CSCI 3022

intro to data science with probability & statistics

Lecture 15
March 7, 2018

Introduction to Statistical Inference & Confidence Intervals

NOTE: It was L'Hospital who happily stole Johann Bernoulli's work & published it as his own. The next time you take an indeterminant limit, remember that & call it "Bernoulli's Rule" instead!

TONY WUZ HERE

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Last time on CSCI 3022

- **Proposition**: If X is a normally distributed random variable with mean μ and standard deviation σ , then Z is a standard normal distribution if

$$Z = \frac{X - \mu}{\sigma} \quad \text{or} \quad X = \sigma Z + \mu$$

- **The Central Limit Theorem**: Let $X_1, X_2, \ldots, X_n$ be i.i.d. draws from some distribution. Then as n becomes large

non-normal or normal

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- A $100(1-\alpha)\%$ confidence interval for the mean μ when the value of σ is known is given by:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Statistical Inference

- **Goal**: Want to extract properties of an underlying population by analyzing sampled data
- **Last time we saw**:
  - How to determine a confidence interval for the population mean
  - How to determine a confidence interval for the population proportion
- **This time we'll see**:
  - How to put a confidence interval on the difference between means of two populations
  - How to put a confidence interval on the difference between proportions of two populations
  - How we can get a good numerical estimate of a CI using something called the Bootstrap

Difference between population means

- How do two sub-populations compare? In particular, are their means the same?
- **Classic Motivating Examples**:
  - Is a drug's effectiveness the same in children and adults? ✓
  - Does cigarette brand A contain more nicotine that cigarette brand B?
  - Does a class perform better when Professor C (his) teaches it or Professor D? (an)
  - Does email ad E generate more customers than email ad F?

Difference between population means

- How do two sub-populations compare? In particular, are their means the same?
- **Solution Process**:
  - Collect samples from both sub-populations, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$

Difference between population means

- How do two sub-populations compare? In particular, are their means the same?
- **Solution Process**:
  - Collect samples from both sub-populations, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$
- **Basic Assumptions**:
  - $X_1, X_2, \ldots, X_m$ is a random sample from a distribution with mean μ_1 and sd σ_1
  - $Y_1, Y_2, \ldots, Y_n$ is a random sample from a distribution with mean μ_2 and sd σ_2
  - The X and Y samples are independent of each other.

Difference between population means

- The natural estimator of $\mu_1 - \mu_2$ is the difference of the sample means $\bar{x} - \bar{y}$
- Is $\bar{x} - \bar{y}$ a *good* estimator for $\mu_1 - \mu_2$?
- The expected value of $\bar{X} - \bar{Y}$ is given by

$\bar{X}$ & $\bar{Y}$ r.v.

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$$ ✓✓

- The standard deviation of $\bar{X} - \bar{Y}$ is given by

$$SD[\bar{X} - \bar{Y}] = \sqrt{Var[\bar{X} - \bar{Y}]} = \sqrt{Var[\bar{X}] + Var[\bar{Y}]} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$ ✓

$(-1)^2$

Normal populations with known SDs

- If both populations are normal, then both $\bar{X}$ and $\bar{Y}$ are normally distributed.
- Independence of the two samples implies that the sample means are independent.
- Therefore, the difference between the means is normally distributed, for any sample sizes, with:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

estimator

expected value of est.

variance

Confidence Interval for the difference

- Standardizing $\bar{X} - \bar{Y}$ gives a standard normal random variable

messy z

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

But oh so nice! z

$$\sim N(0, 1)$$

- And so, we can compute a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$

$$\bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

central est. $\pm$ z-score * SD

Large sample CIs for the difference

- **Not surprisingly**, if both *m* and *n* are large, then our friend, the CLT, kicks in, and our confidence interval for the difference of means is valid, even when the populations are *not* normally distributed!

- **Furthermore**, if *m* and *n* are large, and we don't know the standard deviations, we can replace them with the sample standard deviations:

$$\sigma_1^2 \mapsto s_1^2 = \frac{1}{m-1}\sum_i (x_i - \bar{x})^2$$

$$\sigma_2^2 \mapsto s_2^2 = \frac{1}{n-1}\sum_j (y_j - \bar{y})^2$$

Confidence Interval for the Difference

$[-0.508, 0.068]$

- **Example**: Suppose you run two different email Ad campaigns over many days and record the amount of traffic driven to your website on days that each ad is sent. In particular, suppose that Ad 1 is sent on 50 different days and generates an average of 2 million page views per day with an sd of 1 million views, and Ad 2 is sent on 40 different days and generates an average of 2.25 million page views per day with an sd of a half million views. Find a 95% confidence interval for the difference in average page views per day (in units of millions of views).

$\bar{x} = 2$, $m = 50$, $s_1 = 1$

$\bar{y} = 2.25$, $n = 40$, $s_2 = 0.5$

$z_{\alpha/2} = z_{.05/2} = 1.96$

$$95\%\ CI = (\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

$$= (2 - 2.25) \pm 1.96 \cdot \sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}} = [-0.508, 0.068]$$

Confidence Interval for the Difference

- **Example**: Suppose you run two different email Ad campaigns over many days and record the amount of traffic driven to your website on days that each ad is sent. In particular, suppose that Ad 1 is sent on 50 different days and generates an average of 2 million page views per day with an sd of 1 million views, and Ad 2 is sent on 40 different days and generates an average of 2.25 million page views per day with an sd of a half million views. Find a 95% confidence interval for the difference in average page views per day (in units of millions of views).

Confidence Interval for the Difference

HERE

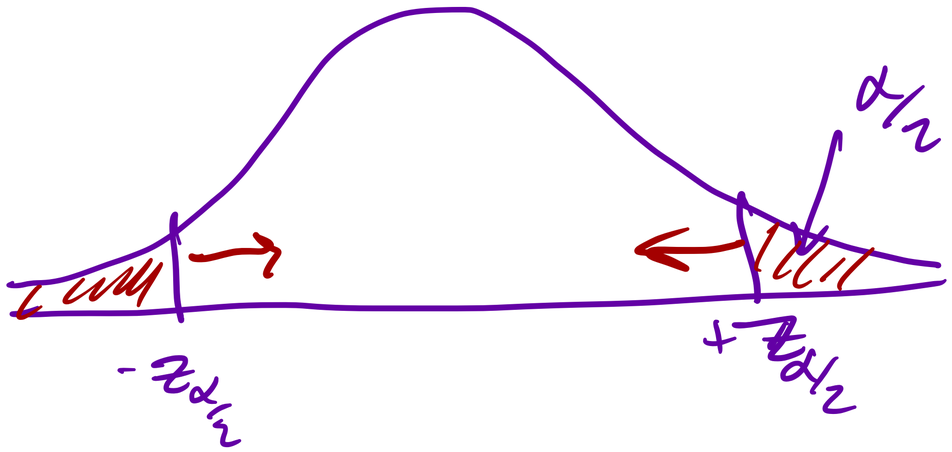
CI width: $2 \cdot z_{\alpha/2} \cdot SD$

- **Looking forward to interpretation**: What does our confidence interval tell us ~~about the effectiveness~~ of the two advertisements?

$[-0.5, +0.1]$ (ish)

$\bar{X} < \bar{Y} \rightarrow$ $\bar{Y}$ is better?

contains 0 → so no statistically significant difference at $\alpha = 0.05$ confidence level



What happens if we increase α?

$\alpha \uparrow \Rightarrow z_{\alpha/2} \downarrow \Rightarrow$ CI width $\downarrow \Rightarrow$ 0 gets locked out

CSCI 3022

intro to data science with probability & statistics

Lecture 15
March 7, 2018

Introduction to Statistical Inference & Confidence Intervals

NOTE: It was L'Hospital who happily stole Johann Bernoulli's work & published it as his own. The next time you take an indeterminant limit, remember that & call it "Bernoulli's Rule" instead!

TONY WUZ HERE

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Last time on CSCI 3022

- **Proposition**: If X is a normally distributed random variable with mean μ and standard deviation σ , then Z is a standard normal distribution if

$$Z = \frac{X - \mu}{\sigma} \quad \text{or} \quad X = \sigma Z + \mu$$

- **The Central Limit Theorem**: Let $X_1, X_2, \ldots, X_n$ be i.i.d. draws from some distribution. Then as n becomes large

non-normal or normal

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- A $100(1-\alpha)\%$ confidence interval for the mean μ when the value of σ is known is given by:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Statistical Inference

- **Goal**: Want to extract properties of an underlying population by analyzing sampled data
- **Last time we saw**:
  - How to determine a confidence interval for the population mean
  - How to determine a confidence interval for the population proportion
- **This time we'll see**:
  - How to put a confidence interval on the difference between means of two populations
  - How to put a confidence interval on the difference between proportions of two populations
  - How we can get a good numerical estimate of a CI using something called the Bootstrap

Difference between population means

- How do two sub-populations compare? In particular, are their means the same?
- **Classic Motivating Examples**:
  - Is a drug's effectiveness the same in children and adults? ✓
  - Does cigarette brand A contain more nicotine that cigarette brand B?
  - Does a class perform better when Professor C (his) teaches it or Professor D? (an)
  - Does email ad E generate more customers than email ad F?

Difference between population means

- How do two sub-populations compare? In particular, are their means the same?
- **Solution Process**:
  - Collect samples from both sub-populations, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$

Difference between population means

- How do two sub-populations compare? In particular, are their means the same?
- **Solution Process**:
  - Collect samples from both sub-populations, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$
- **Basic Assumptions**:
  - $X_1, X_2, \ldots, X_m$ is a random sample from a distribution with mean μ_1 and sd σ_1
  - $Y_1, Y_2, \ldots, Y_n$ is a random sample from a distribution with mean μ_2 and sd σ_2
  - The X and Y samples are independent of each other.

Difference between population means

- The natural estimator of $\mu_1 - \mu_2$ is the difference of the sample means $\bar{x} - \bar{y}$
- Is $\bar{x} - \bar{y}$ a *good* estimator for $\mu_1 - \mu_2$?
- The expected value of $\bar{X} - \bar{Y}$ is given by

$\bar{X}$ & $\bar{Y}$ r.v.

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$$ ✓✓

- The standard deviation of $\bar{X} - \bar{Y}$ is given by

$$SD[\bar{X} - \bar{Y}] = \sqrt{Var[\bar{X} - \bar{Y}]} = \sqrt{Var[\bar{X}] + Var[\bar{Y}]} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$ ✓

$(-1)^2$

Normal populations with known SDs

- If both populations are normal, then both $\bar{X}$ and $\bar{Y}$ are normally distributed.
- Independence of the two samples implies that the sample means are independent.
- Therefore, the difference between the means is normally distributed, for any sample sizes, with:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

estimator

expected value of est.

variance

Confidence Interval for the difference

- Standardizing $\bar{X} - \bar{Y}$ gives a standard normal random variable

messy z

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

But oh so nice! z

$$\sim N(0, 1)$$

- And so, we can compute a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$

$$\bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

central est. $\pm$ z-score * SD

Large sample CIs for the difference

- **Not surprisingly**, if both *m* and *n* are large, then our friend, the CLT, kicks in, and our confidence interval for the difference of means is valid, even when the populations are *not* normally distributed!

- **Furthermore**, if *m* and *n* are large, and we don't know the standard deviations, we can replace them with the sample standard deviations:

$$\sigma_1^2 \mapsto s_1^2 = \frac{1}{m-1}\sum_i (x_i - \bar{x})^2$$

$$\sigma_2^2 \mapsto s_2^2 = \frac{1}{n-1}\sum_j (y_j - \bar{y})^2$$

Confidence Interval for the Difference

$[-0.508, 0.068]$

- **Example**: Suppose you run two different email Ad campaigns over many days and record the amount of traffic driven to your website on days that each ad is sent. In particular, suppose that Ad 1 is sent on 50 different days and generates an average of 2 million page views per day with an sd of 1 million views, and Ad 2 is sent on 40 different days and generates an average of 2.25 million page views per day with an sd of a half million views. Find a 95% confidence interval for the difference in average page views per day (in units of millions of views).

$\bar{x} = 2$, $m = 50$, $s_1 = 1$

$\bar{y} = 2.25$, $n = 40$, $s_2 = 0.5$

$z_{\alpha/2} = z_{.05/2} = 1.96$

$$95\%\ CI = (\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

$$= (2 - 2.25) \pm 1.96 \cdot \sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}} = [-0.508, 0.068]$$

Confidence Interval for the Difference

- **Example**: Suppose you run two different email Ad campaigns over many days and record the amount of traffic driven to your website on days that each ad is sent. In particular, suppose that Ad 1 is sent on 50 different days and generates an average of 2 million page views per day with an sd of 1 million views, and Ad 2 is sent on 40 different days and generates an average of 2.25 million page views per day with an sd of a half million views. Find a 95% confidence interval for the difference in average page views per day (in units of millions of views).

Confidence Interval for the Difference

HERE

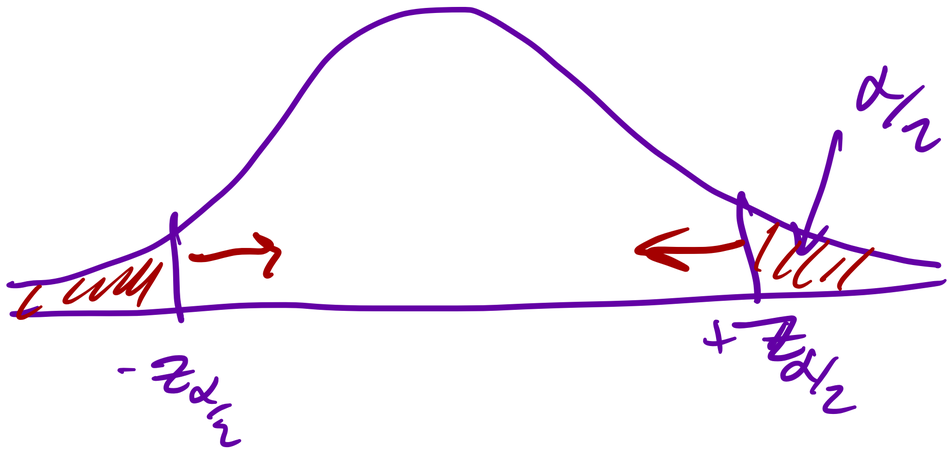
CI width: $2 \cdot z_{\alpha/2} \cdot SD$

- **Looking forward to interpretation**: What does our confidence interval tell us ~~about the effectiveness~~ of the two advertisements?

$[-0.5, +0.1]$ (ish)

$\bar{X} < \bar{Y} \rightarrow$ $\bar{Y}$ is better?

contains 0 → so no statistically significant difference at $\alpha = 0.05$ confidence level



What happens if we increase α?

$\alpha \uparrow \Rightarrow z_{\alpha/2} \downarrow \Rightarrow$ CI width $\downarrow \Rightarrow$ 0 gets locked out

Difference Between Population Proportions

- What if we want to compare population proportions?
- Suppose that a sample of size *m* is selected from the first population and a sample of size *n* is selected from the second population.
- Let X denote the number of units with the characteristic in population 1 (number of "successes") and Y denote the number of units with the characteristic in population 2.
- Reasonable estimators for the population proportions are: $\hat{p}_1 = \frac{X}{m}$ $\hat{p}_2 = \frac{Y}{n}$
- The natural estimator for the difference between population proportions $p_1 - p_2$ is

$\hat{p}_1 - \hat{p}_2$ ← estimate of $p_1 - p_2$

$p_1 - p_2$ ↑ real thing wanted

Difference Between Population Proportions

- Now, let $\hat{p}_1 = \frac{X}{m}$ and $\hat{p}_2 = \frac{Y}{n}$ where $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$

- Assuming that X and Y are independent, we can show that

$$E[\hat{p}_1 - \hat{p}_2] = E[\hat{p}_1] - E[\hat{p}_2] = E\left[\frac{X}{m}\right] - E\left[\frac{Y}{n}\right] = \frac{1}{m} m p_1 - \frac{1}{n} n p_2 = p_1 - p_2$$

- The standard deviation is approximated well by

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

I know $var\left(\hat{p}_1 - \hat{p}_2\right) = var\left(\hat{p}_1\right) + var\left(\hat{p}_2\right)$

next

Difference Between Population Proportions

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \text{var}(\hat{p}_1) + \text{var}(\hat{p}_2)$$

$$= \text{var}\left(\frac{X}{m}\right) + \text{var}\left(\frac{Y}{n}\right)$$

$$= \frac{1}{m^2}\text{var}(X) + \frac{1}{n^2}\text{var}(Y)$$

$$= \frac{1}{m^2} m\, p_1(1-p_1) + \frac{1}{n^2} n\, p_2(1-p_2)$$

$$= \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}$$

$$\text{St.dev} = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$$

CIs for the Difference of Proportions

- The $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is then given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$$

CIs for the Difference of Proportions

$z_{0.005} = 2.576$

$\frac{76}{154} \approx 0.494$

$\frac{98}{164} \approx 0.598$

- **Example**: A study was published in the New Engl. J. of Med. in 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemo and radiation. Of 154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least 15 years. **What is the 99% confidence interval for this difference of proportions?**

(1) (2)

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

$\alpha = 0.01$ $\quad z_{0.005} = 2.576$

chemo only: $\frac{76}{154} = \hat{p}_1 \approx 0.494$ $\quad m = 154$

hybrid: $\frac{98}{164} = \hat{p}_2 \approx 0.598$ $\quad n = 164$

$$0.494 - 0.598 \pm 2.576\sqrt{\frac{0.494(1-0.494)}{154} + \frac{0.598(1-0.598)}{164}}$$

Writing an Autograder

- Suppose you're a TA for Intro Data Science, and your professor-boss has tasked you with writing an autograder for a homework assignment which asks students to write a simulation to estimate the expected winnings in the game of Chuck-a-Luck.

