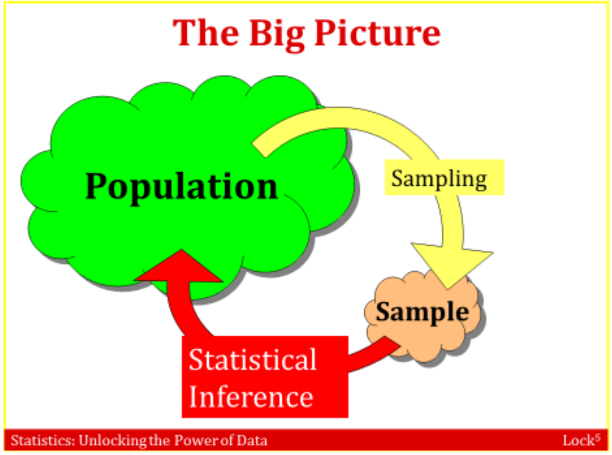
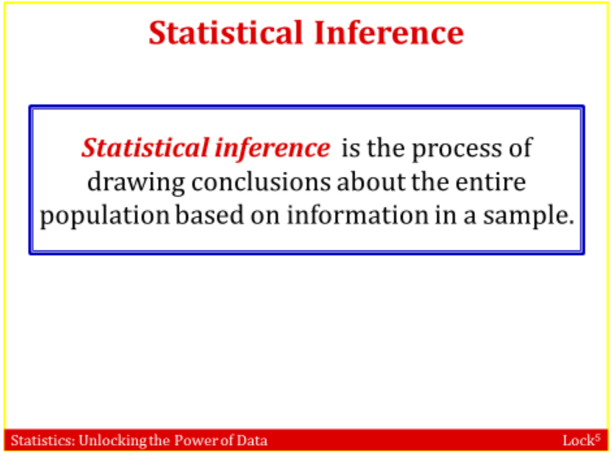
Chapter 3

3.1 **Sampling Ditributions**



The Big Picture

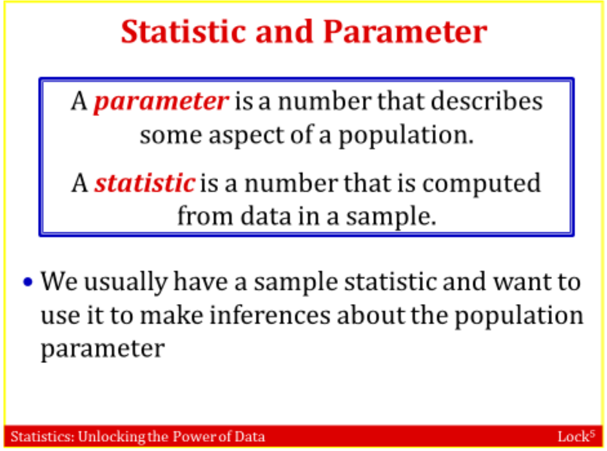
Statistics: Unlocking the Power of Data Lock^5



Statistical Inference

Statistical inference is the process of drawing conclusions about the entire population based on information in a sample.

Statistics: Unlocking the Power of Data Lock^5



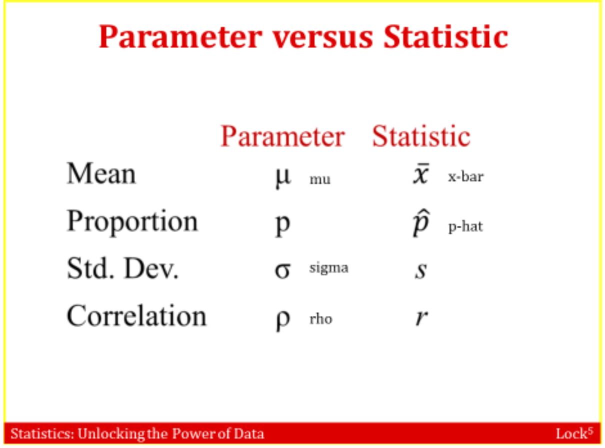
Statistic and Parameter

A ***parameter*** is a number that describes some aspect of a population.

A ***statistic*** is a number that is computed from data in a sample.

- We usually have a sample statistic and want to use it to make inferences about the population parameter

Statistics: Unlocking the Power of Data Lock^5



Parameter versus Statistic

	Parameter	Statistic
Mean	μ mu	$\bar{x}$ x-bar
Proportion	p	$\hat{p}$ p-hat
Std. Dev.	σ sigma	s
Correlation	ρ rho	r

Statistics: Unlocking the Power of Data Lock^5

Example:

For each of the following, state whether the quantity described is a parameter or a statistic, and give the correct notation.

- Average household income for all houses in the US, using data from the US census Parameter, (symbol μ)
- Correlation between height and weight for players on the 2014 Brazil world cup team, using data from all 23 players on the roster Parameter (symbol ρ)
- Proportion of people who use an electric toothbrush, using data from a sample of 300 adults Statistic (symbol $\hat{p}$)

Example 1: Using Search Engines on the Internet

A 2012 survey of a random sample of 2253 US adults found that 1,329 of them reported using a search engine (such as Google) every day to find information on the Internet.

a). Find the relevant proportion and give the correct notation with it. $\hat{p} = \frac{1329}{2253} = 0.590$

b). Is your answer to part (a) a parameter or a statistic? *Statistic* c). Give notation for and define the population parameter that we estimate using the Result of part (a).

p = the proportion of all US adults that would report that they use an Internet search engine every day

Example 2: Number of Books Read in a Year

A survey of 2,986 Americans ages 16 and older found that 80% of them read at least one book in the last year. Of these book readers, the mean number of books read in the last year is 17 while the median number of books read in the last year is 8.

a). How many "book readers" (defined as reading at least one book in the past year) were included in the sample?

Number of book readers is 0.80(2986) = 2388.8 or about 2389 book readers.

b). Why might the mean and median be so different? Using the information given about the mean and median number of books read in a year, what is the likely shape of the distribution of number of books read in a year by book readers?

Since the mean is so much larger than the median, it is likely that there are some very large outliers. The distribution is probably skewed to the right.

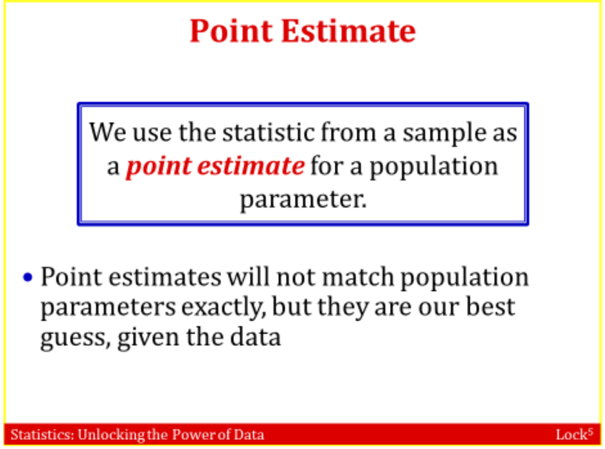
c). Give the correct notation for the value "17" in the information above. Is this value a parameter or a statistic?

It is a statistic and the correct notation for a sample mean is $\bar{x}$.

d). Give notation for and define the population parameter that we estimate using the result of part (c).

μ = the mean number of books read last year by all Americans ages 16 and older who have read at least one book. [Emphasize that we are estimating the mean of the population.]

We have a problem. Usually we want to know the value of the parameter but we can only obtain the statistic. We need a way to use the statistic to tell us something about the parameter.

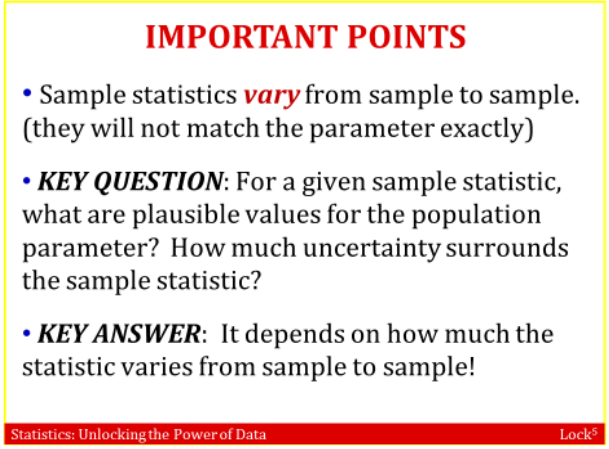


Point Estimate

We use the statistic from a sample as a ***point estimate*** for a population parameter.

- Point estimates will not match population parameters exactly, but they are our best guess, given the data

Statistics: Unlocking the Power of Data Lock^5



IMPORTANT POINTS

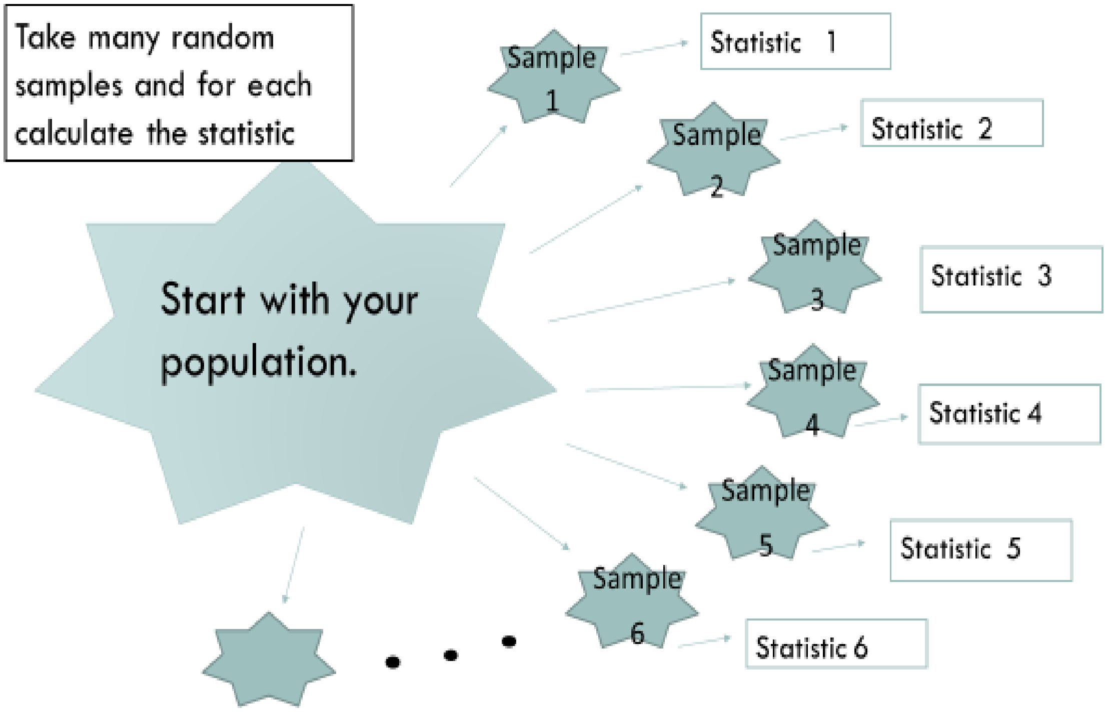
- Sample statistics ***vary*** from sample to sample. (they will not match the parameter exactly)

- ***KEY QUESTION***: For a given sample statistic, what are plausible values for the population parameter? How much uncertainty surrounds the sample statistic?

- ***KEY ANSWER***: It depends on how much the statistic varies from sample to sample!

