Inference and hypothesis testing

EDS 222

Tamma Carleton
Fall 2023

Announcements/check-in

- Increasing divergence between course material and IMS textbook (including today's lecture)

Announcements/check-in

- Increasing divergence between course material and IMS textbook (including today's lecture)
- Change in OH this week (*today* 3:15-4:15pm, Pine Room)

Announcements/check-in

- Increasing divergence between course material and IMS textbook (including today's lecture)
- Change in OH this week (*today* 3:15-4:15pm, Pine Room)
- Assignment 4 posted this week, likely due 12/01 but will depend on...

Announcements/check-in

- Increasing divergence between course material and IMS textbook (including today's lecture)
- Change in OH this week (*today* 3:15-4:15pm, Pine Room)
- Assignment 4 posted this week, likely due 12/01 but will depend on...
- ...the next few weeks. We might need 2.5 weeks for inference + time series. We will focus on going slow enough to fit it all in (we have slack time built in)

Today

Remember week 1? ...why are we we here?

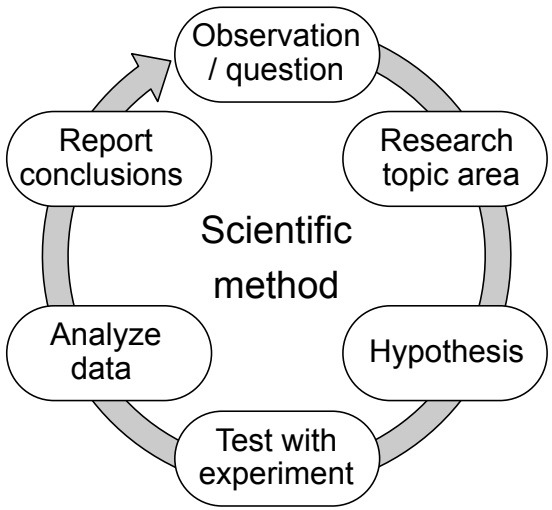
Today

Remember week 1? ...why are we we here?

Statistics:

The science of **collecting**, **manipulating**, and **analyzing** empirical data

Statistics enables us to use environmental data to follow the **scientific method**



Today

Thinking about uncertainty

Sampling distributions

Today

Thinking about uncertainty

Sampling distributions

Hypothesis testing: conceptual foundations

Null hypotheses, alternative hypotheses, rejecting the null

Today

Thinking about uncertainty

Sampling distributions

Hypothesis testing: conceptual foundations

Null hypotheses, alternative hypotheses, rejecting the null

Hypothesis testing: in practice

The Central Limit Theorem, standard errors, Z-scores, p-values

Today

Thinking about uncertainty

Sampling distributions

Hypothesis testing: conceptual foundations

Null hypotheses, alternative hypotheses, rejecting the null

Hypothesis testing: in practice

The Central Limit Theorem, standard errors, Z-scores, p-values

Confidence

Confidence intervals

Thinking about uncertainty

Why does uncertainty matter?

All our sample statistics (e.g., sample means, regression parameters) are uncertain

Why does uncertainty matter?

All our sample statistics (e.g., sample means, regression parameters) are uncertain

- We have a *randomly drawn sample* and are trying to learn about the population from our sample

Why does uncertainty matter?

All our sample statistics (e.g., sample means, regression parameters) are uncertain

- We have a *randomly drawn sample* and are trying to learn about the population from our sample
- But our sample statistics would have been different had we randomly drawn a different set of observations!

Why does uncertainty matter?

All our sample statistics (e.g., sample means, regression parameters) are uncertain

- We have a *randomly drawn sample* and are trying to learn about the population from our sample
- But our sample statistics would have been different had we randomly drawn a different set of observations!
- This is **natural variability** and it means that all our sample statistics are uncertain estimates of population parameters, even if they are unbiased (e.g., no convenience sampling, no systematic non-response, etc.)

Why does uncertainty matter?

Key question: Is our estimate indicating anything more than sampling variability or "noise"?

- This is the question **statistical inference** and **hypothesis testing** are trying to answer

Why does uncertainty matter?

Example: Gender wage gap

- We collect data on annual earnings and sex for 100 Bren alumni. We are interested in whether the population of all Bren alumni exhibit a gender wage gap.

Why does uncertainty matter?

Example: Gender wage gap

- We collect data on annual earnings and sex for 100 Bren alumni. We are interested in whether the population of all Bren alumni exhibit a gender wage gap.
- We see a mean difference between men and women in our 100-observation sample of $4,500 per year, but a wide range of earnings across both men and women

Why does uncertainty matter?

Example: Gender wage gap

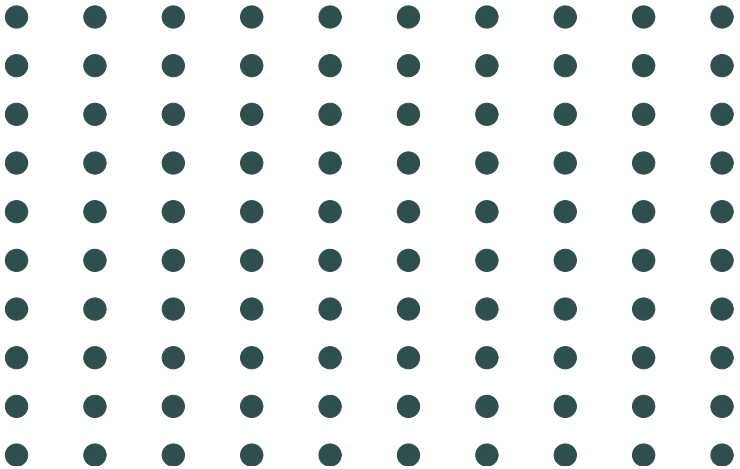
- We collect data on annual earnings and sex for 100 Bren alumni. We are interested in whether the population of all Bren alumni exhibit a gender wage gap.
- We see a mean difference between men and women in our 100-observation sample of $4,500 per year, but a wide range of earnings across both men and women
- Does this mean there is a gender wage gap, or did we just *happen* to get a few high-earning men and a few low-earning women in this group?

Why does uncertainty matter?

Example: Gender wage gap

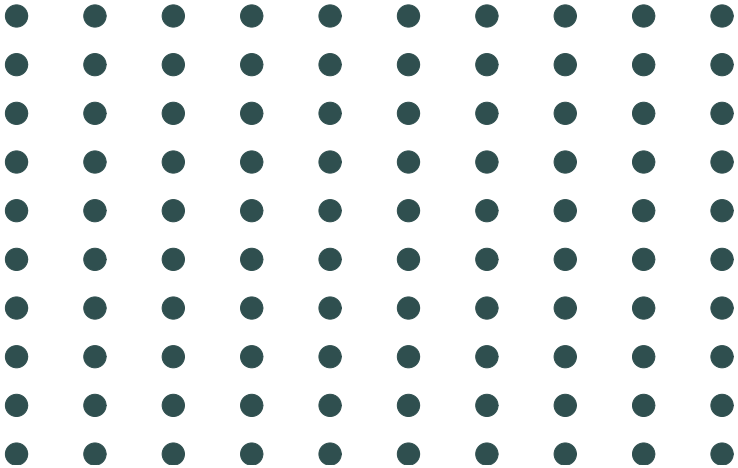
- We collect data on annual earnings and sex for 100 Bren alumni. We are interested in whether the population of all Bren alumni exhibit a gender wage gap.
- We see a mean difference between men and women in our 100-observation sample of $4,500 per year, but a wide range of earnings across both men and women
- Does this mean there is a gender wage gap, or did we just *happen* to get a few high-earning men and a few low-earning women in this group?
- If we collected another independent sample of 100, would the gap be the same?

