Lecture 8: Precision of OLS and Hypothesis Testing

ECON 480 - Econometrics - Fall 2018

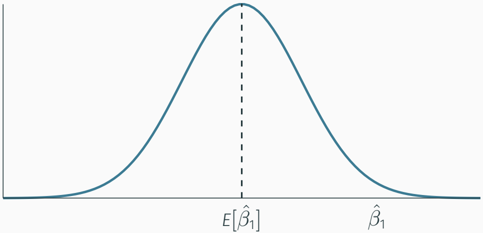
Ryan Safner

September 26, 2018

The Precision of OLS

$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$

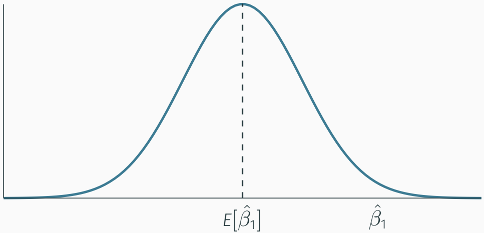
- We want to know:



Recall: The Sampling Distribution of $\hat{\beta}_1$

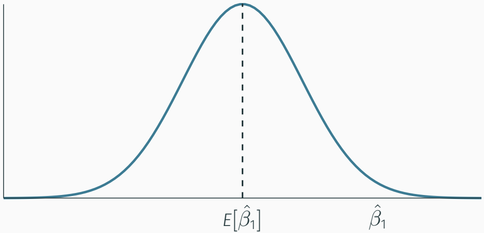
$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$

- We want to know:
    - $E[\hat{\beta}_1]$; what is the center of the distribution?

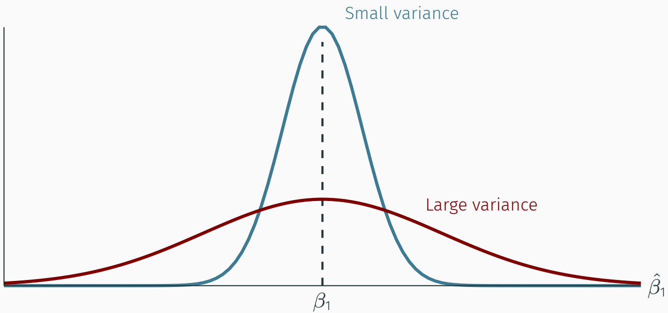


$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$

- We want to know:
    - $E[\hat{\beta}_1]$; what is the center of the distribution?
    - $\sigma_{\hat{\beta}_1}$; how precise is our estimate?

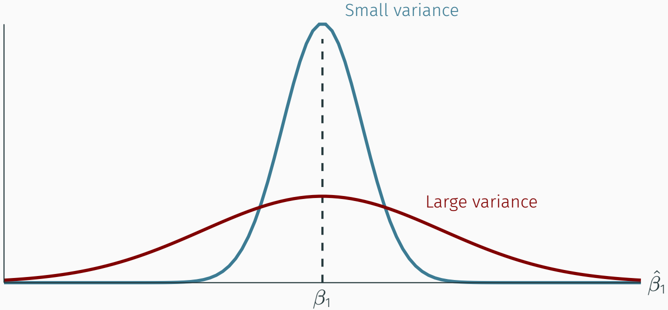


- How precise is our estimate $\hat{\beta}_1$?



[1]The "standard **error**" is the analogue of standard *deviation* for a sample statistic's sampling distribution. Recall the sampling distribution is the distribution of a statistic, like $\bar{X}$ or $\hat{\beta}_1$ over many potential samples.

- How precise is our estimate $\hat{\beta}_1$?
- We can talk of the variance $(\sigma^2_{\hat{\beta}_1})$ or the standard error $(\sigma_{\hat{\beta}_1})$ of $\hat{\beta}_1$[1]



[1] The "standard **error**" is the analogue of standard *deviation* for a sample statistic's sampling distribution. Recall the sampling distribution is the distribution of a statistic, like $\bar{X}$ or $\hat{\beta}_1$ over many potential samples.

- The variance of $\hat{\beta}_1$ is

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

where SER is the standard error of the regression (again):

$$SER = \frac{SSE}{n-2}$$

HOOD COLLEGE

- The variance of $\hat{\beta}_1$ is

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

where SER is the standard error of the regression (again):

$$SER = \frac{SSE}{n-2}$$

- The standard error of $\hat{\beta}_1$ is simply the square root of the variance

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)}$$

What Affects Variance

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

- Variance is affected by three things

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

- Variance is affected by three things
  1. **Goodness of fit of the model:** SER

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

- Variance is affected by three things
  1. **Goodness of fit of the model:** *SER*
     - Larger *SER*, larger $var(\hat{\beta}_1)$

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

- Variance is affected by three things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_1)$
  2. **Sample size,** n

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

- Variance is affected by three things

  1. **Goodness of fit of the model:** *SER*
     - Larger *SER*, larger $var(\hat{\beta}_1)$
  2. **Sample size,** n
     - Larger n, lower $var(\hat{\beta}_1)$

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

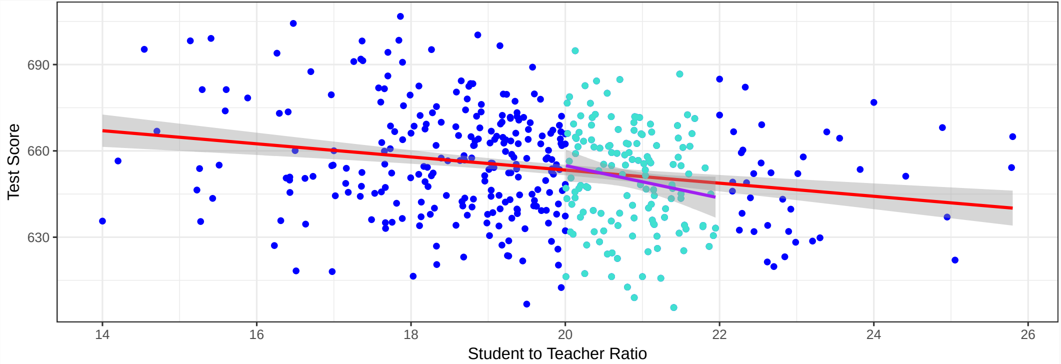
- Variance is affected by three things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_1)$
  2. **Sample size,** n
     - Larger n, lower $var(\hat{\beta}_1)$
  3. **Variation in** X

