Confidence Intervals: Bootstrap Distribution

SECTIONS 3.3, 3.4
- Bootstrap distribution (3.3)
- 95% CI using standard error (3.3)
- Percentile method (3.4)

Reality

One small problem...

... WE ONLY HAVE ONE SAMPLE!!!!

• How do we know how much sample statistics vary, if we only have one sample?!!?

BOOTSTRAP!

“Population”

• Imagine the “population” is many, many copies of the original sample

• (What do you have to assume?)

ONE Reese’s Pieces Sample

Where might the “true” p be?

Reese’s Pieces “Population”

Sample repeatedly from this “population”
**Sampling with Replacement**

- To simulate a sampling distribution, we can just take repeated random samples from this "population" made up of many copies of the sample.
- In practice, we can't actually make infinite copies of the sample...
- ...but we can do this by sampling with replacement from the sample we have (each unit can be selected more than once).

Suppose we have a random sample of 6 people:

**Bootstrap Sample**: Sample with replacement from the original sample, using the same sample size.

**Reese's Pieces**

- How would you take a bootstrap sample from your sample of Reese's Pieces?
**Bootstrap Sample**

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21, 22

a) Yes  
b) No

---

**Bootstrap Sample**

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

a) Yes  
b) No

---

**Bootstrap Sample**

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

a) Yes  
b) No

---

**Bootstrap Sample**

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

a) Yes  
b) No

---

**Bootstrap**

A *bootstrap sample* is a random sample taken with replacement from the original sample, of the same size as the original sample.

A *bootstrap statistic* is the statistic computed on a bootstrap sample.

A *bootstrap distribution* is the distribution of many bootstrap statistics.

---

**StatKey**

[link](http://lock5stat.com/statkey/)
### Bootstrap Sample

You have a sample of size $n = 50$. You sample with replacement 1000 times to get 1000 bootstrap samples.

What is the sample size of each bootstrap sample?

(a) 50  
(b) 1000

---

### Bootstrap Distribution

You have a sample of size $n = 50$. You sample with replacement 1000 times to get 1000 bootstrap samples.

How many bootstrap statistics will you have?

(a) 50  
(b) 1000

---

### Why “bootstrap”?  

“Pull yourself up by your bootstraps”

- Lift yourself in the air simply by pulling up on the laces of your boots
- Metaphor for accomplishing an “impossible” task without any outside help

---

### Bootstrap Distribution

What can we do with just one seed?

Grow a NEW tree!

Estimate the distribution and variability (SE) of $x$'s from the bootstraps

---

### Sampling Distribution

BUT, in practice we don’t see the “tree” or all of the “seeds” – we only have ONE seed

---

### Golden Rule of Bootstrapping

Bootstrap statistics are to the original sample statistic as the original sample statistic is to the population parameter
Center

- The sampling distribution is centered around the population parameter.
- The bootstrap distribution is centered around the
  a) population parameter
  b) sample statistic
  c) bootstrap statistic
  d) bootstrap parameter
- Luckily, we don’t care about the center... we care about the variability!

Standard Error

- The variability of the bootstrap statistics is similar to the variability of the sample statistics.
- The standard error of a statistic can be estimated using the standard deviation of the bootstrap distribution!

Confidence Intervals

- The variability of the bootstrap statistics is similar to the variability of the sample statistics.
- The standard error of a statistic can be estimated using the standard deviation of the bootstrap distribution!

Reese’s Pieces

Based on this sample, give a 95% confidence interval for the true proportion of Reese’s Pieces that are orange.

a) (0.47, 0.57)
b) (0.42, 0.62)
c) (0.41, 0.51)
d) (0.36, 0.56)
e) I have no idea

The Magic of Bootstrapping

- We can use bootstrapping to assess the uncertainty surrounding ANY sample statistic!
- If we have sample data, we can use bootstrapping to create a 95% confidence interval for any parameter!

(Well, almost...)
Used Mustangs

- What’s the average price of a used Mustang car?
- Select a random sample of \( n = 25 \) Mustangs from a website (autotrader.com) and record the price (in $1,000’s) for each car.

Sample of Mustangs:

\( n = 25 \quad \bar{x} = 15.98 \quad s = 11.11 \)

Our best estimate for the average price of used Mustangs is $15,980, but how accurate is that estimate?

BOOTSTRAP!

Original Sample 1. Bootstrap Sample

- Calculate mean price of bootstrap sample
- Repeat many times!