Statistics: Unlocking the Power of Data Lock^5

To see how statistics, vary from sample to sample, we will take many samples and compute many statistics!



Let’s look at an example to illustrate the process:

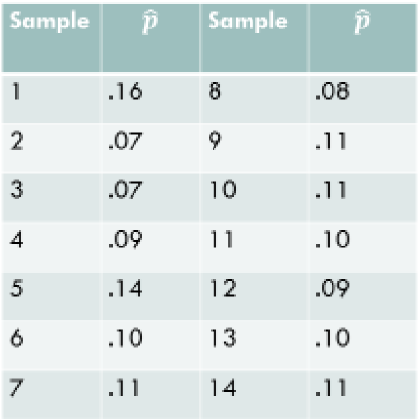
Suppose we are interested in STA 296 students would like to know the proportion of those students that had trouble getting the required books.

Population: All STA 296 Students
Parameter: The proportion of all STA 296 that had trouble getting the book, we can denote it using the letter p.

Suppose I take a random sample of 25 STA 296 students and I find that 4 of the 25 students had trouble. Notice the type of statistic I will calculate is a sample proportion.

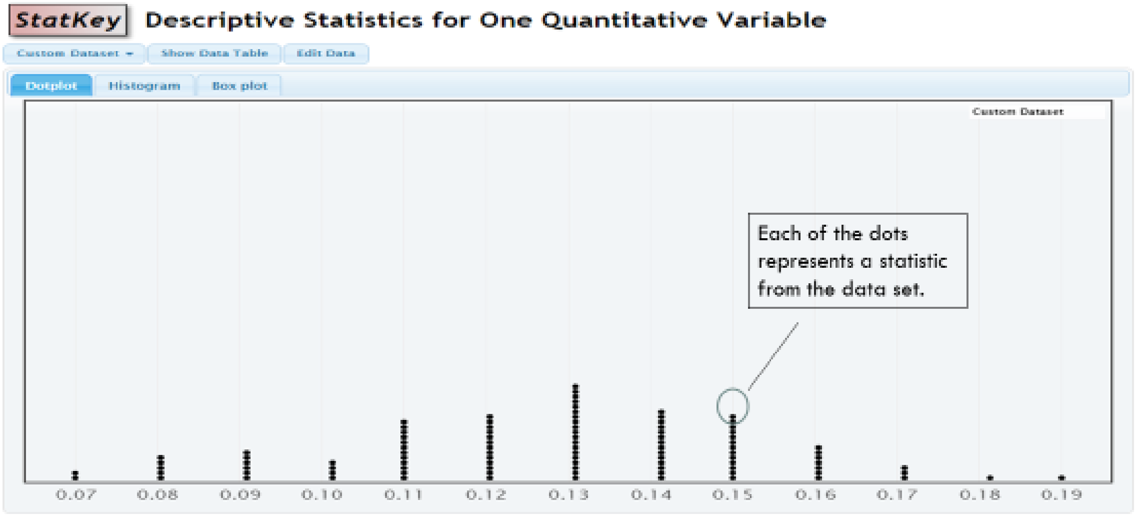
$\hat{p}$ = 4/25=.16
We could also say the point estimate is equal to .16

Suppose I repeat the study 100 times and each time calculate the sample proportion from a new random sample, below are the results for the 1st 14 samples.



Sample	$\hat{p}$	Sample	$\hat{p}$
1	.16	8	.08
2	.07	9	.11
3	.07	10	.11
4	.09	11	.10
5	.14	12	.09
6	.10	13	.10
7	.11	14	.11

- Notice that the value of the statistic varies from sample to sample –Sampling variability
- Each of the values in the data set represents a statistic. IE the data is comprised of statistics.



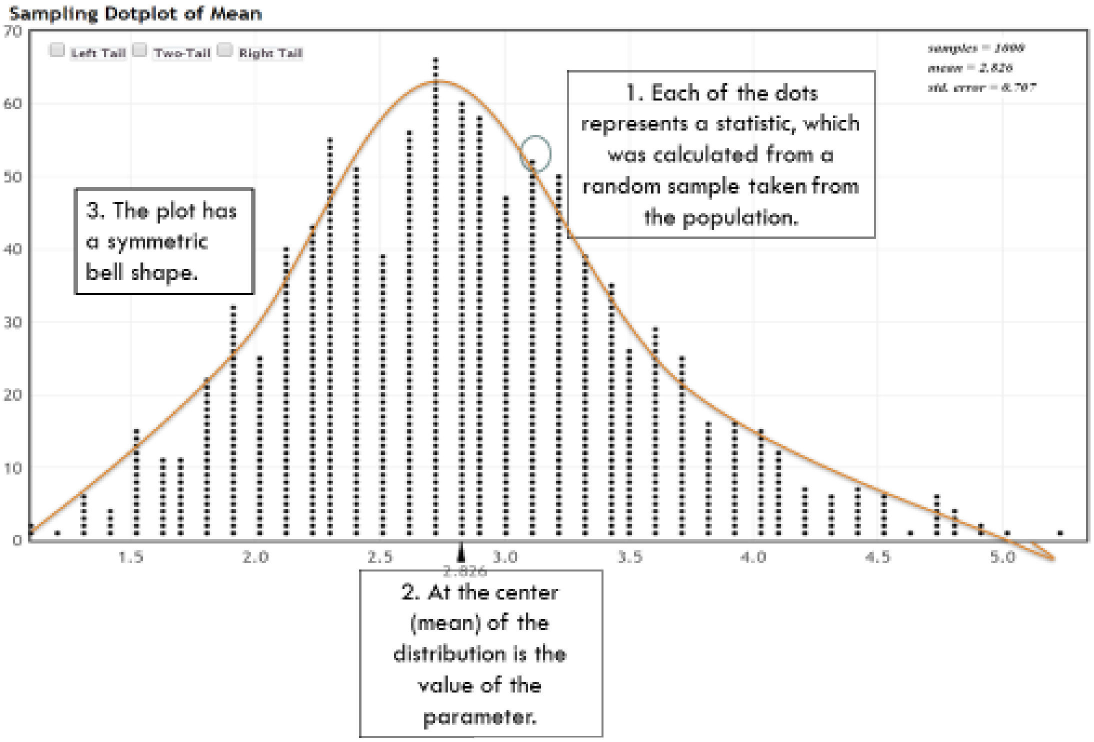
Each of the dots represents a statistic from the data set.

Important: Our data set is comprised of statistics!

A Sampling Distribution is the distribution of sample statistics computed for different samples of the same size from the same population. A sampling distribution shows us how the sample statistic varies from sample to sample.

Properties of Sampling distribution:

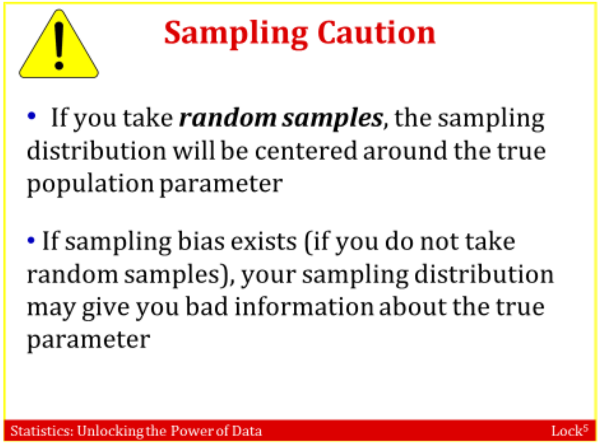
- *Center*: If samples are randomly selected, the sampling distribution will be centered around the population parameter.
- Shape: For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.



1. Each of the dots represents a statistic, which was calculated from a random sample taken from the population.

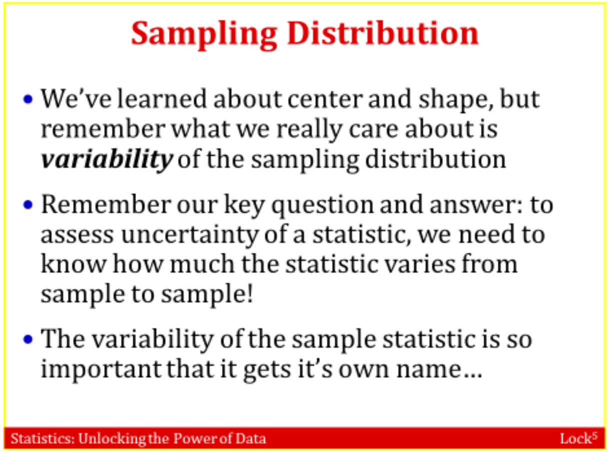
2. At the center (mean) of the distribution is the value of the parameter.

3. The plot has a symmetric bell shape.



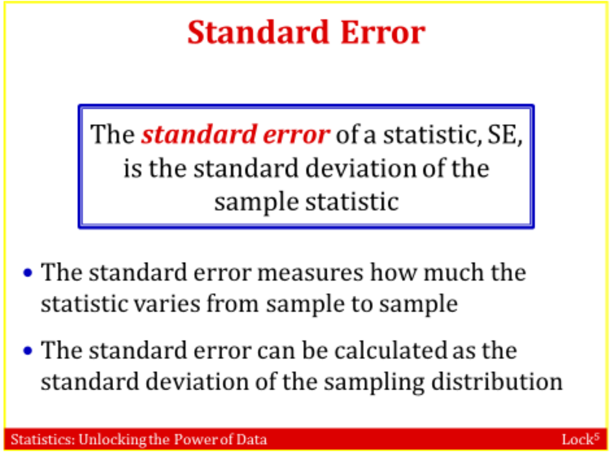
Sampling Caution

- If you take ***random samples***, the sampling distribution will be centered around the true population parameter
- If sampling bias exists (if you do not take random samples), your sampling distribution may give you bad information about the true parameter



Sampling Distribution

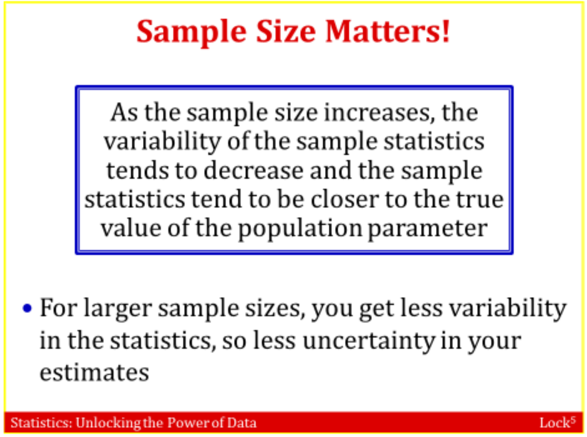
- We've learned about center and shape, but remember what we really care about is ***variability*** of the sampling distribution
- Remember our key question and answer: to assess uncertainty of a statistic, we need to know how much the statistic varies from sample to sample!
- The variability of the sample statistic is so important that it gets it's own name…



Standard Error

The ***standard error*** of a statistic, SE, is the standard deviation of the sample statistic

- The standard error measures how much the statistic varies from sample to sample
- The standard error can be calculated as the standard deviation of the sampling distribution



Sample Size Matters!