$$var(\hat{\beta}_1) = \frac{(SER)^2}{n \times var(X)}$$

- Variance is affected by three things

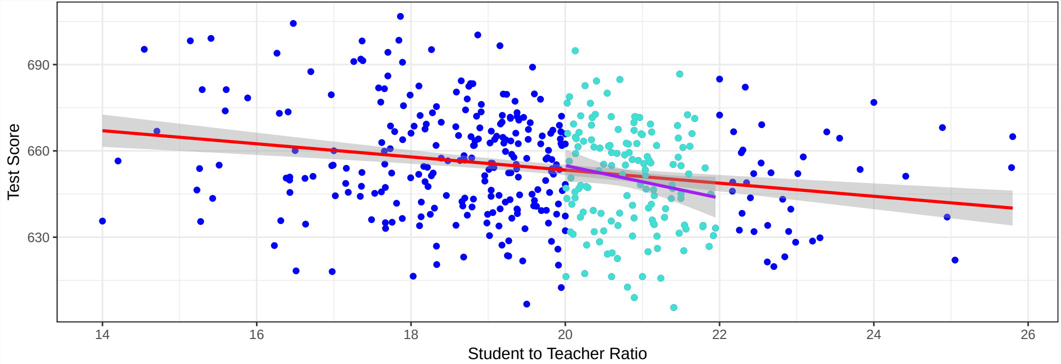
1. **Goodness of fit of the model:** *SER*
    - Larger *SER*, larger $var(\hat{\beta}_1)$
2. **Sample size,** n
    - Larger n, lower $var(\hat{\beta}_1)$
3. **Variation in** X
    - Larger $var(X)$, smaller $var(\hat{\beta}_1)$

THE RELATION BETWEEN VARIANCE OF X AND VARIANCE OF $\hat{\beta}_1$



- Smaller $var(X)$ (light dots only) $\implies$ larger $var(\hat{\beta}_1)$: harder to determine precise slope!

The Relation Between Variance of X and Variance of $\hat{\beta}_1$



- Smaller $var(X)$ (light dots only) $\implies$ larger $var(\hat{\beta}_1)$: harder to determine precise slope!
- Larger $var(X)$ (all dots) $\implies$ smaller $var(\hat{\beta}_1)$: easier to determine precise slope!

Hypothesis Testing About Regression

- We have used statistics (OLS) to estimate a relationship between X and Y

- We have used statistics (OLS) to estimate a relationship between X and Y
    - *Relationship is causal if X is exogenous*

- We have used statistics (OLS) to estimate a relationship between X and Y
  - *Relationship is causal if X is exogenous*
- The "bread and butter" of inferential statistics is testing a hypothesis about (a) population(s) parameter(s)

- We have used statistics (OLS) to estimate a relationship between X and Y
  - *Relationship is causal if X is exogenous*
- The “bread and butter” of inferential statistics is testing a hypothesis about (a) population(s) parameter(s)
  - Does reducing class size actually improve test scores?

- We have used statistics (OLS) to estimate a relationship between X and Y
  - *Relationship is causal if X is exogenous*
- The "bread and butter" of inferential statistics is testing a hypothesis about (a) population(s) parameter(s)
  - Does reducing class size actually improve test scores?
  - Is the gender wage gap between men and women really $0.77?

- We have used statistics (OLS) to estimate a relationship between X and Y
    - *Relationship is causal if X is exogenous*
- The "bread and butter" of inferential statistics is testing a hypothesis about (a) population(s) parameter(s)
    - Does reducing class size actually improve test scores?
    - Is the gender wage gap between men and women really $0.77?
    - What percent of American voters support legalizing marijuana?

- We have used statistics (OLS) to estimate a relationship between X and Y
  - *Relationship is causal if X is exogenous*
- The "bread and butter" of inferential statistics is testing a hypothesis about (a) population(s) parameter(s)
  - Does reducing class size actually improve test scores?
  - Is the gender wage gap between men and women really $0.77?
  - What percent of American voters support legalizing marijuana?
- **All modern science is built upon statistical hypothesis testing, so understand it well!**

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
    - A population mean μ (e.g. average height of adults)

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
    - A population mean μ (e.g. average height of adults)
    - A population proportion p (e.g. percent of voters who voted for Trump)

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
    - A population mean μ (e.g. average height of adults)
    - A population proportion p (e.g. percent of voters who voted for Trump)
    - A difference in population means $\mu_A - \mu_B$ (e.g. difference in average wages of men vs. women)

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
    - A population mean μ (e.g. average height of adults)
    - A population proportion p (e.g. percent of voters who voted for Trump)
    - A difference in population means $\mu_A - \mu_B$ (e.g. difference in average wages of men vs. women)
    - A difference in population proportions $p_A - p_B$ (e.g. difference in percent of patients reporting symptoms of drug A vs B)

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
    - A population mean μ (e.g. average height of adults)
    - A population proportion p (e.g. percent of voters who voted for Trump)
    - A difference in population means $\mu_A - \mu_B$ (e.g. difference in average wages of men vs. women)
    - A difference in population proportions $p_A - p_B$ (e.g. difference in percent of patients reporting symptoms of drug A vs B)
    - See all the possibilities in glorious detail in the **handout** on Inferential Statistics

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
    - A population mean μ (e.g. average height of adults)
    - A population proportion p (e.g. percent of voters who voted for Trump)
    - A difference in population means $\mu_A - \mu_B$ (e.g. difference in average wages of men vs. women)
    - A difference in population proportions $p_A - p_B$ (e.g. difference in percent of patients reporting symptoms of drug A vs B)
    - See all the possibilities in glorious detail in the **handout** on Inferential Statistics
- We will focus on hypotheses about the population regression slope ($\hat{\beta}_1$) between X and Y

Hypothesis Testing II

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
  - A population mean μ (e.g. average height of adults)
  - A population proportion p (e.g. percent of voters who voted for Trump)
  - A difference in population means $\mu_A - \mu_B$ (e.g. difference in average wages of men vs. women)
  - A difference in population proportions $p_A - p_B$ (e.g. difference in percent of patients reporting symptoms of drug A vs B)
  - See all the possibilities in glorious detail in the **handout** on Inferential Statistics
- We will focus on hypotheses about the population regression slope ($\hat{\beta}_1$) between X and Y
  - i.e. if/when we've done our model right, the causal effect of X on Y

- All scientific inquiries begin with a null hypothesis (H_0) that proposes a specific value of a population parameter

Null and Alternative Hypotheses

- All scientific inquiries begin with a null hypothesis (H_0) that proposes a specific value of a population parameter
    - Notation: add a subscript 0: $\beta_{1,0}$ (or μ_0, p_0, etc)

- All scientific inquiries begin with a null hypothesis (H_0) that proposes a specific value of a population parameter
    - Notation: add a subscript 0: $\beta_{1,0}$ (or μ_0, p_0, etc)
- We suggest an alternative hypothesis (H_a), often the one we hope to verify

- All scientific inquiries begin with a null hypothesis (H_0) that proposes a specific value of a population parameter
    - Notation: add a subscript 0: $\beta_{1,0}$ (or μ_0, p_0, etc)
- We suggest an alternative hypothesis (H_a), often the one we hope to verify
    - Note, can be multiple alternative hypotheses: $H_1, H_2, \ldots, H_n$