(1) We know true mean of Chuck-a-Luck winnings → we calculated it!

(2) Run the student's code n times

(3) Compute a CI for the student's code's mean.

(4) Is the true mean in the CI?

Writing an Autograder

- Now suppose your professor-boss asks you to write an autograder for a simulation of Miniopoly. Specifically, she asks you to check solutions to the function that estimates the probability that a player goes Bankrupt within the first 20 turns of the game. How is this problem different from the Chuck-a-Luck problem? How should you proceed?

(1) This is about proportions.

(2) We don't have true proportion.
→ but we have a correct simulation.

(3) compute $\hat{p}_1$ (student) via n simulations
$\hat{p}_2$ (correct) via n simulations

(4) compute CI for diff in proportions.

(5) does it contain 0?

(6) If not, run codes again.

CSCI 3022

intro to data science with probability & statistics

Lecture 16
March 12, 2018

Introduction to Hypothesis Testing

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

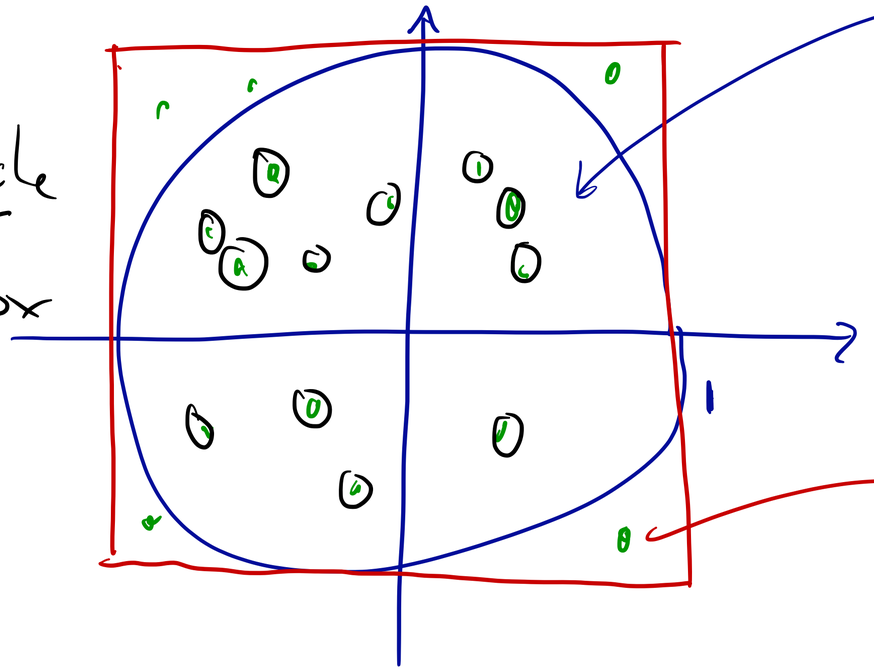
Stuff & Things

- **HW4** due on Friday.

Monte Carlo to be posted on Piazza.

$$P(\text{in circle}) = \frac{\text{Area of circle}}{\text{Area of box}} = \frac{\pi}{4}$$

Sample a jillion times to estimate this



Area $= \pi r^2 = \pi 1^2 = \pi$

Area $= 2 \times 2 = 4$

A thought experiment 🤔

- **Example**: After the introduction of the Euro, Polish mathematicians claimed that the new Belgian 1 Euro coin is not a fair coin. Suppose I hand you a Belgian 1 Euro coin. How could you decide whether or not it is fair?

Flip it a jillion times

[n]

→ record # H, # flips

Goal: anchor this test with stats!

Conclusion: if $\frac{\#H}{\#\text{flips}} \neq 0.5$ then not a fair coin.

Statistical Hypotheses

- **Definition**: A *statistical hypothesis* is a claim about the value of a parameter of a population characteristic.
- **Examples**:
  - Suppose the recovery time of a person suffering from disease D be normally distributed with mean μ_1 and standard deviation σ_1.
    Hypothesis: $\mu_1 > 10$ days.
  - Suppose μ_2 is the recovery time of a person suffering from disease D and given treatment for D. **Hypothesis**: $\mu_2 < \mu_1$
  - Suppose μ_1 is the mean internet speed for Comcast and μ_2 is the mean internet speed for Century Link. **Hypothesis**: $\mu_1 \neq \mu_2$

Null vs Alternative Hypotheses

- In any hypothesis testing problem, there are always two competing hypotheses that we consider:
  1. Null Hypothesis
     H_0
     status quo, default, e.g. coin is fair, $p = 0.5$
  2. Alternative Hypothesis
     H_1
     "research" hypothesis
     what we want to test
     e.g. coin is biased, $p \neq 0.5$
- The **objective** of hypothesis testing is to decide, based on the data that we've sampled, *whether the alternative hypothesis is actually supported by the data.*

The classic jury analogy

- Think about a jury in a criminal trial.

- When a defendant is accused of a crime, the jury is supposed to presume that the defendant is not guilty. **“Not guilty” is the null hypothesis**.

- The jury is then presented with **evidence** (data). If the evidence seems implausible under the assumption of not-guilty, they may **reject** the “not guilty” status, and claim that the defendant is likely guilty.

Null vs Alternative Hypotheses

alternative Hypothesis

- **Is there strong evidence for the alternative**?
- The burden of proof is placed on those that believe the alternative claim, just like in a jury.
- The initially favored claim, written as H_0, will not be rejected in favor of the alternative claim, written as H_1, unless the sample evidence provides a lot of support for the alternative.
- Two possible conclusions:
  1. Reject H_0 in favor of H_1
  2. Fail to reject H_0

Null vs Alternative Hypotheses

- **Why assume the Null Hypothesis**?
  - Sometimes we don't want to accept a particular assertion unless/until data can be shown to strongly support it.
  - Reluctance (measured in cost or time) to change.
- **Example**: A company is considering hiring a new advertising company to help generate traffic to their website. Under their current advertising they get, on average, 200K hits per day. With μ denoting the true average number of hits they'd get per day under the new company's advertising, they would not want to switch companies (because it would be costly) unless evidence strongly suggested that μ exceeds 200K.

Null vs Alternative Hypotheses

- **Example**: A company is considering hiring a new advertising company to help generate traffic to their website. Under their current advertising they get, on average, 200K hits per day. With μ denoting the true average number of hits they'd get per day under the new company's advertising, they would not want to switch companies (because it would be costly) unless evidence strongly suggested that μ exceeds 200K.

- An appropriate problem formulation would involve testing:

H_0: $\mu = 200{,}000$ (status quo) H_1: $\mu > 200{,}000$ (alternative)

- The conclusion that change is justified is identified with the alternative hypothesis and it would take conclusive evidence to justify rejecting H_0 and switching to the new company

"show me enough data to convince me."

Null vs Alternative Hypotheses

- The alternative to the Null Hypothesis $H_0 : \theta = \theta_0$ will look like one of the following assertions (or hypotheses):

①	$\theta > \theta_0$	$p > 0.5$
②	$\theta < \theta_0$	$p < 0.5$
③	$\theta \neq \theta_0$	$p \neq 0.5$

examples

- The equals sign is **always** the Null Hypothesis $\theta = \theta_0$
- The alternative hypothesis is the one for which we are seeking statistical evidence.

Test statistics and evidence

- **Def**: A test statistic is a quantity derived from the sample data and calculated assuming that the Null hypothesis is true. It is used in the decision about whether or not to reject the Null hypothesis.
- **Intuition**:
  - We can think of the test statistics as our evidence about the competing hypotheses.
  - We consider the test statistic under the assumption that H_0 is true by asking: **How likely would we obtain this evidence if the Null were true?**
- **Example**: To determine if the Belgian 1 Euro coin is fair you flip it 100 times and record the number of Heads. What is the test statistic? What are the Null and alternative hypotheses?

Test statistics and evidence

- **Example**: To determine if the Belgian 1 Euro coin is fair you flip it n times and record the number of Heads. What is the test statistic? What are the Null and alternative hypotheses?

Test statistic: $\hat{p} = \frac{\#\text{Heads}}{100}$ proportion of heads in our data.

H_0: $p = 0.5$ Under the null, $\hat{p} = \frac{X}{n}$ $X \sim Bin(n=100, p=0.5)$

H_1: $p \neq 0.5$

$\frac{p(1-p)}{n}$

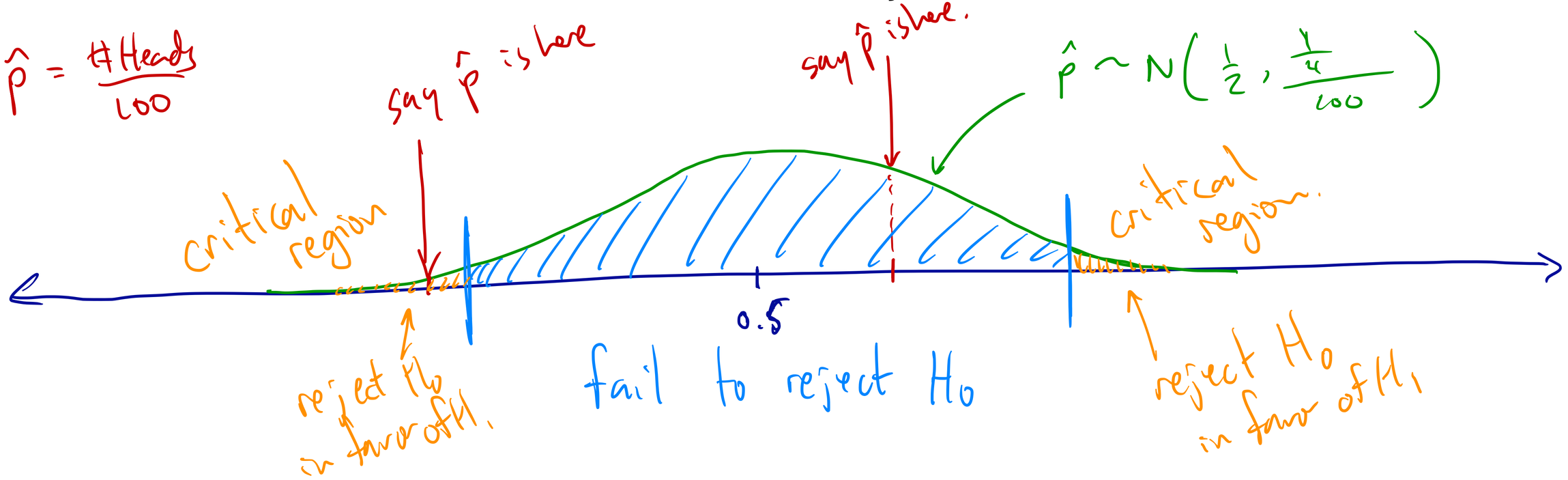
Under the null, $\hat{p} \sim N\left(0.5, \frac{0.5(1-0.5)}{100}\right)$

null!

How likely is it that our actual (observed) $\hat{p}$ occurs under $N\left(0.5, \frac{0.5(1-0.5)}{100}\right)$

Test statistics and evidence

- **Example**: To determine if the Belgian 1 Euro coin is fair you flip it n times and record the number of Heads. What is the test statistic? What are the Null and alternative hypotheses?
- **Question**: What would it take to convince you that the coin is not fair?

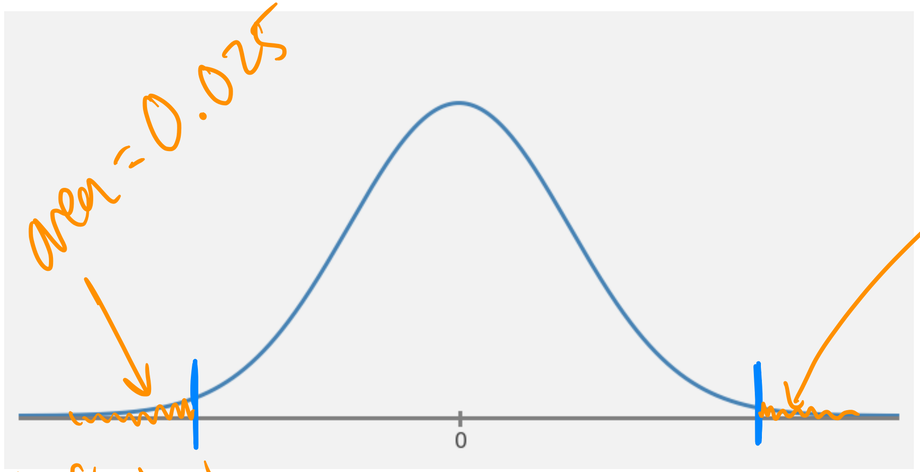


Test statistics and evidence

- **Example**: To determine if the Belgian 1 Euro coin is fair you flip it n times and record the number of Heads. What is the test statistic? What are the Null and alternative hypotheses?
- **Question**: What would it take to convince you that the coin is not fair?

Convert to a std. normal Z. $\alpha = 0.05$

two-tailed test

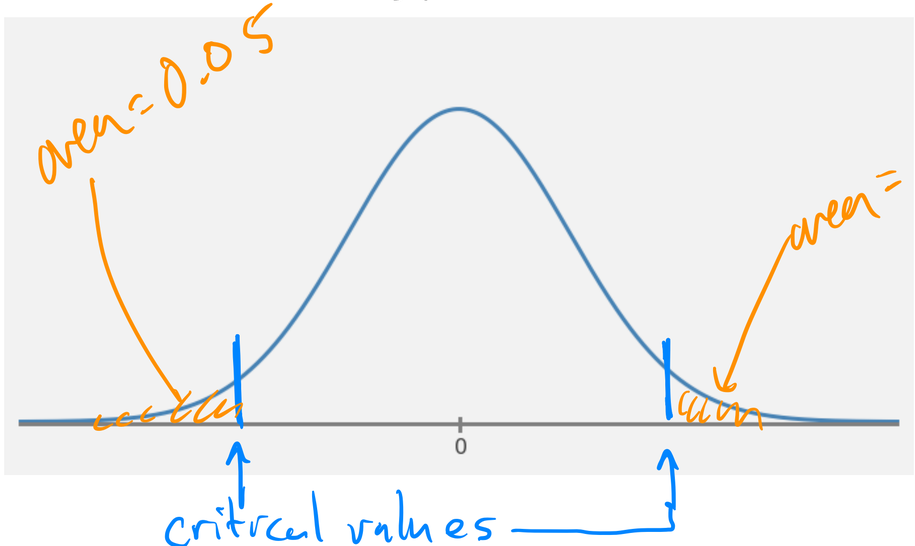


Rejection regions and significance level

- **Example**: To determine if the Belgian 1 Euro coin is fair you flip it n times and record the number of Heads.
- **Def**: The **rejection region** is a range of values of the test statistic that would lead you to **reject** the Null hypothesis.
- **Def**: The **significance level** α indicates the largest probability of the test statistic occurring under the Null hypothesis that would lead you to reject the Null hypothesis.

two-tailed test.

▨ – rejection region



$\alpha = 0.10$

really intuitive!

see next slides

Detecting Biased Coins

- **Example**: To test if the Belgian 1 Euro coin is fair you flip it 100 times and get 38 Heads. Do you reject the Null at the .05 significance level or not?