As the sample size increases, the variability of the sample statistics tends to decrease and the sample statistics tend to be closer to the true value of the population parameter

- For larger sample sizes, you get less variability in the statistics, so less uncertainty in your estimates

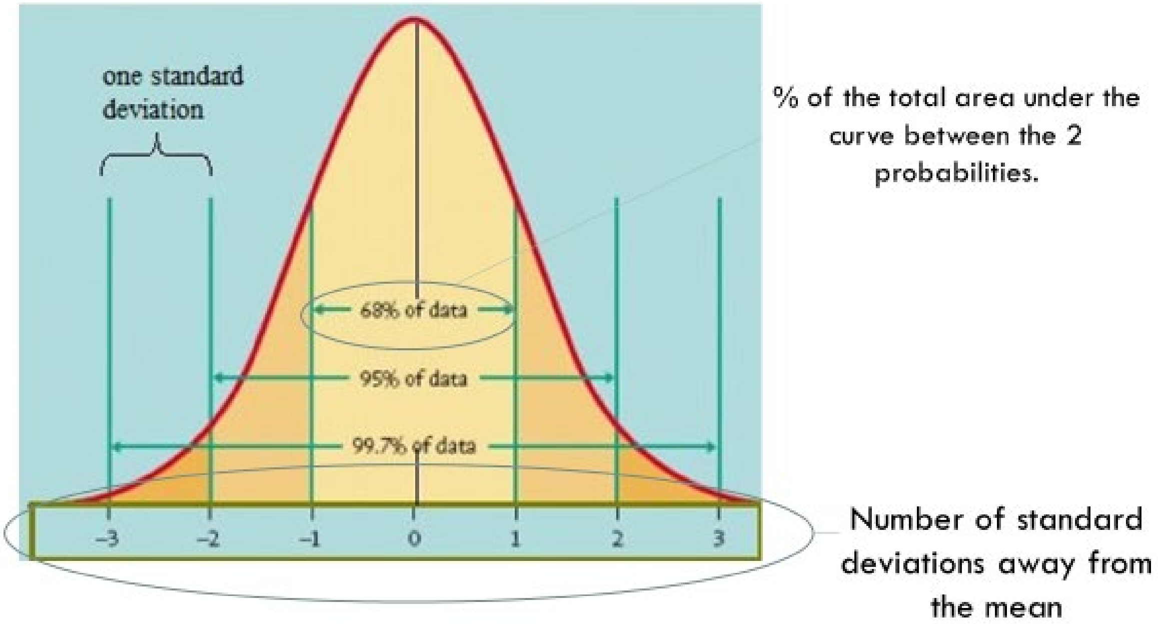
How can do we create sampling distributions?
Use the following steps to plot the data in StatKey.

- Go to: http://www.lock5stat.com/StatKey/
- Click on Sampling Distribution

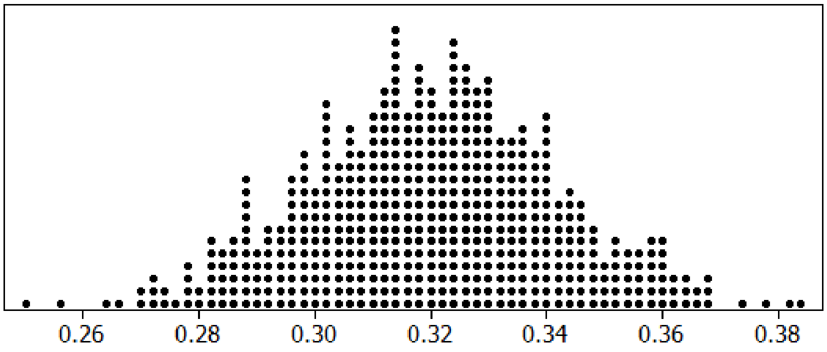
We can estimate the standard error by looking at the sampling distribution by applying the Empirical Rule. Your text book gives an abbreviated version of the empirical rule called the 95% rule.

The empirical rule can be applied to any bell shaped distribution. It states the following ….

- 68 % of data falls within the first standard deviation from the mean.
- 95% fall within two standard deviations.
- 99.7 % fall within three standard deviations.


one standard deviation
% of the total area under the curve between the 2 probabilities.
68% of data
95% of data
99.7% of data
−3
−2
−1
0
1
2
3
Number of standard deviations away from the mean

Example 3: Proportion Never Married
A sampling distribution is shown for the proportion of US citizens over 15 years old who have never been married, using the data from the 2010 US Census and random samples of size n = 500.



a). What does one dot in the dotplot represent?

One dot represents the proportion of people who have never been married in one sample of 500 people.

b). Use the sampling distribution to estimate the proportion of all US citizens over 15 years old who have never been married. Give correct notation for your answer.

We are estimating the population parameter p and we know that the sampling distribution is centered at the population parameter, so we estimate that p ≈ 0.32.

c). If we take a random sample of 500 US citizens over 15 years old and compute the proportion of the sample who have never been married, indicate how likely it is that we will see that result for each sample proportion below. *We see how likely the given sample proportion is to occur in a sample of size 500*

$\hat{p} = 0.30$	$\hat{p} = 0.20$	$\hat{p} = 0.37$	$\hat{p} = 0.74$
Likely to occur	*Very unlikely to occur*	*Not very likely but possible*	*VERY unlikely!*

The idea is just to get the students thinking about what is likely to happen by random chance.

d). Estimate the standard error of the sampling distribution.

We use the 95% rule to give a rough estimate of SE ≈ 0.02.

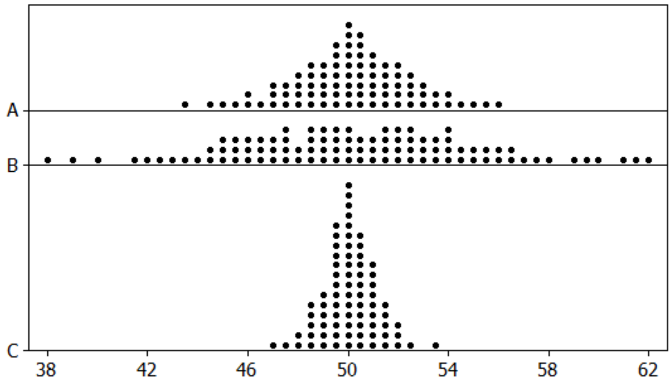
e). If we took samples of size 1000 instead of 500, and used the sample proportions to estimate the population proportion:

Would the estimates be *more accurate* or *less accurate*? More accurate

Would the standard error be *larger* or *smaller*? Smaller

Quick Self-Quiz: Effect of Sample Size

Three different sampling distributions A, B, and C are given for a population with mean 50. One corresponds to samples of size n = 25, one to samples of size n = 100, and one to samples of size n = 400. Match the sampling distributions with the three sample sizes, and estimate the standard error for each.



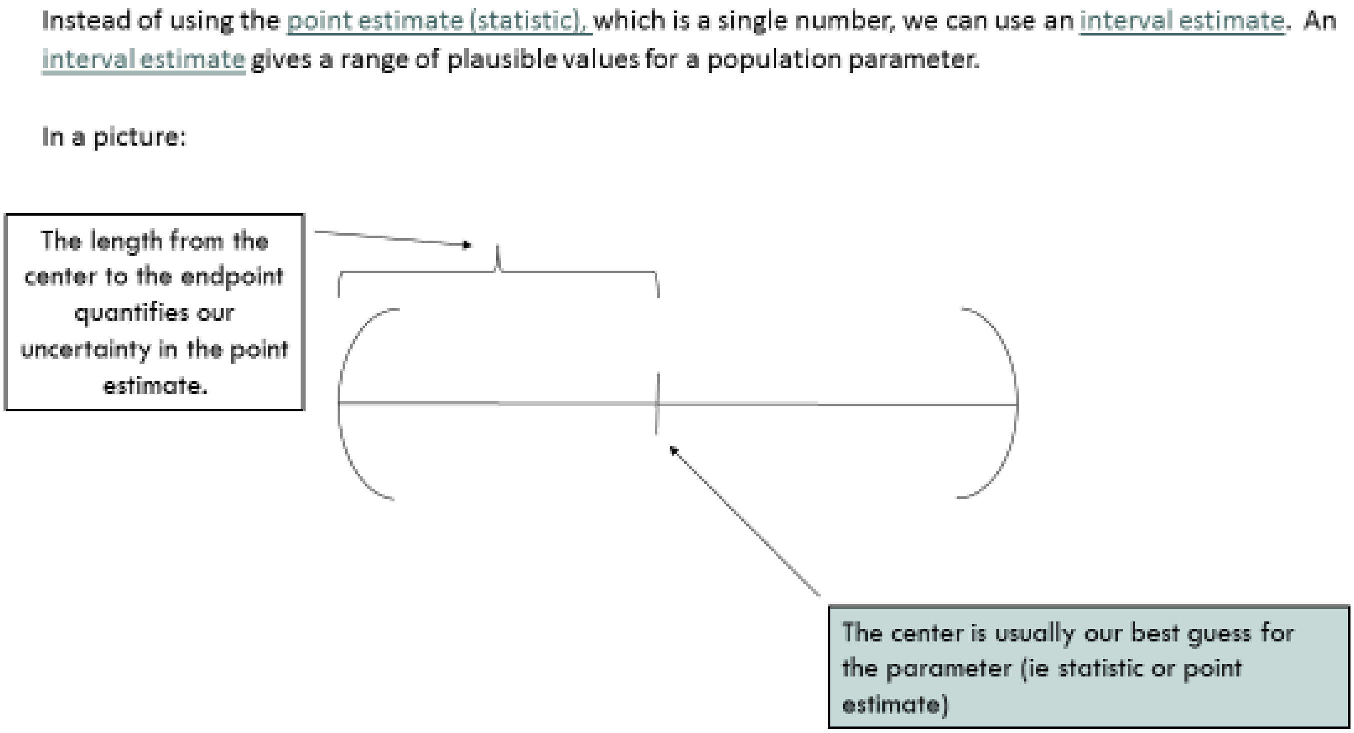
We know that the standard error goes down as the sample size goes up, so we match distribution
A with n = 100
B with n = 25
C with n = 400
We give a rough estimate of the standard error in each case using the 95% rule:
A: SE ≈ 2 *B: SE ≈ 4* *C: SE ≈ 1*

3.2 : Understanding and Interpreting Confidence Intervals

Notice that creating a sampling distribution requires us to take many samples from the population. In the real word we only get to take one sample from the population to help us estimate the parameter. There are 2 issues we need to address:

1. We know that the value of the statistic changes from sample to sample. How can we use a single number to estimate a parameter? –Section 3.2
2. Also since we only get one sample how do we estimate the variability of the statistic? – Section 3.3

- A *point estimate* doesn't tell us how close the estimate is likely to be to the parameter
- An *interval estimate* is more useful

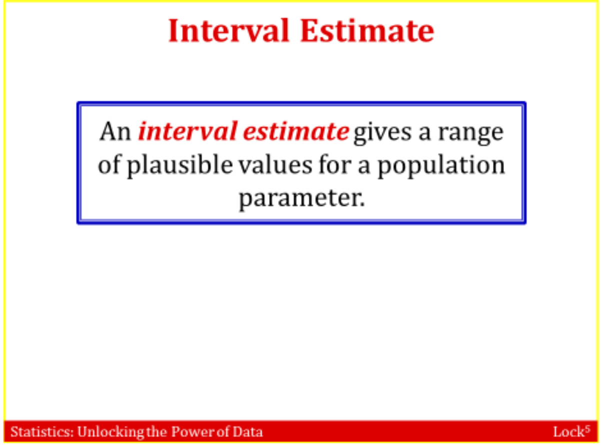


Instead of using the point estimate (statistic), which is a single number, we can use an interval estimate. An interval estimate gives a range of plausible values for a population parameter.

In a picture:

The length from the center to the endpoint quantifies our uncertainty in the point estimate.