- All scientific inquiries begin with a null hypothesis (H_0) that proposes a specific value of a population parameter
  - Notation: add a subscript 0: $\beta_{1,0}$ (or μ_0, p_0, etc)
- We suggest an alternative hypothesis (H_a), often the one we hope to verify
  - Note, can be multiple alternative hypotheses: $H_1, H_2, \ldots, H_n$
- Ask: **“Does our data (sample) give us sufficient evidence to reject H_0 in favor of H_a?”**

Null and Alternative Hypotheses

- All scientific inquiries begin with a null hypothesis (H_0) that proposes a specific value of a population parameter
    - Notation: add a subscript 0: $\beta_{1,0}$ (or μ_0, p_0, etc)
- We suggest an alternative hypothesis (H_a), often the one we hope to verify
    - Note, can be multiple alternative hypotheses: $H_1, H_2, \ldots, H_n$
- Ask: **"Does our data (sample) give us sufficient evidence to reject H_0 in favor of H_a?"**
    - Note: the test is *always* about H_0! See if we have sufficient evidence to reject the status quo

- Null hypothesis assigns a value (or a range) to a population parameter

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!
- Alternative hypothesis must mathematically *contradict* the null hypothesis

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!
- Alternative hypothesis must mathematically *contradict* the null hypothesis
  - e.g. $\beta_1 \neq 2$ or $\beta_1 > 20$ or $\beta_1 \neq 0$

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
    - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
    - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
    - Note: always an equality!
- Alternative hypothesis must mathematically *contradict* the null hypothesis
    - e.g. $\beta_1 \neq 2$ or $\beta_1 > 20$ or $\beta_1 \neq 0$
    - Note: always an inequality!

HOOD COLLEGE

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!
- Alternative hypothesis must mathematically *contradict* the null hypothesis
  - e.g. $\beta_1 \neq 2$ or $\beta_1 > 20$ or $\beta_1 \neq 0$
  - Note: always an inequality!
- Alternative hypotheses come in two forms:

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!
- Alternative hypothesis must mathematically *contradict* the null hypothesis
  - e.g. $\beta_1 \neq 2$ or $\beta_1 > 20$ or $\beta_1 \neq 0$
  - Note: always an inequality!
- Alternative hypotheses come in two forms:
  1. **One-sided alternative**: $\beta_1 > H_0$ or $\beta_1 < H_0$

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!
- Alternative hypothesis must mathematically *contradict* the null hypothesis
  - e.g. $\beta_1 \neq 2$ or $\beta_1 > 20$ or $\beta_1 \neq 0$
  - Note: always an inequality!
- Alternative hypotheses come in two forms:
  1. **One-sided alternative**: $\beta_1 > H_0$ or $\beta_1 < H_0$
  2. **Two-sided alternative**: $\beta_1 \neq H_0$

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common null hypothesis is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!
- Alternative hypothesis must mathematically *contradict* the null hypothesis
  - e.g. $\beta_1 \neq 2$ or $\beta_1 > 20$ or $\beta_1 \neq 0$
  - Note: always an inequality!
- Alternative hypotheses come in two forms:
  1. **One-sided alternative**: $\beta_1 > H_0$ or $\beta_1 < H_0$
  2. **Two-sided alternative**: $\beta_1 \neq H_0$
     - Note this means either $\beta_1 < H_0$ or $\beta_1 > H_0$

- All statistical hypothesis tests have the following components:

- All statistical hypothesis tests have the following components:

1. A **null hypothesis,** H_0

Components of a Valid Hypothesis Test

- All statistical hypothesis tests have the following components:

1. A **null hypothesis,** H_0
2. An **alternative hypothesis,** H_a

- All statistical hypothesis tests have the following components:

1. A **null hypothesis,** H_0
2. An **alternative hypothesis,** H_a
3. A **test statistic** to determine if we reject H_0 when the statistic reaches a "critical value"

- All statistical hypothesis tests have the following components:

1. A **null hypothesis,** H_0
2. An **alternative hypothesis,** H_a
3. A **test statistic** to determine if we reject H_0 when the statistic reaches a "critical value"
    - Beyond the critical value is the "rejection region", sufficient evidence to reject H_0

Components of a Valid Hypothesis Test

- All statistical hypothesis tests have the following components:

1. A **null hypothesis,** H_0
2. An **alternative hypothesis,** H_a
3. A **test statistic** to determine if we reject H_0 when the statistic reaches a "critical value"
    - Beyond the critical value is the "rejection region", sufficient evidence to reject H_0
4. A **conclusion** whether or not to reject H_0 in favor of H_a

HOOD COLLEGE

- Any sample statistic (e.g. $\hat{\beta}_1$) will rarely be exactly equal to the hypothesized population parameter (e.g. β_1)

- Any sample statistic (e.g. $\hat{\beta}_1$) will rarely be exactly equal to the hypothesized population parameter (e.g. β_1)
- Difference between observed statistic and true paremeter could be because:

- Any sample statistic (e.g. $\hat{\beta}_1$) will rarely be exactly equal to the hypothesized population parameter (e.g. β_1)
- Difference between observed statistic and true paremeter could be because:
    1. Parameter is *not* the hypothesized value (H_0 is false)

- Any sample statistic (e.g. $\hat{\beta}_1$) will rarely be exactly equal to the hypothesized population parameter (e.g. β_1)
- Difference between observed statistic and true paremeter could be because:
    1. Parameter is *not* the hypothesized value (H_0 is false)
    2. Parameter is truly the hypothesized value (H_0 is true) but *sampling variability* gave us a different estimate

- Any sample statistic (e.g. $\hat{\beta}_1$) will rarely be exactly equal to the hypothesized population parameter (e.g. β_1)
- Difference between observed statistic and true paremeter could be because:
    1. Parameter is *not* the hypothesized value (H_0 is false)
    2. Parameter is truly the hypothesized value (H_0 is true) but *sampling variability* gave us a different estimate
- **We cannot distinguish between these two possibilities with any certainty**

Type I and Type II Errors II

- We can interpret our estimates probabilistically as commiting one of two types of error:

- We can interpret our estimates probabilistically as commiting one of two types of error:

1. Type I error (false positive): rejecting H_0 when it is in fact true

- We can interpret our estimates probabilistically as commiting one of two types of error:

1. Type I error (false positive): rejecting H_0 when it is in fact true
    - Believing we found an important result when there is truly no relationship

Type I and Type II Errors II

- We can interpret our estimates probabilistically as commiting one of two types of error:

1. Type I error (false positive): rejecting H_0 when it is in fact true
    - Believing we found an important result when there is truly no relationship
2. Type II error (false negative): failing to reject H_0 when it is in fact false

- We can interpret our estimates probabilistically as commiting one of two types of error:

1. Type I error (false positive): rejecting H_0 when it is in fact true
    - Believing we found an important result when there is truly no relationship
2. Type II error (false negative): failing to reject H_0 when it is in fact false
    - Believing we found nothing when there was truly a relationship to find

	H_0 is True	H_0 is False
Reject H_0	Type I Error False Positive	Correct Outcome True Positive
Don't Reject H_0	Correct Outcome True Negative	Type II Error False Negative

Type I and Type II Errors IV

	Defendant is Innocent	Defendant is Guilty
Convict "I think he's guilty"	Type I Error False Positive	Correct Outcome True Positive
Don't Convict "I think he's innocent"	Correct Outcome True Negative	Type II Error False Negative

- Depending on context, committing one type of error may be more serious than the other

Type I and Type II Errors IV

	Defendant is Innocent	Defendant is Guilty
Convict “I think he’s guilty”	Type I Error False Positive	Correct Outcome True Positive
Don’t Convict “I think he’s innocent”	Correct Outcome True Negative	Type II Error False Negative

- Depending on context, committing one type of error may be more serious than the other
- Common law *presumes* the defendant is innocent and a jury judges whether the evidence presented against the defendant would be plausible *if the defendant were in fact innocent*

Example

For each of the following scenarios, identify the Type I error, Type II error, α and β, and which error is of greater consequence?