$H_0: p = 0.5$

$H_1: p \neq 0.5$

$\hat{p} \sim N\left(\frac{1}{2}, \frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{100}\right)$

CLT

$\sigma^2 = \frac{\frac{1}{2}\cdot\frac{1}{2}}{100}$

$\sigma = \frac{\frac{1}{2}}{10} = \frac{1}{20}$

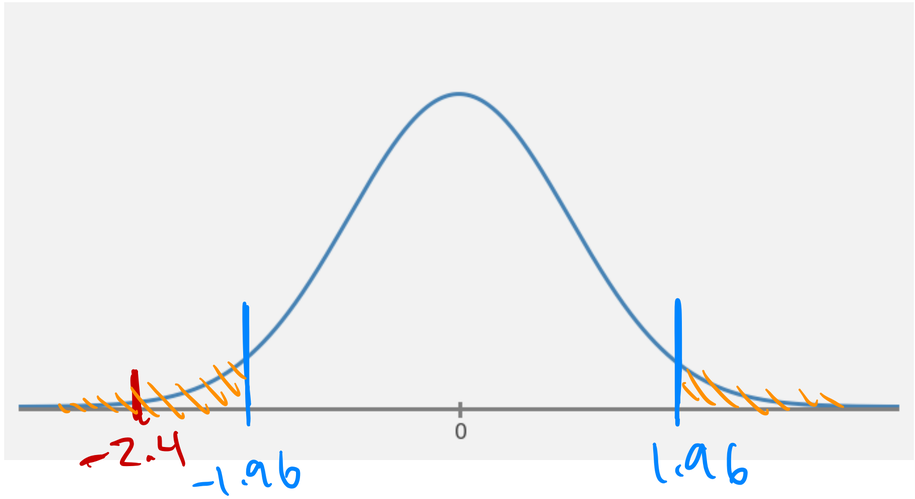
$Z = \frac{X-\mu}{\sigma}$

$= \frac{0.38 - 0.5}{\frac{1}{20}} = -2.4$

$\alpha = 0.05$

$z_{\alpha/2} = 1.96$

$-z_{\alpha/2} = -1.96$



Detecting Biased Coins

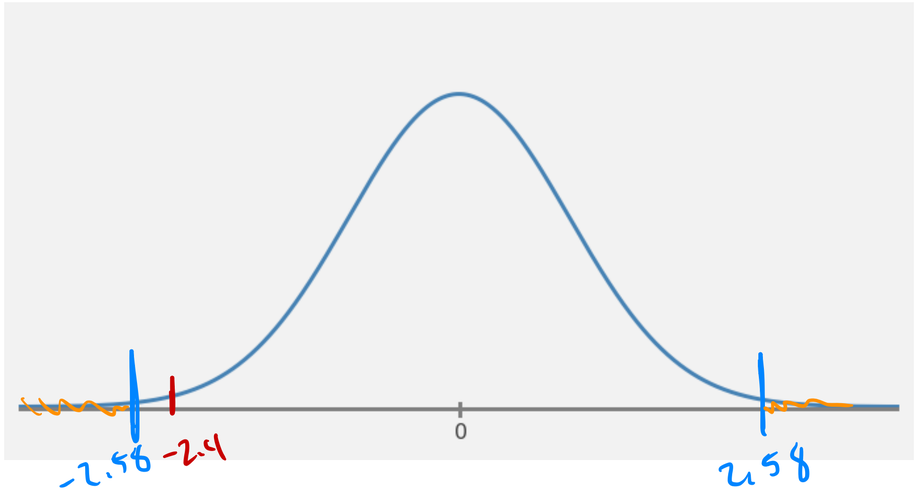
- **Example**: To test if the Belgian 1 Euro coin is fair you flip it 100 times and get 38 Heads. Do you reject the Null at the .01 significance level or not?

$\alpha = 0.01$ $z_{\alpha/2} = z_{0.005} = 2.58$

$-z_{0.005} = -2.58$

prev slide:

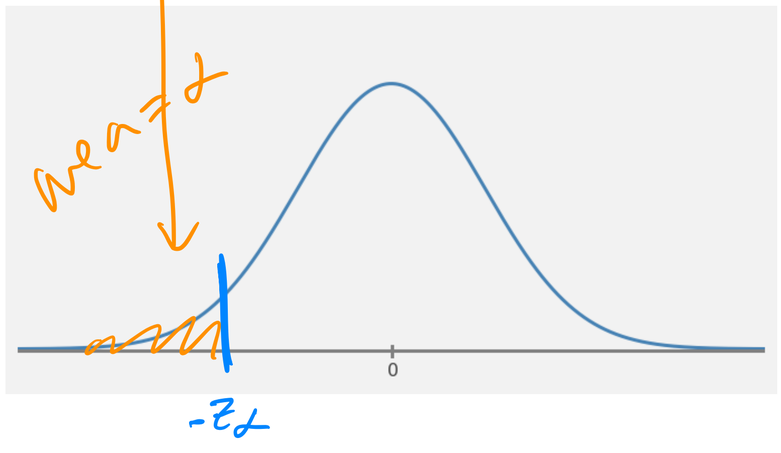
test statistic = -2.4

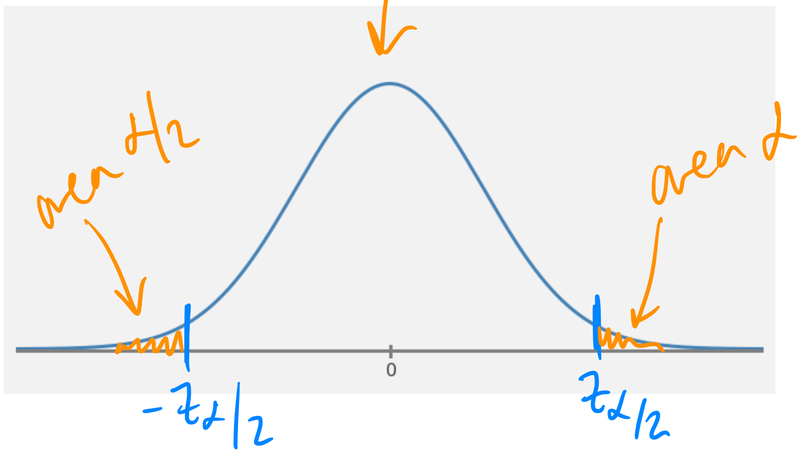


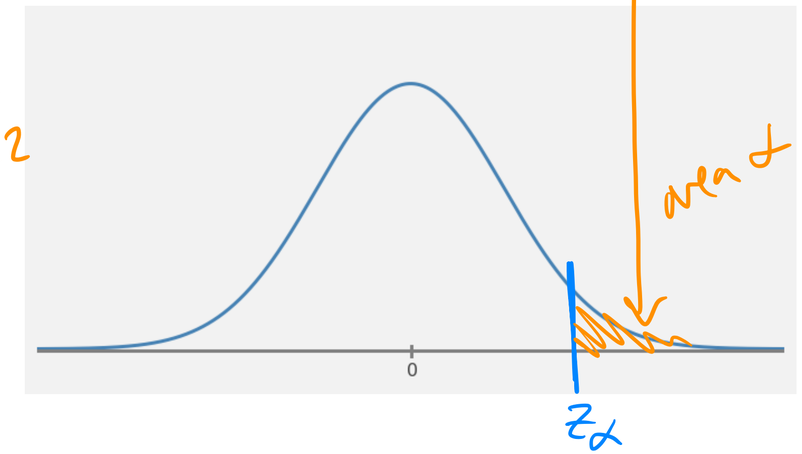
Different tests for different hypotheses

- The coin example was an exampled of a **two-tailed hypothesis test**, because we would have rejected the Null hypothesis had the coin been been biased towards heads OR tails.

Alternative Hypothesis	Rejection Region
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$
$H_1 : \theta \neq \theta_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$







Switching advertising strategies

$H_0: \mu = 200$ $H_1: \mu > 200$

- **Example**: Suppose a company is considering hiring a new outside advertising company to help generate traffic to their website. Under their current advertising they get, on average, 200 thousand hits per day with a standard deviation of 50 thousand hits per day. You decide to hire the new ad company for a 30 day trial. During those 30 days, your website gets 210 thousand hits per day. Perform a hypothesis test to determine if the new ad campaign outperforms the old one at the .05 significance level.

$z_\alpha = z_{0.05} = 1.645$

CLT $N\left(\mu, \frac{\sigma^2}{n}\right)$

If null H_0 were true

$\bar{X} \sim N\left(200, \frac{50^2}{30}\right)$

$Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$

$\frac{210 - 200}{50/\sqrt{30}} = 1.1$

CSCI 3022

intro to data science with probability & statistics

Lecture 17
March 14, 2018

Introduction to *p*-values and hypothesis testing

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

SafeTrek™

Get Paid to Code

Build on SafeTrek's life-saving API for some extra cash and professional experience.

SafeTrek is a tech company on a mission to help users feel safe and protected so that they can live life freely. SafeTrek recently released their connected safety API and are looking for students who can use it to build integrations that could make a big impact in people's lives.

But wait, there's more! SafeTrek is offering the opportunity to make over $300, complete freedom to build whatever you want, and reference letters to boost your portfolio.

Sound good?

Contact **Benjamin at SafeTrek** today to sign up and get started.

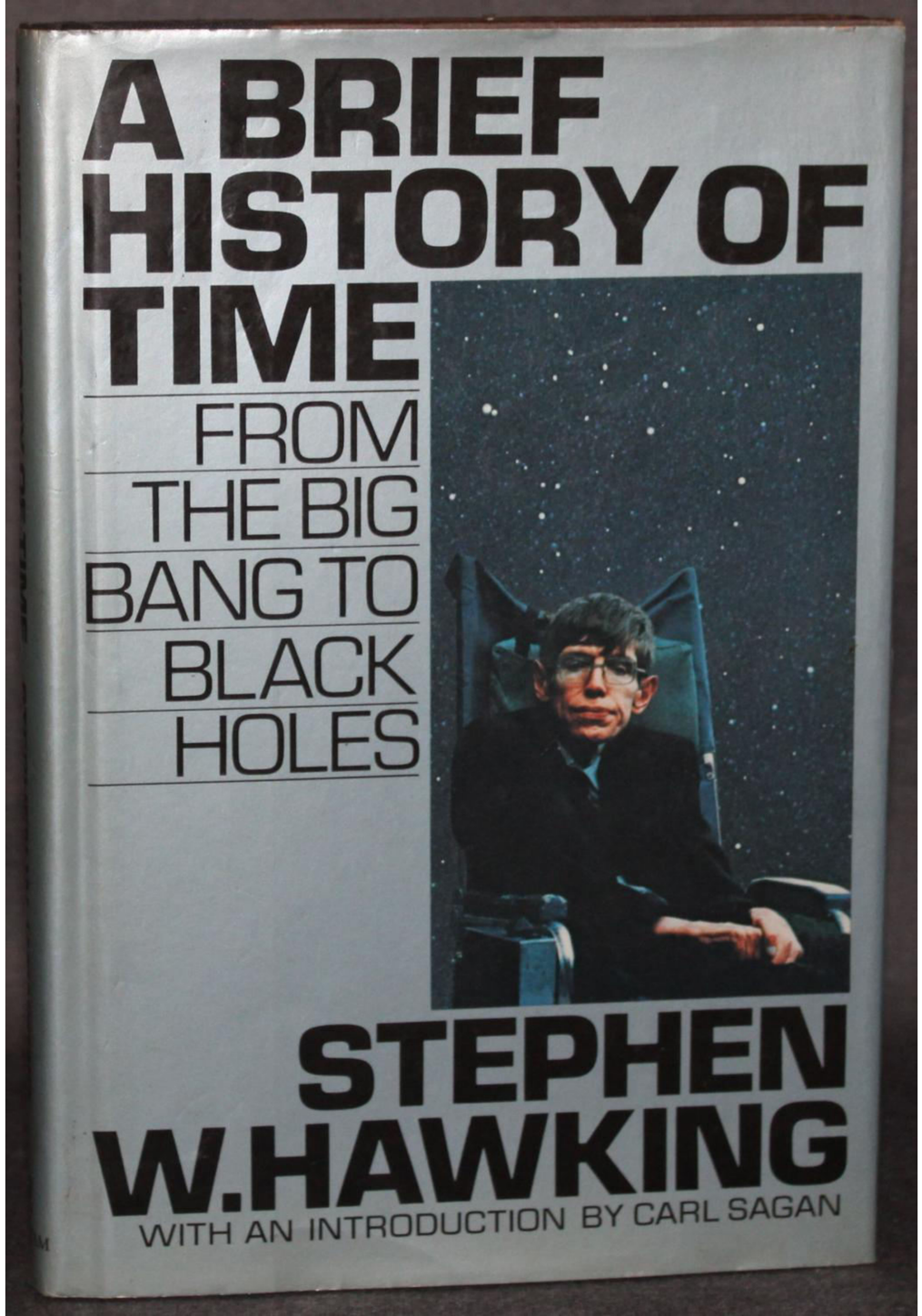
Space is limited, so act fast! 🏃‍♀️🏃‍♂️

"We recently opened our connected safety API to CU students and are offering a Junior Developer Program to pay those who complete integration projects with it.

These will be paid and students will gain experience with REST APIs and OAuth2 flows, and will own their own code."

- **Link 1**
- **Link2**

rest in peace
1942-2018



re c.

Switching advertising strategies

$H_0: \mu = 200$ $H_1: \mu > 200$

- **Example**: Suppose a company is considering hiring a new outside advertising company to help generate traffic to their website. Under their current advertising they get, on average, 200 thousand hits per day with a standard deviation of 50 thousand hits per day. You decide to hire the new ad company for a 30 day trial. During those 30 days, your website gets 210 thousand hits per day. Perform a hypothesis test to determine if the new ad campaign outperforms the old one at the .05 significance level.

$z_\alpha = z_{0.05} = 1.645$

CLT $N\left(\mu, \frac{\sigma^2}{n}\right)$

If null H_0 were true

$\bar{X} \sim N\left(200, \frac{50^2}{30}\right)$

$Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$

$\frac{210 - 200}{50/\sqrt{30}} = 1.1$

Important assumptions

- **Question**: What assumptions did we make in the previous example?

(1) Assumed that CLT would hold. $n=30$ samples (days)

(2) Assumed that we can represent the involved distributions as Normal.

Errors in hypothesis testing

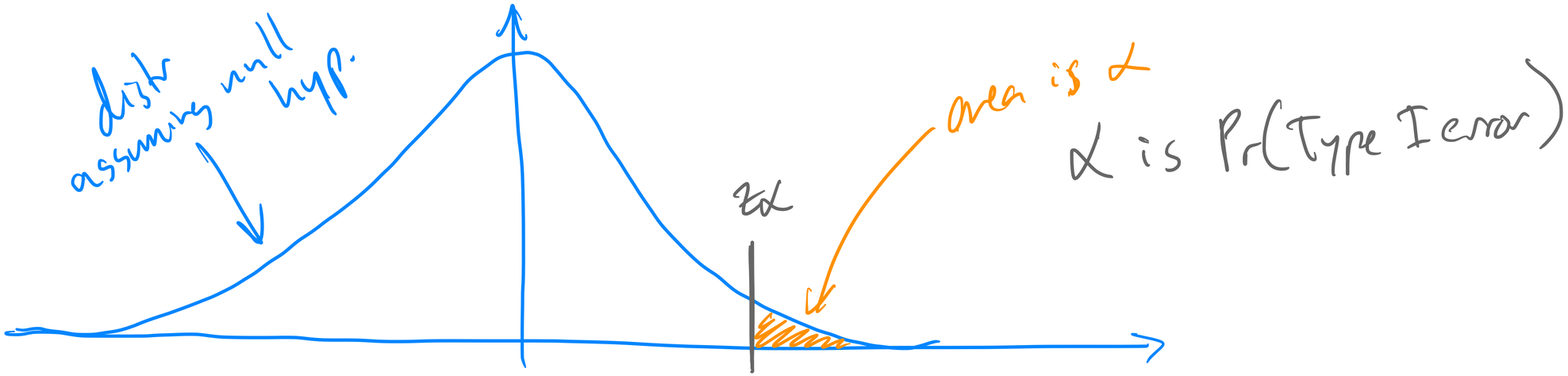
- **Definitions**:
- A **Type I Error** occurs when the Null hypothesis is rejected, but the Null hypothesis is in fact true ("**False Positive**")

Ads are same.
we conclude: different.

- A **Type II Error** occurs when the Null hypothesis is not rejected, but the Null hypothesis is in fact false ("**False Negative**")

Ads are diff.
we conclude: same.

- **Question**: What is the probability that we commit a **Type I Error**?



Errors in hypothesis testing

- **Definitions**:
- A **Type I Error** occurs when the Null hypothesis is rejected, but the Null hypothesis is in fact true (**False Positive**)
- A **Type II Error** occurs when the Null hypothesis is not rejected, but the Null hypothesis is in fact false (**False Negative**)
- **Question**: What is the probability that we commit a **Type I Error**?
- **Answer**: this is exactly the significance level α
- **Consequence**: choose α by considering willingness to risk a Type I error.

Rejection region refresher

- **Example**: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300K miles and standard deviation 150K miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250K miles.
- **Question 1**: What are the Null hypothesis and alternative hypothesis to test the claim that there is statistical evidence that 1999 Jettas made in Mexico have a smaller life expectancy that those made in Germany?

$H_0: \mu_{MX} = 300,000$

$H_1: \mu_{MX} < 300,000$

Rejection region refresher

$H_0: \mu = 300$

$H_1: \mu < 300$

1-tailed test.

- **Example**: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300K miles and standard deviation 150K miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250K miles.