Population *vs.* sample

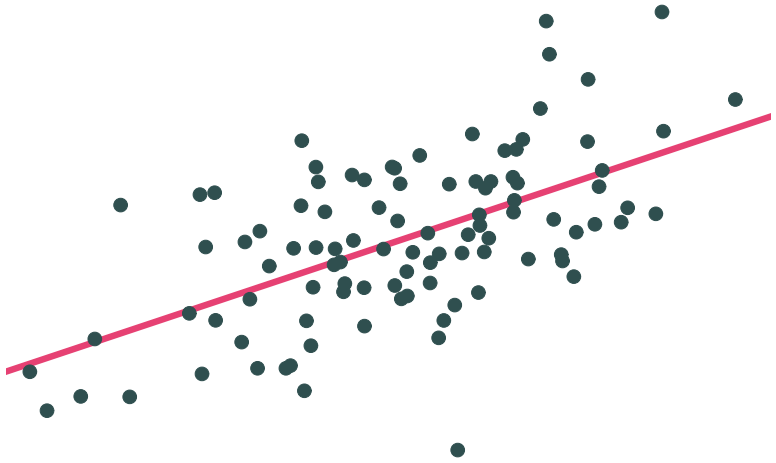


Population

Population *vs.* sample



Population

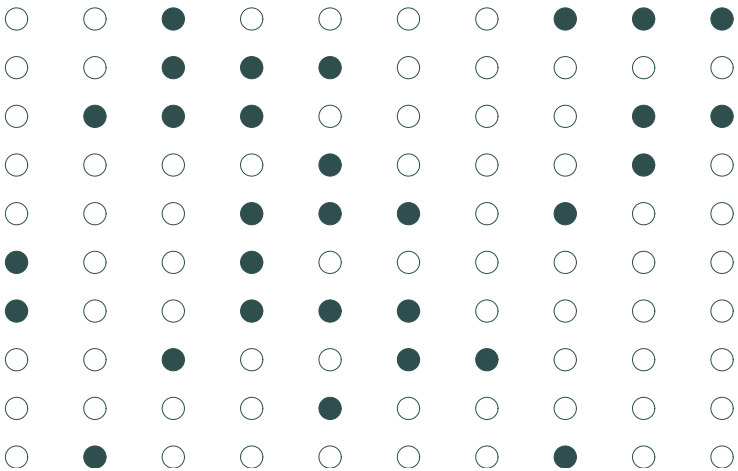


Population relationship

$$y_i = 2.53 + 0.57 x_i + u_i$$

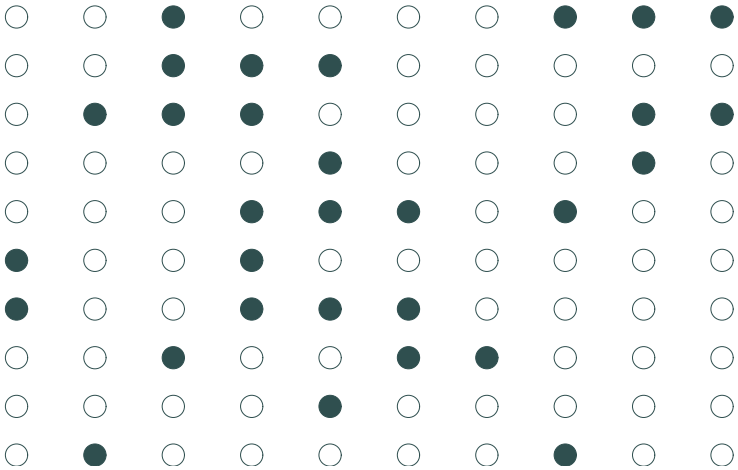
$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Population *vs.* sample