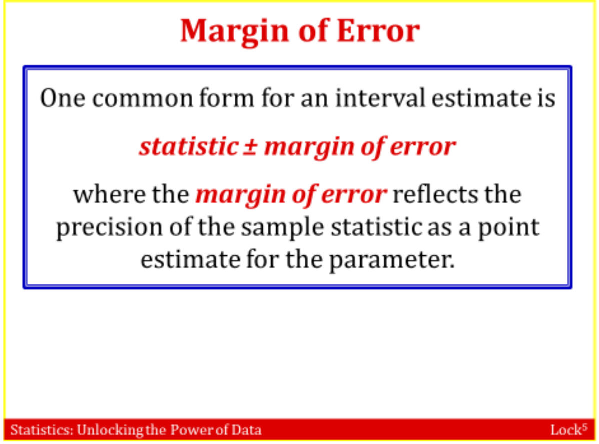
The center is usually our best guess for the parameter (ie statistic or point estimate)



Interval Estimate

An ***interval estimate*** gives a range of plausible values for a population parameter.

Statistics: Unlocking the Power of Data Lock^5



Margin of Error

One common form for an interval estimate is

statistic ± margin of error

where the ***margin of error*** reflects the precision of the sample statistic as a point estimate for the parameter.

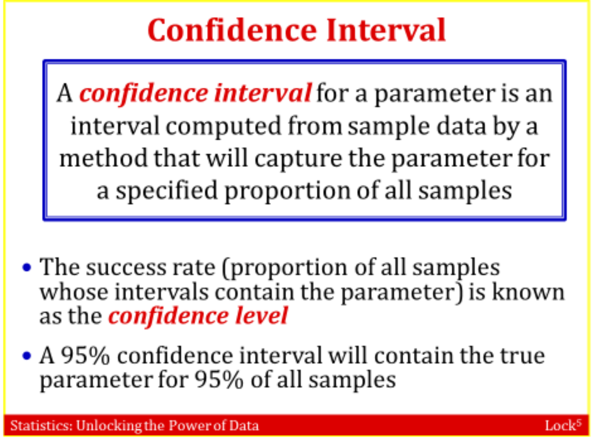
Statistics: Unlocking the Power of Data Lock^5

The *margin of error* measures how accurate the point estimate is likely to be in Estimating a parameter

Example: "Forty-five percent of American adults reported getting their health insurance from an employer…" Later in the report, it says "the maximum margin of sampling error is ±1 percentage point"

Interval estimate: 45% ± 1% = (44%, 46%)

The percentage of American adults getting their health insurance from an employer is probably between 44% and 46%

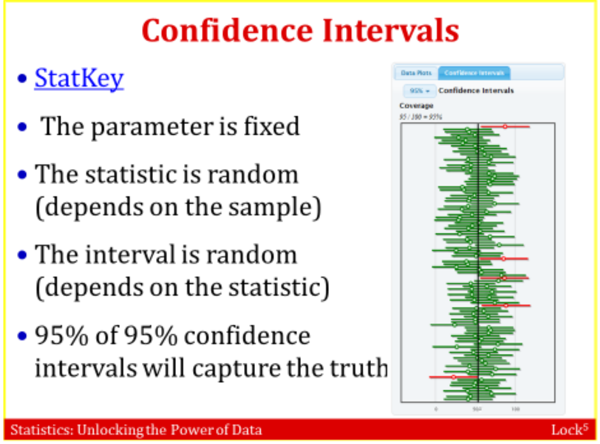


Confidence Interval

A ***confidence interval*** for a parameter is an interval computed from sample data by a method that will capture the parameter for a specified proportion of all samples

- The success rate (proportion of all samples whose intervals contain the parameter) is known as the ***confidence level***
- A 95% confidence interval will contain the true parameter for 95% of all samples

Statistics: Unlocking the Power of Data Lock^5



Confidence Intervals

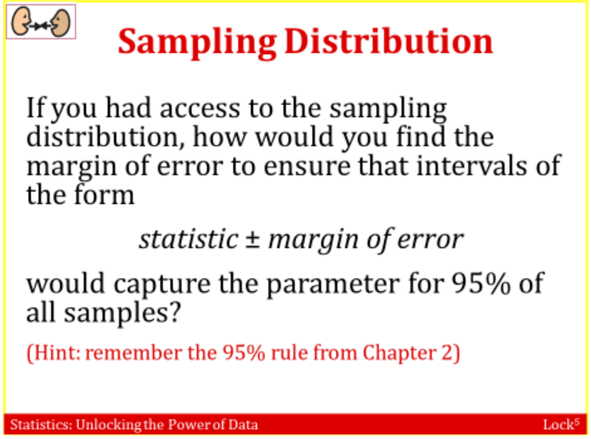
- StatKey
- The parameter is fixed
- The statistic is random (depends on the sample)
- The interval is random (depends on the statistic)
- 95% of 95% confidence intervals will capture the truth

95% Confidence Intervals

Coverage

95 / 100 = 95%

Statistics: Unlocking the Power of Data Lock^5



Sampling Distribution

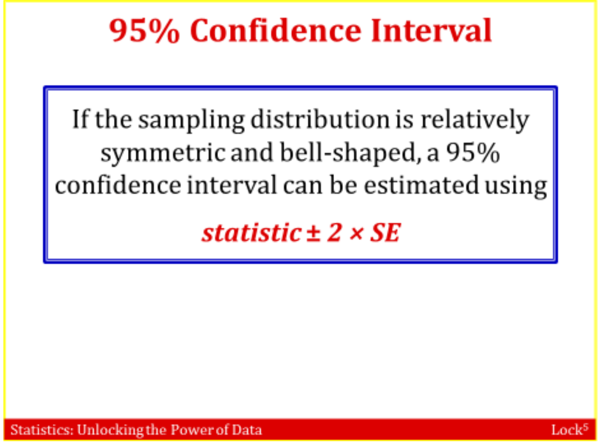
If you had access to the sampling distribution, how would you find the margin of error to ensure that intervals of the form

statistic ± margin of error

would capture the parameter for 95% of all samples?

(Hint: remember the 95% rule from Chapter 2)

Statistics: Unlocking the Power of Data Lock^5



95% Confidence Interval

If the sampling distribution is relatively symmetric and bell-shaped, a 95% confidence interval can be estimated using

statistic ± 2 × SE

Statistics: Unlocking the Power of Data Lock^5

Example: A survey of 1,502 Americans in January 2012 found that 86% consider the economy a "top priority" for the president and congress.The standard error for this statistic is 0.01.

What is the 95% confidence interval for the true proportion of all Americans that considered the economy a "top priority" at that time?

statistic ± 2×SE= 0.86 ± 2×0.01=0.86 ± 0.02=(0.84, 0.88)

Interpreting a Confidence Interval:

- 95% of all samples yield intervals that contain the true parameter, so we say we are "95% sure" or "95% confident" that one interval contains the truth.
- *"We are 95% confident that the true proportion of all Americans that considered the economy a 'top priority' in January 2012 is between 0.84 and 0.88"*

Example: Constructing Confidence Intervals

For each of the following, use the information to construct a 95% confidence interval and give notation for the quantity being estimated.

(a) $\hat{p} = 0.72$With standard error 0.04

Estimating p. Interval is 0.72 ± 2(0.04), giving 0.64 to 0.80.

(b) $\bar{x} = 27$with standard error 3.2.

Estimatingμ.. Interval is 27 ± 2(3.2), giving 20.6 to 33.4.

(c) $\hat{p}_1 - \hat{p}_2 = 0.05$With margin of error for 95% confidence of 0.02

Estimating the difference in proportions $p_1 - p_2$. Interval is 0.05 ± 0.02, giving 0.03 to 0.07.

Example: Adopting a Child in the US

A survey of 1,000 American adults conducted in January 2013 stated, "44% say it's too hard to adopt a child in the US." The survey goes on to say that, "The margin of sampling error is +/- 3 percentage points with a 95% level of confidence."

(a) What is the relevant sample statistic? Give appropriate notation and the value of the statistic.

$\hat{p} = 0.44$

(b) What population parameter are we estimating with this sample statistic?

p = the proportion of all American adults who say it is too hard to adopt a child in the US

(c) Use the margin of error to give a confidence interval for the estimate.

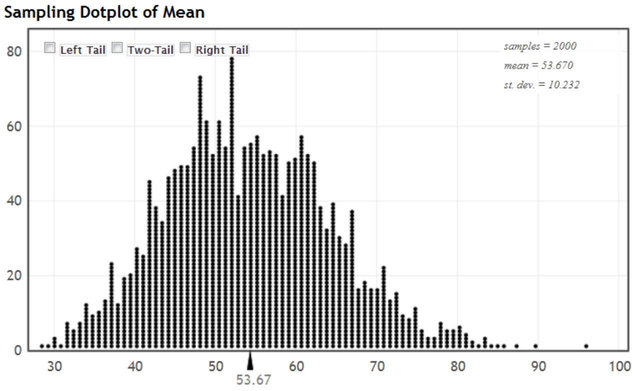
0.44 ± 0.03 gives an interval from 0.41 to 0.47

(d) Is 0.42 a plausible value of the population proportion? Is 0.50 a plausible value?

0.42 lies in the interval so is a plausible value, but 0.50 is not a plausible value.

Example: Budgets of Hollywood Movies

A sampling distribution is shown for mean budget (in millions of dollars) of movies to come out of Hollywood in 2011, using samples of size n = 20. We see that the standard error is about 10.23. Find the following sample means in the distribution and use the standard error 10.23 to find the 95% confidence interval given by each of the sample means listed. Indicate which of the confidence intervals successfully capture the true population mean of 53.48 million dollars (for all 134 movies).



(a) $\bar{x} = 40$

40 ±2(10.23) gives

19.54 to 60.46. Does contain population mean.

(b) $\bar{x} = 70$

70 ±2(10.23) gives

49.54 to 90.46. Does contain population mean.

(c) $\bar{x} = 84$

84 ±2(10.232) gives

63.54 to 104.46. Does not contain population mean.[Note that $\bar{x} = 84$ *is quite far in the tail of the sampling distribution.]*

Interpretation of a Confidence Interval

- 95% of all samples yield intervals that contain the true parameter, so we say we are "95% sure" or "95% confident" that one interval contains the truth.
- "We are 95% confident that the true proportion of all Americans that considered the economy a 'top priority' in January 2012 is between 0.84 and 0.88"

Common misinterpretation of a Confidence Interval

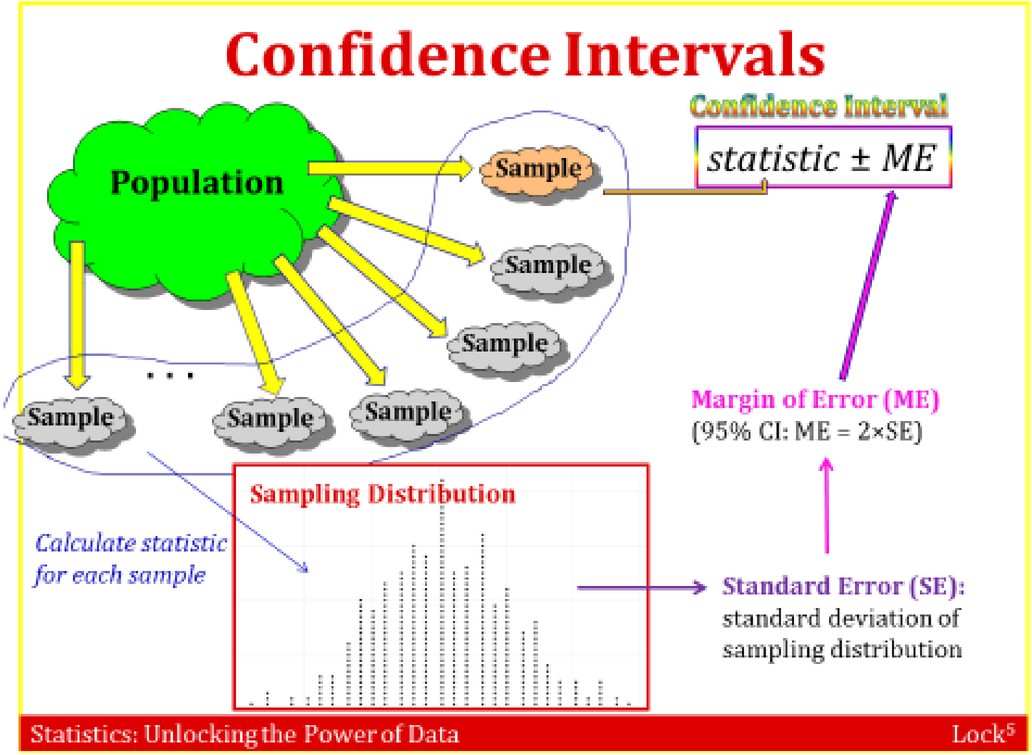
- Misinterpretation 1: "*A 95% confidence interval contains 95% of the data in the population*"

  The interpretation needs to refer to the location of the parameter not the data.
- Misinterpretation 2: "*I am 95% sure that the mean of a sample will fall within a 95% confidence interval for the mean*"

  The interpretation needs to refer to the location of the parameter not the statistic.

