SOC245: Visualizing Data
Lecture 9: Central Limit Theorem

Chris Felton and Vikram Ramaswamy

Freshman Scholars Institute
Princeton University

Aug 3, 2022

Where we are and where we're going

Before:

- Computing estimates on a sample.

Where we are and where we're going

Before:

- Computing estimates on a sample.

Last class: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution

Where we are and where we're going

Before:

- Computing estimates on a sample.

Last class: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.

Where we are and where we're going

Before:

- Computing estimates on a sample.

Last class: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping

Where we are and where we're going

Before:

- Computing estimates on a sample.

Last class: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping

Where we are and where we're going

Before:

- Computing estimates on a sample.

Last class: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping

Today:

- Getting guarantees about confidence intervals (with more assumptions)

Review: Sampling Distribution

- Suppose we're interested in Biden's approval rating (like yesterday).

Review: Sampling Distribution

- Suppose we're interested in Biden's approval rating (like yesterday).
- We pick a sample of 100 people, ask them how they feel about Biden, and compute the mean (this is our sample mean).

Review: Sampling Distribution

- Suppose we're interested in Biden's approval rating (like yesterday).
- We pick a sample of 100 people, ask them how they feel about Biden, and compute the mean (this is our sample mean).
- What's the sampling distribution?

Review: Sampling Distribution

- Suppose we're interested in Biden's approval rating (like yesterday).
- We pick a sample of 100 people, ask them how they feel about Biden, and compute the mean (this is our sample mean).
- What's the sampling distribution?
- Suppose we repeat this sampling procedure 10000 times

Review: Sampling Distribution

- Suppose we're interested in Biden's approval rating (like yesterday).
- We pick a sample of 100 people, ask them how they feel about Biden, and compute the mean (this is our sample mean).
- What's the sampling distribution?
- Suppose we repeat this sampling procedure 10000 times
- We pick a sample of 100, compute the sample mean for this, and repeat.

Review: Sampling Distribution

- Suppose we're interested in Biden's approval rating (like yesterday).
- We pick a sample of 100 people, ask them how they feel about Biden, and compute the mean (this is our sample mean).
- What's the sampling distribution?
- Suppose we repeat this sampling procedure 10000 times
- We pick a sample of 100, compute the sample mean for this, and repeat.
- We now have this set of values, and we can analyze the distribution.

Review: Sampling Distribution

- Suppose we're interested in Biden's approval rating (like yesterday).
- We pick a sample of 100 people, ask them how they feel about Biden, and compute the mean (this is our sample mean).
- What's the sampling distribution?
- Suppose we repeat this sampling procedure 10000 times
- We pick a sample of 100, compute the sample mean for this, and repeat.
- We now have this set of values, and we can analyze the distribution.
- This is called the **sampling distribution**

Pop Quiz!

Can we compute the sampling distribution in practice?

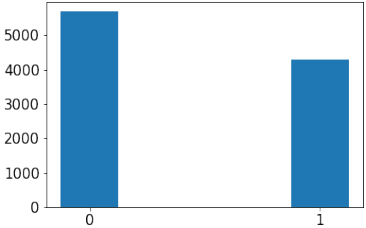
Pop Quiz!

Can we compute the sampling distribution in practice?

No! It's too time consuming and expensive to keep picking samples from the population.

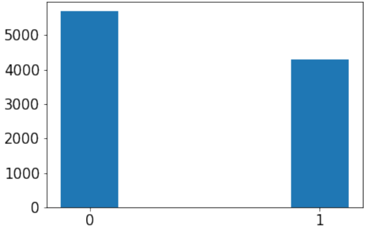
Outline

What do sampling distributions look like?

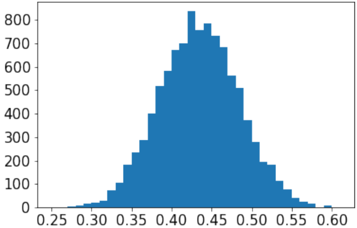


Population distribution

What do sampling distributions look like?



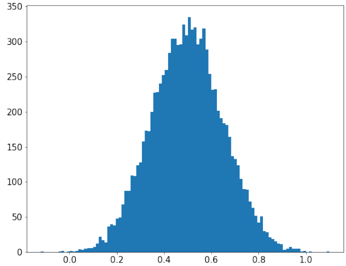
Population distribution



Distribution of sample mean

What do sampling distributions look like?

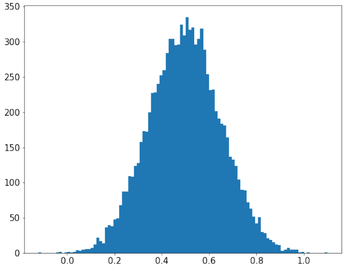
Now, instead of having a population that has a 0 / 1 rating, suppose we have people expressing their fraction of support for Biden, and the population distribution looks like this:



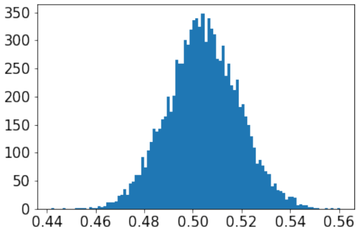
Population distribution

What do sampling distributions look like?

Now, instead of having a population that has a 0 / 1 rating, suppose we have people expressing their fraction of support for Biden, and the population distribution looks like this:



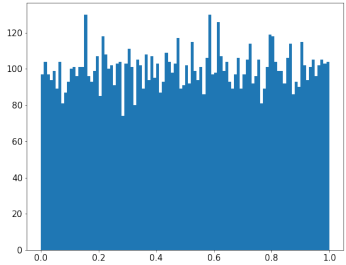
Population distribution



Distribution of sample mean

What do sampling distributions look like?