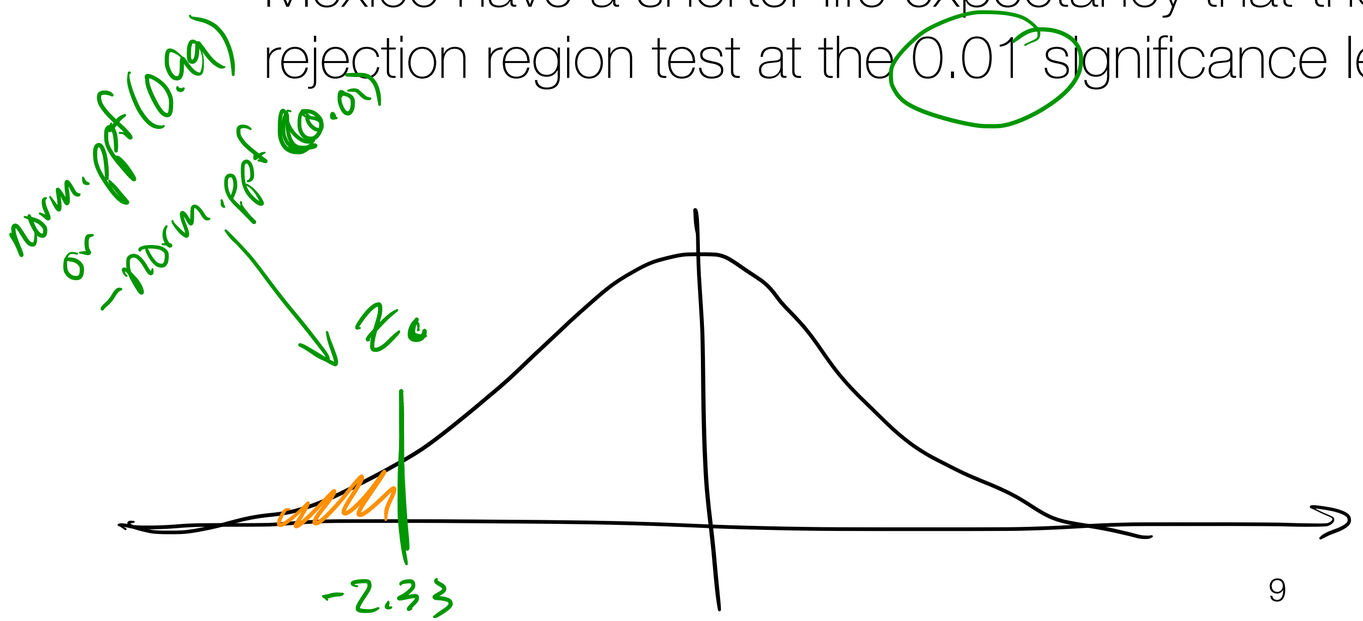
- **Question 2**: Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out a rejection region test at the 0.01 significance level.

$\alpha = 0.01$ $\quad n = 100$

$\mu = 300$

$\sigma = 150$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



Rejection region refresher

- **Example**: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300K miles and standard deviation 150K miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250K miles.

- **Question 2**: Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out a rejection region test at the 0.01 significance level.

data

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

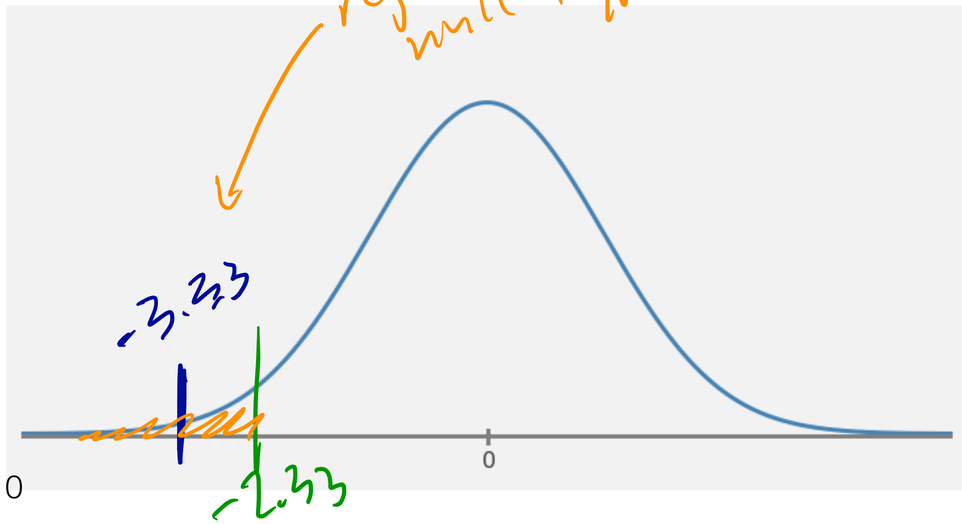
CLT

null hypothesis

sample size

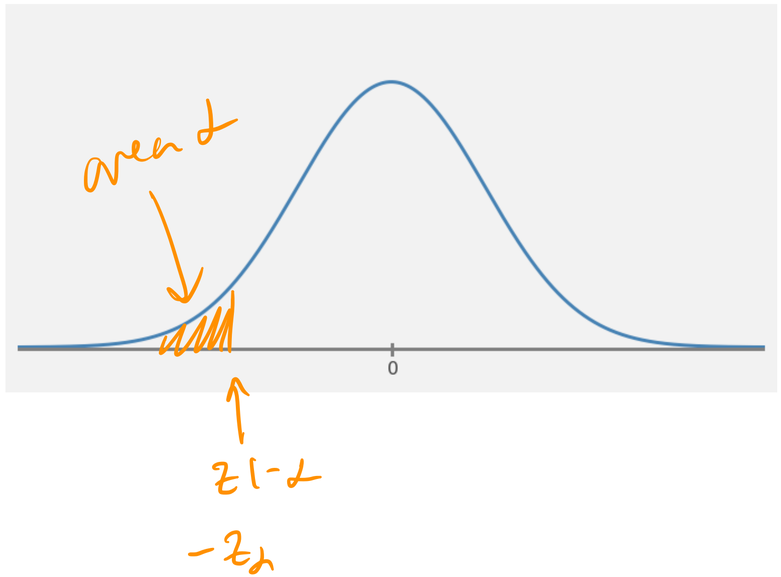
$$\frac{250 - 300}{150/\sqrt{100}} = -3.33$$

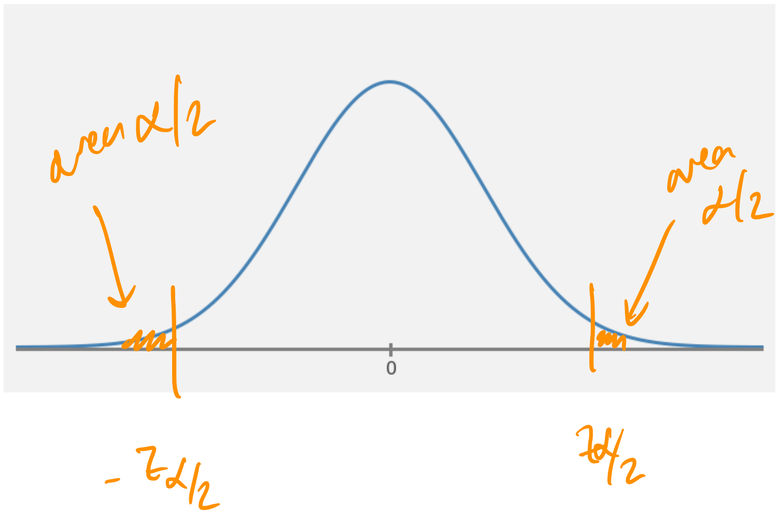
$$z_c = -2.33$$

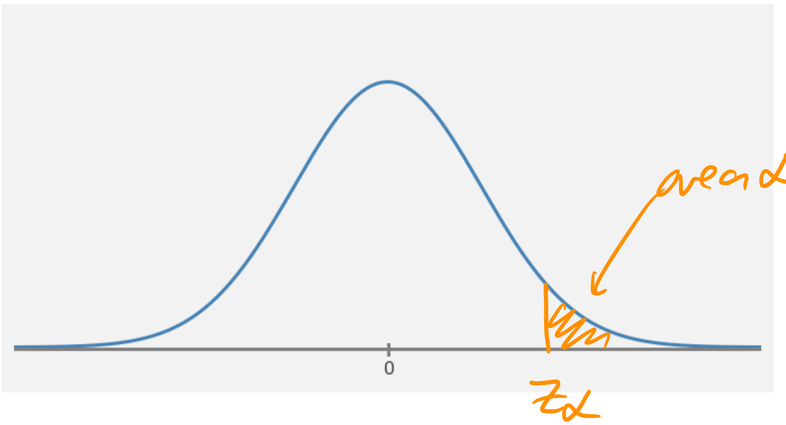


Rejection region & critical value summary

Alternative Hypothesis	Rejection Region
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$
$H_1 : \theta \neq \theta_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$







Critical region HT summary

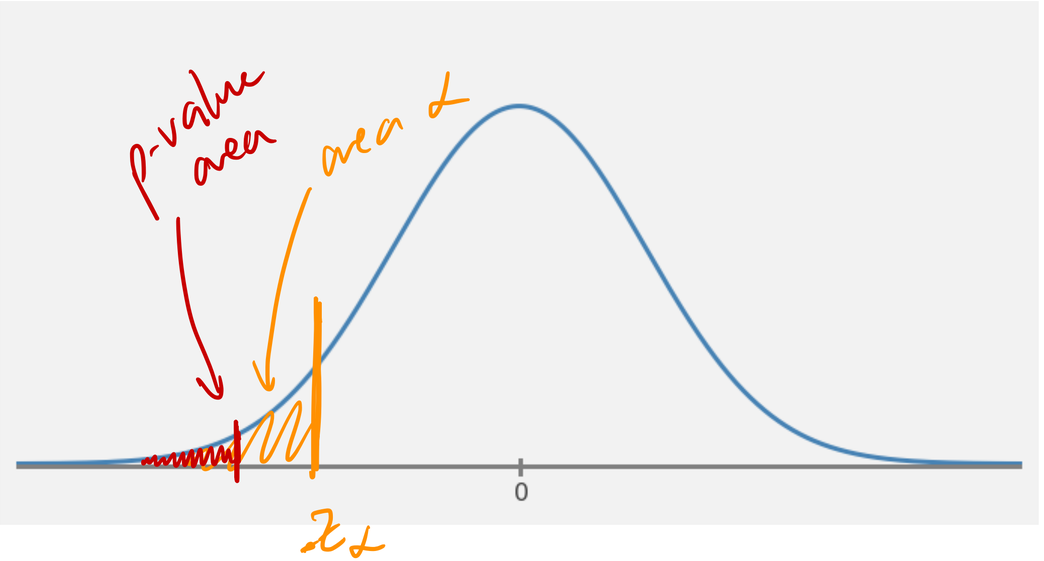
- **Critical Region** is region where test statistic has low probability under Null Hypothesis.
- Requires normally distributed data, or large enough sample for Central Limit Theorem.
- Under these assumptions we call this a Z-Test
- Rejecting the Null when the Null is true is called a Type I Error
- The probability of committing a Type I Error is α , the significance level of the test.
- Failing to reject the Null when the Null is false is called a Type II Error

Introduction to p-values

- Another way to view the critical region hypothesis test is through a so-called p-value
- This framework for HT is very popular in scientific study and reporting
- **Example**: Consider a lower-tail critical region test with the following hypotheses.

$$H_0: \quad \mu = \mu_0$$
$$H_1: \quad \mu < \mu_0$$

- The critical region test is:



p-values for various hypothesis tests

- **Def**: A p-value is the probability, under the Null hypothesis, that we would get a test statistic at least as extreme as the one we calculated.
- **Def**: For a lower-tailed test with test statistic x, the p-value is equal to $P(X \leq x|H_0)$
- Intuition: The p-value assesses the extremeness of the test statistic. The smaller the p-value, the more evidence we have against the Null hypothesis
- **Important Notes**:
  - The p-value is calculated under the assumption that the Null hypothesis is true
  - The p-value is always a value between 0 and 1
  - The p-value is NOT the probability that the Null is true!!

The p-value decision rule

- As before, select a significance level α before performing the hypothesis test
- Then the decision rule is:
  - If p-value $\leq \alpha$ then reject the Null hypothesis
  - If p-value $> \alpha$ then fail to reject the Null hypothesis
- Thus if the p-value exceeds the selected significance level then we cannot reject the Null hypothesis.

e.g. if $p = 0.1$ and $\alpha = 0.05$, we cannot reject.

- Note: The p-value can be thought of as the smallest significance level at which the Null hypothesis can be rejected.

Jetta life expectancy with p-values

- **Example**: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300K miles and standard deviation 150K miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250K miles.
- Is there sufficient evidence to conclude that 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out the p-value test at the 0.01 SL.

test statistic $\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{250 - 300}{150/\sqrt{100}} = -3.33$

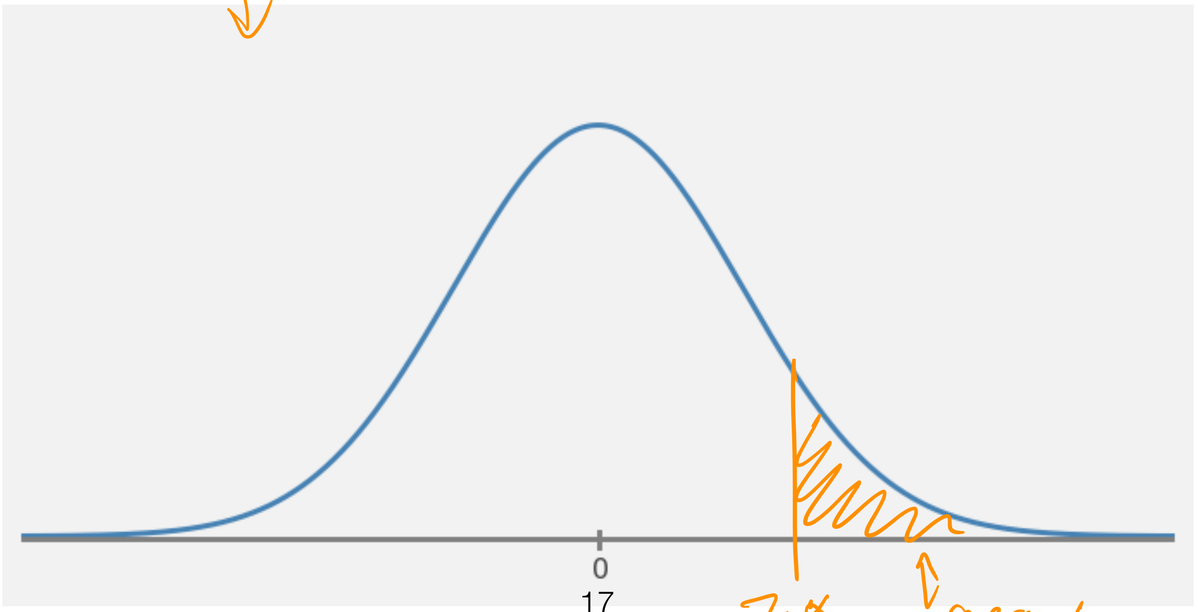
α

$P(Z \leq -3.33) \rightarrow$ CDF! p-value $= \Phi(-3.33) = 0.00043 \leq 0.01$

Reject Null Hypothesis

p-values for different z-tests

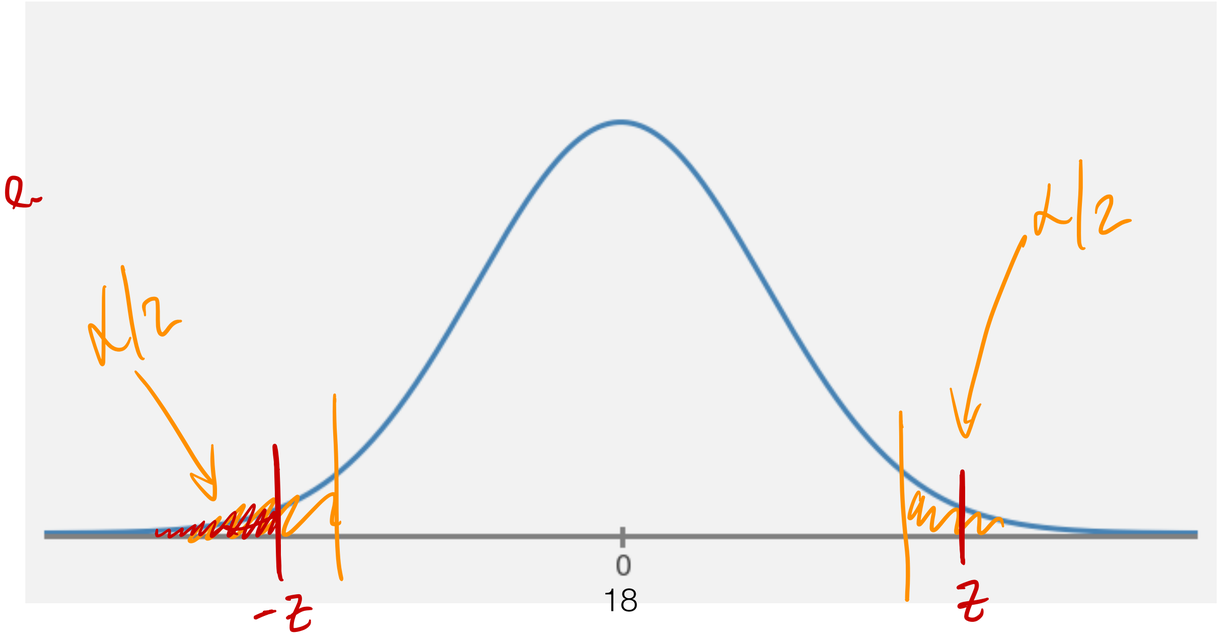
Alternative Hypothesis	Critical Region Level α Test	p-value Level α Test
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$	If $1-\Phi(z) \leq \alpha$
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$	If $\Phi(z) \leq \alpha$
$H_1 : \theta \neq \theta_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	Next slide.