- Misinterpretation 3: "*The probability that the population parameter is in this particular 95% confidence interval is 0.95*"

  The parameter is a single fixed number, this is saying the parameter is random



Quick Self-Quiz: Interpreting a Confidence Interval

Using a sample of 24 deliveries described in "Diary of a Pizza Girl" on the Slice website, we find a 95% confidence interval for the mean tip given for a pizza delivery to be $2.18 to $3.90. Which of the following is a correct interpretation of this interval? Indicate all that are correct interpretations.

a). I am 95% sure that all pizza delivery tips will be between $2.18 and $3.90.

Incorrect. The interval is about a parameter (a summary of a population, mean here), not the individual values in the population (which is what this interpretation is referring to).

b). 95% of all pizza delivery tips will be between $2.18 and $3.90.

Incorrect. The interval is about the mean, not individual tips.

c). I am 95% sure that the mean pizza delivery tip for this sample will be between $2.18 and $3.90.

Incorrect. The interval is about the mean of the population, not a sample mean.

d). I am 95% sure that the mean tip for all pizza deliveries in this area will be between $2.18 and $3.90. *Correct!*

e). I am 95% sure that the confidence interval for the mean pizza delivery tip will be between $2.18 and $3.90.

Incorrect. The confidence is in where the population mean is, not where the interval itself is.

3.3: Constructing Bootstrap Confidence Intervals

Summary

- To create a plausible range of values for a parameter:
  - Take many random samples from the population, and compute the sample statistic for each sample
  - Compute the standard error as the standard deviation of all these statistics
  - Use statistic ± 2×SE
- One small problem…

Reality

… WE ONLY HAVE ONE SAMPLE!!!!

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

BOOTSTRAP!

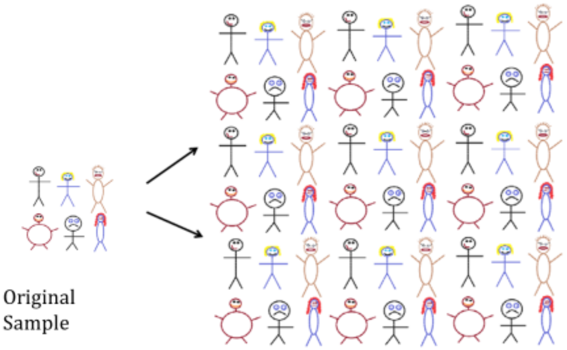
"Population"

- Imagine the "population" is many, many copies of the original sample
- (What do you have to assume?)

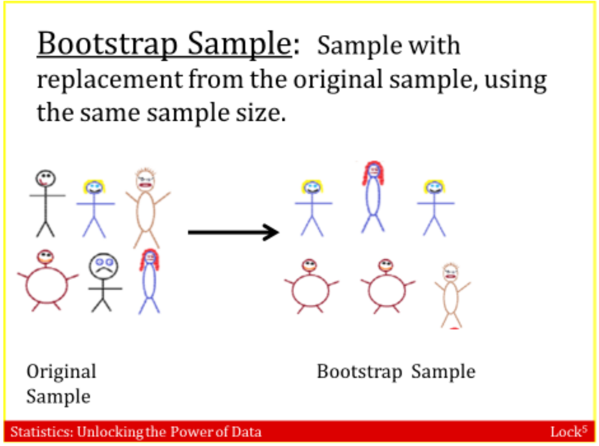
Sampling with Replacement

- 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:



A simulated "population" to sample from

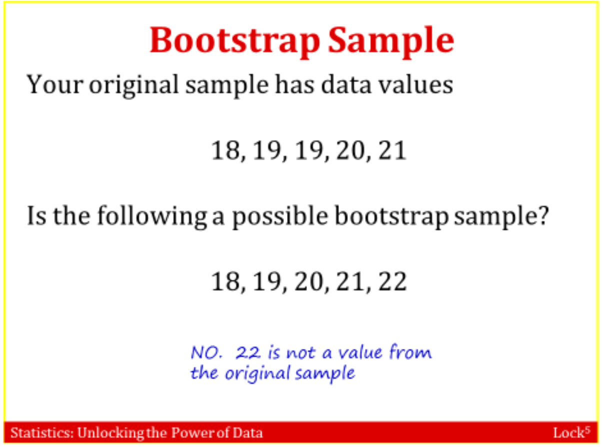


Bootstrap Sample: Sample with replacement from the original sample, using the same sample size.

Original Sample

Bootstrap Sample

Statistics: Unlocking the Power of Data Lock^5



Bootstrap Sample

Your original sample has data values

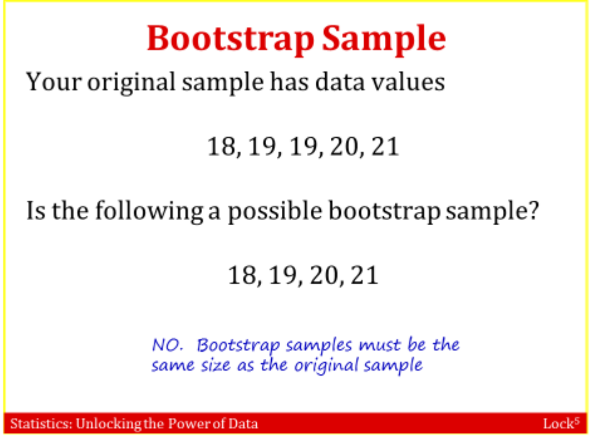
18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21, 22

NO. 22 is not a value from the original sample

Statistics: Unlocking the Power of Data Lock^5



Bootstrap Sample

Your original sample has data values

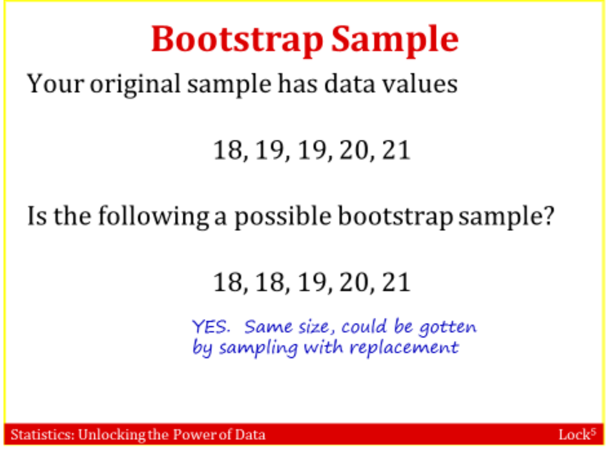
18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

NO. Bootstrap samples must be the same size as the original sample

Statistics: Unlocking the Power of Data Lock^5



Bootstrap Sample

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 18, 19, 20, 21

YES. Same size, could be gotten by sampling with replacement

Statistics: Unlocking the Power of Data Lock^5

Example 1: Textbook Prices

Prices of a random sample of 10 textbooks (rounded to the nearest dollar) are shown:

$132 $87 $185 $52 $23 $147 $125 $93 $85 $72

a). What is the sample mean? $\bar{x} = 100.1$

b). Describe carefully how we could use cards to create one bootstrap statistic from this sample. Be specific.

We use 10 cards and write the 10 sample values on the cards. We then mix them up, draw one, record the value on it, and put it back. Mix them up again, draw another, record the value, and put it back. Do this 10 times. Then compute the sample mean of this bootstrap sample.

c). Where will be bootstrap distribution be centered? What shape do we expect it to have?

It will be centered approximately at the sample mean of 100.1 and we expect it to be bell-shaped.

Example 2: Reese's Pieces

We wish to estimate the proportion of Reese's Pieces that are orange, and we have one package of Reese's Pieces containing 55 pieces. Describe carefully how we can use this one sample to create a bootstrap statistic. Be specific.

Mix up the pieces and pull one out and record whether or not it is orange and then put it back. Mix them up again, pull one out again and record whether or not it is orange. Put it back and continue the process until we have recorded the result for 55 pieces sampled this way. Compute the proportion of orange candies from this bootstrap sample.

Quick Self-Quiz: Bootstrap Samples

A sample consists of the following values: 8, 4, 11, 3, 7.

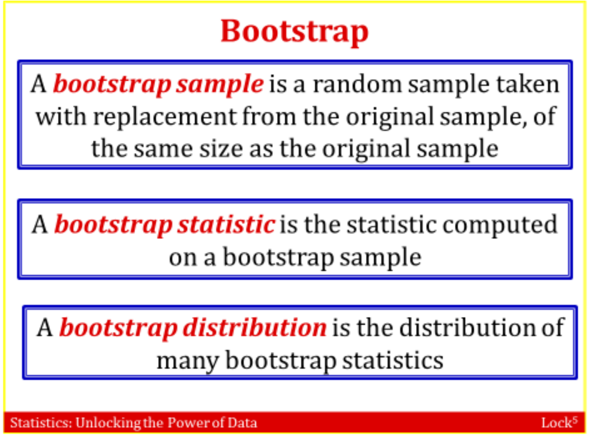
Which of the following are possible bootstrap samples from this sample?

a). 8, 3, 7, 11 *No because the sample size does not match.*

b). 4, 11, 4, 3, 3 *Yes.*

c). 3, 4, 5, 7, 8 *No, because 5 is not in the original sample.*

d). 7, 8, 8, 3, 4 *Yes.*

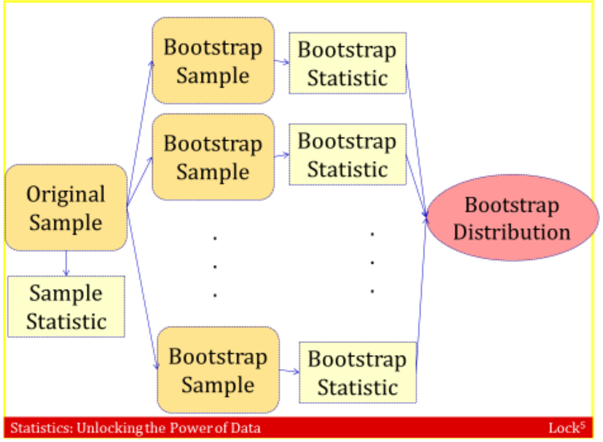


Bootstrap

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



Original Sample

Sample Statistic

Bootstrap Sample

Bootstrap Statistic

Bootstrap Sample

Bootstrap Statistic

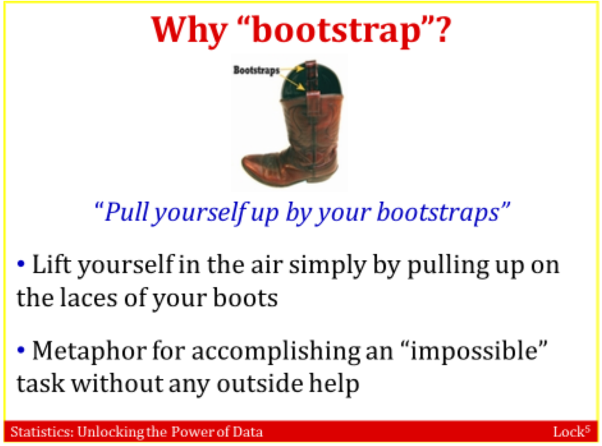
. . .