Example

For each of the following scenarios, identify the Type I error, Type II error, α and β, and which error is of greater consequence?

- H_0 : the victim of an automobile accident is alive when he arrives at the Emergency Room

Example

For each of the following scenarios, identify the Type I error, Type II error, α and β, and which error is of greater consequence?

- H_0 : the victim of an automobile accident is alive when he arrives at the Emergency Room
- H_0: a rock climber's equipment is safe

Example

For each of the following scenarios, identify the Type I error, Type II error, α and β, and which error is of greater consequence?

- H_0 : the victim of an automobile accident is alive when he arrives at the Emergency Room
- H_0: a rock climber's equipment is safe
- H_0: a woman is not pregnant

Example

For each of the following scenarios, identify the Type I error, Type II error, α and β, and which error is of greater consequence?

- H_0 : the victim of an automobile accident is alive when he arrives at the Emergency Room
- H_0: a rock climber's equipment is safe
- H_0: a woman is not pregnant
- H_0: a highway project will cost no more than $10 million

Example

For each of the following scenarios, identify the Type I error, Type II error, α and β, and which error is of greater consequence?

- H_0 : the victim of an automobile accident is alive when he arrives at the Emergency Room
- H_0: a rock climber's equipment is safe
- H_0: a woman is not pregnant
- H_0: a highway project will cost no more than \$10 million
- H_0: an experimental cancer drug has a cure rate of at least 75%

- The significance level, α, is the probability of a **Type I error**

$$\alpha = P\big(\text{Reject } H_0 \big| H_0 \text{ is true}\big)$$

$$\beta = P\big(\text{Don't reject} H_0 \big| H_0 \text{ is false}\big)$$

Significance Level, α

- The significance level, α, is the probability of a **Type I error**

$$\alpha = P\big(\text{Reject } H_0 \big| H_0 \text{ is true}\big)$$

- The confidence level is defined as $1 - \alpha$

$$\beta = P\big(\text{Don't reject} H_0 \big| H_0 \text{ is false}\big)$$

Significance Level, α

- The significance level, α, is the probability of a **Type I error**

$$\alpha = P\left(\text{Reject } H_0 \middle| H_0 \text{ is true}\right)$$

- The confidence level is defined as $1 - \alpha$
- We often specify in advance an α-level (0.10,0.05,0.01) with associated confidence level (90%, 95%, 99%)

$$\beta = P\left(\text{Don't reject} H_0 \middle| H_0 \text{ is false}\right)$$

Significance Level, α

- The significance level, α, is the probability of a **Type I error**

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is true})$$

- The confidence level is defined as $1 - \alpha$
- We often specify in advance an α-level (0.10,0.05,0.01) with associated confidence level (90%, 95%, 99%)
- The probability of a **Type II error** is defined as β:

$$\beta = P(\text{Don't reject} H_0 | H_0 \text{ is false})$$

α AND β

	H_0 is True	H_0 is False
Reject H_0	Type I Error α	Correct Outcome $(1-\beta)$
Don't Reject H_0	Correct Outcome $(1-\alpha)$	Type II Error β

- The statistical power of the test is $1 - \beta$, the probability of correctly rejecting H_0 when H_0 is in fact false (e.g. not convicting an innocent person)

$$\text{Power} = 1 - \beta = P\big(\text{Reject } H_0 \big| H_0 \text{ is false}\big)$$

Power and p-values

- The statistical power of the test is $1 - \beta$, the probability of correctly rejecting H_0 when H_0 is in fact false (e.g. not convicting an innocent person)

$$\text{Power} = 1 - \beta = P\big(\text{Reject } H_0 \big| H_0 \text{ is false}\big)$$

- The p-value or significance probability is the probability that, given the null hypothesis is true, the test statistic from a random sample will be at least as extreme as the test statistic of our sample

Power and p-values

- The statistical power of the test is $1-\beta$, the probability of correctly rejecting H_0 when H_0 is in fact false (e.g. not convicting an innocent person)

$$\text{Power} = 1-\beta = P(\text{Reject } H_0 | H_0 \text{ is false})$$

- The p-value or significance probability is the probability that, given the null hypothesis is true, the test statistic from a random sample will be at least as extreme as the test statistic of our sample
- If $p < \alpha$, the results are "**statistically significant**"

HOOD COLLEGE

- After running our test, we need to make a *decision* between the competing hypotheses

- After running our test, we need to make a *decision* between the competing hypotheses
- Compare p-value with *pre-determined* α (commonly, $\alpha = 0.05$, 95% confidence level)

- After running our test, we need to make a *decision* between the competing hypotheses
- Compare p-value with *pre-determined* α (commonly, $\alpha = 0.05$, 95% confidence level)
    - If $p < \alpha$: **statistically significant** evidence sufficient to *reject* H_0 in favor of H_a

- After running our test, we need to make a *decision* between the competing hypotheses
- Compare p-value with *pre-determined* α (commonly, $\alpha = 0.05$, 95% confidence level)
    - If $p < \alpha$: **statistically significant** evidence sufficient to *reject* H_0 in favor of H_a
    - If $p \geq \alpha$: *insufficient* evidence to reject H_0

HOOD COLLEGE

- After running our test, we need to make a *decision* between the competing hypotheses
- Compare p-value with *pre-determined* α (commonly, $\alpha = 0.05$, 95% confidence level)
  - If $p < \alpha$: **statistically significant** evidence sufficient to *reject* H_0 in favor of H_a
  - If $p \geq \alpha$: *insufficient* evidence to reject H_0
    - Note this does **not** mean H_0 is true! We merely have *failed* to *reject* H_0

Digression: p-Values and the Philosophy of Science

"The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

(1931). *The Design of Experiments*

Sir Ronald A. Fisher

(1890-1962)

- Modern philosophy of science is largely based off of hypothesis testing and **falsifiability**, which form the ”Scientific Method”

- Modern philosophy of science is largely based off of hypothesis testing and **falsifiability**, which form the ”Scientific Method”
- For something to be ”scientific”, it must be *falsifiable*, or at least *testable*