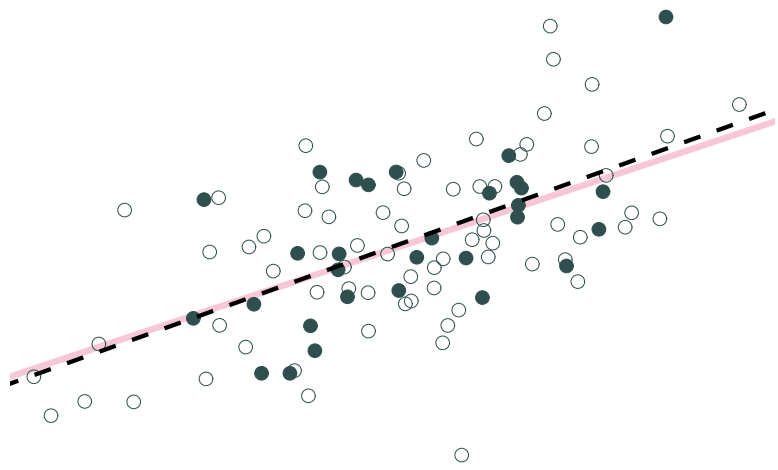


Sample 1: 30 random individuals

Population *vs.* sample



Sample 1: 30 random individuals



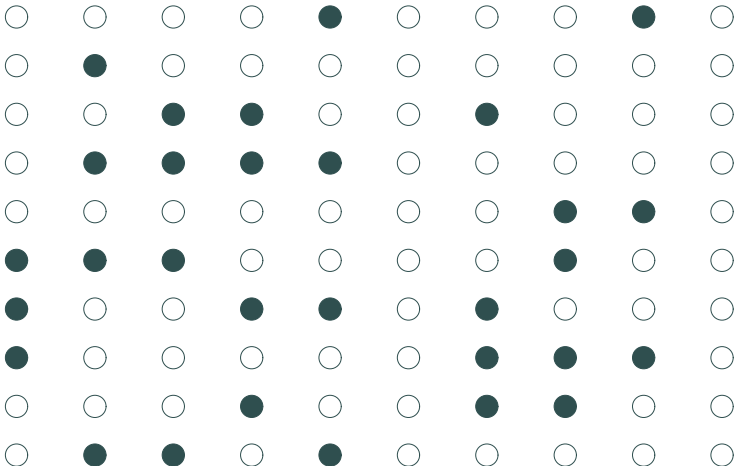
Population relationship

$$y_i = 2.53 + 0.57x_i + u_i$$

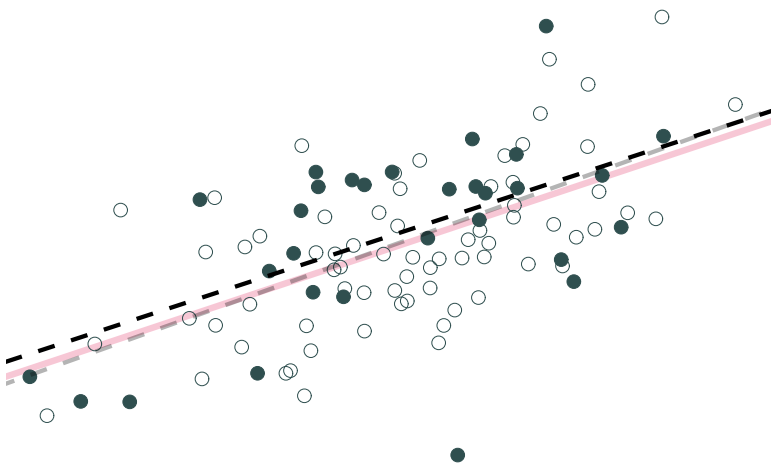
Sample relationship

$$\hat{y}_i = 2.36 + 0.61x_i$$

Population *vs.* sample



Sample 2: 30 random individuals



Population relationship

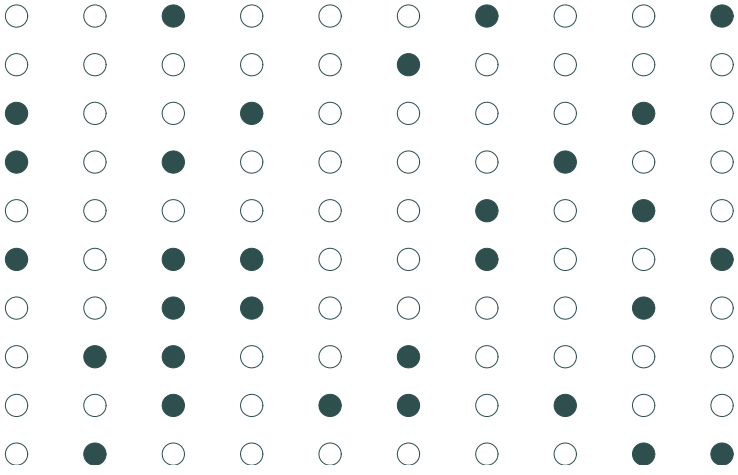
$$y_i = 2.53 + 0.57x_i + u_i$$

Sample relationship

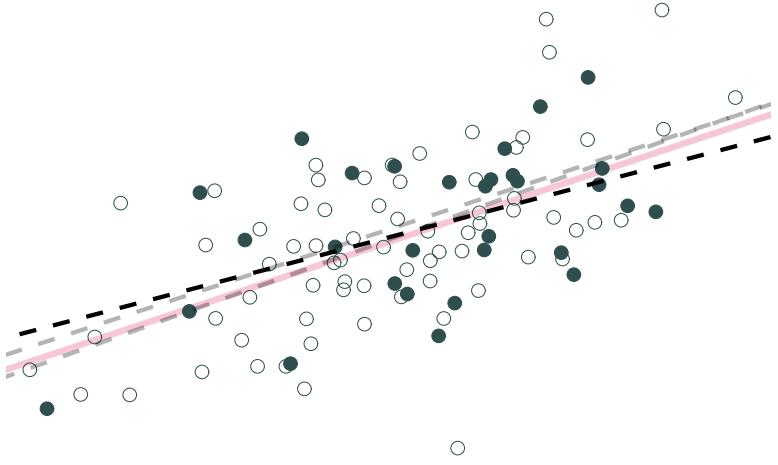
$$\hat{y}_i = 2.79 + 0.56x_i$$

Population *vs.* sample

count: false



Sample 3: 30 random individuals



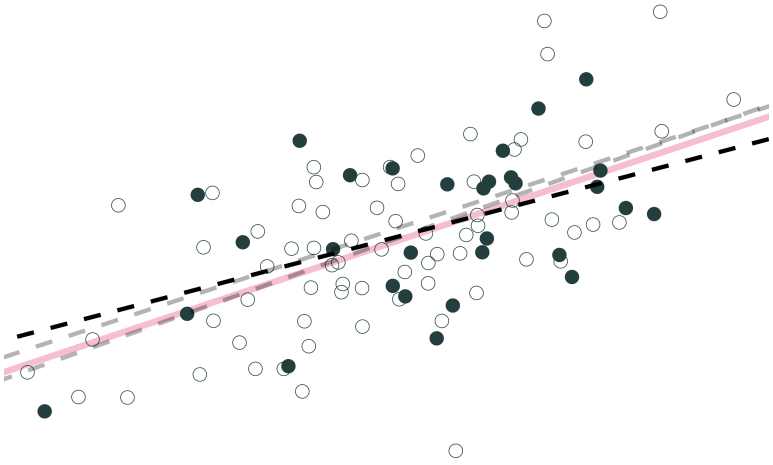
Population relationship

$$y_i = 2.53 + 0.57x_i + u_i$$

Sample relationship

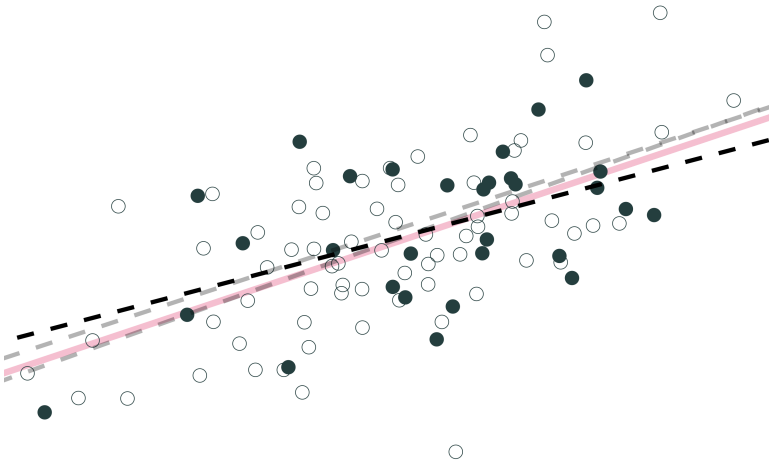
$$\hat{y}_i = 3.21 + 0.45x_i$$

Population *vs.* sample



- On **average**, our regression lines match the population line very nicely.
- However, **individual lines** (samples) can really miss the mark.
- Differences between individual samples and the population lead to **uncertainty** for the statistican

Population *vs.* sample



- On **average**, our regression lines match the population line very nicely.
- However, **individual lines** (samples) can really miss the mark.
- Differences between individual samples and the population lead to **uncertainty** for the statistican

> Keeping track of uncertainty allows us to test hypotheses about the population using just our sample

Hypothesis testing: conceptual foundations

Hypothesis testing

A **hypothesis test** is a statistical method used to evaluate competing claims about population parameters *based on a sample of data*

Hypothesis testing

A **hypothesis test** is a statistical method used to evaluate competing claims about population parameters *based on a sample of data*

- H_0: **Null hypothesis** A default hypothesis that the measured quantity (e.g., sample mean, difference in means, regression parameters) is zero. In other words: whatever I recovered in my sample is due to random chance.

Hypothesis testing

A **hypothesis test** is a statistical method used to evaluate competing claims about population parameters *based on a sample of data*

- H_0: **Null hypothesis** A default hypothesis that the measured quantity (e.g., sample mean, difference in means, regression parameters) is zero. In other words: whatever I recovered in my sample is due to random chance.

- H_A: **Alternative hypothesis** A hypothesis that the measured quantity is nonzero. In other words: whatever I recovered in my sample is due to true population differences or effects.

Hypothesis testing: Example

Example: Are ducks moving north?

There are reports that Midwestern duck populations are moving northwards in response to climate change.

Hypothesis testing: Example

Example: Are ducks moving north?

There are reports that Midwestern duck populations are moving northwards in response to climate change.

You have a random sample of 100 tagged ducks in Minnesota from 2010 and 2023.

- Mean 2010 latitude: 44.27 degrees N
- Mean 2023 latitude: 46.12 degrees N
- Standard deviation 2023 latitude: 0.92 degrees

Hypothesis testing: Example

Example: Are ducks moving north?

There are reports that Midwestern duck populations are moving northwards in response to climate change.

You have a random sample of 100 tagged ducks in Minnesota from 2010 and 2023.

- Mean 2010 latitude: 44.27 degrees N
- Mean 2023 latitude: 46.12 degrees N
- Standard deviation 2023 latitude: 0.92 degrees

- H_0: The average latitude was the same in the two years. That is, $\mu_{2023} - \mu_{2010} = 0$

- H_A: The average latitude was not the same in the two years. That is, $\mu_{2023} - \mu_{2010} \neq 0$

Hypothesis testing: Example

Example: Are ducks moving north?

We call the calculated statistic of interest the **point estimate**

Hypothesis testing: Example

Example: Are ducks moving north?

We call the calculated statistic of interest the **point estimate**

Here, the difference in mean latitude between the 2023 sample and the 2010 sample is:

- 46.12 - 44.27 = 1.85

Hypothesis testing: Example

Example: Are ducks moving north?

We call the calculated statistic of interest the **point estimate**

Here, the difference in mean latitude between the 2023 sample and the 2010 sample is:

- 46.12 - 44.27 = 1.85

> Hypothesis test asks if this point estimate is actually different from zero once we account for sampling variability

Hypothesis testing: Rejecting the null

How do we choose between H_0 and H_A?

Hypothesis testing: Rejecting the null

How do we choose between H_0 and H_A?

If the data conflict so much with H_0 that the null cannot be deemed reasonable we **reject the null**

Hypothesis testing: Rejecting the null

How do we choose between H_0 and H_A?

If the data conflict so much with H_0 that the null cannot be deemed reasonable we **reject the null**

For example:

Hypothesis testing: Rejecting the null

How do we choose between H_0 and H_A?

If the data conflict so much with H_0 that the null cannot be deemed reasonable we **reject the null**

For example:

- The distribution of 2023 duck latitudes are so far from the 2010 distribution that we can reject the means are the same

Hypothesis testing: Rejecting the null

How do we choose between H_0 and H_A?

If the data conflict so much with H_0 that the null cannot be deemed reasonable we **reject the null**

For example:

- The distribution of 2023 duck latitudes are so far from the 2010 distribution that we can reject the means are the same
- Wages across a random sample of 100 Bren alumni are so strongly differentiated across men and women that we reject a gender wage gap of zero

Hypothesis testing: Rejecting the null

How do we choose between H_0 and H_A?

If the data conflict so much with H_0 that the null cannot be deemed reasonable we **reject the null**

For example:

- The distribution of 2023 duck latitudes are so far from the 2010 distribution that we can reject the means are the same
- Wages across a random sample of 100 Bren alumni are so strongly differentiated across men and women that we reject a gender wage gap of zero

Rejecting the null involves both a **point estimate** and a measure of **uncertainty** or spread in your data

Hypothesis testing: in practice

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses

2. Collect data and compute the **point estimate of the statistic**

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**
4. Given this variability, **quantify the probability that your sample statistic differs from the null** by the observed amount, if the null were true

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**
4. Given this variability, **quantify the probability that your sample statistic differs from the null** by the observed amount, if the null were true
5. Based on #4, either **reject** or **fail to reject** the null

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses

2. Collect data and compute the **point estimate of the statistic**

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**

We already know all about these two steps.

- Null and alternative hypotheses will depend entirely on the statistical question of interest.
- Data collection and point estimates (e.g., means, regression parameters, variances, etc.) we have studied at length in this class

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**

Ack! What is this? Something about how much noise there is in a sample statistic in any given sample...

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**

Ack! What is this? Something about how much noise there is in a sample statistic in any given sample...

Let's turn to some definitions.

Sampling distribution

A **sampling distribution** is the distribution of all possible values of a sample statistic from samples of a given size from a given population.

Sampling distribution

A **sampling distribution** is the distribution of all possible values of a sample statistic from samples of a given size from a given population.

- The sampling distribution describes how sample statistics (e.g., mean, regression parameters) vary from one sample (or study) to the next

Sampling distribution

A **sampling distribution** is the distribution of all possible values of a sample statistic from samples of a given size from a given population.

