Bootstrapping

Javier E. Flores

February 18, 2019

Bootstrapping

In our third lab, we introduced and explored the utility of the seemingly magical **bootstrap** procedure.

We learned that, using the bootstrap, we can approximate the sampling distribution of a statistic even when we only have a single sample to work with!

Using this approximate sampling distribution, we can then construct interval estimates one of two ways depending on the shape of the approximated distribution.

Given this incredible feature of the bootstrap, when it comes to interval estimation we might be tempted to immediately say...

😱 William James O'Reilly Jr.!

However...

In this lecture, we'll learn when using the bootstrap is a good idea (i.e. when it actually works) and when it doesn't work.

When It Works

In using the bootstrap, the most important determinant for success is how representative our sample is of the population.

Remember, in practice, the contents of our sample are all the information we have available about the population of interest.

In collecting our sample, if we made sure to sample as a good statistician would (i.e. use simple random sampling), it's likely that our sample would be representative.

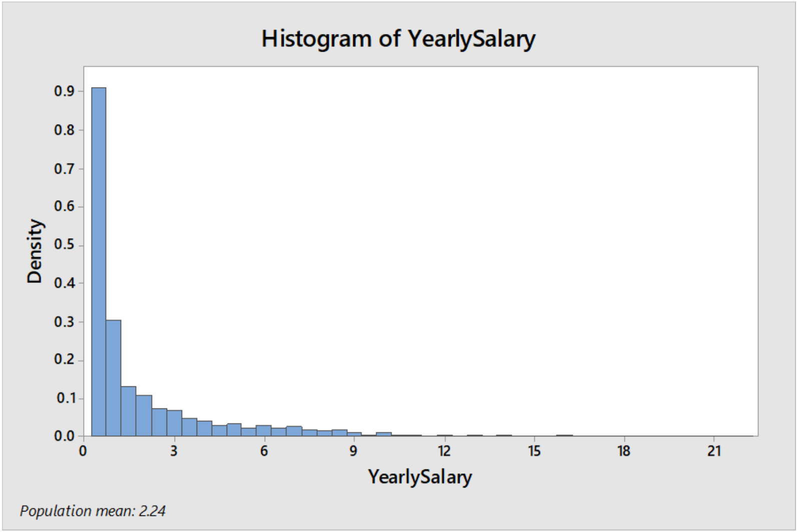
As a result, when using the bootstrap to approximate the sampling distribution, we can expect the bootstrap distribution to closely match the true sampling distribution of our statistic.

See For Yourself!

With your groups,

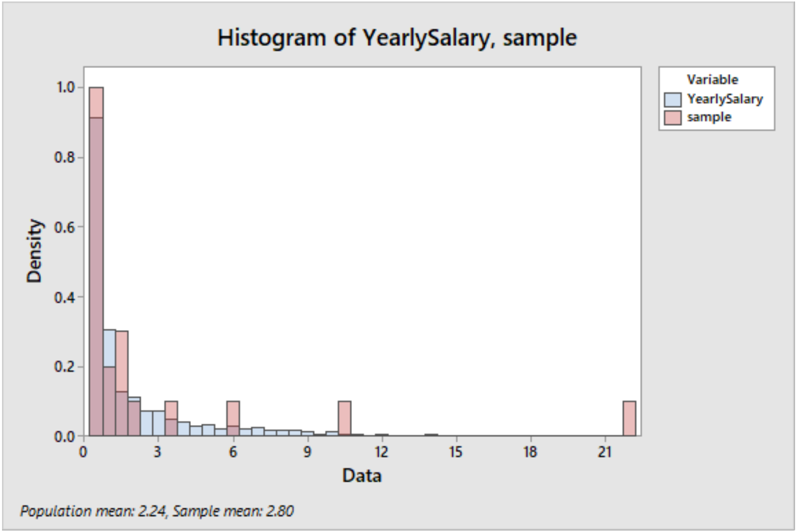
1) Using the 2015 NFL Contracts dataset, generate a random sample of size $n = 20$ containing data on yearly salaries.
2) Create a histogram of the entire data and then overlay a histogram of your sample data. Based on this figure, would you say your population was representative of the population data? Be sure to set the "Y-Scale Type" to "Density" when creating your plot.
3) Import your sample data to StatKey and generate the bootstrap sampling distribution (use 1000 bootstrap samples).
4) In a separate window, use StatKey to generate the actual sampling distribution of the mean. How does this true sampling distribution compare to what was approximated by the bootstrap?