. . .

Bootstrap Sample

Bootstrap Statistic

Bootstrap Distribution

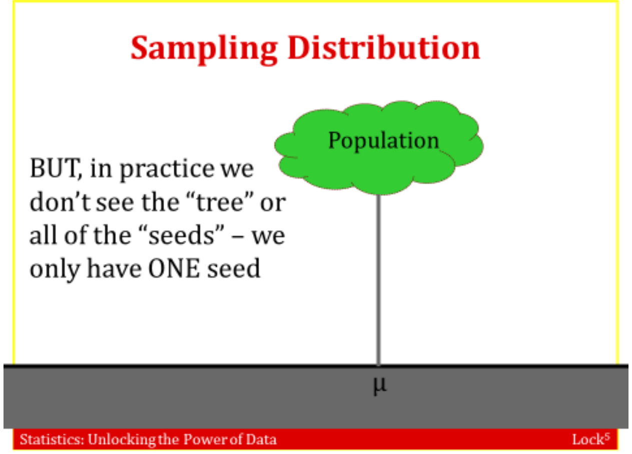


Why "bootstrap"?

Bootstraps

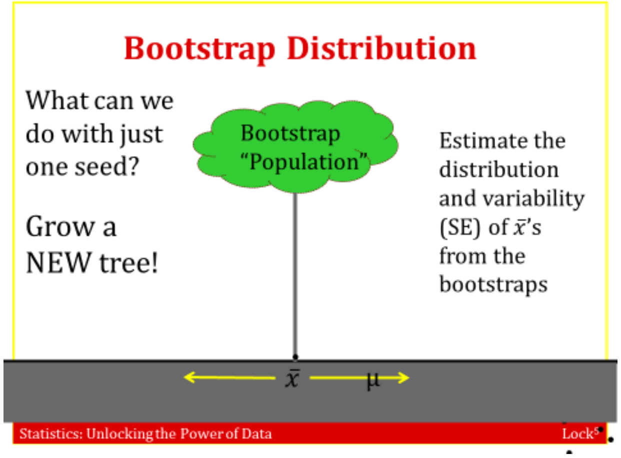
"Pull yourself up by your bootstraps"

- 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



Sampling Distribution

BUT, in practice we don't see the "tree" or all of the "seeds" – we only have ONE seed

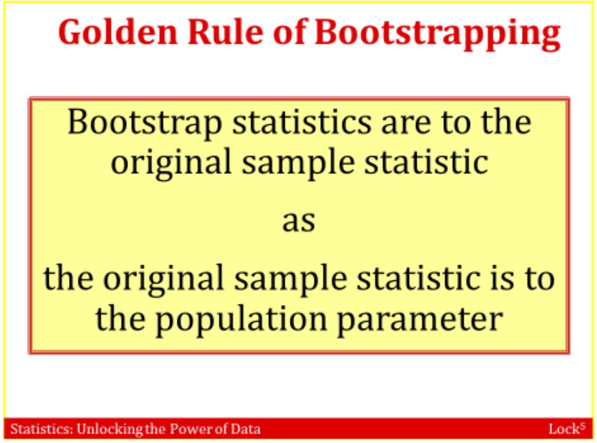


Bootstrap Distribution

What can we do with just one seed?

Grow a NEW tree!

Estimate the distribution and variability (SE) of $\bar{x}$'s from the bootstraps

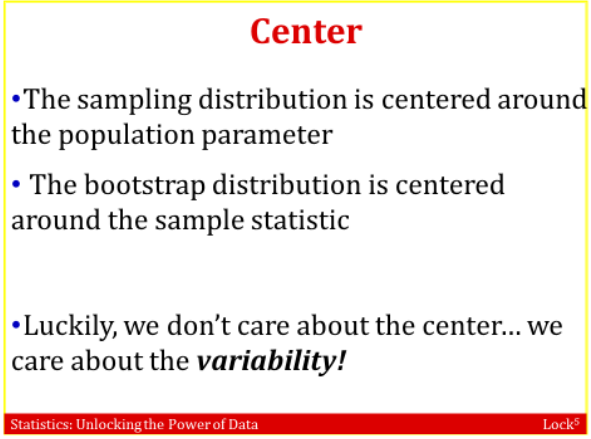


Golden Rule of Bootstrapping

Bootstrap statistics are to the original sample statistic

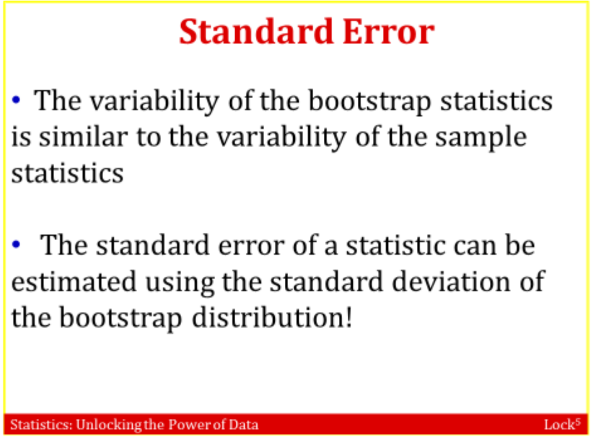
as

the original sample statistic is to the population parameter



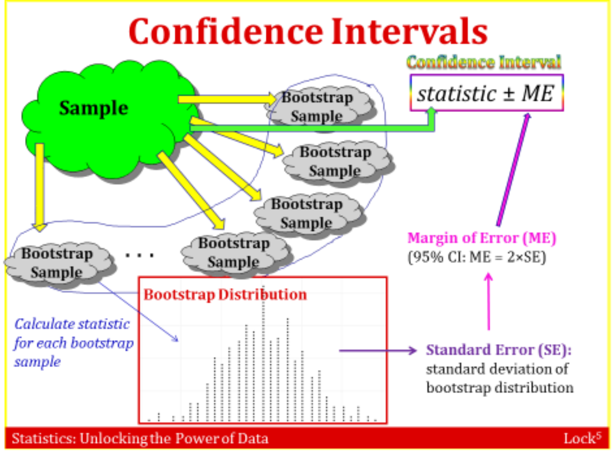
Center

- The sampling distribution is centered around the population parameter
- The bootstrap distribution is centered around the sample statistic
- Luckily, we don't care about the center… we care about the ***variability!***

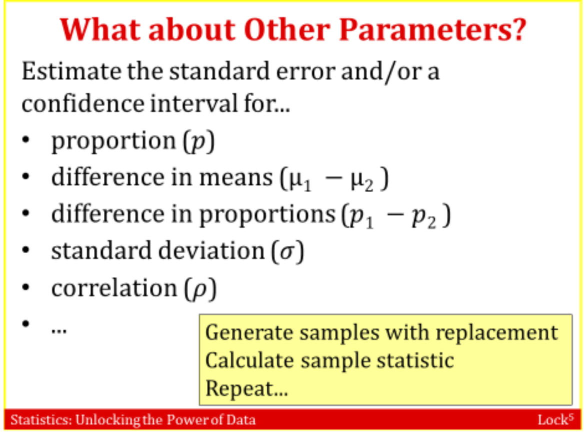


Standard Error

- 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

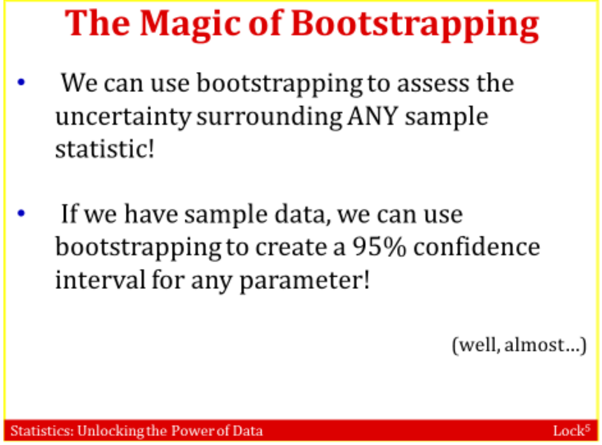


What about Other Parameters?

Estimate the standard error and/or a confidence interval for...

- proportion (p)
- difference in means ($\mu_1 - \mu_2$)
- difference in proportions ($p_1 - p_2$)
- standard deviation (σ)
- correlation (ρ)
- …

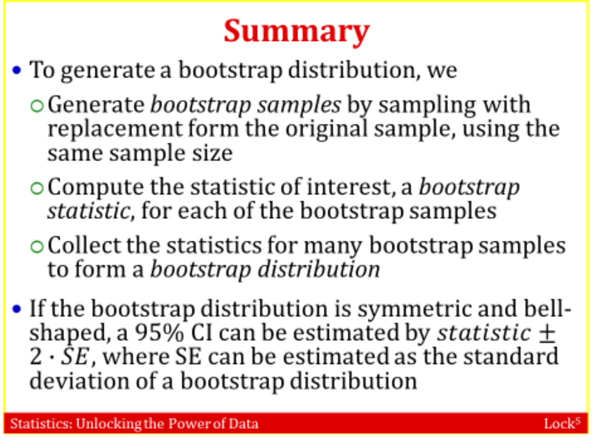
Generate samples with replacement
Calculate sample statistic
Repeat…



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…)