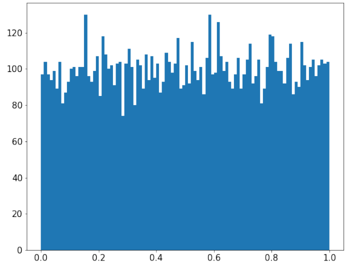
Now, instead of having a population that has a 0 / 1 rating, suppose we have people expressing their fraction of support for Biden, and the population distribution looks like this:



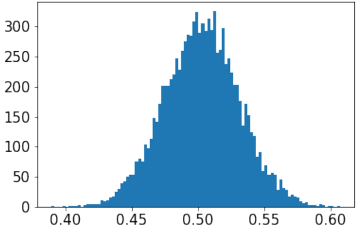
Population distribution

What do sampling distributions look like?

Now, instead of having a population that has a 0 / 1 rating, suppose we have people expressing their fraction of support for Biden, and the population distribution looks like this:



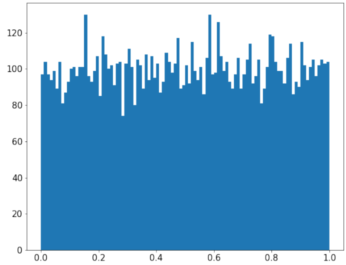
Population distribution



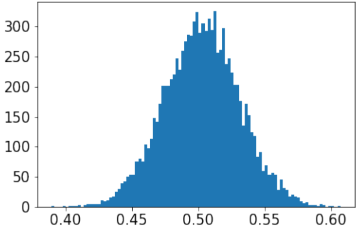
Distribution of sample mean

What do sampling distributions look like?

Now, instead of having a population that has a 0 / 1 rating, suppose we have people expressing their fraction of support for Biden, and the population distribution looks like this:



Population distribution



Distribution of sample mean

What do you notice about these sampling distributions?

What do sampling distributions look like?

- They all appear to have the same shape!

What do sampling distributions look like?

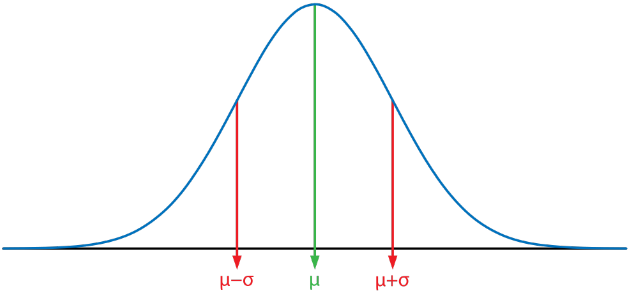
- They all appear to have the same shape!
- This shape is called a **Normal** or a **Gaussian** distribution.

What do sampling distributions look like?

- They all appear to have the same shape!
- This shape is called a **Normal** or a **Gaussian** distribution.
- You might have also seen it called a bell-curve (but that covers other distributions as well).

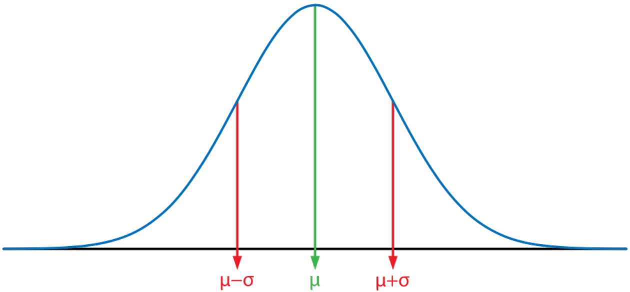
Normal distributions

Suppose we have a Normal distribution with mean μ and standard deviation σ.



Normal distributions

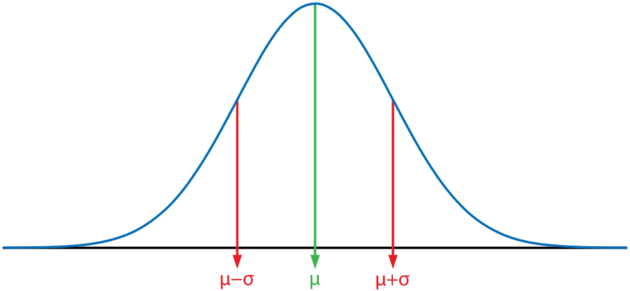
Suppose we have a Normal distribution with mean μ and standard deviation σ.



- It's symmetric around the mean value

Normal distributions

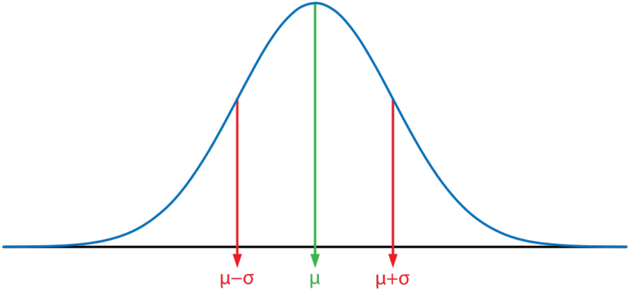
Suppose we have a Normal distribution with mean μ and standard deviation σ.



- The mean, median and mode of this distibution is μ

Normal distributions

Suppose we have a Normal distribution with mean μ and standard deviation σ.



- This distribution occurs naturally pretty often: the height of people is a Normal distribution within each gender.

An amazing theorem in Statistics

The sampling distribution of the sample means of n independent, identically distributed random variables from a population with mean μ and variance σ^2 approaches a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size gets larger

An amazing theorem in Statistics

The sampling distribution of the sample means of n independent, identically distributed random variables from a population with mean μ and variance σ^2 approaches a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size gets larger

This is called the **Central Limit Theorem**.

An amazing theorem in Statistics

The sampling distribution of the sample means of n **independent**, identically distributed random variables from a population with mean μ and variance σ^2 approaches a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size gets larger

This is defining how we sample: drawing an observation should not affect any observations we draw after that.