Solution

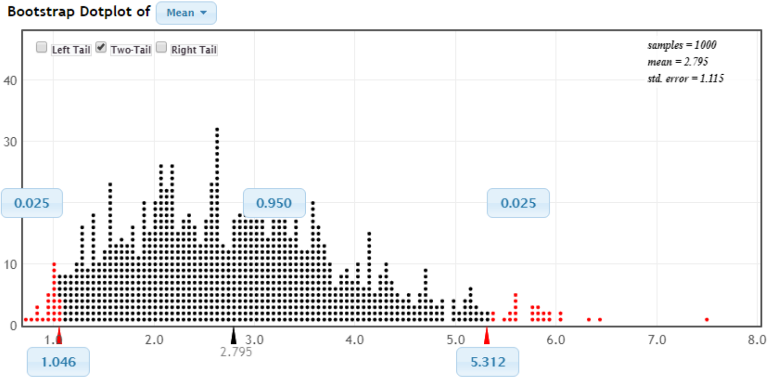


Population mean: 2.24

Solution


Histogram of YearlySalary, sample
Density
1.0
0.8
0.6
0.4
0.2
0.0
0
3
6
9
12
15
18
21
Data
Variable
YearlySalary
sample
Population mean: 2.24, Sample mean: 2.80

Solution

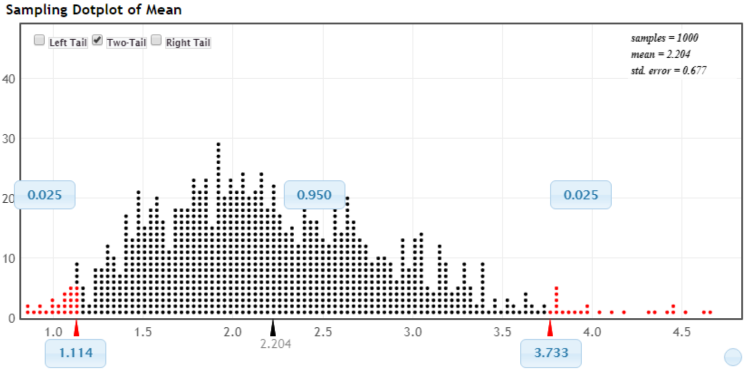


Bootstrap Dotplot of Mean ▾

□ Left Tail ☑ Two-Tail □ Right Tail

samples = 1000
mean = 2.795
std. error = 1.115

Solution


Sampling Dotplot of Mean

☐ Left Tail ☑ Two-Tail ☐ Right Tail

samples = 1000
mean = 2.204
std. error = 0.677

Solution

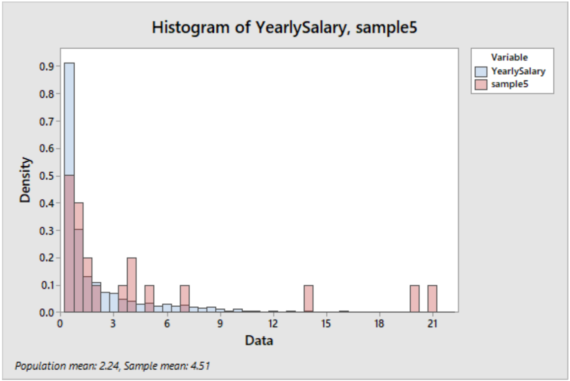
While not perfect, our random sample tracks relatively well with the actual population. The major shortcoming of our sample is the apparent oversampling of higher-salaried players.

As a result of this oversampling, our sample mean is higher than the population mean. This bias propogates to the bootstrap distribution, as we see it centered at 2.795. The true sampling distribution is centered at 2.204.

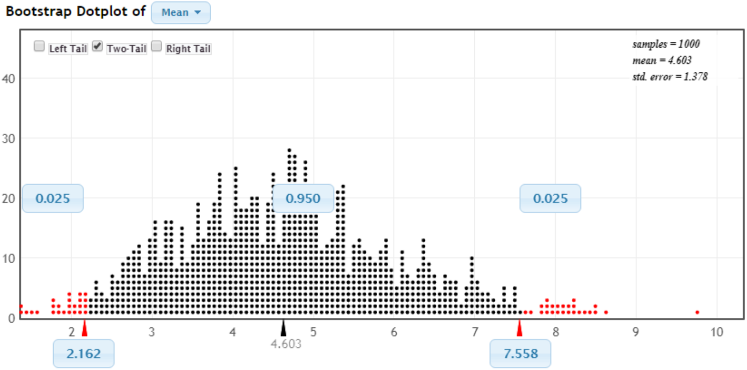
Question: Both the true sampling distribution and bootstrap distribution were generated with samples of size 20. Why is the confidence interval using the bootstrap distribution wider than the true sampling distribution CI?