- Modern philosophy of science is largely based off of hypothesis testing and **falsifiability**, which form the ”Scientific Method”
- For something to be ”scientific”, it must be *falsifiable*, or at least *testable*
- Hypotheses can be corroborated with evidence, but always *tentative* until falsified by data in suggesting an alternative hypothesis

- Modern philosophy of science is largely based off of hypothesis testing and **falsifiability**, which form the ”Scientific Method”
- For something to be ”scientific”, it must be *falsifiable*, or at least *testable*
- Hypotheses can be corroborated with evidence, but always *tentative* until falsified by data in suggesting an alternative hypothesis
    - e.g. ”**All swans are white**” is a hypothesis rejected upon discovery of a single black swan

- Modern philosophy of science is largely based off of hypothesis testing and **falsifiability**, which form the "Scientific Method"
- For something to be "scientific", it must be *falsifiable*, or at least *testable*
- Hypotheses can be corroborated with evidence, but always *tentative* until falsified by data in suggesting an alternative hypothesis
    - e.g. "**All swans are white**" is a hypothesis rejected upon discovery of a single black swan
- Note: economics is a very different kind of "science" with a different methodology!

Caution

It is easy to misinterpret what statistical significance and p-values mean. THE FOLLOWING ARE FALSE:

Caution
It is easy to misinterpret what statistical significance and p-values mean. THE FOLLOWING ARE FALSE:

- p is the probability that the alternative hypothesis is true (We can never *prove* an alternative hypothesis, only tentatively reject a null hypothesis)

Caution

It is easy to misinterpret what statistical significance and p-values mean. THE FOLLOWING ARE FALSE:

- p is the probability that the alternative hypothesis is true (We can never *prove* an alternative hypothesis, only tentatively reject a null hypothesis)
- p is the probability that the null hypothesis is false (We are not proving the null hypothesis false, only saying that it is very unlikely that under the null hypothesis, we obtain an event as rare as our sample)

Caution

It is easy to misinterpret what statistical significance and p-values mean. **THE FOLLOWING ARE FALSE**:

- p is the probability that the alternative hypothesis is true (We can never *prove* an alternative hypothesis, only tentatively reject a null hypothesis)
- p is the probability that the null hypothesis is false (We are not proving the null hypothesis false, only saying that it is very unlikely that under the null hypothesis, we obtain an event as rare as our sample)
- p is the probability that the observed effects were produced purely by random chance (p is computed under a specific model (assuming H_0 is true)

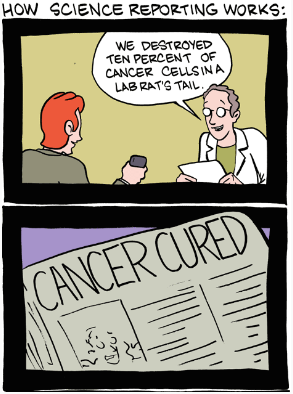
Caution
It is easy to misinterpret what statistical significance and p-values mean. THE FOLLOWING ARE FALSE:

- p is the probability that the alternative hypothesis is true (We can never *prove* an alternative hypothesis, only tentatively reject a null hypothesis)
- p is the probability that the null hypothesis is false (We are not proving the null hypothesis false, only saying that it is very unlikely that under the null hypothesis, we obtain an event as rare as our sample)
- p is the probability that the observed effects were produced purely by random chance (p is computed under a specific model (assuming H_0 is true)
- p tells us how significant our finding is (p tells us nothing about the *size* or the *real world significance* of any effect deemed "statistically significant")

- Again, **p is the probability that, assuming the null hypothesis is true, we obtain (by pure random chance) a test statistic at least as extreme as the one we estimated for our sample**
    - This will make more sense in context, when we discuss the nature of our test statistics

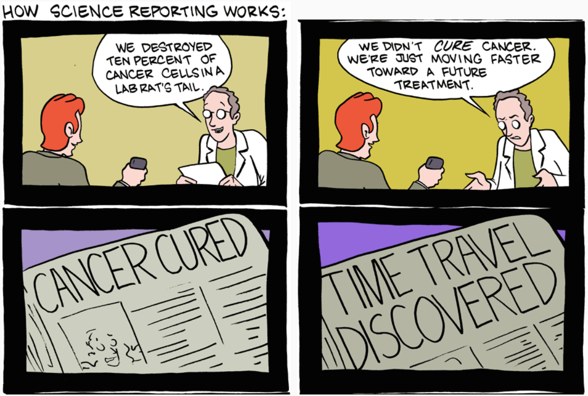
Statistical Significance: Cautions II

- Again, **p is the probability that, assuming the null hypothesis is true, we obtain (by pure random chance) a test statistic at least as extreme as the one we estimated for our sample**
  - This will make more sense in context, when we discuss the nature of our test statistics
- Remember a low p-value means **either** that the null hypothesis is true and a highly improbable event has occurred or that the null hypothesis is false (we don't know which!)