p-values for different z-tests

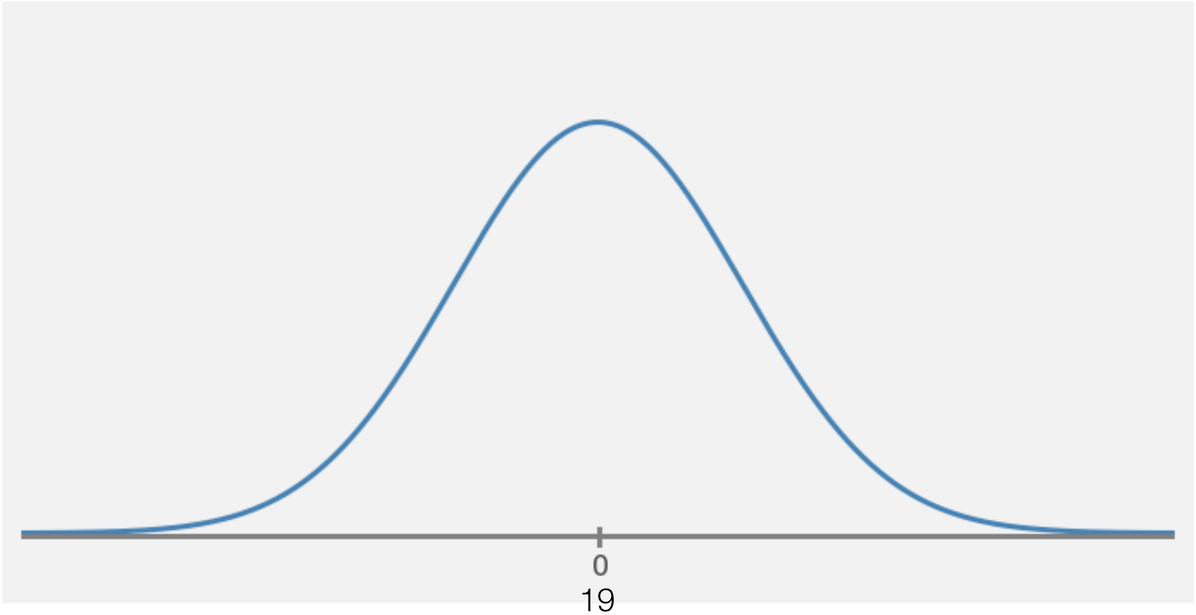
Alternative Hypothesis	Critical Region Level α Test	p-value Level α Test
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$	
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$	
$H_1 : \theta \neq \theta_0$	$z \leq -z_\alpha$ or $z \geq z_\alpha$	p-value: $2 \times \Phi(-\lvert z \rvert) \leq \alpha$

z is test statistic
z comes from data
α is our Type I error tolerance



p-values for different z-tests

Alternative Hypothesis	Critical Region Level α Test	p-value Level α Test
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$	
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$	
$H_1 : \theta \neq \theta_0$	$z \leq -z_\alpha$ or $z \geq z_\alpha$	



Is the Belgian 1 Euro biased?

- Example: To test if the Belgian 1 Euro coin is fair you flip it 100 times and observe 38 Heads. Perform a p-value Z-test at the .05 significance level.

$H_0: p = 0.5$ fair

$H_1: p \neq 0.5$ biased

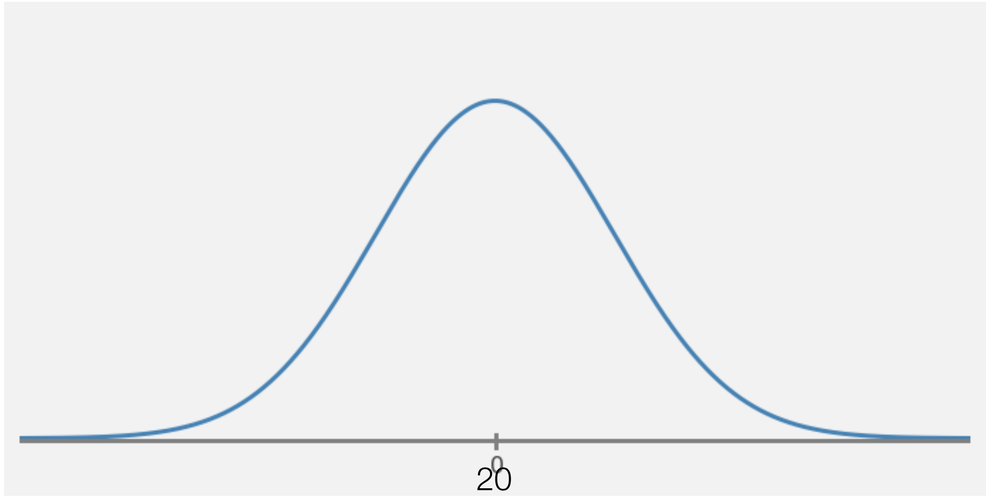
$\alpha = 0.05$ two sided

$2 \times \Phi(-|-2.4|)$

$= 0.0164$ p-value

$\hat{p} = 0.38$ (data)

$$z = \frac{0.38 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = -2.4$$



$0.0164 \leq 0.05$ (α)

Reject H_0!

Coin is not fair.

Two-Sample Testing for Difference of Means

- Suppose we want to test whether or not the difference of sample means from two populations is nonzero, or equal to a particular value.
- **Question**: What kinds of Null and alternative hypotheses might we want to test?

$H_0: \mu_1 - \mu_2 = C$ ← some constant

$H_1: \mu_1 - \mu_2 \neq C$

$H_1: \mu_1 - \mu_2 < C$

$H_1: \mu_1 - \mu_2 > C$

CLT

$$\frac{(\mu_1 - \mu_2) - C}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

Two-Sample Testing for Difference of Means

- Suppose we want to test whether or not the difference of sample means from two populations is nonzero, or equal to a particular value. c
- Assuming that our sample sizes are large enough, we can standardize our test statistics as:

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - c}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

- We can then compute an appropriate p-value in the usual way! Yay!

Two-Sample Testing for Difference of Means

$$z = \frac{(\mu_1 - \mu_2) - c}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

- **Example**: Data on calorie intake both for a sample of teens that reported that they do not typically eat fast food and another sample of teens who said they did usually eat fast food is as follows:

Fast Food	Sample Size	Sample Mean	Sample SD
No (pop 2)	663 n	2258 μ_2	1519 s_2
Yes (pop 1)	413 m	2637 μ_1	1138 s_1

- Does this data provide statistical evidence at the 0.05 (α) significance level that true average calorie intake for teens who typically eat fast food exceeds that of teens who do not typically eat fast food by more than 200 cals per day?

$H_0: \mu_1 - \mu_2 = 200$

$H_1: \mu_1 - \mu_2 > 200$

$$z = \frac{(2637 - 2258) - 200}{\sqrt{\frac{1138^2}{413} + \frac{1519^2}{663}}} = 2.20$$

p-value = ??

CSCI 3022

intro to data science with probability & statistics

Lecture 18
March 16, 2018

More *p*-values and hypothesis testing

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

Two-Sample Testing for Difference of Means

$$z = \frac{(\mu_1 - \mu_2) - c}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

- **Example**: Data on calorie intake both for a sample of teens that reported that they do not typically eat fast food and another sample of teens who said they did usually eat fast food is as follows:

Fast Food	Sample Size	Sample Mean	Sample SD
No (pop 2)	663 n	2258 μ_2	1519 s_2
Yes (pop 1)	413 m	2637 μ_1	1138 s_1

α

- Does this data provide statistical evidence at the 0.05 significance level that true average calorie intake for teens who typically eat fast food exceeds that of teens who do not typically eat fast food by more than 200 cals per day?

$H_0: \mu_1 - \mu_2 = 200$

$H_1: \mu_1 - \mu_2 > 200$

$$z = \frac{(2637 - 2258) - 200}{\sqrt{\frac{1138^2}{413} + \frac{1519^2}{663}}} \approx 2.20$$

p-value = ??

Two-Sample Testing for Difference of Means

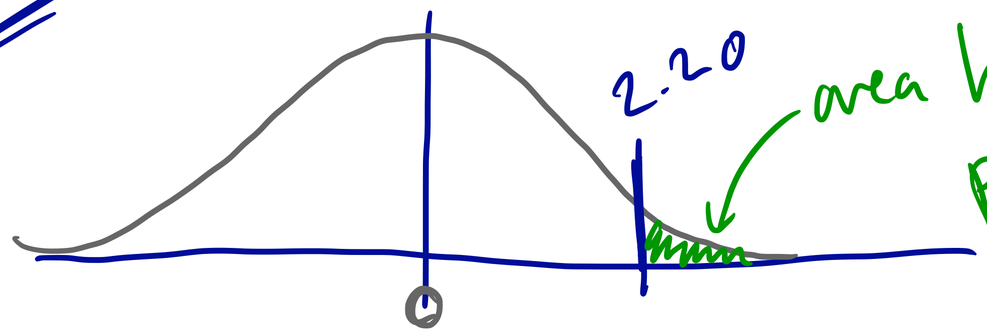
- **Example**: Data on calorie intake both for a sample of teens that reported that they do not typically eat fast food and another sample of teens who said they did usually eat fast food is as follows:

Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

$\alpha = 0.05$

- Does this data provide statistical evidence at the 0.05 significance level that true average calorie intake for teens who typically eat fast food exceeds that of teens who do not typically eat fast food by more than 200 cals per day?

$z = 2.20$



area here = p value

$Pr(Z \geq 2.20)$

Is p-value $< \alpha$

$Pr(Z \geq 2.20) < 0.05$

Common p-value misunderstandings

- **Misconception #1**: If p = 0.05, the Null hypothesis only has a 5% chance of being true.

WRONG.

p-value is $Pr(\text{obs our data} \mid H_0)$

Common p-value misunderstandings

- **Misconception #2**: If p is very small then your alt hypothesis is very likely to be significant.

Significance at what value of α?

Type I error rate

Nope

Common p-value misunderstandings

- **Misconception #3**: A statistically significant effect is equivalent to a substantial effect

we can tell from the data

Reject H_0 in favor of alt. Hyp. H_1

$\theta = \theta_0$

$\theta > \theta_0$

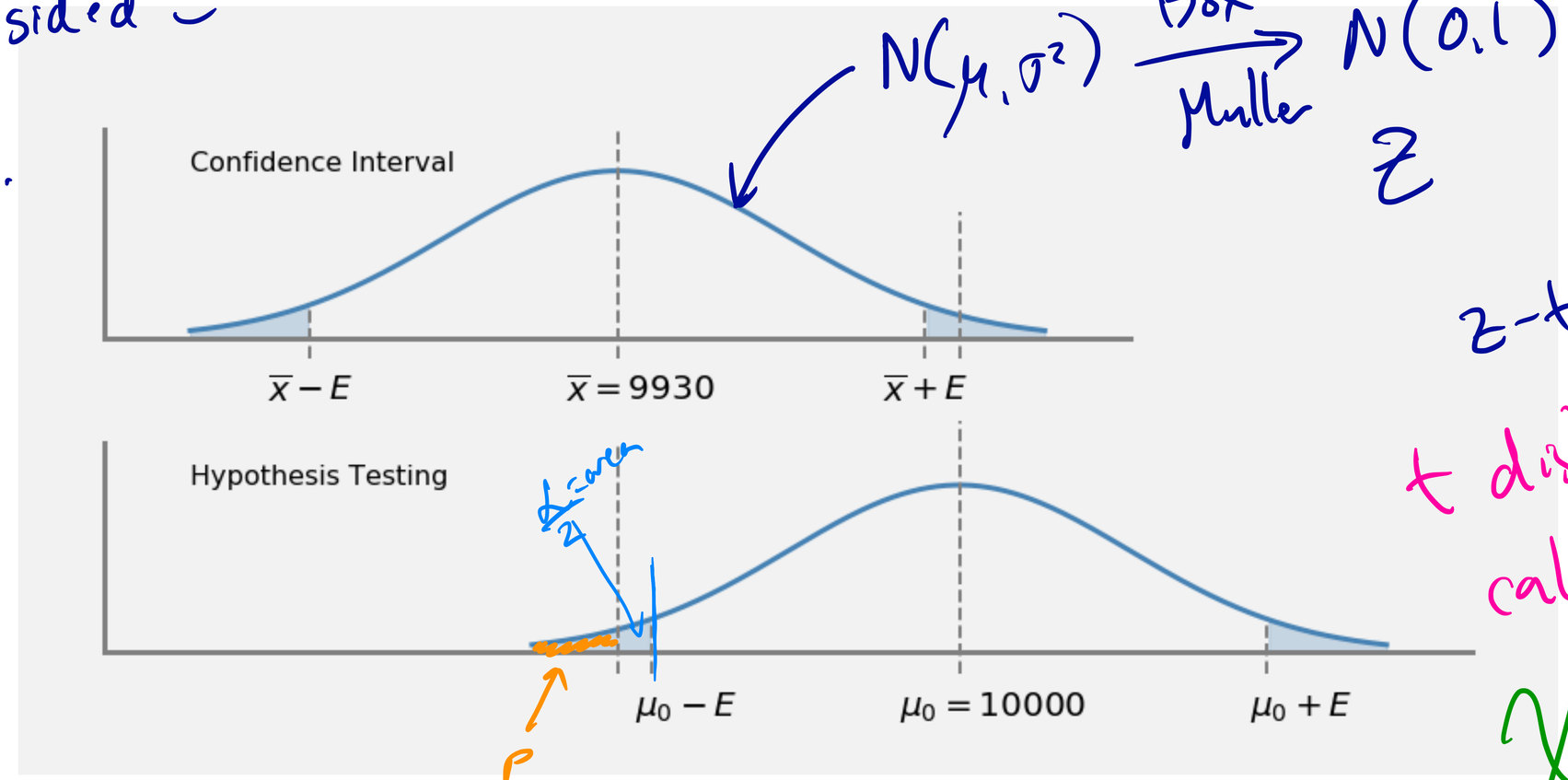
Nope

large effect.

"effect size" vs. "effect significance"

CIs vs Critical Regions vs P-Values

- Confidence Intervals, Critical Regions, and P-Values are three sides to the same coin.



Let’s notebooks!

CSCI 3022

intro to data science with probability & statistics

Lecture 19
March 19, 2018

Small sample size hypothesis testing

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

CSCI 3022

intro to data science with probability & statistics

Lecture 19
March 19, 2018

Small sample size hypothesis testing

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- HW5 posted tonight. Due the Friday *after* Spring Break.
- Dan's OH cancelled this Weds & Fri.

Previously on CSCI 3022

- Statistical inference for population mean **when data is normal** and n is large and…

"z tests"

- σ is known: $\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right) \sim N(0,1)$

$\bar{x} \pm z_{\alpha/n}\,\sigma/\sqrt{n}$

center — window

- σ is unknown: $\left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right) \sim N(0,1)$

"empirical std. dev."

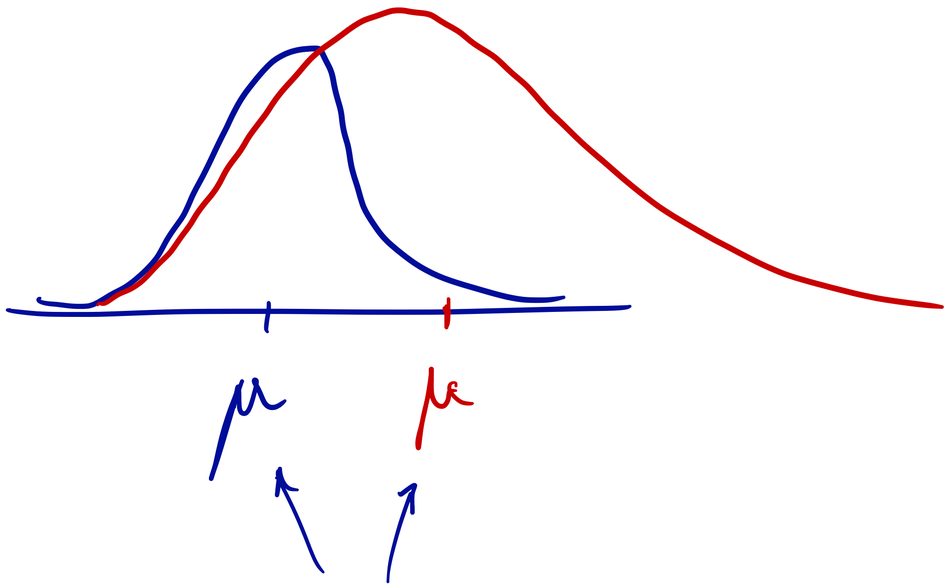
Previously on CSCI 3022

- Statistical inference for population mean **when data is NOT normal** and n is large and…

- σ is known: $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) \sim N(0,1)$

- σ is unknown: $\left(\frac{\bar{X}-\mu}{s/\sqrt{n}}\right) \sim N(0,1)$

"Thanks, CLT!"