An amazing theorem in Statistics

The sampling distribution of the sample means of n independent, **identically distributed** random variables from a population with mean μ and variance σ^2 approaches a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size gets larger

This says that all my observations come from the same underlying population.

An amazing theorem in Statistics

The sampling distribution of the sample means of n independent, identically distributed random variables from **a population with mean μ and variance σ^2** approaches a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size gets larger

Notice that we have no other constraints on the population: the distribution of the population can be anything!

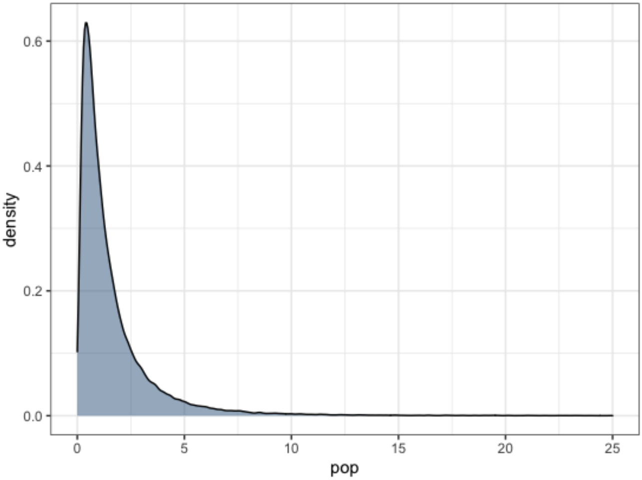
An amazing theorem in Statistics

The sampling distribution of the sample means of n independent, identically distributed random variables from a population with mean μ and variance σ^2 approaches a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ **as the sample size gets larger**

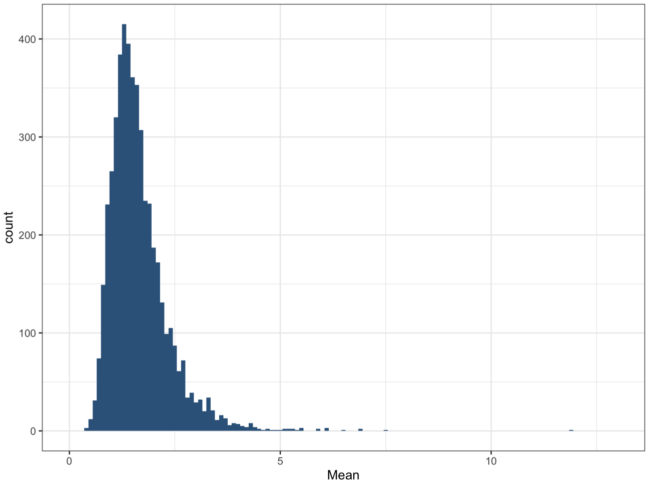
The text is very important. Let's see some examples.

Sample size is important for the Central limit Theorem

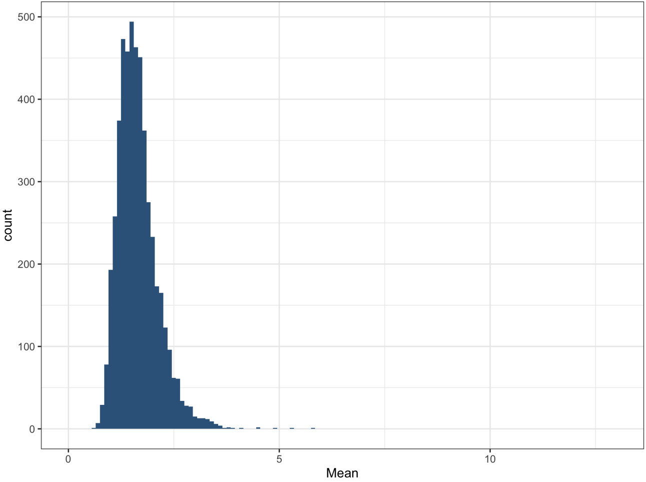
Suppose we're drawing samples from a population that looks like this, and we want to compute the mean.