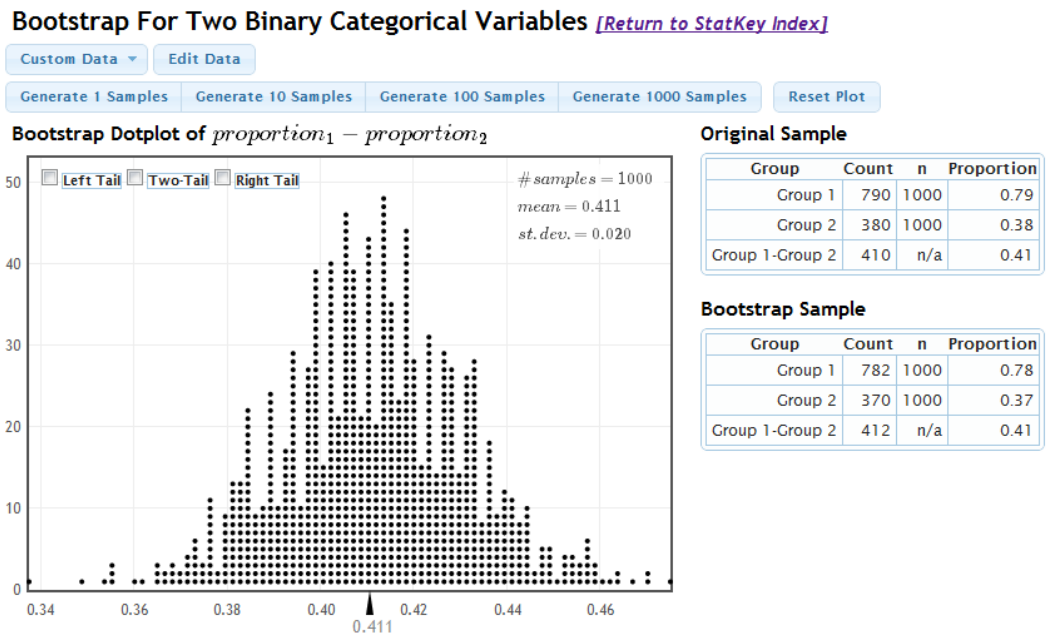


Summary

- To generate a bootstrap distribution, we
  - ○ Generate *bootstrap samples* by sampling with replacement form the original sample, using the same sample size
  - ○ Compute the statistic of interest, a *bootstrap statistic*, for each of the bootstrap samples
  - ○ Collect the statistics for many bootstrap samples to form a *bootstrap distribution*
- If the bootstrap distribution is symmetric and bell-shaped, a 95% CI can be estimated by $statistic \pm 2 \cdot SE$, where SE can be estimated as the standard deviation of a bootstrap distribution

Example : Global Warning by Political Party

Does belief in global warming differ by political party? When the question *"Is there solid evidence of global warming?"* was asked, the sample proportion answering "yes" was 79% among Democrats and 38% among Republicans. A bootstrap distribution for the difference in proportions ($\hat{p}_D - \hat{p}_R$) is shown (assuming samples of size 1000 from each party). Use the information there to give a 95% CI for the difference in proportions.



Bootstrap For Two Binary Categorical Variables [Return to StatKey Index]

Original Sample

Group	Count	n	Proportion
Group 1	790	1000	0.79
Group 2	380	1000	0.38
Group 1-Group 2	410	n/a	0.41

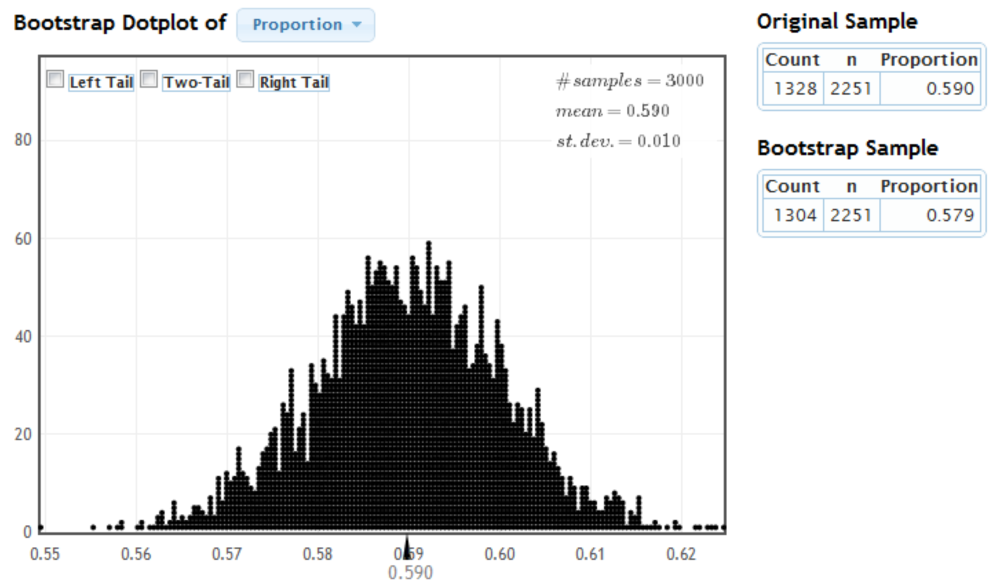
Bootstrap Sample

Group	Count	n	Proportion
Group 1	782	1000	0.78
Group 2	370	1000	0.37
Group 1-Group 2	412	n/a	0.41

The sample difference in proportions is 0.79 – 0.38 = 0.41 and the standard error from the bootstrap distribution is 0.020 so we compute the 95% confidence interval using 0.41 ± 2(0.020), giving an interval of 0.37 to 0.45. We are 95% confident that the proportion of Democrats who believe there is solid evidence of global warming is between 0.37 and 0.45 higher than the proportion of Republicans who believe this.

Example : Global Warming

What percentage of Americans believe in global warming? A survey on 2,251 randomly selected individuals conducted in October 2010 found that 1,328 answered Yes to the question *"Is there solid evidence of global warming?"* A bootstrap distribution for this data is shown. Use the information there to give and interpret a 95% CI for the proportion of Americans who believe there is solid evidence of global warming.



Original Sample

Count	n	Proportion
1328	2251	0.590

Bootstrap Sample

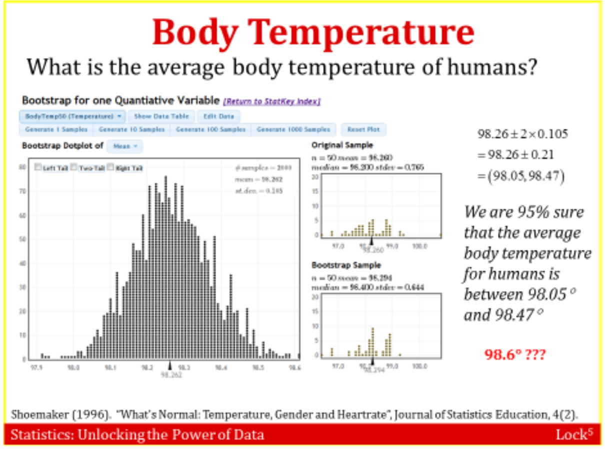
Count	n	Proportion
1304	2251	0.579

The sample proportion is 0.590 and the standard error from the bootstrap distribution is 0.010 so we compute the 95% confidence interval using 0.590 ± 2(0.010), giving an interval of 0.57 to 0.61. We are 95% confident that the proportion of Americans who believe there is solid evidence of global warming is between 0.57 and 0.61.

3.4 : Bootstrap Confidence Intervals using Percentiles

In the previous sections we looked at how to construct a 95% confidence interval, where we estimated the SE by using the sampling distribution or the bootstrap distribution. In this section we are going to look at how to create a confidence interval for any level of confidence using a bootstrap distribution.

.



Body Temperature

What is the average body temperature of humans?

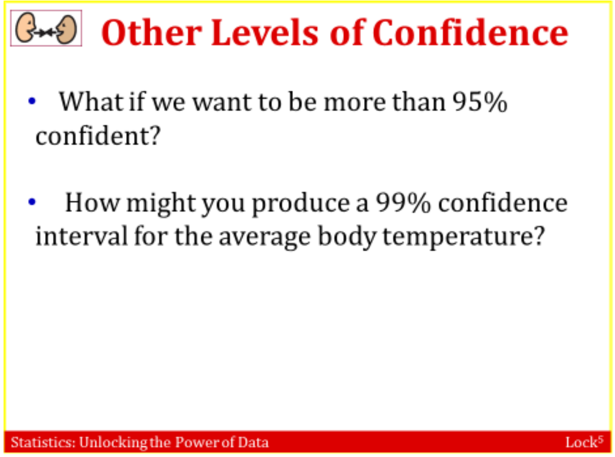
$$98.26 \pm 2 \times 0.105$$
$$= 98.26 \pm 0.21$$
$$= (98.05, 98.47)$$

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

98.6° ???

Shoemaker (1996). "What's Normal: Temperature, Gender and Heartrate", Journal of Statistics Education, 4(2).

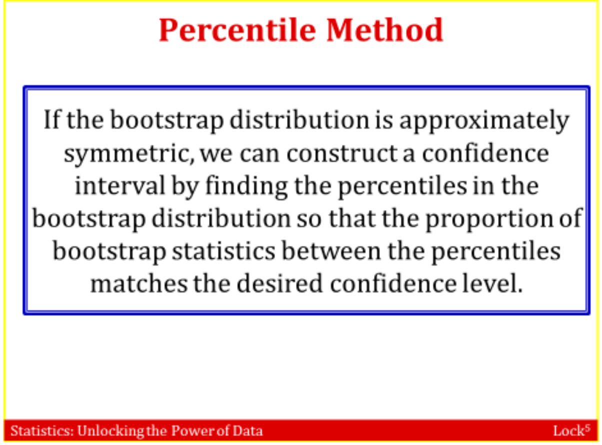
Statistics: Unlocking the Power of Data Lock^5



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?

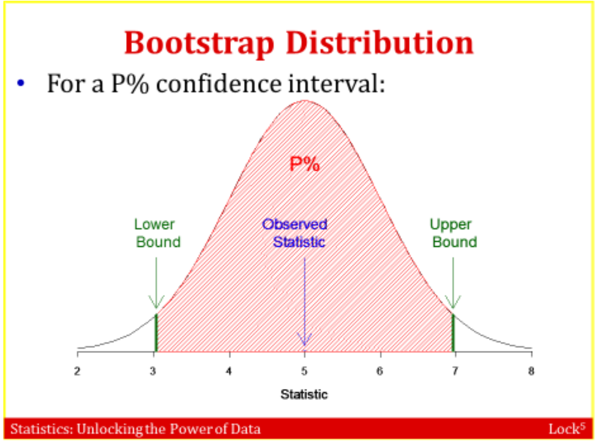
Statistics: Unlocking the Power of Data Lock^5



Percentile Method

If the bootstrap distribution is approximately symmetric, we can construct a confidence interval by finding the percentiles in the bootstrap distribution so that the proportion of bootstrap statistics between the percentiles matches the desired confidence level.

Statistics: Unlocking the Power of Data Lock^5



Bootstrap Distribution

- For a P% confidence interval:

Statistics: Unlocking the Power of Data Lock^5

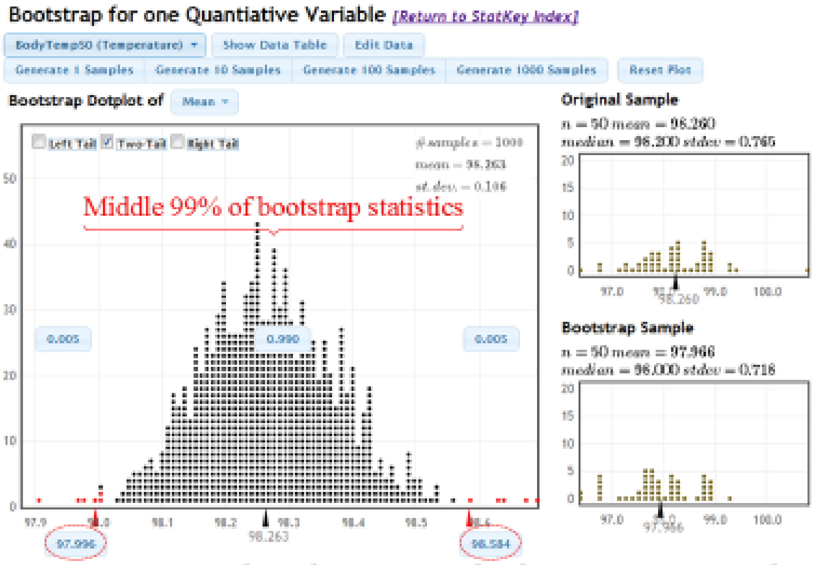
Percentile Method

- 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

  (0.5th percentile, 99.5th percentile)

where the percentiles refer to the bootstrap distribution.