Previously on CSCI 3022

- Statistical inference for population mean **when data is normal** and n is small and…

$n < 30$

- σ is known:

$$\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right) \sim N(0,1)$$

- σ is unknown:

???

The story so far, for means

- Thus far, we've talked about Hypothesis Testing & Confidence Intervals for the mean of a population in the following cases:

	"n is large" $n \geq 30$	"n is small" $n < 30$
Normal Data / Known σ		
Normal Data / Unknown σ (use s)		
Non-Normal Data / Known σ		
Non-Normal Data / Unknown σ		

- z-test
- t test (TODAY!)
- Bootstrap (after Spring Break)

Small-sample tests

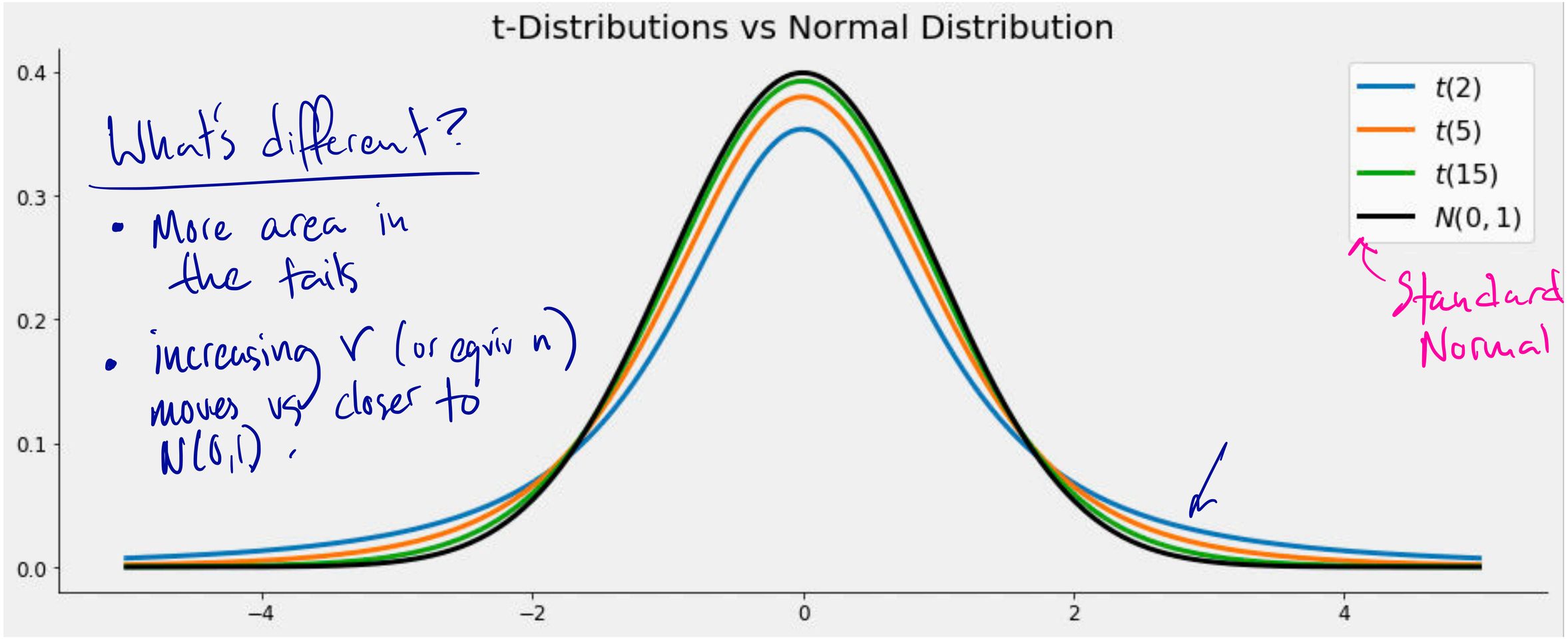
- When n is small we cannot invoke the Central Limit Theorem
- When n is small and the variance is unknown we need to do something else …
- When $\bar{X}$ is the sample mean of a random sample of size *n* from a normal distribution with mean μ, the random variable

$$\left(\frac{\bar{X}-\mu}{S/\sqrt{n}}\right)$$

follows a probability distribution called a **t-Distribution** with parameter $\nu = n - 1$ degrees of freedom.

The t-Distribution

- The following figure shows the pdf of some members of the family of t-Distributions



- What do you notice about these t-Distributions, compared with the Standard Normal curve?

Properties of t-Distributions

- Let t_ν denote the t-Distribution with parameter ν degrees of freedom

$\nu = n-1$

(Aside: \nu in LaTeX.)

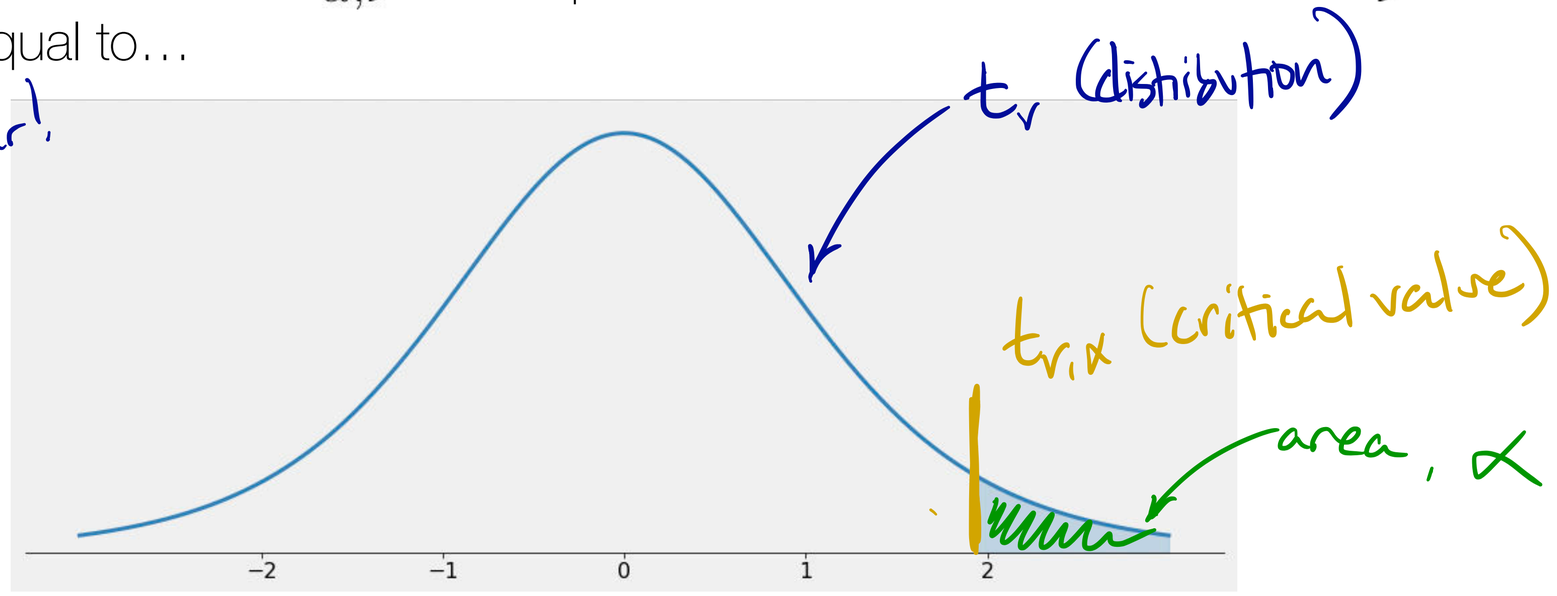
- Each t_ν -curve is bell-shaped and centered at 0
- Each t_ν-curve is more spread out than the standard normal distribution
- As ν increases, the spread of the corresponding t_ν-curve decreases
- As $\nu \to \infty$ the sequence of t_ν-curves approaches the standard normal curve

The t-critical value

- We can extend all of our inferential mechanics to the small-sample case by introducing the so-called t-critical value, which we denote $t_{\alpha,\nu}$

- **Definition**: the t-critical value $t_{\alpha,\nu}$ is the point such that the area under the t_ν -curve to the right of $t_{\alpha,\nu}$ is equal to…

This should look familiar! Very similar to our old friend, the z-test, and using z to critical values



- Example: $t_{0.05,6}$ is the t-critical value that captures the upper-tail area of 0.05 under the t curve with 6 degrees of freedom.

The t-confidence interval for the mean

- Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from the results of a random sample with of size n from a normal population with mean μ.

- Then a $100(1-\alpha)\%$ t-confidence interval for the mean μ is given by:

$$\left[\bar{x} - t_{\alpha/2,n-1}\, s/\sqrt{n},\ \bar{x} + t_{\alpha/2,n-1}\, s/\sqrt{n}\right]$$

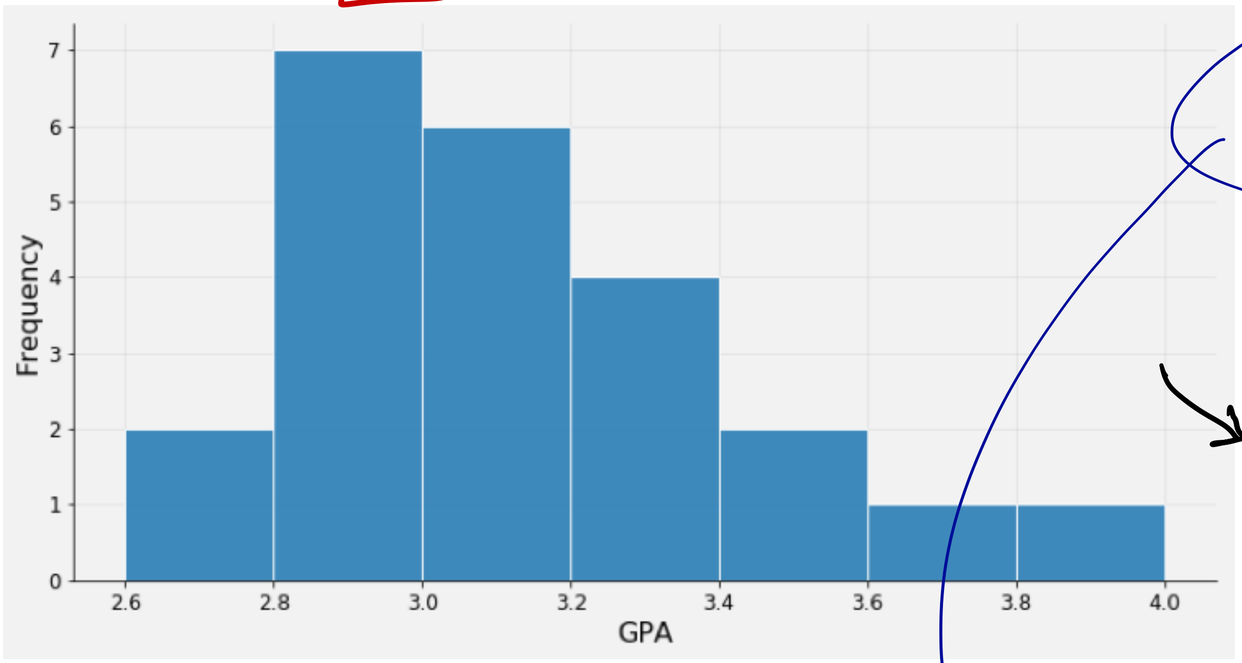
- Or more compactly:

$$\bar{x} \pm t_{\alpha/2,n-1}\, s/\sqrt{n}$$

CI

t-confidence interval example

- **Example**: Suppose the GPAs for 23 students have a histogram that looks as follows:



$n = 23$

$\bar{x} = 3.146$

$s = 0.308$

$\alpha = 0.1$

$\alpha/2 = 0.05$

$CI = \bar{x} \pm t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$

$t_{\alpha/2, n-1}$

stats.t.ppf(0.95, 23-1)

$= 1.717$

$n-1$

- The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Find a 90% confidence interval for the mean GPA.

$\alpha = 0.1$

"$(1-\alpha) \cdot 100$"% CI

$3.146 \pm 1.717 \cdot \frac{0.308}{\sqrt{23}}$

$\Rightarrow [3.033, 3.259]$

The t-Test, Critical Regions and P-Values

$H_0 : \quad \theta = \theta_0$

Alternative Hypothesis	Critical Region Level α Test
$H_1 : \quad \theta > \theta_0$	$t \geq t_{\alpha,\nu}$
$H_1 : \quad \theta < \theta_0$	$t \leq t_{\alpha,\nu}$
$H_1 : \quad \theta \neq \theta_0$	$(t \leq -t_{\alpha/2,\nu})$ or $(t \geq t_{\alpha/2,\nu})$

t test statistic looks just like z test statistics!

confidence

degrees of freedom

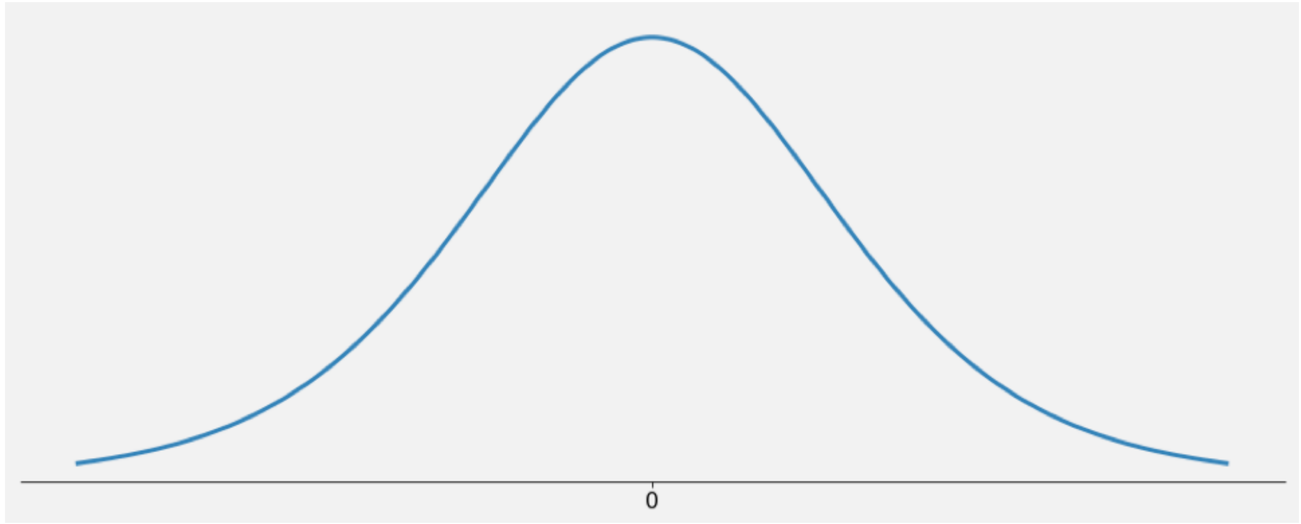
The only difference ... is n is small ($n < 30$)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

"standardized statistic"

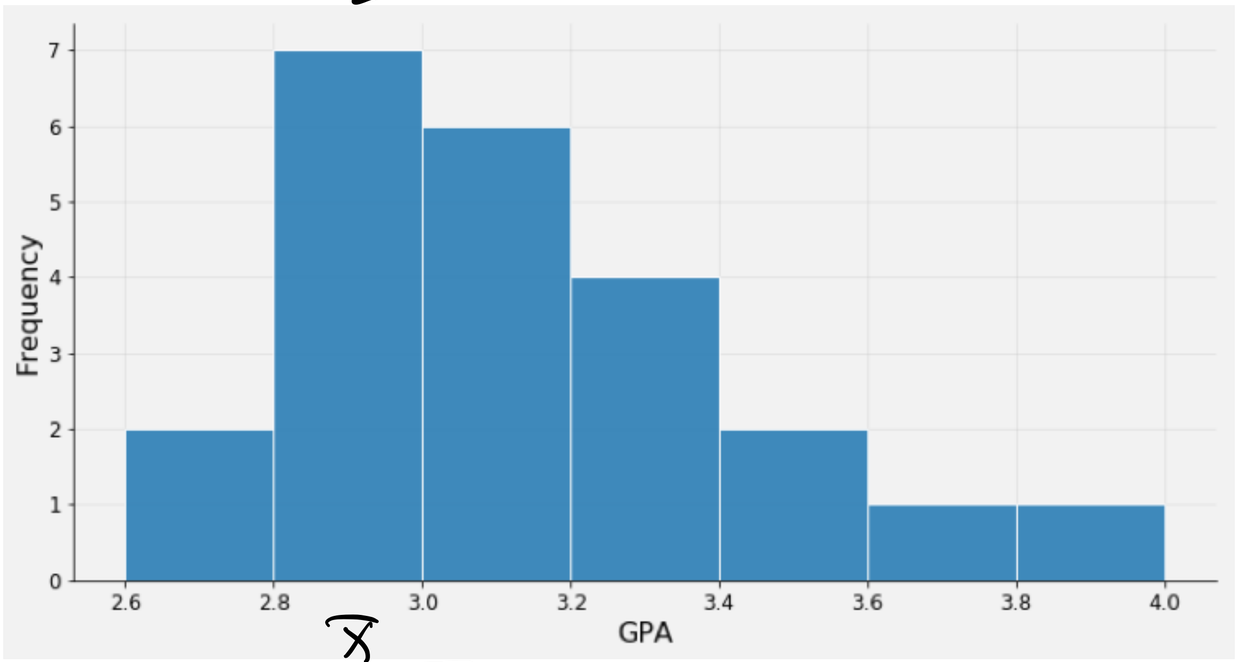
The t-Test, Critical Regions and P-Values

Alternative Hypothesis	P-Value Level α Test
$H_1 : \quad \theta > \theta_0$	$P(T \geq t \mid H_0) \leq \alpha$
$H_1 : \quad \theta < \theta_0$	$P(T \leq t \mid H_0) \leq \alpha$
$H_1 : \quad \theta \neq \theta_0$	$2 \min \{P(T \leq t \mid H_0),\ P(T \geq t \mid H_0)\} \leq \alpha$



t-Test example (p-value method)

- **Example**: Suppose the GPAs for 23 students have a histogram that looks as follows:



- The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Determine if there is sufficient evidence to conclude at the 0.10 significance level that the mean GPA is not equal to 3.30.