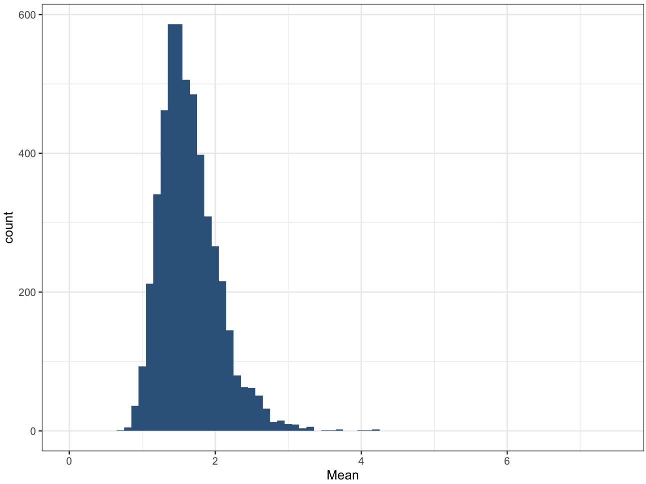
Sample size = 10



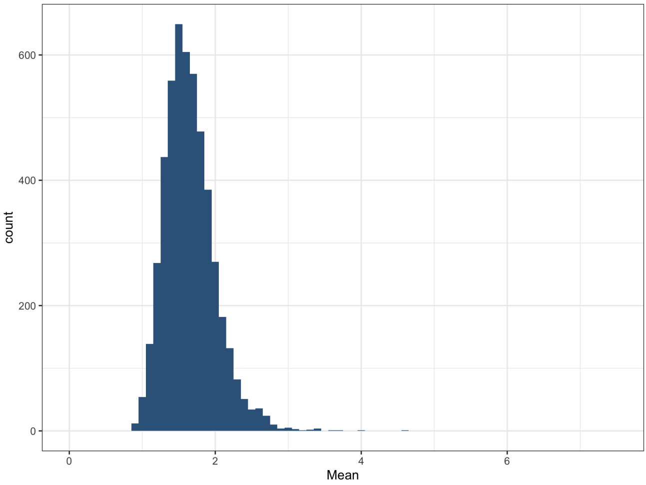
Sample size = 20



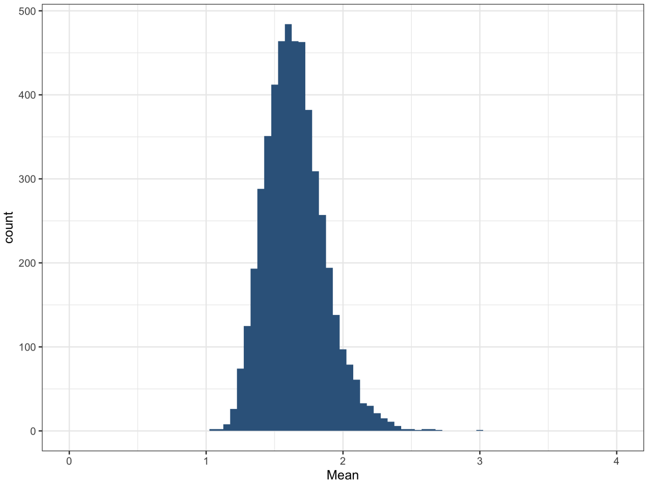
Sample size = 30



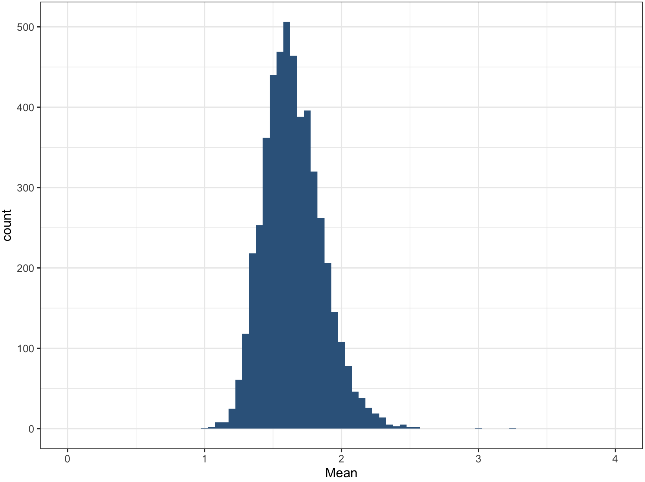
Sample size = 40



Sample size = 50



Sample size = 100



What do you notice?

- As the sample size increases, the distribution becomes more Normal.

What do you notice?

- As the sample size increases, the distribution becomes more Normal.
- The spread of the sampling distribution reduces.

Pop quiz

True or False: As we increase the sample size, the distribution of the sample becomes Normal.

Pop quiz

True or False: As we increase the sample size, the distribution of the sample becomes Normal.

False. The **sampling distribution** of the sample means approximates a Normal.

Outline

Key Assumption

Our sample size is large enough so that the central limit theorem holds, i.e, the sampling distribution is close to being a Normal distribution.

Key Assumption

Our sample size is large enough so that the central limit theorem holds, i.e, the sampling distribution is close to being a Normal distribution.

We can compute percentiles of the assumed distribution to get confidence bounds.

Percentiles of a Normal Distribution

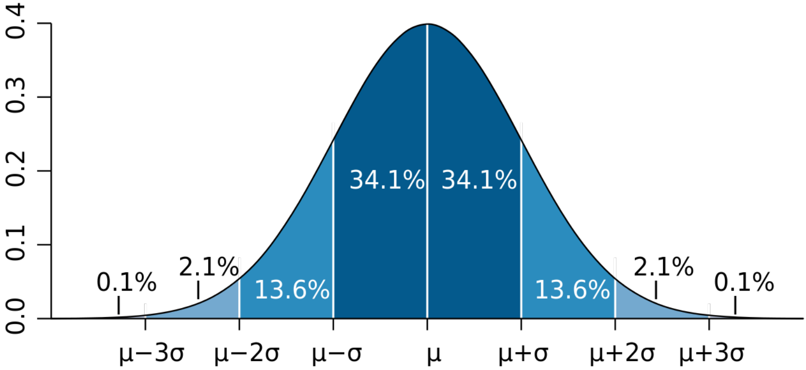
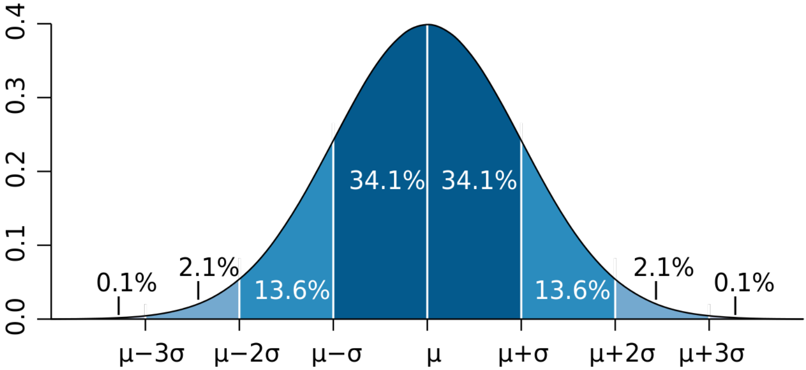


Image from Wikipedia: `https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Standard_deviation_diagram_micro.svg`