HOW SCIENCE REPORTING WORKS:
WE DESTROYED TEN PERCENT OF CANCER CELLS IN A LAB RAT'S TAIL.
CANCER CURED

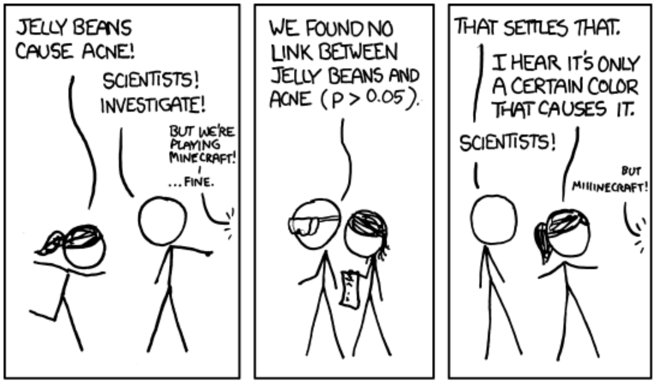
SMBC 1623

Statistical Significance and p-Values



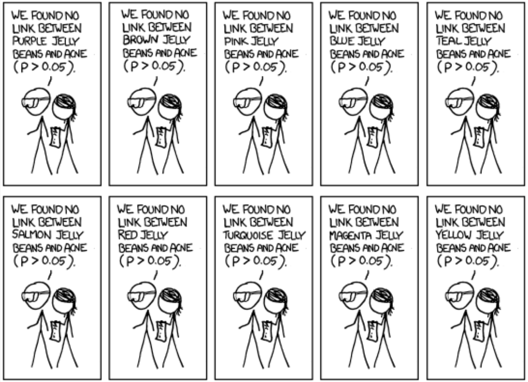
SMBC 1623

Statistical Significance and p-Values II



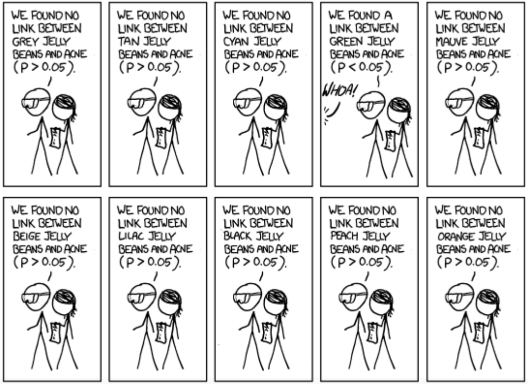
XKCD 882

Statistical Significance and p-Values III

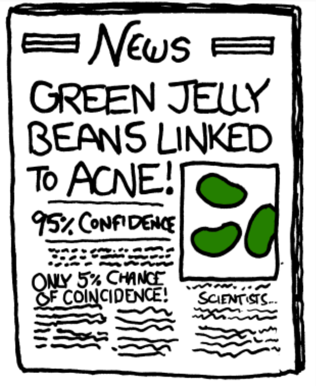


XKCD 882

Statistical Significance and p-Values IV

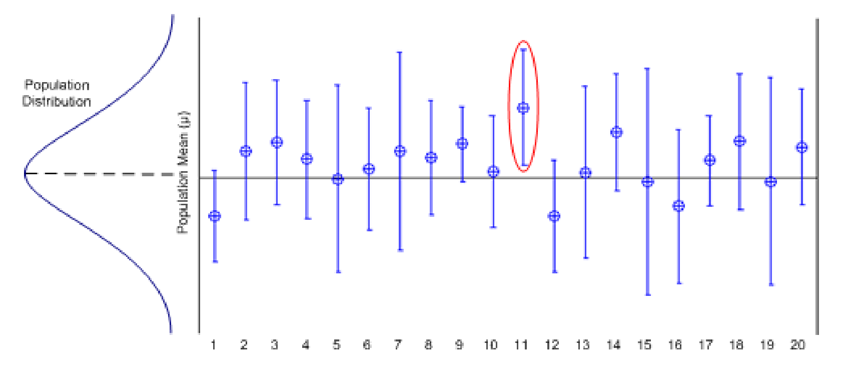


XKCD 882

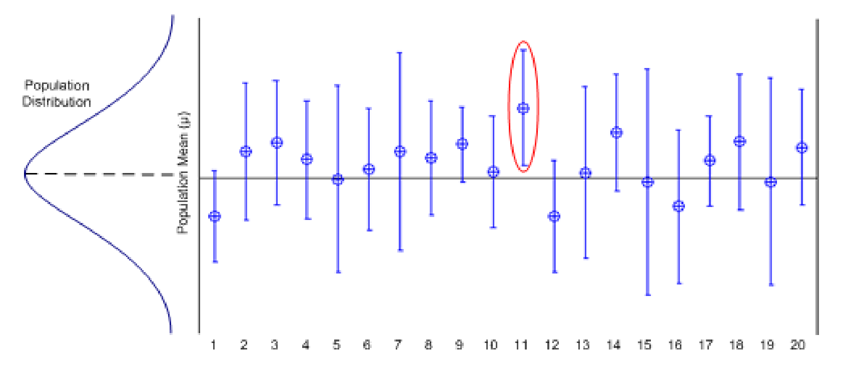

NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

XKCD 882

- Consider what "95% significance" or $\alpha = 0.05$ means:



- Consider what "95% significance" or $\alpha = 0.05$ means:
    - If we repeat a procedure 20 times, we should *expect* 1/20 (5%) to produce a fluke result!



"The widespread use of "statistical significance" (generally interpreted as ($p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process."

Wasserstein, Ronald L. and Nicole A. Lazar, (2016). "The ASA's Statement on p-Values: Context, Process, and Purpose" *The American Statistician* 30(2): 129-133.

Morning Mix

How, and why, a journalist tricked news outlets into thinking chocolate makes you thin

By Sarah Kaplan May 28, 2015

Most Read

1 Black man fatally shot by Tulsa police was unarmed, chief says, as 'disturbing' video is released

2 Scientists uncovered a skeleton from the ancient world's most famous — and mysterious — shipwreck

3 Father of suspected bomber Ahmad Rahami says he had called the FBI about him

4 'You can sleep tonight knowing the Klan is awake.' Fliers like these are showing up on lawns across the U.S.

5 Aren't more white people than black people killed by police? Yes, but no.

Our Online Games

Play right from this page

Washington Post: How, and why, a journalist tricked news outlets into thinking chocolate makes you thin

Back to Our Hypothesis Test: The Test-Statistic

- We next consider the population distribution **assuming H_0 is true** and calculate a test statistic, which takes the following form:

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the statistic}}$$

- We next consider the population distribution **assuming H_0 is true** and calculate a test statistic, which takes the following form:

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the statistic}}$$

- We then compare our test statistic against a **critical value** to determine if we can reject H_0

Distribution of H_0

- We next consider the population distribution **assuming H_0 is true** and calculate a test statistic, which takes the following form:

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the statistic}}$$

- We then compare our test statistic against a **critical value** to determine if we can reject H_0
- Essentially: test to see how likely a sample statistic at least as extreme as our discovery is if H_0 were true

Distribution of H_0 II

- We are testing our estimated $\hat{\beta}_1$ against a null hypothesis, e.g. $\beta_{1,0} = 0$

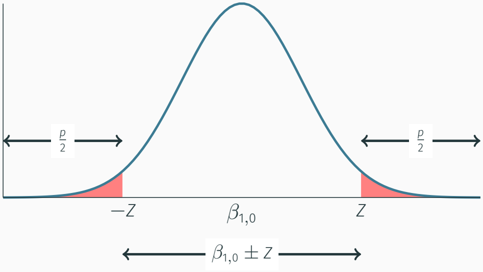
- We are testing our estimated $\hat{\beta}_1$ against a null hypothesis, e.g. $\beta_{1,0} = 0$
- It would be nice if we could use normal distribution, our test statistic would just be Z-score:

$$Z = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- We are testing our estimated $\hat{\beta}_1$ against a null hypothesis, e.g. $\beta_{1,0} = 0$
- It would be nice if we could use normal distribution, our test statistic would just be Z-score:

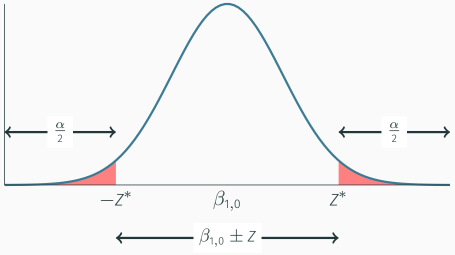
$$Z = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- *p*-**value**: area in the tail(s) of the distr. of $\hat{\beta}_1$ under H_0 beyond our Z score



Distribution of β_1 under $H_0 : \beta_1 = 0$; p-value is the area in red (two-tailed test)

- The **critical value** Z^* is determined by our α level (e.g. 0.05)

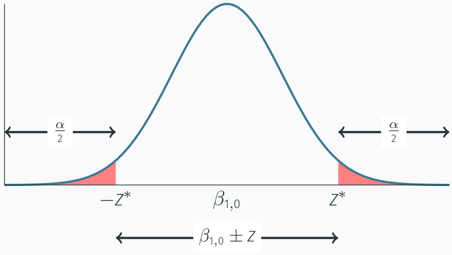


Critical values of Z^* with rejection regions in red

[2] As you can see, the empirical 68-95-99.7% rule is very close, but not perfect!

