

## 14 Constructing Confidence Intervals

In this lab, we will construct and visualize confidence intervals for different sample sizes and different levels of confidence.

### Sampling Data

Please load the following dataset into *Stata*.

```
. use http://www.stat.ucla.edu/labs/datasets/cilab.dta
```

We have thirty variables regarding stocks listed in the Standard and Poor's 500. We will focus on the percentage return during a twenty-six week period (*pctchg26wks*).

```
. summarize pctchg26wks
```

Looking at the statistical summary of this variable, we can see that the average stock value in the Standard and Poor's 500 dropped 12.6% during this 26 week period, with a maximum loss of 92% to a maximum gain of 99%.

We are going to use this data set and this variable *pctchg26wks* to test the theory of confidence intervals. In general, the way we construct a confidence interval is by obtaining a sample and then constructing an interval around the mean of that sample. This interval is intended to give us an estimate of the true population mean, which in general is unknown. We are going to use a sort of reverse psychology. We know the true population mean for the percentage return for the S & P's 500 stocks over this particular 26 week period and we know the true population standard deviation (23.6) of this variable. The dataset we loaded contains the entire population! We can test the theory of confidence intervals by taking samples of the dataset and constructing confidence intervals around the means of those samples. Then we can check and see if these confidence intervals, that we estimated, contain the true mean or not.

We will use a statistical technique called *bootstrapping*. This simply means we will take simple random samples of the dataset with replacement. For the purposes of constructing confidence intervals, we are not really interested

in the sample itself, we are more interested in the mean of the sample. The **bs** command will bypass the output of individual samples and give us the sample mean for each sample it draws. The **bs** command takes as many samples as you want of the specified variable and then takes the mean of each of those samples. It outputs a new dataset, where each observation represents a single sample from the original population.

This new file should be saved into your home directory. To ensure this occurs, change your working directory by clicking on the File menu and selecting “Set Working Folder.” The default is your “Documents” folder. This is the correct location, so click on the “Choose” button. Now any files you create will show up in your documents folder.

Now to ensure we truly are taking random samples, we want to randomly set a seed for *Stata* to start from.

```
. set seed YOUR STUDENT ID NUMBER
```

Finally, we are ready to start sampling.

```
. bs "summarize pctchg26wks" "r(mean)", reps(100) size(16)
dots saving (cidata) replace
```

This command will take 100 random samples of size 16 and calculate the mean of the variable *pctchg26wks* for each sample. It will then save this information in a file called *cidata.dta*.

Now we want to open up this dataset we just created and construct a confidence interval for each of the 100 samples.

```
. use cidata, clear
```

When we issue the command

```
. list
```

we see one variable *bs1* (short for bootstrap one), which is a list of 100 means from the 100 samples of size 16 that we selected from the S&P’s 500 stocks. We can look at the distribution of these sample means. (Remember the Central Limit Theorem tells us that as the size of our samples increase, the distribution of the sample means becomes more and more normally distributed.)

```
. graph bs1, xlabel ylabel norm bin(10)
```

## Constructing & Visualizing Confidence Intervals

For each sample mean, we want to generate an appropriate confidence interval. Recall the formula for constructing a confidence interval when the standard deviation  $\sigma$  is known.

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

We know  $\sigma$  (the true population standard deviation) is 23.66127. We also know that  $n$  (the size of the sample) is 16. We want to explore what happens to confidence intervals when we change our level of confidence.

We will create 68%, 90%, and 95% confidence intervals for each of our 100 samples. The technique we will use is to create the lower bound of the confidence interval  $(\bar{x} - z^* \frac{\sigma}{\sqrt{n}})$  and then the upper bound of the confidence interval  $(\bar{x} + z^* \frac{\sigma}{\sqrt{n}})$ .

For a 68% confidence interval,  $z^* = 1.00$ . The variable `bs1` contains all our sample means ( $\bar{x}$ ).

```
. generate lower68 = bs1 - 1.00*23.66127/sqrt(16)
. generate upper68 = bs1 + 1.00*23.66127/sqrt(16)
```

*Note: If you make a mistake when generating your new variables, use the **drop** command to remove variables from your dataset and reissue the correct command. For example, if I messed up and typed **generate upper68 = bs1 + 1.00\*23.66127/sqrt(160)**, then I could remove this variable by typing **drop upper68** and then regenerate the variable correctly.*

```
. list
```

Look at the data. As you can see, the variables `lower68` and `upper68` form an interval surrounding the `bs1` variable.

*Question 1: We know that the true mean for this population is -12.6%. According to the theory of confidence intervals, how many of these confidence intervals should contain the true mean?*

We can actually determine exactly how many of our confidence intervals contain the true mean.

```
. count if lower68 <= -12.567 & if upper68 >= -12.567
```

This command counts the value only if the lower bound is below the true population mean **and** the upper bound is above the true population mean.

*Question 2: How many of your 68% confidence intervals captured the true population mean? Does this number surprise you or does it seem about right?*

We can visualize the confidence intervals by plotting them side by side. To do this we must create a helper variable...

```
. gen num = _n
```

Then issue the graph command.

```
. graph upper68 lower68 bs1 num, connect (||.) symbol(iio)
yline(-12.567) xlabel ylabel ti("100 68% confidence intervals
from samples n=16")
```

Each of your confidence intervals pop up in the resulting graph. Each line corresponds to exactly one confidence interval and the dot in the middle corresponds to the  $\bar{x}$  for that confidence interval. We inserted a line going through the confidence intervals at the true population mean of -12.567. As you can see, some of the confidence intervals capture this true mean and others don't. This is the caveat of confidence intervals. In a real life setting, we have no way of knowing if the one sample we have is one of the cases that does not capture the true mean! This is why large sample sizes and high levels of confidence are so important.

*Question 3: What is the length of a 68% confidence interval in this setting?*

Next we repeat the process by generating 90% and then 95% confidence intervals. For a 90% confidence interval,  $z^*$  is equal to 1.645.

*Question 4: What number will you be adding and subtracting to “bs1” to obtain a 90% confidence interval? Based on this, what will be the length of a 90% confidence interval?*

Construct variables *lower90* and *upper90* similarly to the way you constructed *lower68* and *upper68*. Be sure you make the appropriate change in the formula.

*Question 5: According to theory, how many of your 90% confidence intervals are expected to capture the true population mean? How many of your 90% confidence intervals actually capture the true population mean?*

Create a graph, with the appropriate title, of your 90% confidence intervals.

Repeat the process for 95% confidence intervals. Use a  $z^*$  value of 1.96.

*Question 6: According to theory, how many of your 95% confidence intervals are expected to capture the true population mean? How many of your 95% confidence intervals actually capture the true population mean?*

## Assignment

What happens to the length of confidence intervals if we change the sample size  $n$ ?

Reload the original dataset and run a new bootstrap command, this time with samples of size 36. (You do not need to reset the seed, this only needs to be done once.)

*Question 7: Are corresponding confidence intervals using samples of size 36 longer or shorter than those for samples of size 16?*

What happens if we use samples of size 100? Reload the original dataset and run a new bootstrap command using samples of size 100.

*Question 8: Are corresponding confidence intervals using samples of size 100 longer or shorter than those for samples of size 36?*

*Question 9: As we increase the confidence level from 68% to 90% to 95%, what happens to the length of our intervals?*