- The sampling distribution describes how sample statistics (e.g., mean, regression parameters) vary from one sample (or study) to the next
- This is *not* the same as the **data distribution**!
    - Distribution of your data = distribution within one sample (e.g., gives you *one* sample mean)
    - Sampling distribution = distribution across samples (e.g., gives you *many* sample means)

Sampling distribution

For example, recall our regression above, where the population model is:

$$y_i = \mathbf{2.53} + \mathbf{0.57}x_i + u_i$$

Sampling distribution

For example, recall our regression above, where the population model is:

$$y_i = 2.53 + 0.57x_i + u_i$$

- A regression using one sample gives us *one* set of coefficients, called the **point estimates**. For example, $\hat{\beta}_0 = 2.36$ and $\hat{\beta}_1 = 0.61$

Sampling distribution

For example, recall our regression above, where the population model is:

$$y_i = 2.53 + 0.57x_i + u_i$$

- A regression using one sample gives us *one* set of coefficients, called the **point estimates**. For example, $\hat{\beta}_0 = 2.36$ and $\hat{\beta}_1 = 0.61$
- If we could collect 1000 samples and run that regression 1,000 times, we would recover the **sampling distribution** for each coefficient

Sampling distribution

Why do we need a sampling distribution?

Sampling distribution

Why do we need a sampling distribution?

Tells us how certain we are that our sample statistics are informative of a population parameter

Sampling distribution

Why do we need a sampling distribution?

Tells us how certain we are that our sample statistics are informative of a population parameter

- Wide sampling distribution = high uncertainty = hard to prove anything about the population

Sampling distribution

Why do we need a sampling distribution?

Tells us how certain we are that our sample statistics are informative of a population parameter

- Wide sampling distribution = high uncertainty = hard to prove anything about the population

But how do we obtain one of these?

You only have one sample of data! Where does the sampling distribution come from?

We derive the sampling distribution from applying the Central Limit Theorem

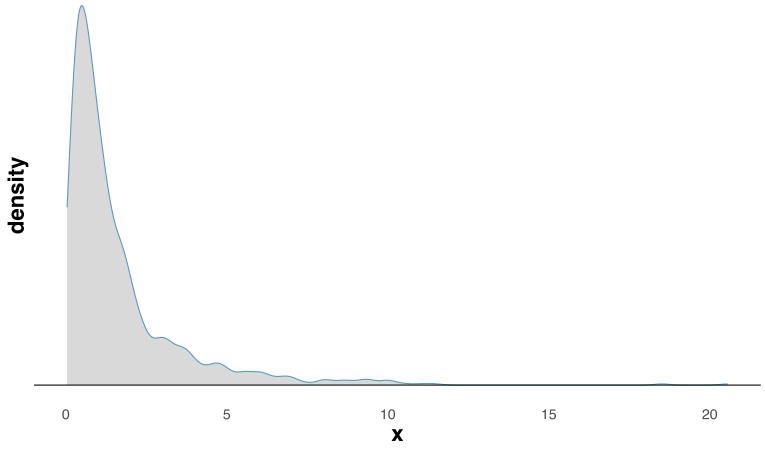
Central Limit Theorem

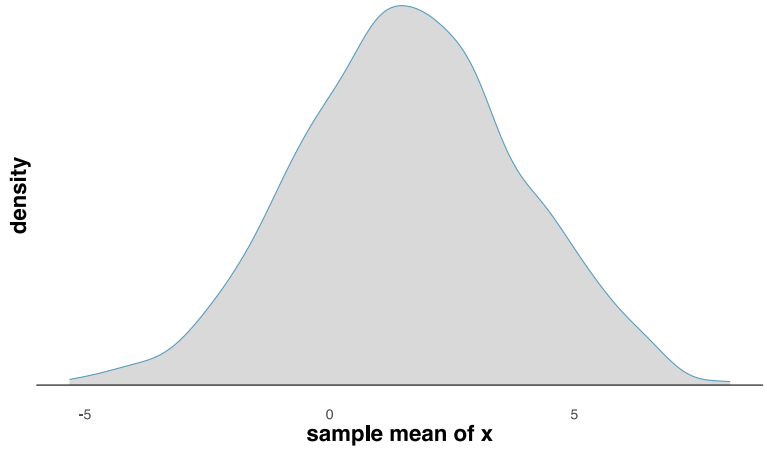
The **Central Limit Theorem (CLT)** establishes that the sampling distribution of a population parameter is **normal** if the sample size n is sufficiently large and observations are drawn randomly and independently.

Central Limit Theorem

The **Central Limit Theorem (CLT)** establishes that the sampling distribution of a population parameter is **normal** if the sample size n is sufficiently large and observations are drawn randomly and independently.

This is true *even if* the underlying data are not normally distributed!





Central Limit Theorem

The **Central Limit Theorem (CLT)** tells us our sampling distribution is normal, but only if n is big enough.

Central Limit Theorem

The **Central Limit Theorem (CLT)** tells us our sampling distribution is normal, but only if n is big enough.

Question: How big does our sample need to be?

Central Limit Theorem

The **Central Limit Theorem (CLT)** tells us our sampling distribution is normal, but only if *n* is big enough.

Question: How big does our sample need to be?

Answer: Rule of thumb is $n \geq 30$

But this is not a hard and fast rule! Be cautious about hypothesis testing and inference with small sample sizes.

Standard errors

So we know the sample statistic is drawn from a normal distribution...

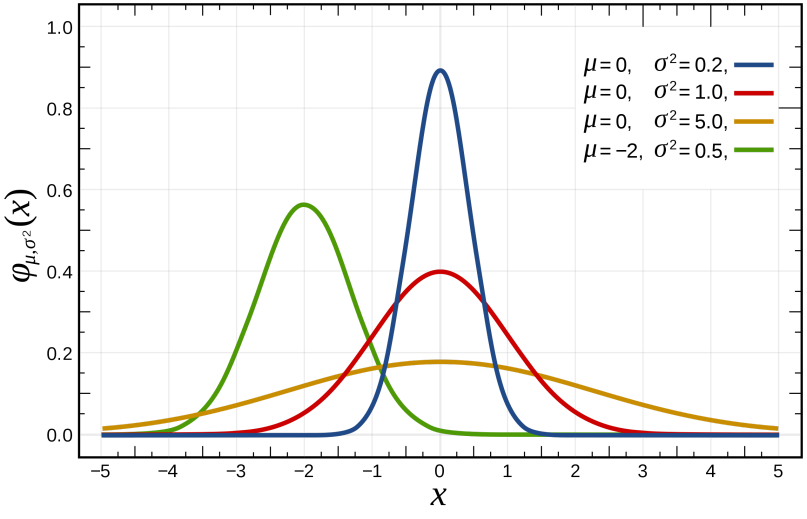
but there are so many normal distributions!

Standard errors

So we know the sample statistic is drawn from a normal distribution...

but there are so many normal distributions!

We need to know the μ and σ of our sampling distribution in order to fully model the variability of our statistic.



Standard errors

We are testing the likelihood that the null is true

Standard errors

We are testing the likelihood that the null is true

Therefore, the mean of our sampling distribution is defined by the null.

Standard errors

We are testing the likelihood that the null is true

Therefore, the mean of our sampling distribution is defined by the null.

For example:

Standard errors

We are testing the likelihood that the null is true

Therefore, the mean of our sampling distribution is defined by the null.

For example:

- H_0: Duck latitudes in 2023 have the same mean as in 2010.
  $\mu_{2023} - \mu_{2010} = 0$

Standard errors

We are testing the likelihood that the null is true

Therefore, the mean of our sampling distribution is defined by the null.

For example:

- H_0: Duck latitudes in 2023 have the same mean as in 2010.
  $\mu_{2023} - \mu_{2010} = 0$
- H_0: Male and female wages have a mean *difference* of zero.
  $\mu_{men} - \mu_{women} = 0$.

Standard errors

We are testing the likelihood that the null is true

Therefore, the mean of our sampling distribution is defined by the null.

For example:

- H_0: Duck latitudes in 2023 have the same mean as in 2010.
  $\mu_{2023} - \mu_{2010} = 0$
- H_0: Male and female wages have a mean *difference* of zero.
  $\mu_{men} - \mu_{women} = 0$.
- H_0: There is *no effect* of neonicotinoid use on colony collapse disorder.
  $\beta_1 = 0$. (Note that linear regression parameters are conditional means -- mean of y conditional on a level of x)

Standard error of the sample mean

The standard deviation of your sampling distribution is called the **standard error**

Standard error of the sample mean

The standard deviation of your sampling distribution is called the **standard error**

How you calculate the standard error depends on the research question.