Distribution of H_0 III

- The **critical value** Z^* is determined by our α level (e.g. 0.05)
- For a 2-sided alternative and $\alpha = 0.05$, $Z^* = 1.96$[2]

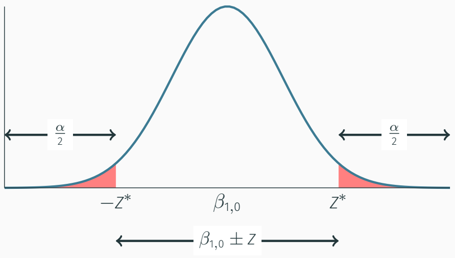


Critical values of Z^* with rejection regions in red

[2] As you can see, the empirical 68-95-99.7% rule is very close, but not perfect!

Distribution of H_0 III

- The **critical value** Z^* is determined by our α level (e.g. 0.05)
- For a 2-sided alternative and $\alpha = 0.05$, $Z^* = 1.96$[2]
- Any Z-score beyond ± 1.96 is in rejection region, sufficient evidence to reject H_0



Critical values of Z^* with rejection regions in red

[2] As you can see, the empirical 68-95-99.7% rule is very close, but not perfect!

Distribution of H_0 IV

- It would be nice if we *could* just use the normal distribution, and run a Z-test, as described above

- It would be nice if we *could* just use the normal distribution, and run a Z-test, as described above
- **Central Limit Theorem** lets us if $n \geq 30$ and we know the population distribution μ, σ

Distribution of H_0 IV

- It would be nice if we *could* just use the normal distribution, and run a Z-test, as described above
- **Central Limit Theorem** lets us if $n \geq 30$ and we know the population distribution μ, σ
- We almost never know them...

- Worked at Guinness testing beer quality

William Sealy Gosset

(1876-1937)

- Worked at Guinness testing beer quality
- Using normal distributions with small sample sizes did not yield accurate estimates

William Sealy Gosset

(1876-1937)

- Worked at Guinness testing beer quality
- Using normal distributions with small sample sizes did not yield accurate estimates
- Developed a new distribution, using the pseudonym "Student," to publish, the Student's t-distribution

William Sealy Gosset

(1876-1937)

- Instead of Z-scores, we use the t-**score**, which has the same intuition (# of standard deviations above/below the mean)

$$t \sim t_{n-1}$$

- Instead of Z-scores, we use the t-**score**, which has the same intuition (# of standard deviations above/below the mean)

$$t \sim t_{n-1}$$

- t-scores follow a Student's t-distribution with $n - k - 1$ degrees of freedom

- Instead of Z-scores, we use the t-**score**, which has the same intuition (# of standard deviations above/below the mean)

$$t \sim t_{n-1}$$

- t-scores follow a Student's t-distribution with $n - k - 1$ degrees of freedom
  - k: number of variables (e.g. X's)

- Instead of Z-scores, we use the t-**score**, which has the same intuition (# of standard deviations above/below the mean)

$$t \sim t_{n-1}$$

- t-scores follow a Student's t-distribution with $n - k - 1$ degrees of freedom
    - k: number of variables (e.g. X's)
    - Formally, **degrees of freedom** (df or ν) are the number of independent values used for the calculation of a statistic minus the number of other statistics used as intermediate steps

t-DISTRIBUTION

- Instead of Z-scores, we use the t-**score**, which has the same intuition (# of standard deviations above/below the mean)

$$t \sim t_{n-1}$$

- t-scores follow a Student's t-distribution with $n - k - 1$ degrees of freedom
  - k: number of variables (e.g. X's)
  - Formally, **degrees of freedom** (df or ν) are the number of independent values used for the calculation of a statistic minus the number of other statistics used as intermediate steps
  - Here, we *first* had to calculate two statistics ($\hat{\beta}_0, \hat{\beta}_1$) with our sample *before* estimating the sampling distribution of $\hat{\beta}_1$ for hypothesis testing, thus $df = n - 2$

- Instead of Z-scores, we use the t-**score**, which has the same intuition (# of standard deviations above/below the mean)

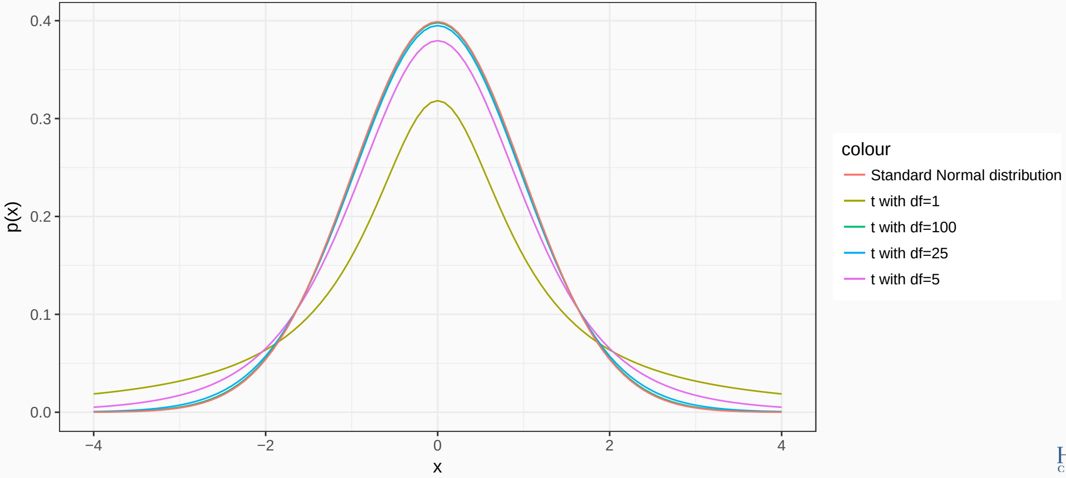
$$t \sim t_{n-1}$$

- t-scores follow a Student's t-distribution with $n-k-1$ degrees of freedom
  - k: number of variables (e.g. X's)
  - Formally, **degrees of freedom** (df or ν) are the number of independent values used for the calculation of a statistic minus the number of other statistics used as intermediate steps
  - Here, we *first* had to calculate two statistics ($\hat{\beta}_0, \hat{\beta}_1$) with our sample *before* estimating the sampling distribution of $\hat{\beta}_1$ for hypothesis testing, thus $df = n-2$
- t-distribution looks normal-ish (symmetric, unimodal, mean=0), but more mass in the tails