Body Temperature



Bootstrap for one Quantiative Variable [Return to StatKey Index]

# samples = 1000
mean = 98.263
st. dev. = 0.106

Middle 99% of bootstrap statistics

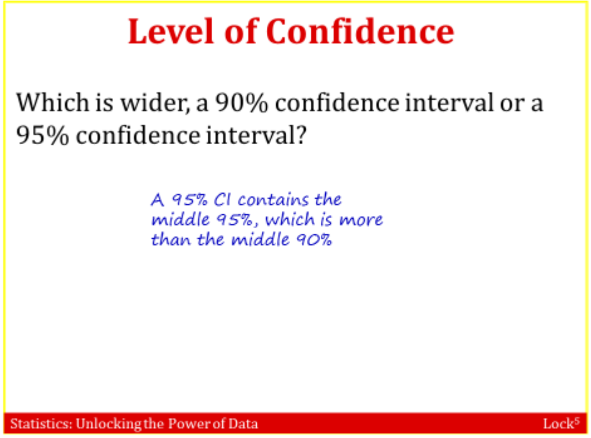
Original Sample

$n = 50$ mean = 98.260
median = 98.200 stdev = 0.765

Bootstrap Sample

$n = 50$ mean = 97.966
median = 98.000 stdev = 0.718

We are 99% sure that the average body temperature is between 98.00° and 98.58°.

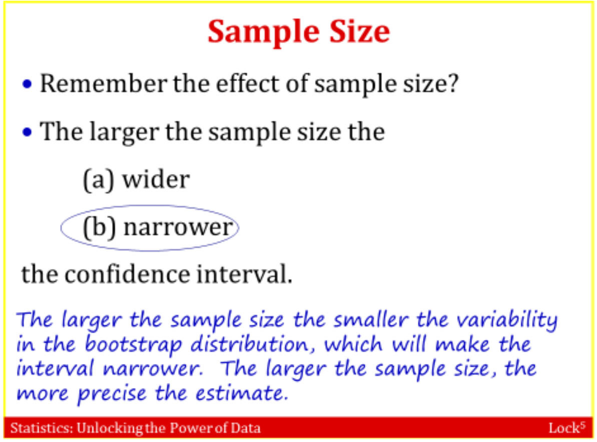


Level of Confidence

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

A 95% CI contains the middle 95%, which is more than the middle 90%

Statistics: Unlocking the Power of Data Lock^5



Sample Size

- Remember the effect of sample size?
- The larger the sample size the

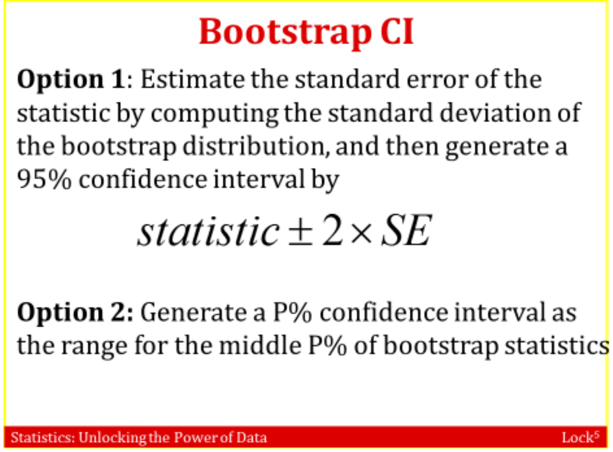
(a) wider

[✓ (b) narrower]

the confidence interval.

The larger the sample size the smaller the variability in the bootstrap distribution, which will make the interval narrower. The larger the sample size, the more precise the estimate.

Statistics: Unlocking the Power of Data Lock^5



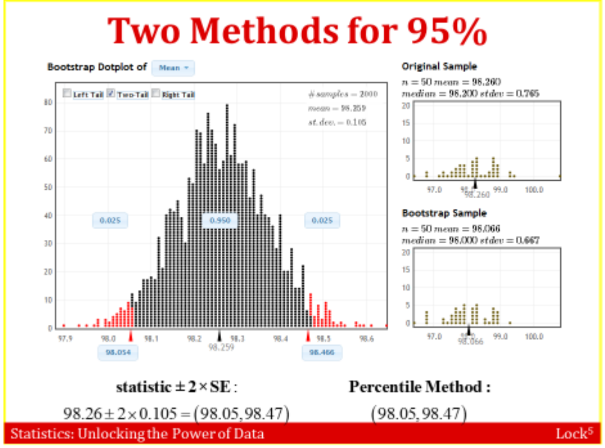
Bootstrap CI

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

$$statistic \pm 2 \times SE$$

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

Statistics: Unlocking the Power of Data Lock^5



Two Methods for 95%

statistic ± 2 × SE :

98.26 ± 2 × 0.105 = (98.05, 98.47)

Percentile Method :

(98.05, 98.47)

Statistics: Unlocking the Power of Data Lock^5

Example: Body Temperature

Is normal body temperature really 98.6 ◦F? A sample of body temperature for 50 healthy individuals was taken. Find this dataset in StatKey under “Confidence Interval for a Mean” or as BodyFat50 in the text’s datasets.

(a) What is the sample mean? What is the sample standard deviation?

Mean = 98:26 and standard deviation = 0:765

(b) Generate a bootstrap distribution, using at least 1000 simulated statistics. What is the standard error?

SE ≈ 0:105. Answers will vary slightly with different simulations.

(c) Use the standard error to find a 95% confidence interval. Show your work. Is 98.6 in the interval?

x ±2 · *SE* = 98:26 ±2(0:105), which gives the interval 98.05 to 98.47. We see that 98.6 is not in the interval.

(d) Using the same distribution, find a 95% confidence interval using the “Two-tail” option on StatKey (or other technology to give percentiles from the bootstrap distribution).

98.05 to 98.47. Answers will vary slightly with different simulations

(e) Compare the two 95% confidence intervals you found. Are they similar?

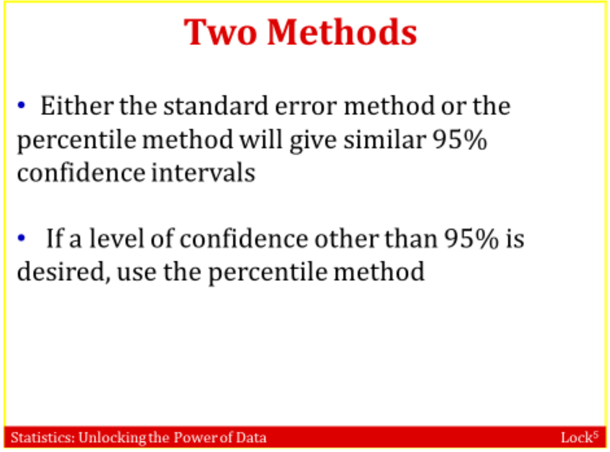
Yes, very similar.

(f) Still using the same bootstrap distribution, give a 99% confidence interval. 97.98 to 98.54

(g) Is the 99% confidence interval wider or narrower than the 95% confidence interval? Wider

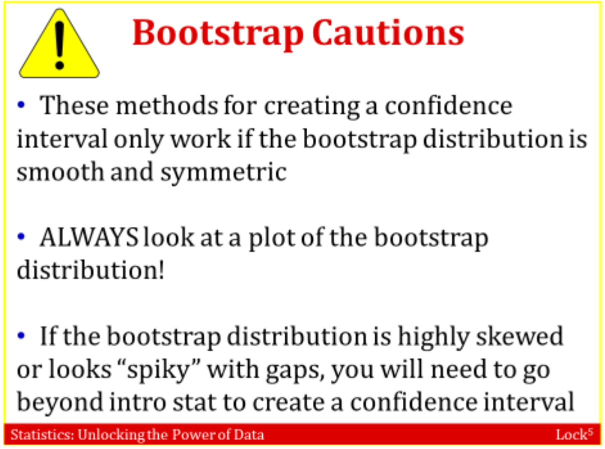
(h) Clearly interpret the 99% confidence interval

We are 99% sure that the mean body temperature for all healthy individuals is between 98.017 and 98.544.



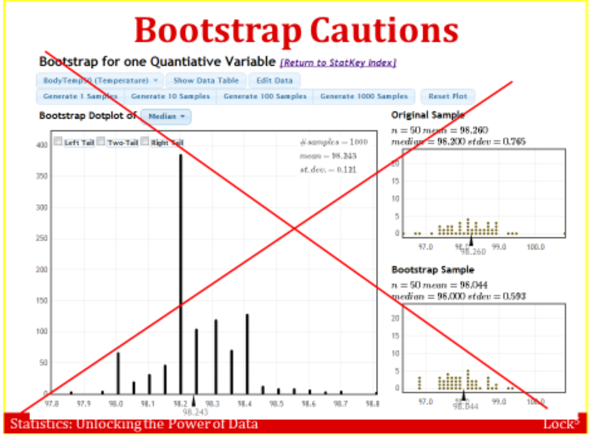
Two Methods

- Either the standard error method or the percentile method will give similar 95% confidence intervals
- 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

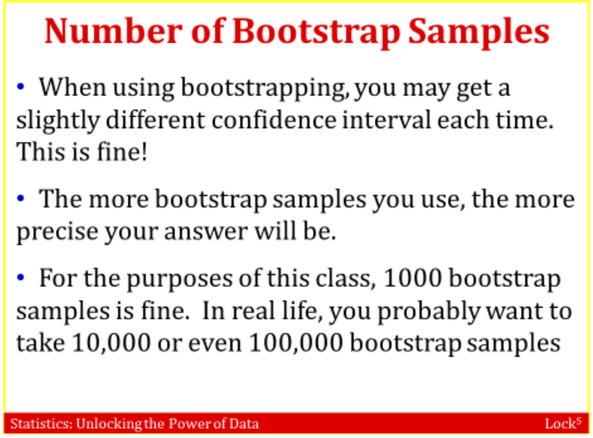


Bootstrap Cautions

Bootstrap for one Quantiative Variable [Return to StatKey index]

Original Sample
$n = 50$ $mean = 98.260$
$median = 98.200$ $stdev = 0.765$

Bootstrap Sample
$n = 50$ $mean = 98.044$
$median = 98.000$ $stdev = 0.593$



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.
- 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

Example: Problems with Bootstrap Distributions
If a bootstrap distribution is not relatively symmetric, it is not appropriate to use the methods of this chapter to construct a confidence interval. Consider the following data set: 5, 6, 7, 8, 25, 100
(a) What is the standard deviation of this dataset? $s = 37{:}413$

(b) Use StatKey (or other technology) to create a bootstrap distribution for the standard deviation of this dataset. Describe the distribution. Is the distribution symmetric and bell-shaped?
It is not at all symmetric or bell-shaped. There are four separate sharp peaks and the distribution is strongly skewed to the right.
(c) Is it appropriate to use the methods of this section to find a bootstrap confidence interval for this standard deviation?
No, we should only use the methods of this section when the bootstrap distribution is roughly symmetric.
(d) Discuss with a neighbor why the bootstrap distribution might look the way it does.
There are two outliers in this dataset, and the standard deviation from a bootstrap sample will be dramatically different depending on how many copies of each outlier get selected.