Standard error of the sample mean

The standard deviation of your sampling distribution is called the **standard error**

How you calculate the standard error depends on the research question.

For example, if we are interested in a sample mean, our friend the **Central Limit Theorem** tells us that:

$$SE = \frac{s^2}{\sqrt{n}}$$

where s is the sample standard deviation and n is the sample size.

> Q: What happens to the standard error as sample size increases? Why?

Standard error for regression slope

The standard deviation of your sampling distribution is called the **standard error**

How you calculate the standard error depends on the research question.

For example, if we are interested in a regression slope, the CLT plus some algebra tell us that:

$$SE = \sqrt{var(\hat{\beta}_1)} = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}$$

where s^2 is the sample variance of model errors $\hat{y}_i - y_i$

> Q: What happens to the standard error as sample size increases? Why?

SE for comparing two means

The standard deviation of your sampling distribution is called the **standard error**

How you calculate the standard error depends on the research question.

For example, if we are interested in the *difference* between two means, the CLT plus some algebra tell us that:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where s_k is the sample standard deviation in each of the two samples and n_k is the sample size in each sample

> In all these cases, the SE is the standard deviation of the sampling distribution!

Summary: Standard errors

If we could collect many samples from the population, and we computed our statistic for each sample (e.g., mean, slope coefficient), we could construct the **sampling distribution.**

Summary: Standard errors

If we could collect many samples from the population, and we computed our statistic for each sample (e.g., mean, slope coefficient), we could construct the **sampling distribution.**

The **standard error** is our estimate of the the standard deviation of the sampling distribution. We can never actually collect hundreds of independent samples, so we use our single sample to approximate the true sampling distribution standard deviation, leveraging the **Central Limit Theorem**

Summary: Standard errors

If we could collect many samples from the population, and we computed our statistic for each sample (e.g., mean, slope coefficient), we could construct the **sampling distribution.**

The **standard error** is our estimate of the the standard deviation of the sampling distribution. We can never actually collect hundreds of independent samples, so we use our single sample to approximate the true sampling distribution standard deviation, leveraging the **Central Limit Theorem**

Standard error measures how dispersed our sample statistic is around the population parameter of interest (highly dispersed = large SE = a lot of uncertainty about the population parameter from our one sample)

Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**
4. Given this variability, **quantify the probability that your sample statistic differs from the null** by the observed amount, if the null were true

Hypothesis testing in five steps

Step 4: quantify the probability that your sample statistic differs from the null by the observed amount, if the null were true

- I know how that my sample statistic is drawn from a normal distribution with mean μ and an estimated standard deviation given by SE.

Hypothesis testing in five steps

Step 4: quantify the probability that your sample statistic differs from the null by the observed amount, if the null were true

- I know how that my sample statistic is drawn from a normal distribution with mean μ and an estimated standard deviation given by SE.
- This should tell me something about **how unlikely it was** that I happened to draw my point estimate if the null were true, right?

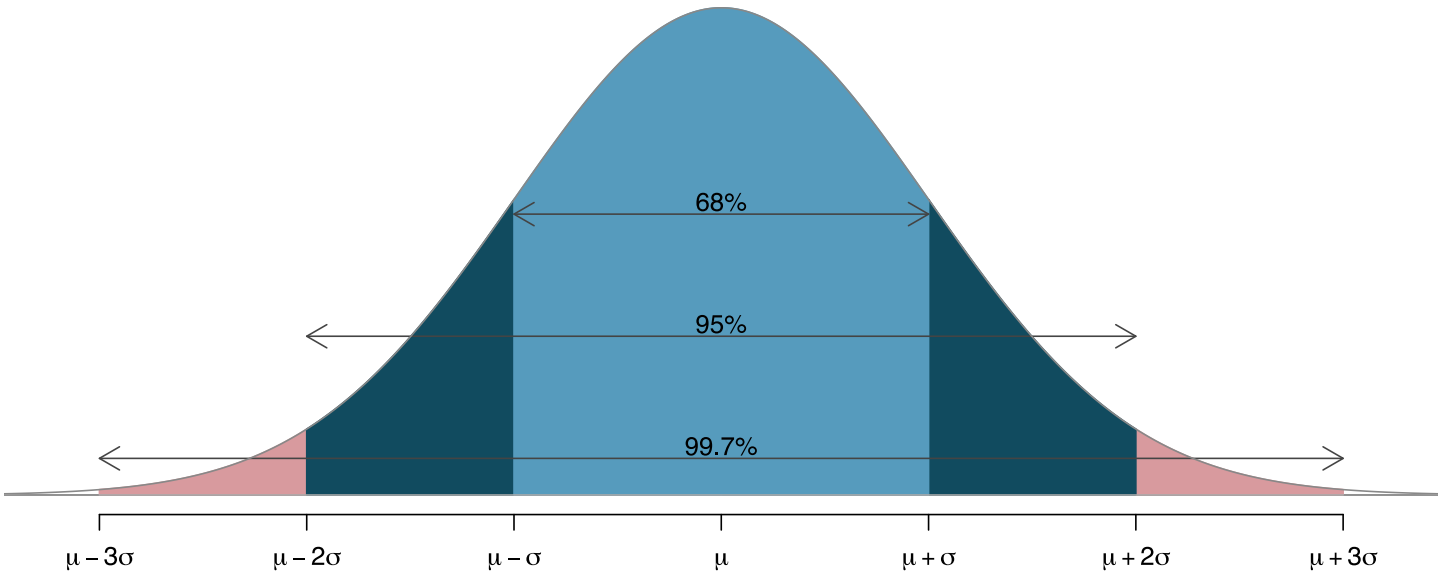
Hypothesis testing in five steps

Step 4: quantify the probability that your sample statistic differs from the null by the observed amount, if the null were true

- I know how that my sample statistic is drawn from a normal distribution with mean μ and an estimated standard deviation given by SE.
- This should tell me something about **how unlikely it was** that I happened to draw my point estimate if the null were true, right?
- Yep! But we need a couple more definitions to get all the way there.

The 68-95-99.7 rule

For a normal distribution:



Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

Z-score

- The 68-95-99.7 rule is helpful if your point estimate (sample statistic) is exactly 1, 2, or 3 standard deviations from the mean (i.e., the null_

Z-score

- The 68-95-99.7 rule is helpful if your point estimate (sample statistic) is exactly 1, 2, or 3 standard deviations from the mean (i.e., the null_
- But what about all the other values?

Z-score

- The 68-95-99.7 rule is helpful if your point estimate (sample statistic) is exactly 1, 2, or 3 standard deviations from the mean (i.e., the null_
- But what about all the other values?

Z-score: How many standard deviations is a value from the mean?

$$z = \frac{x_i - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation

Z-score for hypothesis testing

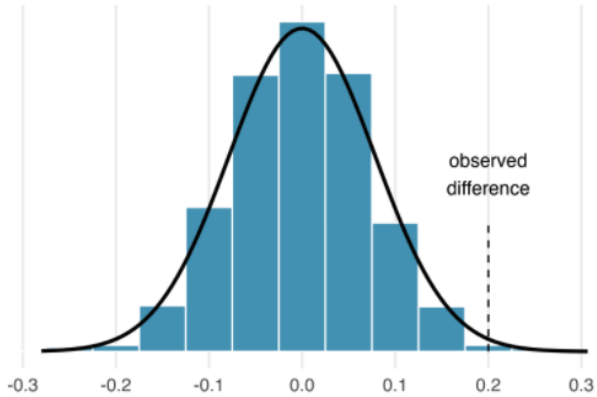
- When testing hypotheses, we care about how far our point estimate is from the null.

Z-score for hypothesis testing

- When testing hypotheses, we care about how far our point estimate is from the null.

Z-score for hypothesis testing: How many standard deviations is a point estimate from the null?

$$z = \frac{\text{point estimate} - \text{null value}}{SE}$$



Quantifying probabilities: *p*-value

- The Z-score is also called the **test statistic**

Quantifying probabilities: *p*-value

- The Z-score is also called the **test statistic**
- The Z-score/test statistic allows us to compute the ***p*-value**:

Quantifying probabilities: *p*-value

- The Z-score is also called the **test statistic**
- The Z-score/test statistic allows us to compute the ***p*-value**:

***p*-value** is the probability of getting a point estimate *at least as extreme* as ours **if the null hypothesis were true.**

$$p - value = Pr(Z < -|z| \text{ or } Z > |z|) = 2 * Pr(Z > |z|)$$

where z is the z-score computed using your point estimate.