(Extremely) Non-representative Sample

In contrast to our previous sample, this second sample (also drawn randomly) is not at all representative of our population.



(Extremely) Non-representative Sample



(Extremely) Non-representative Sample

This sample is characterized by an even more substantial overrepresentation of higher salaries than our first sample.

An obvious consequence of this is the terribly inaccurate confidence interval produced by the bootstrap distribution - it just barely captures the true population mean!

In obtaining both samples, random sampling was used. These abnormalities in our sample are therefore entirely due to random chance. Both examples serve to show that while bootstrapping is a wonderfully useful procedure, it isn't perfect.

Summary

The accuracy of any bootstrap confidence interval is highly dependent on the quality of your sample: "Garbage in, garbage out."

As we've seen in the past, bias and sample size remain key factors in how well we can estimate population parameters.

If working with a small and biased sample (e.g. the second sample in our example), there is no amount of bootstrap resampling that will address either shortcoming.

CI Methods

Assuming we do trust our sample to yield bootstrap interval estimates, we also learned that there are two different approaches to obtaining bootstrap CI's:

- The standard error (SE) approach, in which we make use of the standard error of the bootstrap distribution to obtain a confidence interval
- The percentile approach, in which we make use of the bootstrap distribution percentiles in order to obtain a confidence interval

Recall that the SE approach relies on the assumption that the bootstrap distribution is symmetric and bell-shaped.

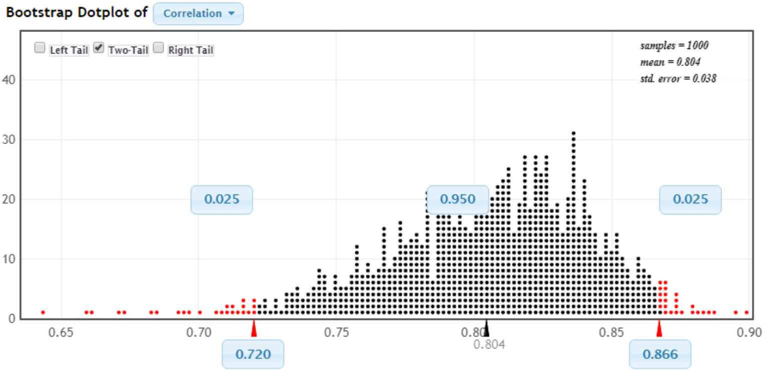
How does the SE approach perform when this assumption is violated?

Exercise

With your groups,

1) Import the Atlanta Commute Time dataset into StatKey. Use only the variables "Distance" and "Time".
2) In StatKey, generate a bootstrap distribution for the correlation coefficient between commute time and commute distance. Is the resulting distribution symmetric?
3) Compute a 95% confidence interval for the correlation coefficient using the standard error approach.
4) In StatKey, check the boxes for "Left Tail" and "Right Tail" and enter in the bottom two boxes the lower and upper bounds of the confidence interval you found in 3), respectively. Does this interval cover 95% of the distribution (as it is intended)?

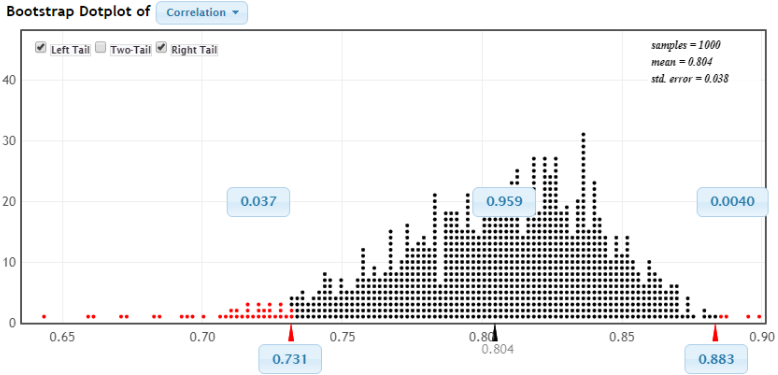
Solution


Bootstrap Dotplot of Correlation ▾

☐ Left Tail ☑ Two-Tail ☐ Right Tail

samples = 1000
mean = 0.804
std. error = 0.038

Solution


Bootstrap Dotplot of
Correlation
☑ Left Tail
☐ Two-Tail
☑ Right Tail
samples = 1000
mean = 0.804
std. error = 0.038
0
10
20
30
40
0.65
0.70
0.75
0.80
0.85
0.90
0.037
0.959
0.0040
0.731
0.804
0.883

Solution

The bootstrap distribution of the correlation is skewed left. This means our assumption of symmetry is violated.