Percentiles of a Normal Distribution

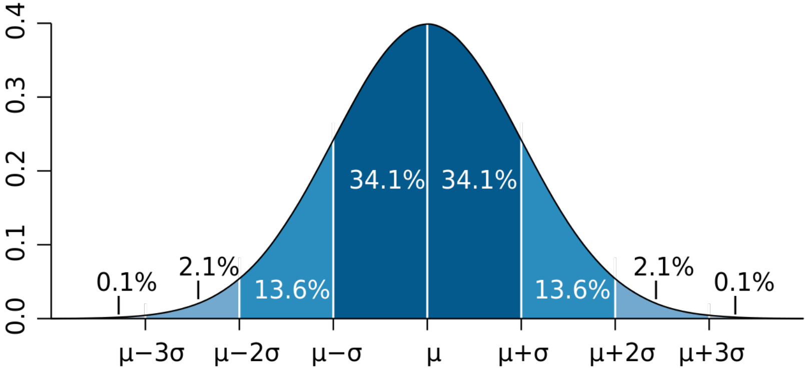


For a Normal distribution with mean μ and variance σ,

$$\Pr(X \geq \mu - \sigma \text{ and } X < \mu + \sigma) = 0.682$$

Image from Wikipedia: `https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Standard_deviation_diagram_micro.svg`

Percentiles of a Normal Distribution

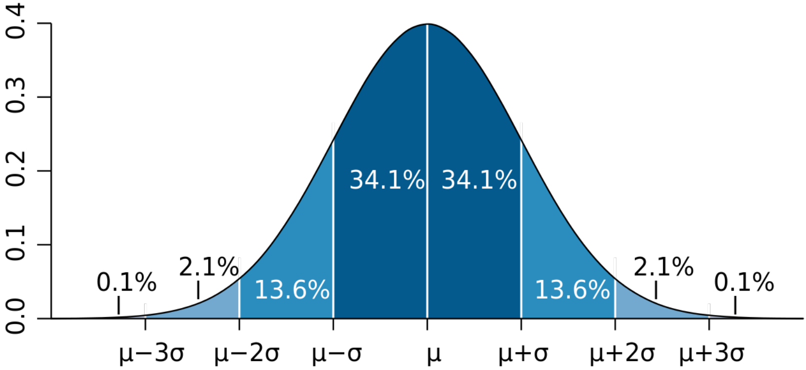


For a Normal distribution with mean μ and variance σ,

$$\Pr(X \geq \mu - 2\sigma \text{ and } X < \mu + 2\sigma) = 0.9558$$

Image from Wikipedia: `https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Standard_deviation_diagram_micro.svg`

Percentiles of a Normal Distribution



For a Normal distribution with mean μ and variance σ,

$$\Pr(X \geq \mu - 3\sigma \text{ and } X < \mu + 3\sigma) > 0.99$$

Image from Wikipedia: `https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Standard_deviation_diagram_micro.svg`

Computing confidence intervals assuming Normality

- Suppose we knew the variance σ^2 of the population.

Computing confidence intervals assuming Normality

- Suppose we knew the variance σ^2 of the population.
- Now, given a sample, we want to ask what the probability of this sample mean $\overline{X}$ lying far away from the mean of the sampling distribution μ is.

Computing confidence intervals assuming Normality

- Suppose we knew the variance σ^2 of the population.
- Now, given a sample, we want to ask what the probability of this sample mean $\bar{X}$ lying far away from the mean of the sampling distribution μ is.
- Let's say that we want to compute 95% confidence intervals. Thus, we want to compute the 2.5 and 97.5 percentiles for a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. (We'll show you how to do this for different percentile values tomorrow in precept)

Computing confidence intervals assuming Normality

- Suppose we knew the variance σ^2 of the population.
- Now, given a sample, we want to ask what the probability of this sample mean $\overline{X}$ lying far away from the mean of the sampling distribution μ is.
- Let's say that we want to compute 95% confidence intervals. Thus, we want to compute the 2.5 and 97.5 percentiles for a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. (We'll show you how to do this for different percentile values tomorrow in precept)
- The 2.5 percentile for this distribution is $\mu - 1.96\frac{\sigma}{\sqrt{n}}$.

Computing confidence intervals assuming Normality

- Suppose we knew the variance σ^2 of the population.
- Now, given a sample, we want to ask what the probability of this sample mean $\overline{X}$ lying far away from the mean of the sampling distribution μ is.
- Let's say that we want to compute 95% confidence intervals. Thus, we want to compute the 2.5 and 97.5 percentiles for a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. (We'll show you how to do this for different percentile values tomorrow in precept)
- The 2.5 percentile for this distribution is $\mu - 1.96\frac{\sigma}{\sqrt{n}}$.
- The 97.5 percentile for this distribution is $\mu + 1.96\frac{\sigma}{\sqrt{n}}$.

Computing confidence intervals assuming Normality

- Suppose we knew the variance σ^2 of the population.
- Now, given a sample, we want to ask what the probability of this sample mean $\overline{X}$ lying far away from the mean of the sampling distribution μ is.
- Let's say that we want to compute 95% confidence intervals. Thus, we want to compute the 2.5 and 97.5 percentiles for a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. (We'll show you how to do this for different percentile values tomorrow in precept)
- The 2.5 percentile for this distribution is $\mu - 1.96\frac{\sigma}{\sqrt{n}}$.
- The 97.5 percentile for this distribution is $\mu + 1.96\frac{\sigma}{\sqrt{n}}$.
- Thus, 95% of the sampling distribution lies between $[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}]$.

Computing confidence intervals assuming Normality

- Thus, 95% of the sampling distribution lies between $[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}]$.

Computing confidence intervals assuming Normality

- Thus, 95% of the sampling distribution lies between $[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}]$.
- We can write this out mathematically:

$$\Pr(\overline{X} \geq \mu - 1.96\frac{\sigma}{\sqrt{n}} \text{ and } \overline{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

with $\overline{X}$ representing the sample mean.