Quantifying probabilities: *p*-value

- The Z-score is also called the **test statistic**
- The Z-score/test statistic allows us to compute the ***p*-value**:

***p*-value** is the probability of getting a point estimate *at least as extreme* as ours **if the null hypothesis were true.**

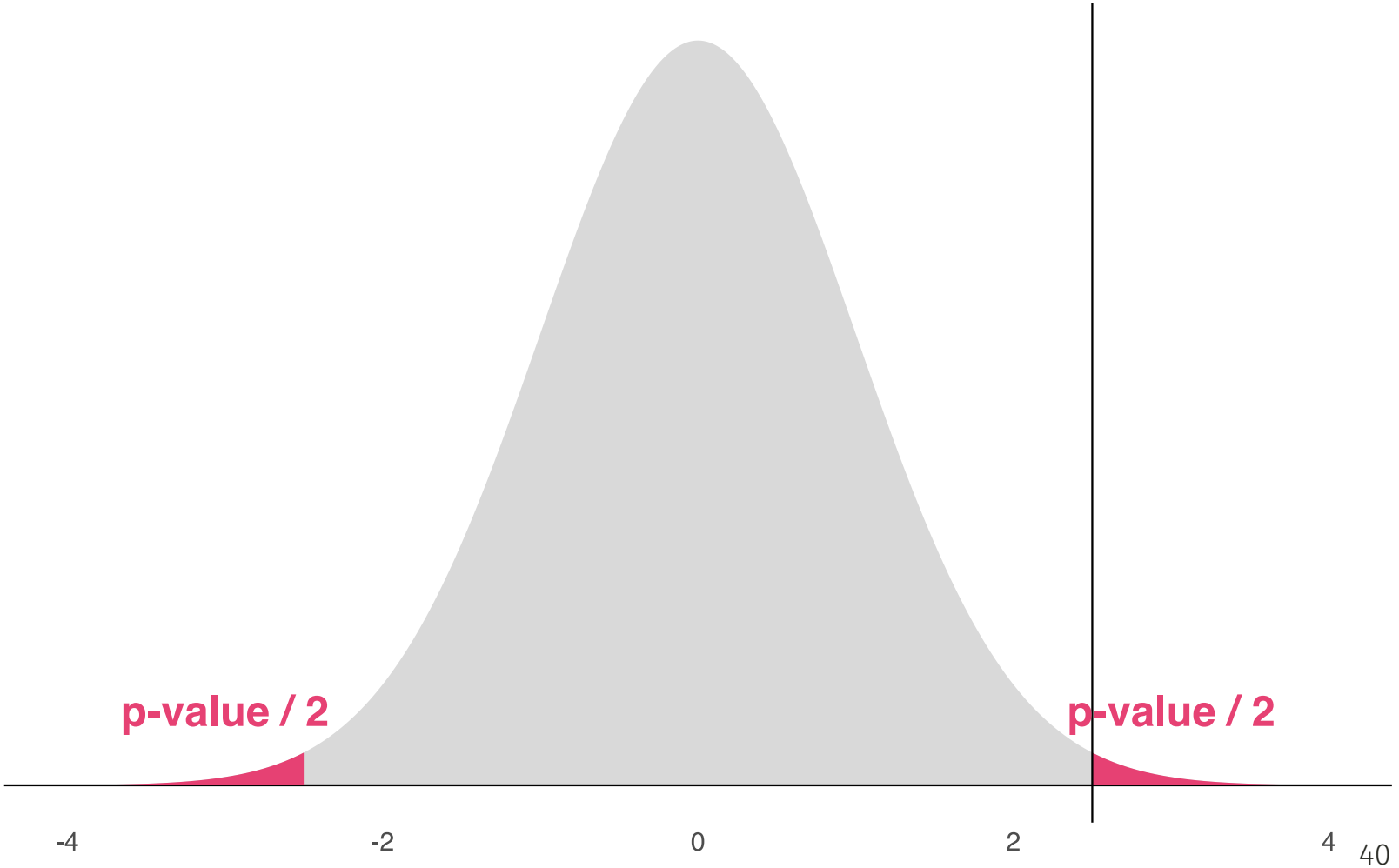
$$p - value = Pr(Z < -|z| \text{ or } Z > |z|) = 2 * Pr(Z > |z|)$$

where z is the z-score computed using your point estimate.

> Question: What feature of the normal distribution lets us simplify this to $2 * Pr(Z > |z|)$?

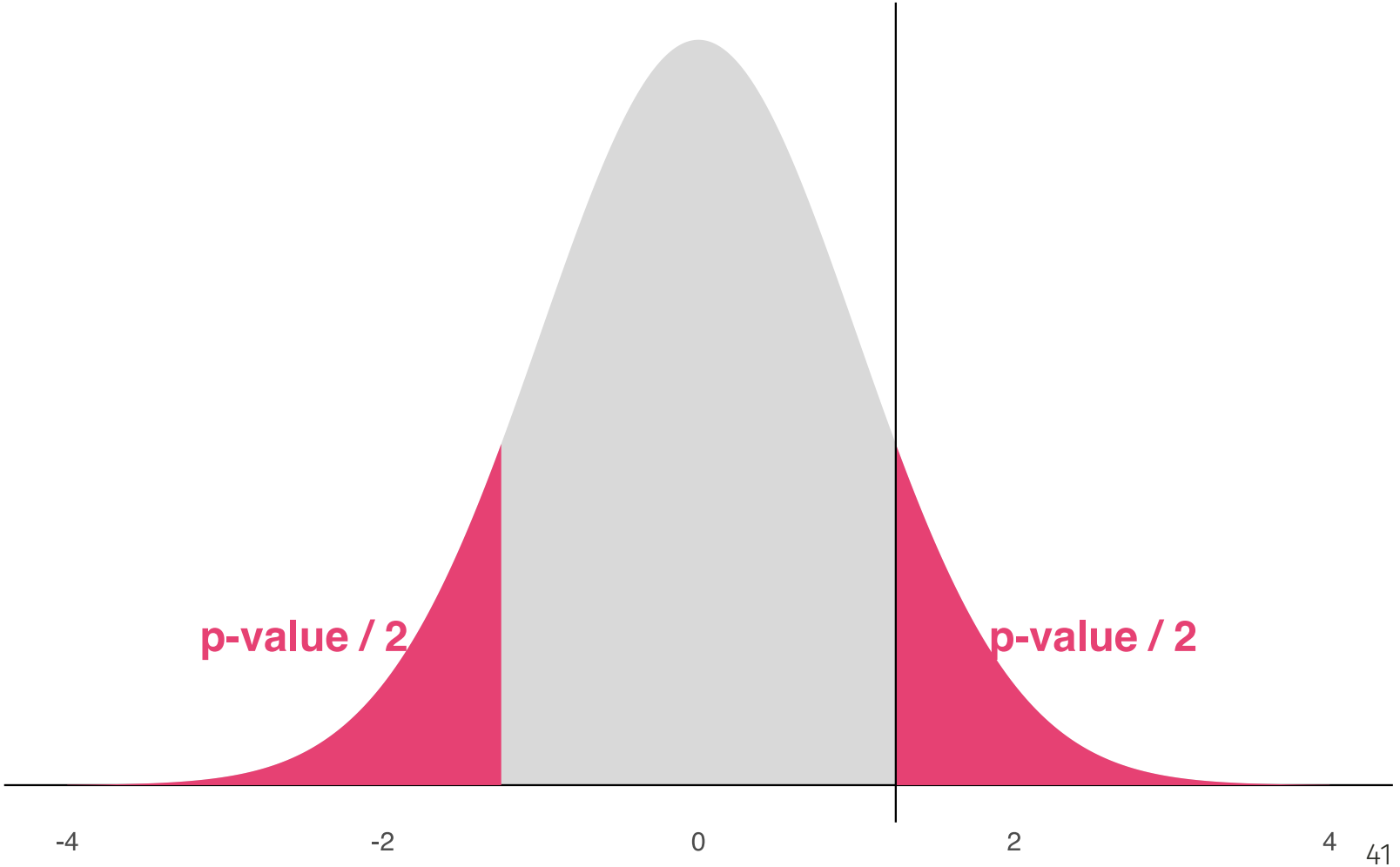
Quantifying probabilities: *p*-value

$$p - value = Pr(Z < -|z| \text{ or } Z > |z|) = 2 * Pr(Z > |z|)$$



Quantifying probabilities: *p*-value

$$p - value = Pr(Z < -|z| \text{ or } Z > |z|) = 2 * Pr(Z > |z|)$$



Hypothesis testing with *p*-values

- *p*-value is the probability of observing a point estimate as extreme as yours if the null were true

Hypothesis testing with p-values

- p-value is the probability of observing a point estimate as extreme as yours if the null were true
- p-value is the area under the sampling distribution to the right and to the left of the absolute value of your test statistic (z-score, z)