H_1: GPA $\neq$ 3.30

$\alpha = 0.1$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3.146 - 3.30}{0.308/\sqrt{23}} = -2.398$$

t-Test example (p-value method)

$2 \times$ stats.t.cdf(-2.398, 22) = 0.0254 < 0.10

-2.398: test statistic; 22: dof; 0.0254: p-value; 0.10: α

CSCI 3022

intro to data science with probability & statistics

Lecture 20
April 2, 2018

Small sample size hypothesis testing
and The Bootstrap

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

Stuff & Things

- HW5 due **this Friday**. ✓
- **New notetaker needed please**! Done
  - 1. Take notes as you normally would.
  - 2. Scan (with smartphone) after class and email to two of your peers.
- Questions about your HW grades? The graders are happy to explain!
  - sudeep.galgali@colorado.edu ✓
  - ajay.kedia@colorado.edu ✓

C+P
Overhead in OH
→ ASCII

H.C. 0 or F ✓
Course policy!

Previously on CSCI 3022

The story so far, for means

- Thus far, we've talked about Hypothesis Testing & Confidence Intervals for the mean of a population in the following cases:

	"n is large" $n \geq 30$	"n is small" $n < 30$
Normal Data / Known σ		
Normal Data / Unknown σ (use s)		
Non-Normal Data / Known σ		
Non-Normal Data / Unknown σ		

- z-test
- t test (TODAY!)
- Bootstrap (after Spring Break)

The t-Test, Critical Regions and P-Values

$H_0 : \quad \theta = \theta_0$

Alternative Hypothesis	Critical Region Level α Test
$H_1 : \quad \theta > \theta_0$	$t \geq t_{\alpha,\nu}$
$H_1 : \quad \theta < \theta_0$	$t \leq t_{\alpha,\nu}$
$H_1 : \quad \theta \neq \theta_0$	$(t \leq -t_{\alpha/2,\nu})$ or $(t \geq t_{\alpha/2,\nu})$

t test statistic looks just like z test statistics!

confidence

degrees of freedom = n - 1

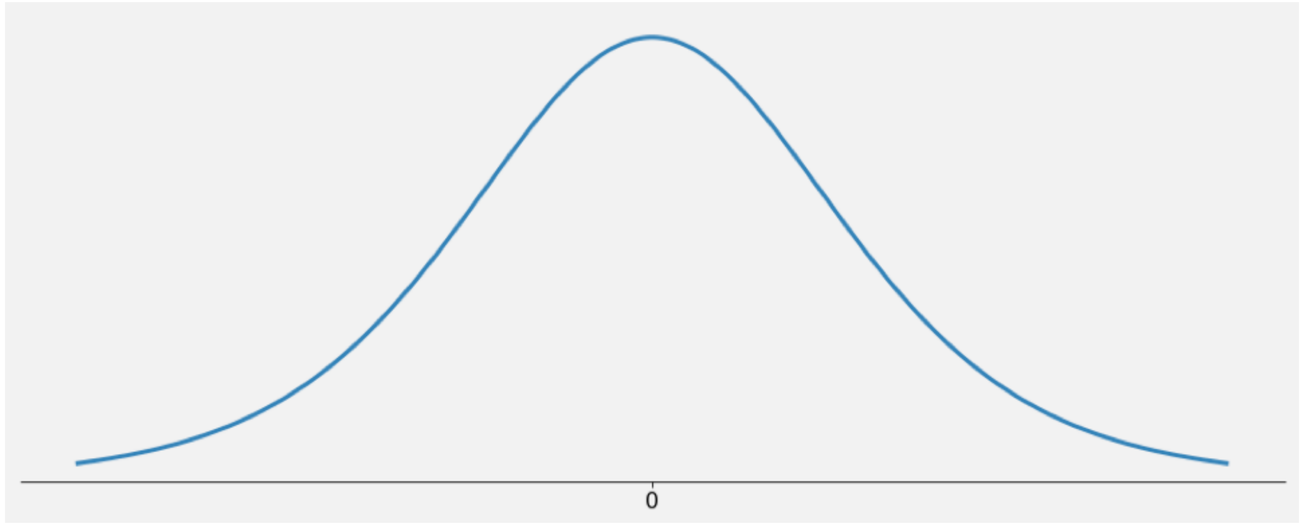
The only difference ... is n is small (n < 30)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

"standardized statistic"

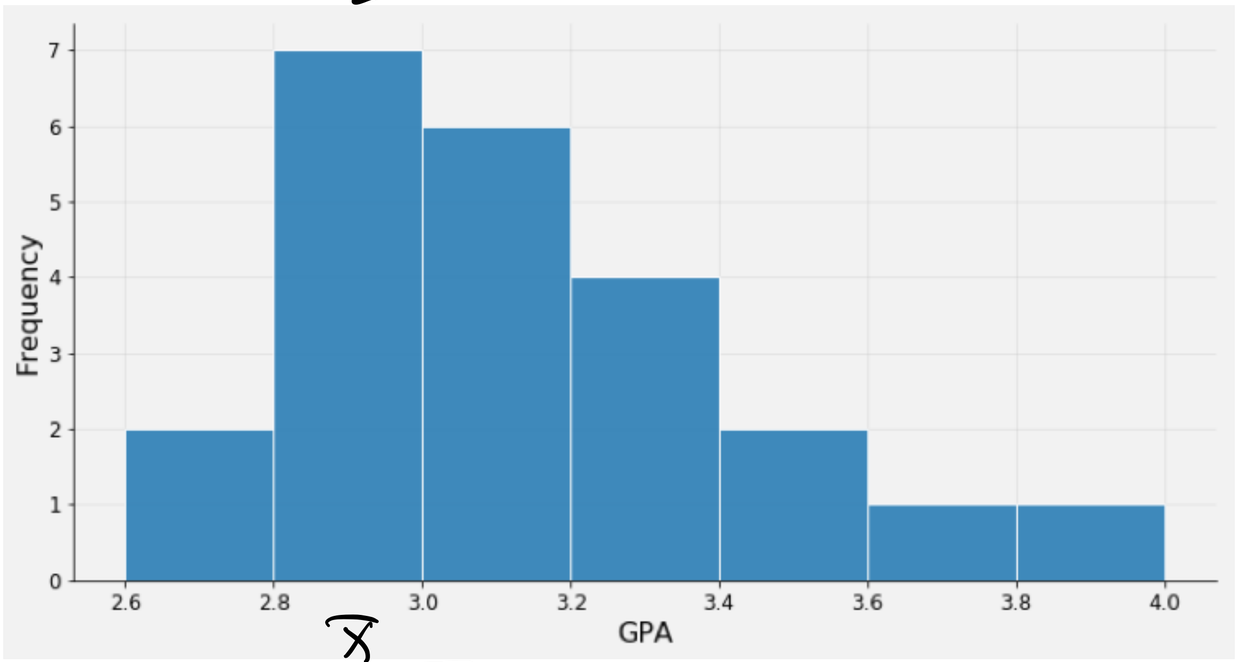
The t-Test, Critical Regions and P-Values

Alternative Hypothesis	P-Value Level α Test
$H_1 : \quad \theta > \theta_0$	$P(T \geq t \mid H_0) \leq \alpha$
$H_1 : \quad \theta < \theta_0$	$P(T \leq t \mid H_0) \leq \alpha$
$H_1 : \quad \theta \neq \theta_0$	$2 \min \{P(T \leq t \mid H_0),\ P(T \geq t \mid H_0)\} \leq \alpha$



t-Test example (p-value method)

- **Example**: Suppose the GPAs for 23 students have a histogram that looks as follows:



- The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Determine if there is sufficient evidence to conclude at the 0.10 significance level that the mean GPA is not equal to 3.30.

H_1: GPA $\neq$ 3.30

$\alpha = 0.1$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3.146 - 3.30}{0.308/\sqrt{23}} = -2.398$$

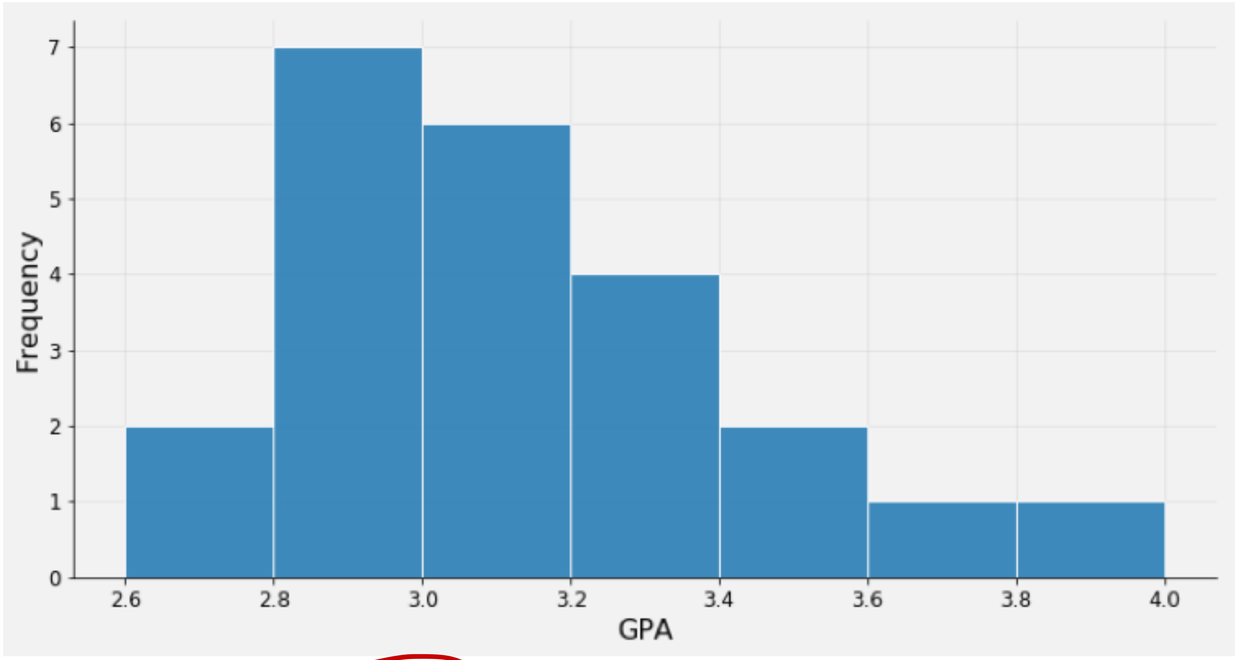
t-Test example (p-value method)

$2 \times$ stats.t.cdf(−2.398, 22) $= 0.0254 < 0.10$

−2.398: test statistic; 22: dof; 0.0254: p-value; 0.10: α

t-Test example (rejection region method)

- **Example**: Suppose the GPAs for 23 students have a histogram that looks as follows:



- The sample mean of the data is 3.146 and the sample standard deviation is 0.308. Determine if there is sufficient evidence to conclude at the 0.10 significance level that the mean GPA is not equal to 3.30.

$n = 23$

$s = 0.308$

$\bar{x} = 3.146$

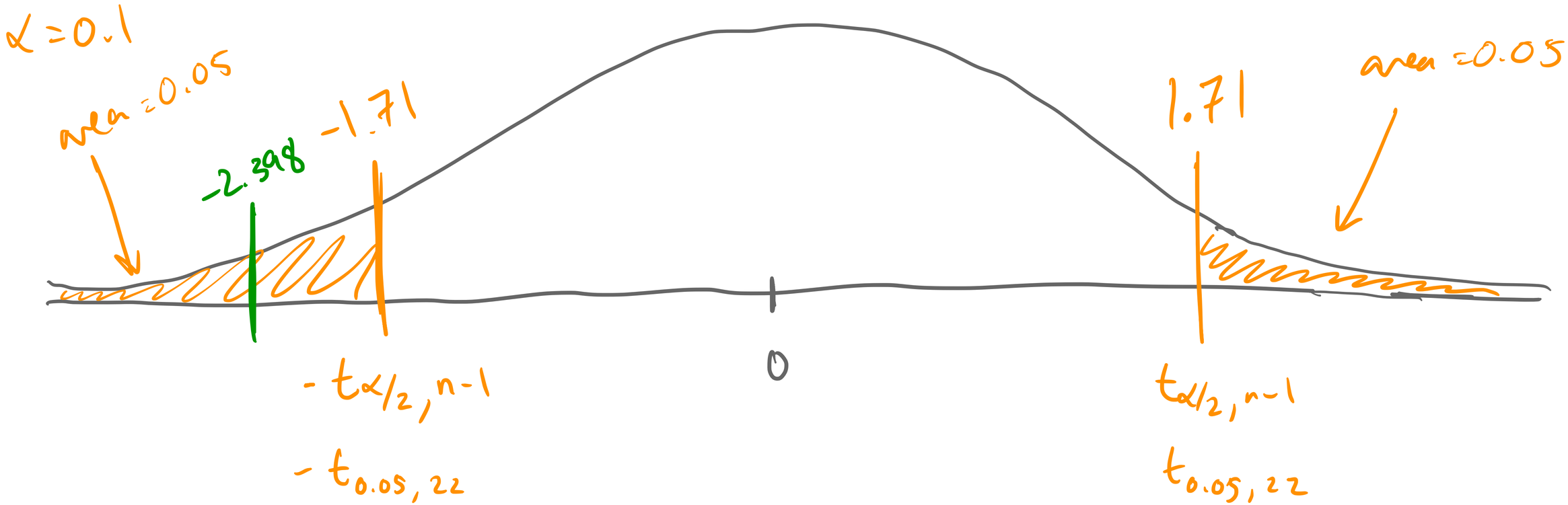
$\alpha = 0.1$

$\mu = 3.30$

$H_0: \mu = 3.30$

$H_1: \mu \neq 3.30$

t-Test example (rejection region method)



stats.t.ppf(0.95, 22) = 1.71

Prev. Slides $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = -2.398$ "test statistic"

In the rejection region! REJECT H_0.