As a result, when we use the SE approach to forming a confidence interval, the resulting interval coverage is biased.

The SE confidence interval is wider and less precise than the percentile confidence interval when the bell-shape/symmetry assumption is violated.

Both the SE and percentile approaches will produce the same results when the bell-shape/symmetry assumption holds.

"Testing" with Confidence Intervals

Now that we have a thorough understanding of the bootstrap and its utility for interval estimation, consider next what we might do with these intervals.

Our motivation for obtaining interval estimates in the first place is often derived from the need to "test" a claim made about data.

Think back to our first lab when we were interested in comparing study times between introverts and extroverts.

At the time, we were limited to a simple visual comparison between boxplots generated for each group.

Now, we can use bootstrapping and confidence intervals to better answer this question.

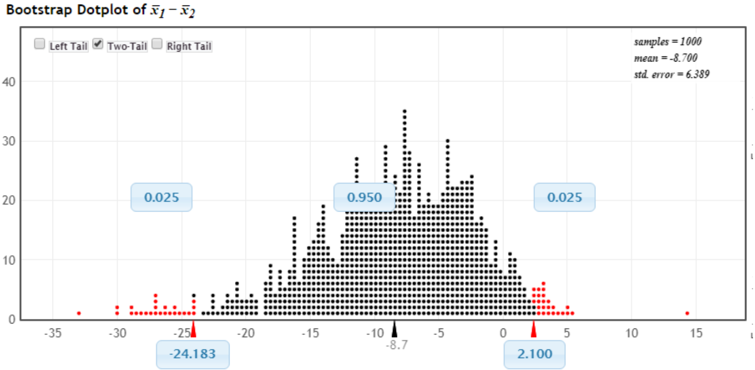
Relative to our crude visual comparison, confidence intervals allow us to quantify whether a real population level difference actually exists.

Exercise

With your groups,

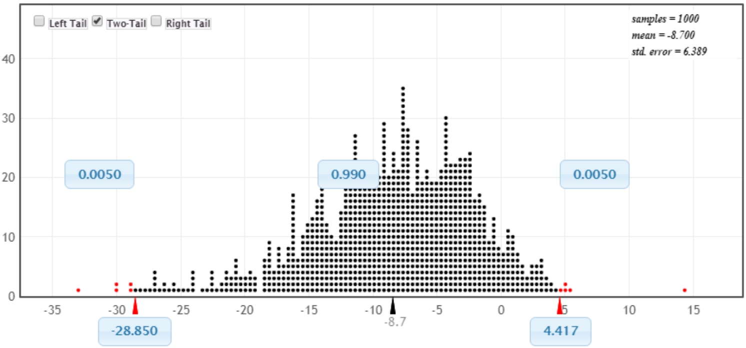
1) Load the Class Survey data into Minitab.
2) Separate the columns "IntroExtro" and "StudyTime" from the rest of the variables.
3) Load those columns into StatKey and construct a 95% confidence interval for the difference in study time between introverts and extroverts.
4) Does your interval support the idea that the mean study time is different between these groups? Does this change if you compute a 99% interval?

Solution


Bootstrap Dotplot of $\bar{x}_1 - \bar{x}_2$
□ Left Tail ☑ Two-Tail □ Right Tail
samples = 1000
mean = -8.700
std. error = 6.389
0
10
20
30
40
0.025
0.950
0.025
-35
-30
-25
-20
-15
-10
-5
0
5
10
15
-8.7
-24.183
2.100

Solution

Bootstrap Dotplot of $\bar{x}_1 - \bar{x}_2$



□ Left Tail ☑ Two-Tail □ Right Tail

samples = 1000
mean = -8.700
std. error = 6.389

Solution

If there were no difference in study time between introverts and extroverts, we'd expect the mean difference to be centered at 0.

From the previous two figures, this is clearly not the case. In fact, the majority of our interval (in either case) is below 0!

Since we are taking the average difference in study time between extroverts and introverts, this makes a strong case for the claim that introverts study (on average) more than extroverts.

Wrap-Up

Right now, you should...

- Understand bootstrapping and when it works well
- Know the differences between the SE and percentile approaches to bootstrap confidence intervals
- Be able to obtain random samples and construct bootstrap confidence intervals for a variety of parameters using StatKey
- Understand how confidence intervals may be used to assess claims about data

These notes cover sections 3.3-3.4 of the textbook. Please read through each section and examples along with any links provided in this lecture.