Computing confidence intervals assuming Normality

- Thus, 95% of the sampling distribution lies between $[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}]$.
- We can write this out mathematically:

$$\Pr(\overline{X} \geq \mu - 1.96\frac{\sigma}{\sqrt{n}} \text{ and } \overline{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

  with $\overline{X}$ representing the sample mean.
- Let's rearrange this statement:

$$\Pr(\overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \geq \mu \text{ and } \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu) = 0.95$$

Computing confidence intervals assuming Normality

- Thus, 95% of the sampling distribution lies between $[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}]$.
- We can write this out mathematically:

$$\Pr(\overline{X} \geq \mu - 1.96\frac{\sigma}{\sqrt{n}} \text{ and } \overline{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

  with $\overline{X}$ representing the sample mean.
- Let's rearrange this statement:

$$\Pr(\overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \geq \mu \text{ and } \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu) = 0.95$$

- Or,

$$\Pr(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Computing confidence intervals assuming Normality

- Thus, 95% of the sampling distribution lies between $[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}]$.
- We can write this out mathematically:

$$\Pr(\overline{X} \geq \mu - 1.96\frac{\sigma}{\sqrt{n}} \text{ and } \overline{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

  with $\overline{X}$ representing the sample mean.
- Let's rearrange this statement:

$$\Pr(\overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \geq \mu \text{ and } \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu) = 0.95$$

- Or,

$$\Pr(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Computing confidence intervals assuming Normality

- Thus, 95% of the sampling distribution lies between $[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}]$.
- We can write this out mathematically:

$$\Pr(\overline{X} \geq \mu - 1.96\frac{\sigma}{\sqrt{n}} \text{ and } \overline{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

  with $\overline{X}$ representing the sample mean.
- Let's rearrange this statement:

$$\Pr(\overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \geq \mu \text{ and } \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu) = 0.95$$

- Or,

$$\Pr(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Problem: We actually don't know σ (the standard deviation of the population).

Computing confidence intervals assuming Normality

$$\Pr(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

- σ is the **population** standard deviation.

Computing confidence intervals assuming Normality

$$\Pr(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

- σ is the **population** standard deviation.
- We can approximate σ using the **sample** standard deviation instead.

Computing confidence intervals assuming Normality

$$\Pr(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

- σ is the **population** standard deviation.
- We can approximate σ using the **sample** standard deviation instead.
- What is the formula for the sample variance?

Computing confidence intervals assuming Normality

$$\Pr(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

- σ is the **population** standard deviation.
- We can approximate σ using the **sample** standard deviation instead.
- What is the formula for the sample variance?

Computing confidence intervals assuming Normality

$$\Pr(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

- σ is the **population** standard deviation.
- We can approximate σ using the **sample** standard deviation instead.
- What is the formula for the sample variance?

$$\hat{\sigma^2} = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$$

- Thus, we get a 95% confidence interval of :

$$[\overline{X} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}, \overline{X} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}]$$

What does this look like in practice?

Let's say we're calculating Biden's popularity rating using a sample of size 100.

What does this look like in practice?

Let's say we're calculating Biden's popularity rating using a sample of size 100.
Within the sample, we find that the mean is 0.38 and the sample standard deviation is 0.4878.

What does this look like in practice?

Let's say we're calculating Biden's popularity rating using a sample of size 100.

Within the sample, we find that the mean is 0.38 and the sample standard deviation is 0.4878.

Let's say we want to estimate 95% confidence intervals. How would we do that?

What does this look like in practice?

Let's say we're calculating Biden's popularity rating using a sample of size 100.
Within the sample, we find that the mean is 0.38 and the sample standard deviation is 0.4878.
Let's say we want to estimate 95% confidence intervals. How would we do that?

- Let's assume that the sample size is large enough for the sampling distribution to approximate a Normal.

What does this look like in practice?

Let's say we're calculating Biden's popularity rating using a sample of size 100.
Within the sample, we find that the mean is 0.38 and the sample standard deviation is 0.4878.
Let's say we want to estimate 95% confidence intervals. How would we do that?

- Let's assume that the sample size is large enough for the sampling distribution to approximate a Normal.
- What's the standard deviation of the sampling distribution?

What does this look like in practice?

Let's say we're calculating Biden's popularity rating using a sample of size 100.
Within the sample, we find that the mean is 0.38 and the sample standard deviation is 0.4878.
Let's say we want to estimate 95% confidence intervals. How would we do that?

- Let's assume that the sample size is large enough for the sampling distribution to approximate a Normal.
- What's the standard deviation of the sampling distribution?

What does this look like in practice?

Let's say we're calculating Biden's popularity rating using a sample of size 100.

Within the sample, we find that the mean is 0.38 and the sample standard deviation is 0.4878.

Let's say we want to estimate 95% confidence intervals. How would we do that?

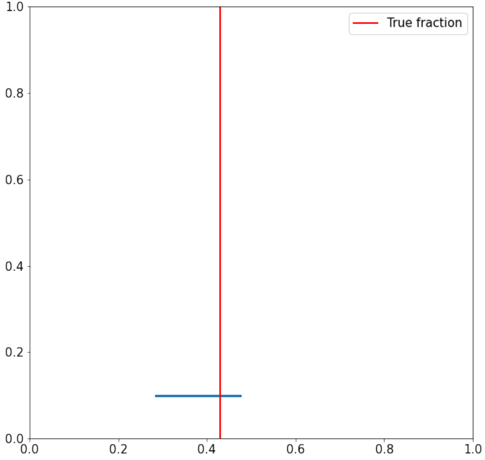
- Let's assume that the sample size is large enough for the sampling distribution to approximate a Normal.
- What's the standard deviation of the sampling distribution? We don't know. But we can estimate it using $\frac{\hat{\sigma}}{\sqrt{100}} = 0.04878$.
- Thus, we can estimate 95% confidence intervals as $[0.38 - 1.96 \times 0.04878, 0.38 + 1.96 \times 0.04878] = [0.2844, 0.4756]$.

Frequentist interpretation of this confidence interval

If we were to repeat this sampling many many times, the mean would lie within the confidence interval 95% of the time.