- Instead of Z-scores, we use the t-**score**, which has the same intuition (# of standard deviations above/below the mean)

$$t \sim t_{n-1}$$

- t-scores follow a Student's t-distribution with $n - k - 1$ degrees of freedom
  - k: number of variables (e.g. X's)
  - Formally, **degrees of freedom** (df or ν) are the number of independent values used for the calculation of a statistic minus the number of other statistics used as intermediate steps
  - Here, we *first* had to calculate two statistics ($\hat{\beta}_0, \hat{\beta}_1$) with our sample *before* estimating the sampling distribution of $\hat{\beta}_1$ for hypothesis testing, thus $df = n - 2$
- t-distribution looks normal-ish (symmetric, unimodal, mean=0), but more mass in the tails
- Exact shape of t depends on df: as $\uparrow df$, $t \rightarrow$ Normal distribution

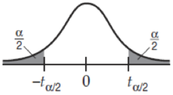


0.4
0.3
0.2
0.1
0.0
p(x)
−4
−2
0
2
4
x
colour
Standard Normal distribution
t with df=1
t with df=100
t with df=25
t with df=5

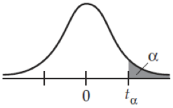
Calculating t-scores: Old-Fashioned Way

Table T

Values of t_α



Two tails



One tail

Two tail probability	0.20	0.10	0.05
One tail probability	0.10	0.05	0.025
df			
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571
6	1.440	1.943	2.447
7	1.415	1.895	2.365
8	1.397	1.860	2.306
9	1.383	1.833	2.262
10	1.372	1.812	2.228
11	1.363	1.796	2.201
12	1.356	1.782	2.179
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120
17	1.333	1.740	2.110
18	1.330	1.734	2.101
19	1.328	1.729	2.093
⋮	⋮	⋮	⋮
∞	1.282	1.645	1.960
Confidence levels	80%	90%	95%

Calculating t-scores: In R

```r
# use pt() command, needs t value and df
pt(2,df=5) #probability of t>2 with 5 df
```

```
## [1] 0.9490303
```

```r
pt(2,df=40)# probability of t>2 with 40 df
```

```
## [1] 0.9738388
```

```r
pt(2, df=100) # probability of t>2 with 100 df
```

```
## [1] 0.9758939
```

```r
pnorm(2, mean=0, sd=1) # compare to normal distribution!
```

- So our test statistic is a t-score (instead of Z-score)

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- So our test statistic is a t-score (instead of Z-score)

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- We then compare t to the critical value of t^* determined by our α-level and the df for our t-distribution $(n - k - 1)$

- So our test statistic is a t-score (instead of Z-score)

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- We then compare t to the critical value of t^* determined by our α-level and the df for our t-distribution $(n-k-1)$
    - Note: there will be a unique critical value for every value of $n-k-1$!

Hypothesis Testing with t-distribution

- So our test statistic is a t-score (instead of Z-score)

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- We then compare t to the critical value of t^* determined by our α-level and the df for our t-distribution $(n - k - 1)$
    - Note: there will be a unique critical value for every value of $n - k - 1$!
    - R determines the critical t^* automatically with regression

- So our test statistic is a t-score (instead of Z-score)

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- We then compare t to the critical value of t^* determined by our α-level and the df for our t-distribution $(n-k-1)$
    - Note: there will be a unique critical value for every value of $n-k-1$!
    - R determines the critical t^* automatically with regression
- p-value$= P(t < T)$

Hypothesis Testing with t-distribution

- So our test statistic is a t-score (instead of Z-score)

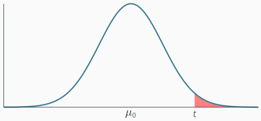
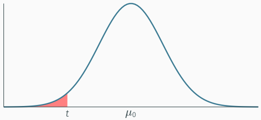
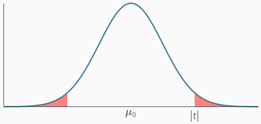
$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

- We then compare t to the critical value of t^* determined by our α-level and the df for our t-distribution $(n-k-1)$
  - Note: there will be a unique critical value for every value of $n-k-1$!
  - R determines the critical t^* automatically with regression
- p-value$= P(t < T)$
- Reject H_0 if p-value $< \alpha$

HOOD COLLEGE

Hypothesis Testing with t-distribution II

Depending on the desired alternative hypothesis:

Alternative	p-value	PDF
$H_a : \beta_1 > \beta_{1,0}$	$P(T \geq t)$	
$H_a : \beta_1 < \beta_{1,0}$	$P(T \leq t)$	
$H_a : \beta_1 \neq \beta_{1,0}$	$2P(T \geq \lvert t \rvert)$	

Example

We have an estimated regression line:

$$\widehat{\text{Test Score}} = \underset{(9.47)}{689.93} - \underset{(0.48)}{2.28}\ \text{STR}$$

- Regression reporting format: Coefficients with their (standard errors) beneath them

Example

We have an estimated regression line:

$$\widehat{\text{Test Score}} = \underset{(9.47)}{689.93} - \underset{(0.48)}{2.28}\,\text{STR}$$

- Regression reporting format: Coefficients with their (standard errors) beneath them
- Let $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ (two-sided alternative)

Example

We have an estimated regression line:

$$\widehat{\text{Test Score}} = 689.93 - 2.28\ \text{STR}$$

$$(9.47) \quad (0.48)$$

- Regression reporting format: Coefficients with their (standard errors) beneath them
- Let $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ (two-sided alternative)
- t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

Example

We have an estimated regression line:

$$\widehat{\text{Test Score}} = 689.93 - 2.28\ \text{STR}$$
$$(9.47) \quad (0.48)$$

- Regression reporting format: Coefficients with their (standard errors) beneath them
- Let $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ (two-sided alternative)
- t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.48}$$

Example

We have an estimated regression line:

$$\widehat{\text{Test Score}} = 689.93 - 2.28\ \text{STR}$$
$$(9.47) \quad (0.48)$$

- Regression reporting format: Coefficients with their (standard errors) beneath them
- Let $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ (two-sided alternative)
- t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.48} = -4.75$$

Example
We have an estimated regression line:

$$\widehat{\text{Test Score}} = 689.93 - 2.28\ \text{STR}$$
$$(9.47) \quad (0.48)$$

- Regression reporting format: Coefficients with their (standard errors) beneath them
- Let $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ (two-sided alternative)
- t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.48} = -4.75$$

```
# calculate p-value for t=-4.75, df=418
2*pt(-4.75,df=418) # x2 because we want both tails!
```

```r
summary(school.regression)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

- If $|\hat{\beta}_k| > 2 \times SE(\hat{\beta}_k)$, the estimate is significant

- If $|\hat{\beta}_k| > 2 \times SE(\hat{\beta}_k)$, the estimate is significant

$$\widehat{\text{Test Score}} = \underset{(9.47)}{689.93} - \underset{(0.48)}{2.28}\ \text{STR}$$

- Since essentially $t = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$ and we roughly want $t \geq 2$ for 95% confidence level (α=0.05)