help facility.

Once you have several values (random means) in column c3, write a batch file to add a bootstrap estimate to column 3 every time you execute the file.

```
MTB > Store 'Jackal.ct1'  
STOR>
```

***Create
control file***

Make sure the batch file is adding a new and slight different estimate to c3 each time the batch file is executed.

Once the batch file is working correctly, then Tape, paste, or write out by hand the batch file:

Bootstrap estimates. Confidence intervals.

Name _____

Now accumulate 500 bootstrap estimates of the mean in c3.

To make sure you have 500 observations

```
MTB > n c3
```

Compute the mean value of the bootstrap estimates. _____

How does this compare to the estimate of 113.4? _____

A bootstrap estimate of a mean is hardly necessary because simply taking the sum and dividing by the number of observations will produce a perfectly good estimate. But what about the confidence limits around the mean? A theoretical distribution may not be available for obtaining a confidence limit, in which case a bootstrap estimate of the confidence limits will be useful.

You will recall that the confidence limit is a range around an estimate that will include the true value of the parameter 95% of the time (or 99% of the time, or whatever).

To obtain a rough idea of the 95% confidence limits around the bootstrap estimate of the mean, construct a histogram of the bootstrap estimates in c3.

Tape or paste your histogram here

Name _____

Based on this histogram, make a rough guess and write out the upper and lower limits that include 95% of the bootstrap estimates (answer does not have to be exact).

Upper limit = _____ mm

Lower limit = _____ mm

Next, try to think of a way to determine the 95% limits exactly, from the 500 bootstrap estimates in c3 (this will be marked as present/absent not on whether it is correct).

There are several ways of computing the 95% limits exactly for the 500 bootstrap estimates in c3. One is to use the percentile function found in many packages, including spreadsheets.

Another is to use the sort function, found in almost all packages.

With the sort function you will need to count downward in the sorted column, to reach the value for which 2.5% of the values are smaller. . You will then need to count upward in the column to reach the value for which 2.5% of the values are larger.

Try the percentile method. If your package does not make the computation, try the sort method.

Compute the upper and lower limits that include 95% of the bootstrap estimates.

Upper limit = _____ mm

Lower limit = _____ mm

Now compute then compare these bootstrap estimates of the 95% confidence limits to the theoretical limits using the t-distribution. Here are the computations:

```
MTB > describe c1
      N      MEAN   MEDIAN  TRMEAN   STDEV   SEMEAN
C1    10    113.40   113.50   113.37    3.72    1.18

MTB > invcdf .95;
SUBC> t 9.
      0.9500    1.8331
```

What is wrong with this invcdf command wrong? _____

Laboratory #11. Bootstrap Estimates

Name _____

```
MTB > invcdf .975;
SUBC> t 9.
      0.9750      2.2622
```

Now compute the theoretically based upper and lower limits using: the mean, the standard deviation, sample number n, and the t-statistic:

```
MTB > let k1 = 113.4 - 2.2622*stdev(c1)/sqrt(10)
MTB > let k2 = 113.4 + 2.2622*stdev(c1)/sqrt(10)
MTB > print k1 k2

K1          110.740
K2          116.060
```

Write up for this laboratory:

1. Work through the bootstrap example shown above, filling in all of the blanks.
2. Use bootstrap methods to set 95% confidence limits on the proportion of ants in the diet of short-horned lizards *Phrynosoma douglassi brevirostre*, from near Bow Island, Alberta during the month of June (data from Manly 1991 p 247). Use 1000 bootstrap estimates to obtain a good estimate of the upper and lower limits.

$p = [33 \quad 100 \quad 43 \quad 100 \quad 14 \quad 64 \quad 40 \quad 0.0 \quad 47] \cdot \%$
Upper limit = _____ mm

Lower limit = _____ mm

Compute the upper and lower limit 95% confidence limits using a t-distribution.

Upper limit = _____ mm

Lower limit = _____ mm

Should you report the first or the second pair of confidence intervals in a thesis or other document? Why?