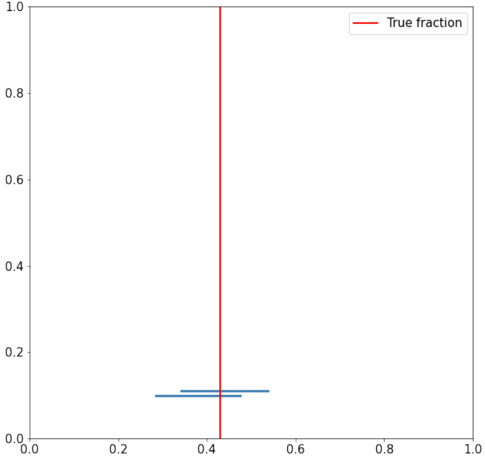
Frequentist interpretation of this confidence interval

If we were to repeat this sampling many many times, the mean would lie within the confidence interval 95% of the time.



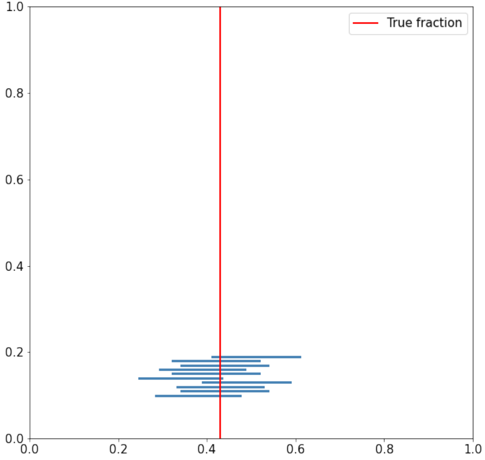
Frequentist interpretation of this confidence interval

If we were to repeat this sampling many many times, the mean would lie within the confidence interval 95% of the time.



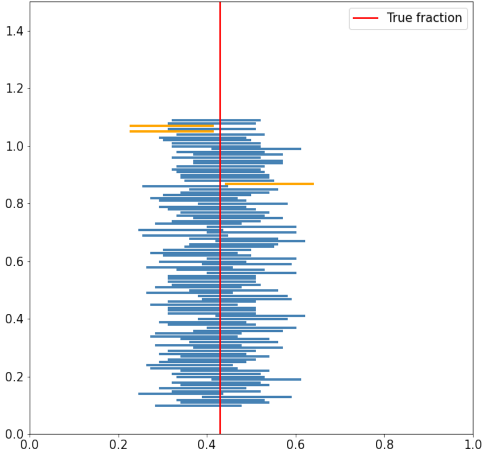
Frequentist interpretation of this confidence interval

If we were to repeat this sampling many many times, the mean would lie within the confidence interval 95% of the time.



Frequentist interpretation of this confidence interval

If we were to repeat this sampling many many times, the mean would lie within the confidence interval 95% of the time.



Outline

Another way to think about uncertainity

- Suppose Biden's true approval rating was 10%

Another way to think about uncertainity

- Suppose Biden's true approval rating was 10%
- And our sample was actually drawn from that distribution.

Another way to think about uncertainity

- Suppose Biden's true approval rating was 10%
- And our sample was actually drawn from that distribution.
- What is the probability that we see sample values as extreme as we do?

Another way to think about uncertainity

- Suppose Biden's true approval rating was 10%
- And our sample was actually drawn from that distribution.
- What is the probability that we see sample values as extreme as we do?

Another way to think about uncertainity

- Suppose Biden's true approval rating was 10%
- And our sample was actually drawn from that distribution.
- What is the probability that we see sample values as extreme as we do?

This probability is called the p-value.

Another example

- Suppose we want to measure Biden's approval rating among Democrats and Republicans (to see if there is a difference)

$$p\text{-value} = \Pr(\text{Sample is this extreme} \mid \text{population distribution of Democrats and Republicans are the same})$$

Another example

- Suppose we want to measure Biden's approval rating among Democrats and Republicans (to see if there is a difference)
- We can pick a sample, and calculate the approval rating among Democrats and the approval rating among Republicans.

$$p\text{-value} = \Pr(\text{Sample is this extreme} \mid \text{population distribution of Democrats and Republicans are the same})$$

Another example

- Suppose we want to measure Biden's approval rating among Democrats and Republicans (to see if there is a difference)
- We can pick a sample, and calculate the approval rating among Democrats and the approval rating among Republicans.
- Suppose the distribution of the ratings among Democrats and Republicans was exactly the same. The p-value measures the probability that we see a sample as extreme as our current sample assuming that the distribution of the ratings is exactly the same.

$$p\text{-value} = \Pr(\text{Sample is this extreme} \mid \text{population distribution of Democrats and Republicans are the same})$$

Another example

- Suppose we want to measure Biden's approval rating among Democrats and Republicans (to see if there is a difference)
- We can pick a sample, and calculate the approval rating among Democrats and the approval rating among Republicans.
- Suppose the distribution of the ratings among Democrats and Republicans was exactly the same. The p-value measures the probability that we see a sample as extreme as our current sample assuming that the distribution of the ratings is exactly the same.
- Using our conditional probability notation:

$$p\text{-value} = \Pr(\text{Sample is this extreme} \mid \text{population distribution of Democrats and Republicans are the same})$$

Outline

Statistical significance

- Very often, in research articles, you'll see the term **statistically significant ($P < 0.05$)**.

Statistical significance

- Very often, in research articles, you'll see the term **statistically significant ($P < 0.05$)**.
- By that, what they mean is the probability of their sample being drawn from the different hypothetical population is lower than 0.05.

Statistical significance

- Very often, in research articles, you'll see the term **statistically significant ($P < 0.05$)**.
- By that, what they mean is the probability of their sample being drawn from the different hypothetical population is lower than 0.05.
- Or a claim that these results are statistically significant because the confidence intervals exclude 0. (for example, if we want to measure the difference in approval ratings for Biden among Democrats and Republicans.)