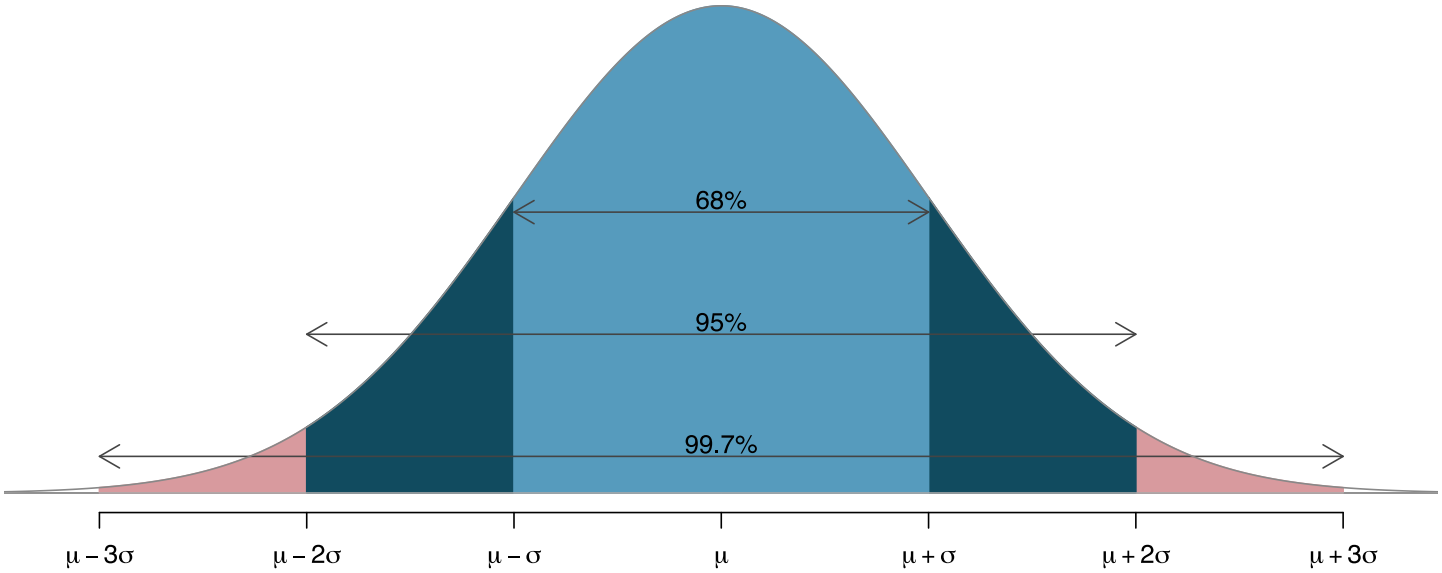
Hypothesis testing with p-values

- p-value is the probability of observing a point estimate as extreme as yours if the null were true
- p-value is the area under the sampling distribution to the right and to the left of the absolute value of your test statistic (z-score, z)

How do I compute a p-value from a test-statistic?

Hypothesis testing with *p*-values

How do I compute a *p*-value from a test-statistic?



Hypothesis testing with *p*-values

- *p*-value is the probability of observing a point estimate as extreme as yours if the null were true

Hypothesis testing with p-values

- p-value is the probability of observing a point estimate as extreme as yours if the null were true
- p-value is the area under the sampling distribution to the right and to the left of the absolute value of your test statistic (z-score, z)

Hypothesis testing with p-values

- p-value is the probability of observing a point estimate as extreme as yours if the null were true
- p-value is the area under the sampling distribution to the right and to the left of the absolute value of your test statistic (z-score, z)

How do I compute a p-value from a test statistic?

Hypothesis testing with *p*-values

- *p*-value is the probability of observing a point estimate as extreme as yours if the null were true
- *p*-value is the area under the sampling distribution to the right and to the left of the absolute value of your test statistic (z-score, z)

How do I compute a *p*-value from a test statistic?

- **In math:** Integrate the sampling distribution's probability density function between $-\infty$ and $-|z|$; multiply by 2
- **In `R`:** `pnorm()`, `t.test()`, `summary(lm())`, ...

A note on the CLT

The **Central Limit Theorem** has done a lot of work for us so far. However, it only holds under the following conditions:

A note on the CLT

The **Central Limit Theorem** has done a lot of work for us so far. However, it only holds under the following conditions:

1. Observations in our sample are **independent**

A note on the CLT

The **Central Limit Theorem** has done a lot of work for us so far. However, it only holds under the following conditions:

1. Observations in our sample are **independent**

2. We have a **large enough sample** (at the very least $n \geq 30$)

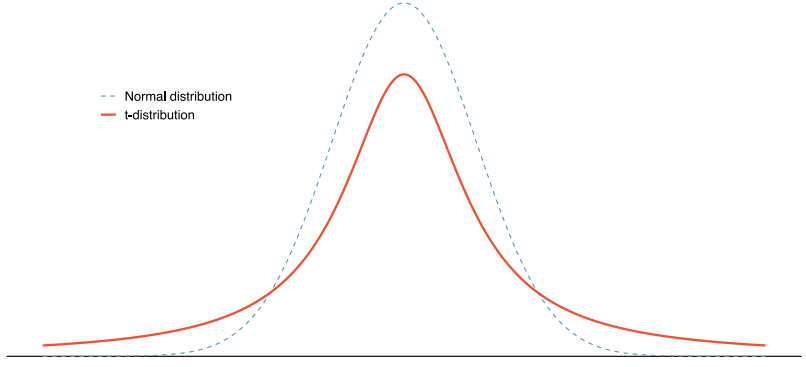
A note on the CLT

The **Central Limit Theorem** has done a lot of work for us so far. However, it only holds under the following conditions:

1. Observations in our sample are **independent**

2. We have a **large enough sample** (at the very least $n \geq 30$)

When n is relatively small, we can still proceed, we just need to use a t-distribution (and T-score -- use `pt()` in `R`) instead of a normal distribution (and Z-score)



Hypothesis testing in five steps

The general framework for implementing a hypothesis test is:

1. **Define the null** and alternative hypotheses
2. Collect data and compute the **point estimate of the statistic**
3. Model the **variability of the statistic**
4. Given this variability, **quantify the probability that your sample statistic differs from the null** by the observed amount, if the null were true
5. Based on #4, either **reject** or **fail to reject** the null

Can we finally test something?

Step 5: Based on the p-value, either **reject** or **fail to reject** the null

- Low p-value $\rightarrow$ very unlikely to see your point estimate if the null were true

Can we finally test something?

Step 5: Based on the p-value, either **reject** or **fail to reject** the null

- Low p-value $\rightarrow$ very unlikely to see your point estimate if the null were true
- High p-value $\rightarrow$ very likely to see your point estimate if the null were true

Can we finally test something?

Step 5: Based on the p-value, either **reject** or **fail to reject** the null

- Low p-value $\rightarrow$ very unlikely to see your point estimate if the null were true
- High p-value $\rightarrow$ very likely to see your point estimate if the null were true

So...what is a low enough p-value to **reject the null**?

Can we finally test something?

What is a low enough p-value to **reject the null**?

> This is a heavily debated question.

Can we finally test something?

What is a low enough *p*-value to **reject the null**?

> This is a heavily debated question.

- Best to **report your *p*-value** alongside any conclusions you reach about your hypothesis

Can we finally test something?

What is a low enough *p*-value to **reject the null**?

> This is a heavily debated question.

- Best to **report your *p*-value** alongside any conclusions you reach about your hypothesis
- Traditionally, we use a **significance level** of $\alpha = 0.05$
  - This says you have a 5% chance of observing your point estimate even if the null were true
  - Reject the null if $p < 0.05$ and $\alpha = 0.05$

Can we finally test something?

What is a low enough *p*-value to **reject the null**?

> This is a heavily debated question.

- Best to **report your *p*-value** alongside any conclusions you reach about your hypothesis
- Traditionally, we use a **significance level** of $\alpha = 0.05$
  - This says you have a 5% chance of observing your point estimate even if the null were true
  - Reject the null if $p < 0.05$ and $\alpha = 0.05$
- In general, reject the null if $p < \alpha$.
  - Other common αs: 0.01, 0.1

Statistical significance

We say a point estimate is "statistically significant" when:

$p < \alpha$

For example:

> "[W]e find a **statistically-significant** effect whereby increases in surface UV intensity lowers subsequent COVID-19 growth rates...we estimate that a 1 kJm−2hr−1 increase in local UV reduces local COVID-19 growth rates by .09 (±.04, p =.01) percentage points over the ensuing 17 days." --- *Carleton et al., 2021*

Hypothesis testing: Rejecting the null

We can **reject the null** hypothesis *or* **fail to reject the null** hypothesis.

Hypothesis testing: Rejecting the null

We can **reject the null** hypothesis *or* **fail to reject the null** hypothesis.

We **never accept the null** hypothesis.