Used Mustangs

- 95% CI:
  
  \[ \text{statistic} \pm 2 \cdot SE \]
  
  $15,980 \pm 2 \cdot $2,178
  
  ($11,624, $20,336)

- We are 95% confident that the average price of a used Mustang on autotrader.com is between $11,624 and $20,336

Global Warming

What percentage of Americans believe in global warming?

A survey on 2,251 randomly selected individuals conducted in October 2010 found that 1328 answered “Yes” to the question “Is there solid evidence of global warming?”

Give and interpret a 95% CI for the proportion of Americans who believe there is solid evidence of global warming.

Global Warming

We are 95% sure that the true percentage of all Americans that believe there is solid evidence of global warming is between 57% and 61%.

0.59 ± 2(0.01) = (0.57, 0.61)

Give and interpret a 95% CI for the proportion of Americans who believe there is solid evidence of global warming.

Global Warming

Does belief in global warming differ by political party?

"Is there solid evidence of global warming?"

The sample proportion answering “yes” was 79% among Democrats and 38% among Republicans. (exact numbers for each party not given, but assume n=1000 for each group)

Give a 95% CI for the difference in proportions.


Global Warming

Based on the data just analyzed, can you conclude with 95% certainty that the proportion of people believing in global warming differs by political party?

(a) Yes
(b) No

Other Levels of Confidence

• What if we want to be more than 95% confident?

• How might you produce a 99% confidence interval for the average body temperature?

Statistics: Unlocking the Power of Data

Body Temperature

What is the average body temperature of humans?

98.26 ± 2(0.105) = 98.26 ± 0.21

We are 95% sure that the average body temperature for humans is between 98.05°C and 98.47°C.


98.6°F !!!!
**Percentile Method**

- For a P% confidence interval, keep the middle P% of bootstrap statistics.
- For a 99% confidence interval, keep the middle 99%, leaving 0.5% in each tail.
- The 99% confidence interval would be (0.5\textsuperscript{th} percentile, 99.5\textsuperscript{th} percentile) where the percentiles refer to the bootstrap distribution.

**Body Temperature**

We are 99% sure that the average body temperature is between $98.00^\circ$ and $98.58^\circ$.

**Level of Confidence**

Which is wider, a 90% confidence interval or a 95% confidence interval?

(a) 90% CI
(b) 95% CI

**Mercury and pH in Lakes**

- For Florida lakes, what is the correlation between average mercury level (ppm) in fish taken from a lake and acidity (pH) of the lake?

\[ r = -0.575 \]

Give a 90% CI for \( r \)

We are 90% confident that the true correlation between average mercury level and pH of Florida lakes is between -0.702 and -0.433.
Bootstrap CI

**Option 1:** Estimate the standard error of the statistic by computing the standard deviation of the bootstrap distribution, and then generate a 95% confidence interval by

\[ \text{statistic} \pm 2 \times \text{SE} \]

**Option 2:** Generate a P% confidence interval as the range for the middle P% of bootstrap statistics

Two Methods

- For a symmetric, bell-shaped bootstrap distribution, using either the standard error method or the percentile method will give similar 95% confidence intervals
- If the bootstrap distribution is not bell-shaped or if a level of confidence other than 95% is desired, use the percentile method

Bootstrap Cautions

- These methods for creating a confidence interval only work if the bootstrap distribution is smooth and symmetric
- ALWAYS look at a plot of the bootstrap distribution!
- If the bootstrap distribution is highly skewed or looks "spiky" with gaps, you will need to go beyond intro stat to create a confidence interval

Number of Bootstrap Samples

- When using bootstrapping, you may get a slightly different confidence interval each time. This is fine!
- The more bootstrap samples you use, the more precise your answer will be
- Increasing the number of bootstrap samples will not change the SE or interval (except for random fluctuation)
- For the purposes of this class, 1000 bootstrap samples is fine. In real life, you probably want to take 10,000 or even 100,000 bootstrap samples
Summary

- The standard error of a statistic is the standard deviation of the sample statistic, which can be estimated from a bootstrap distribution.

- Confidence intervals can be created using the standard error or the percentiles of a bootstrap distribution.

- Confidence intervals can be created this way for any parameter, as long as the bootstrap distribution is approximately symmetric and continuous.

To Do

- Read Sections 3.3, 3.4
- Do HW 3 (due Monday, 2/10)