Statistical significance

- Very often, in research articles, you'll see the term **statistically significant ($P < 0.05$)**.
- By that, what they mean is the probability of their sample being drawn from the different hypothetical population is lower than 0.05.
- Or a claim that these results are statistically significant because the confidence intervals exclude 0. (for example, if we want to measure the difference in approval ratings for Biden among Democrats and Republicans.)
- Much like picking 95% confidence intervals, the threshold for significance is arbitrary. Why $P < 0.05$ or $P < 0.01$? Why not $P < 0.04$?

Statistical significance

- Very often, in research articles, you'll see the term **statistically significant ($P < 0.05$)**.
- By that, what they mean is the probability of their sample being drawn from the different hypothetical population is lower than 0.05.
- Or a claim that these results are statistically significant because the confidence intervals exclude 0. (for example, if we want to measure the difference in approval ratings for Biden among Democrats and Republicans.)
- Much like picking 95% confidence intervals, the threshold for significance is arbitrary. Why $P < 0.05$ or $P < 0.01$? Why not $P < 0.04$?
- Notion of a simple yes-or-no answer for statistical significance doesn't really make sense : design choices like the size of the sample, how the data is collected, how the analysis is done is often a lot more important.

Statistical significance versus significance

- **Statistical significance** and **significance** is often conflated.

Statistical significance versus significance

- **Statistical significance** and **significance** is often conflated.
- In reality, things that are statistically significant might not be significant.

Statistical significance versus significance

- Suppose a drug is measured to lower blood pressure of participatants by 0.1, with $P = 0.01$.

Statistical significance versus significance

- Suppose a drug is measured to lower blood pressure of participatants by 0.1, with $P = 0.01$.
- This means that if our data was drawn from a population where the effect of this drug is negligible, the probability of getting our measurements as extreme as ours is 0.01.

Statistical significance versus significance

- Suppose a drug is measured to lower blood pressure of participatants by 0.1, with $P = 0.01$.
- This means that if our data was drawn from a population where the effect of this drug is negligible, the probability of getting our measurements as extreme as ours is 0.01.
- Is this statistically significant?

Statistical significance versus significance

- Suppose a drug is measured to lower blood pressure of participatants by 0.1, with $P = 0.01$.
- This means that if our data was drawn from a population where the effect of this drug is negligible, the probability of getting our measurements as extreme as ours is 0.01.
- Is this statistically significant?

Statistical significance versus significance

- Suppose a drug is measured to lower blood pressure of participatants by 0.1, with $P = 0.01$.
- This means that if our data was drawn from a population where the effect of this drug is negligible, the probability of getting our measurements as extreme as ours is 0.01.
- Is this statistically significant? **Yes** – the p value is very small.
- Is this actually significant?

Statistical significance versus significance

- Suppose a drug is measured to lower blood pressure of participatants by 0.1, with $P = 0.01$.
- This means that if our data was drawn from a population where the effect of this drug is negligible, the probability of getting our measurements as extreme as ours is 0.01.
- Is this statistically significant? **Yes** – the p value is very small.
- Is this actually significant?

Statistical significance versus significance

- Suppose a drug is measured to lower blood pressure of participatants by 0.1, with $P = 0.01$.
- This means that if our data was drawn from a population where the effect of this drug is negligible, the probability of getting our measurements as extreme as ours is 0.01.
- Is this statistically significant? **Yes** – the p value is very small.
- Is this actually significant? **Probably not!**

Statistical significance versus significance

- Researchers measured the effect of quiet time between 1 and 3 pm at a post-partum ward.

Statistical significance versus significance

- Researchers measured the effect of quiet time between 1 and 3 pm at a post-partum ward.
- They found that this increased the women's exclusive breastfeeding rates by 14%.

Statistical significance versus significance

- Researchers measured the effect of quiet time between 1 and 3 pm at a post-partum ward.
- They found that this increased the women's exclusive breastfeeding rates by 14%.
- The P-value here was 0.39. Is this statistically significant?

Statistical significance versus significance

- Researchers measured the effect of quiet time between 1 and 3 pm at a post-partum ward.
- They found that this increased the women's exclusive breastfeeding rates by 14%.
- The P-value here was 0.39. Is this statistically significant?

Statistical significance versus significance

- Researchers measured the effect of quiet time between 1 and 3 pm at a post-partum ward.
- They found that this increased the women's exclusive breastfeeding rates by 14%.
- The P-value here was 0.39. Is this statistically significant? No.
- But, this is a low-cost, easy to implement solution that benefited some mothers.

Statistical significance versus significance

- Researchers measured the effect of quiet time between 1 and 3 pm at a post-partum ward.
- They found that this increased the women's exclusive breastfeeding rates by 14%.
- The P-value here was 0.39. Is this statistically significant? No.
- But, this is a low-cost, easy to implement solution that benefited some mothers.
- And thus, might still be useful to implement (Polit and Beck (2012))

Where we are and where we're going

Before:

- Computing estimates on a sample.

Where we are and where we're going

Before:

- Computing estimates on a sample.

This week: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution

Where we are and where we're going

Before:

- Computing estimates on a sample.

This week: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.

Where we are and where we're going

Before:

- Computing estimates on a sample.

This week: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping

Where we are and where we're going

Before:

- Computing estimates on a sample.

This week: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping
- Central limit theorem

Where we are and where we're going

Before:

- Computing estimates on a sample.

This week: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping
- Central limit theorem
- Getting guarantees about confidence intervals (with more assumptions)

Where we are and where we're going

Before:

- Computing estimates on a sample.

This week: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping
- Central limit theorem
- Getting guarantees about confidence intervals (with more assumptions)

Where we are and where we're going

Before:

- Computing estimates on a sample.

This week: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using bootstrapping.
- Confidence intervals using bootstrapping
- Central limit theorem
- Getting guarantees about confidence intervals (with more assumptions)

Next week:

- Causality