Hypothesis testing: Rejecting the null

We can **reject the null** hypothesis *or* **fail to reject the null** hypothesis.

We **never accept the null** hypothesis.

Why not?

- Lack of evidence is not proof! If $p > \alpha$, there is so much sampling variability that we cannot distinguish the null from the point estimate.

Hypothesis testing: Rejecting the null

We can **reject the null** hypothesis *or* **fail to reject the null** hypothesis.

We **never accept the null** hypothesis.

Why not?

- Lack of evidence is not proof! If $p > \alpha$, there is so much sampling variability that we cannot distinguish the null from the point estimate.

Think of this as innocent (null is true) until proven guilty (null is rejected).

Hypothesis testing: Rejecting the null

We can **reject the null** hypothesis *or* **fail to reject the null** hypothesis.

We **never accept the null** hypothesis.

Why not?

- Lack of evidence is not proof! If $p > \alpha$, there is so much sampling variability that we cannot distinguish the null from the point estimate.

Think of this as innocent (null is true) until proven guilty (null is rejected).

- Failing to reject the null tells us we do not have sufficient evidence to prove there is an effect or a difference

Constructing confidence intervals

Why use confidence intervals?

- *p*-values are not enough for us to conclude anything meaningful about an analysis

Why use confidence intervals?

- *p*-values are not enough for us to conclude anything meaningful about an analysis
- Effect sizes are important! We care not just about whether a treatment effects an outcome, but by *how much*

Why use confidence intervals?

- *p*-values are not enough for us to conclude anything meaningful about an analysis
- Effect sizes are important! We care not just about whether a treatment effects an outcome, but by *how much*

A **confidence interval** is a range of plausible values where we may find the true population value.

- It tells us something about the magnitude of the parameter of interest, as well as our uncertainty around our estimate

Confidence intervals

When the sampling distribution of a point estimate can be modeled as normal, the point estimate we observe will be within 1.96 standard errors of the true value of interest about 95% of the time (think back to the 68-95-99.7 rule).

Confidence intervals

When the sampling distribution of a point estimate can be modeled as normal, the point estimate we observe will be within 1.96 standard errors of the true value of interest about 95% of the time (think back to the 68-95-99.7 rule).

Thus, a 95% confidence interval for such a point estimate can be constructed:

$$\text{point estimate} \pm 1.96 * SE$$

We can be 95% confident this interval captures the true value.

> Also can see this as: `2*pnorm(-1.96) = .05`

Confidence intervals

You can build a confidence interval for any level of α:

$$\textbf{point estimate} \pm z_{\alpha/2} * SE$$

where $z_{\alpha/2}$ is a "critical value" that varies with significance level α.

$z_{\alpha/2}$ is the z-score at which $\alpha/2$ percent of the sampling distribution exceeds that z-score

Confidence intervals

You can build a confidence interval for any level of α:

$$\text{point estimate} \pm z_{\alpha/2} * SE$$

where $z_{\alpha/2}$ is a "critical value" that varies with significance level α.

$z_{\alpha/2}$ is the z-score at which $\alpha/2$ percent of the sampling distribution exceeds that z-score

For example:

- $\alpha = 0.1$ = 90% confidence interval: $\text{point\_estimate} \pm 1.64 * SE$

Confidence intervals

You can build a confidence interval for any level of α:

$$\textbf{point estimate} \pm z_{\alpha/2} * SE$$

where $z_{\alpha/2}$ is a "critical value" that varies with significance level α.

$z_{\alpha/2}$ is the z-score at which $\alpha/2$ percent of the sampling distribution exceeds that z-score

For example:

- $\alpha = 0.1$ = 90% confidence interval: $\textbf{point\_estimate} \pm 1.64 * SE$
- $\alpha = 0.01$ = 99% confidence interval: $\textbf{point\_estimate} \pm 2.57 * SE$

Confidence intervals

You can build a confidence interval for any level of α:

$$\textbf{point estimate} \pm z_{\alpha/2} * SE$$

where $z_{\alpha/2}$ is a "critical value" that varies with significance level α.

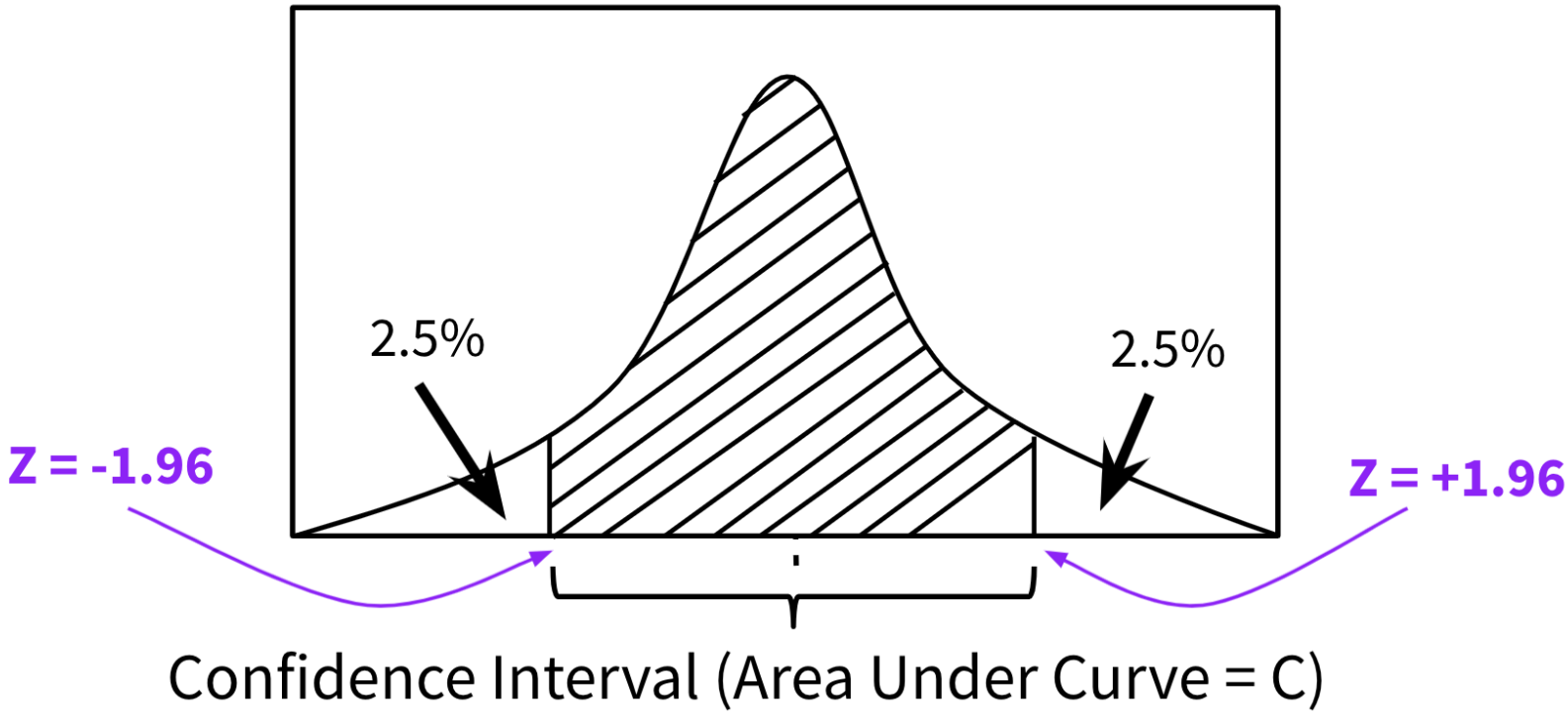
$z_{\alpha/2}$ is the z-score at which $\alpha/2$ percent of the sampling distribution exceeds that z-score

For example:

- $\alpha = 0.1$ = 90% confidence interval: $\textbf{point\_estimate} \pm 1.64 * SE$
- $\alpha = 0.01$ = 99% confidence interval: $\textbf{point\_estimate} \pm 2.57 * SE$

> Why do the z-scores get larger for higher confidence intervals?

A 95% confidence interval visual



Confidence intervals: Interpretation

A 95% confidence interval tells you there is a **95% chance that your interval includes the true population parameter**.

Confidence intervals: Interpretation

A 95% confidence interval tells you there is a **95% chance that your interval includes the true population parameter**.

A very common misinterpretation:

There is a 95% chance the true population parameter falls inside my confidence interval.

Confidence intervals: Interpretation

A 95% confidence interval tells you there is a **95% chance that your interval includes the true population parameter**.

A very common misinterpretation:

There is a 95% chance the true population parameter falls inside my confidence interval.

Why is this a big deal?

The population parameter *is not random*. So it either **is or is not** inside your CI.

Slides created via the R package **xaringan**.

Some slide components were borrowed from Ed Rubin's awesome course materials.