Inference for *variances*

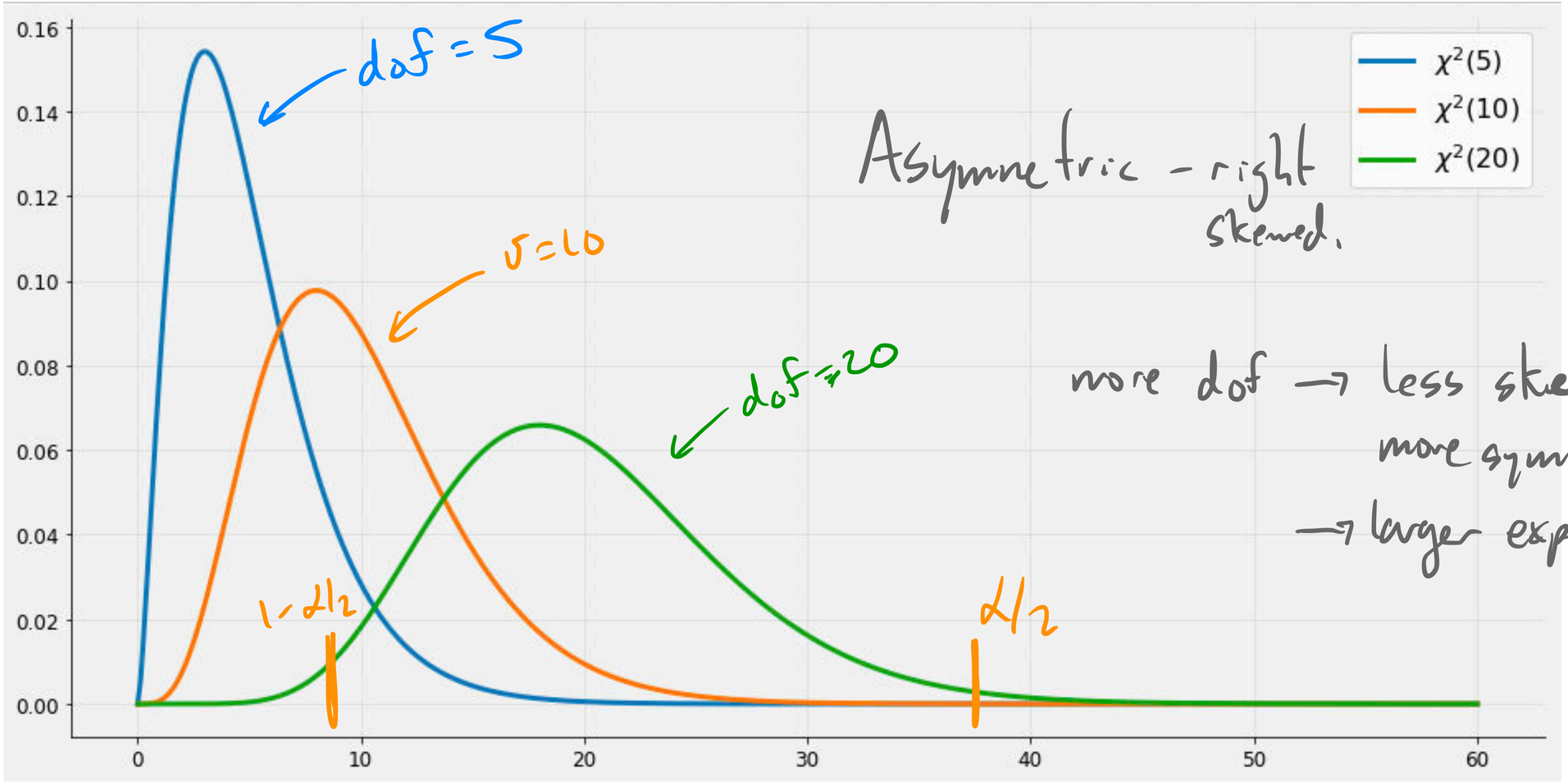
- ~~After Spring Break~~ Today, we'll talk about estimating confidence intervals for the variance of a population using something [wonderful] called **The Bootstrap**.

- But if your population is normally distributed, we have some [wonderful] theory which gives us a better confidence interval and works for both large and small sample sizes!

- **Question**: What does the sampling distribution of the variance look like when the population is **normally distributed**?

The Chi-Squared Distribution

- The chi-squared distribution (χ^2_ν) is also parameterized by degrees of freedom $\nu = n - 1$
- The pdfs of the family of χ^2_ν distributions are gross, so lets just draw them!



A confidence interval for the variance

- Let $X_1, X_2, \ldots X_n$ be IID samples from a normal distribution with mean μ and standard deviation σ. Define the *sample variance* in the usual way as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Then the random variable $(n-1)S^2/\sigma^2$ follows the distribution χ^2_{n-1}.

- Then it follows that

$$P\left(\chi^2_{1-\alpha/2,\,n-1} < \frac{(n-1)\,s^2}{\sigma^2} < \chi^2_{\alpha/2,\,n-1}\right) = 1-\alpha$$

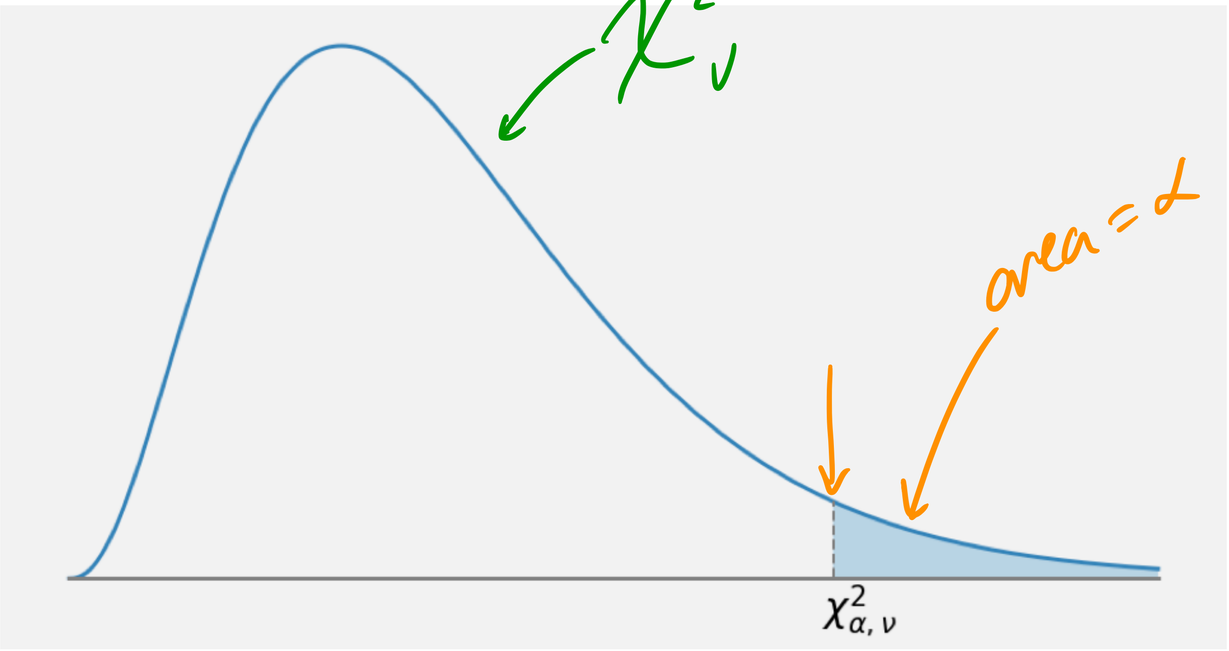
d.o.f.

Left bound

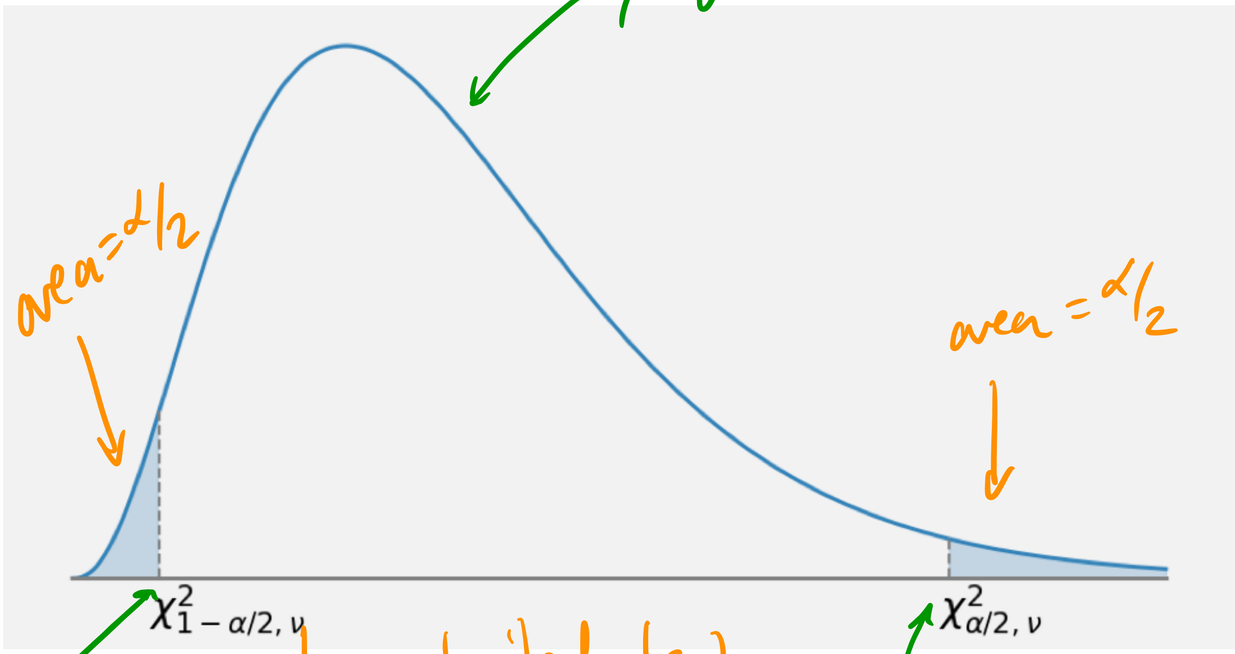
Right bound

The Chi-Squared Dist is Non-Symmetric

- Because the distribution is non-symmetric, we need to use two different critical values.



one-tailed test



two-tailed test

stats.chi2.ppf(1-$\alpha/2$, ν) stats.chi2.ppf($\alpha/2$, ν)

A confidence interval for the variance

$\frac{x}{y} < s$ $\quad x < s_y$ $\quad y > \frac{x}{s}$

- For a $100(1-\alpha)\%$ confidence interval we choose the two critical values $X^2_{1-\alpha/2,n-1}$ and $X^2_{\alpha,n-1}$ which puts $\alpha/2$ probability in each tail. Then, with $100(1-\alpha)\%$ confidence we can say that

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

Solve for this

$$\frac{1}{\chi^2_{\alpha/2,n-1}} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi^2_{1-\alpha/2,n-1}}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

Conf. Interval for variance σ^2

A confidence interval for the variance

- For a $100(1-\alpha)\%$ confidence interval we choose the two critical values $X^2_{1-\alpha/2,n-1}$ and $X^2_{\alpha,n-1}$ which puts $\alpha/2$ probability in each tail. Then, with $100(1-\alpha)\%$ confidence we can say that

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

Question: How can we use this to get a $100(1-\alpha)\%$ confidence interval for the standard deviation?

$$\sqrt{\frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}}} < \sigma < \sqrt{\frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}}}$$

yay!

useless!

- Example: A large candy manufacturer produces packages of candy targeted to weight 52g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation in the produced packages is larger than acceptable. In an attempt to estimate the variance she selects n=10 bags at random and weighs them. The sample yields a sample variance of 4.2g. Find a 95% confidence interval for the variance and a 95% confidence interval for the standard deviation.

$s^2 = 4.2$

$n = 10$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$\chi^2_{0.975,9} = \text{stats.chi2.ppf}(0.025, 9) = 2.70$

$\chi^2_{0.025,9} = \text{stats.chi2.ppf}(0.975, 9) = 19.02$

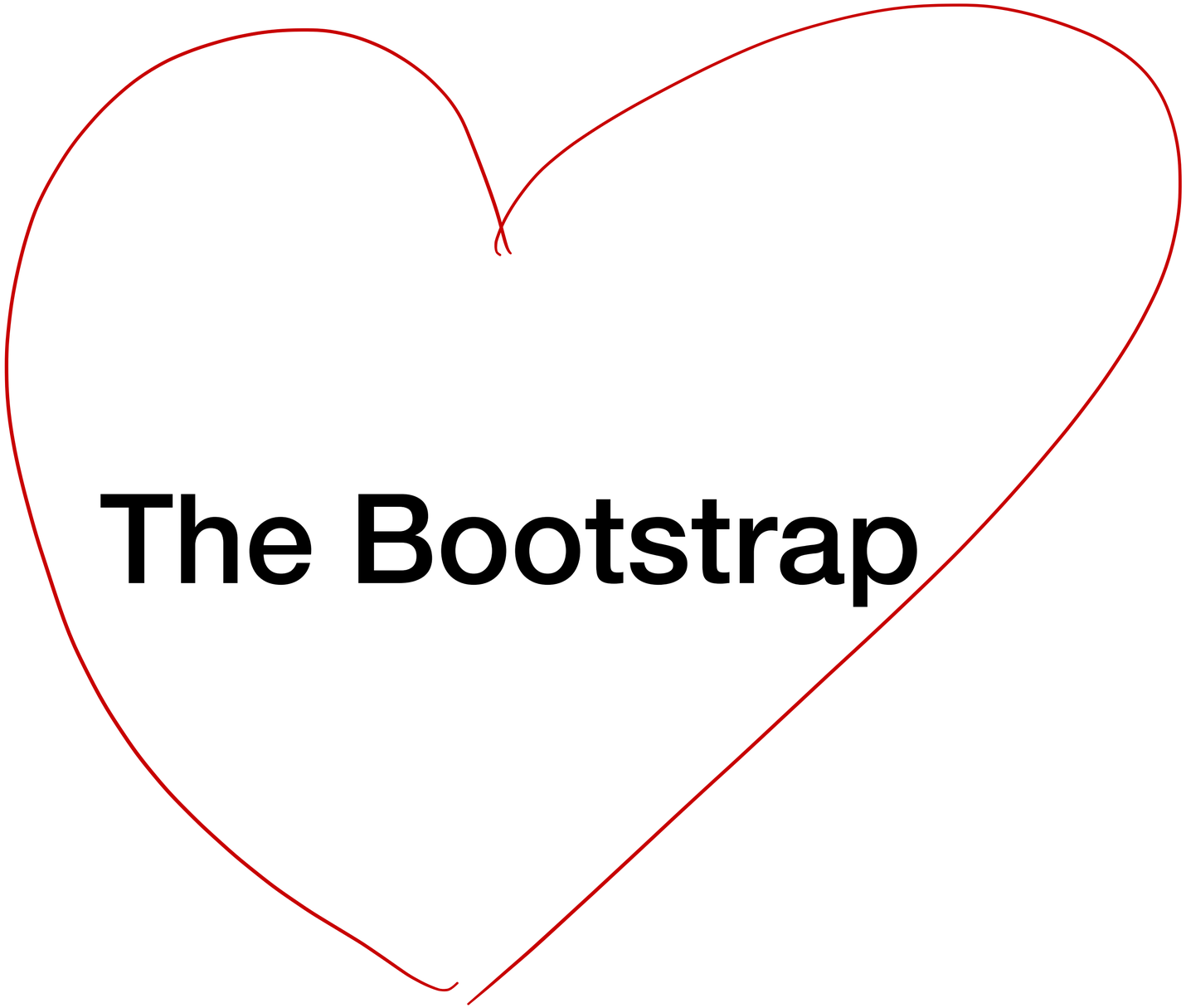
95% CI for σ^2: [1.99, 14.0]

$\frac{(n-1)s^2}{\chi^2}$

L: $\frac{(10-1)4.2}{19.02} = 1.99$

R: $\frac{(10-1)4.2}{2.70} = 14.0$

for σ = [1.41, 3.74]


The Bootstrap

Not all datapoints come cheap…

- In real scenarios, **data can be expensive**…
  - in **money**. For example, data from an aircraft in a wind tunnel.
  - in **time**. For example, polling people in surveys is time consuming.
  - in **privacy tradeoffs**. For example, storing another person's genome in the database incurs ethical risk or cost, even when it does not cost much time or money.

- Today, we'll learn a technique that enables us to learn from small amounts of data to compute confidence intervals: **the bootstrap**

What are bootstraps?

- Bootstraps are the straps that you use to pull your boots on.
- To “pull yourself up by your bootstraps” is to somehow lift yourself upward by pulling on your own shoes. Obviously impossible.
- Now, however, bootstrapping means to accomplish something without aid. To accomplish what you need to with what you’ve got.
- The statistical bootstrap is in this last sense. It allows us to really **make the most of a small dataset** without sacrificing statistical rigor or collecting more $ samples.

A confidence interval for the mean

- **Recall**: if we have *n* samples from a distribution that is normal *or* non-normal, then by the Central Limit Theorem, the confidence interval for the mean is given by $\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ or for an unknown variance $\bar{X} \pm z_{\alpha/2}\sqrt{\frac{s^2}{n}}$

- The bootstrap is a different approach. Consider the same set of samples as above, $X_1, X_2, \ldots, X_n$, but instead of computing a CI analytically from this sample, instead *re*-sample your sample many times and examine (?) those!

- **Definition**: a bootstrapped resample is a set of *n* draws from the original set, sampled *with replacement*.

of n values

A confidence interval for the mean

- **Definition**: a bootstrapped resample is a set of *n* draws from the original dataset (drawn IID from *X*), sampled *with replacement*.
- **Example**: suppose we have the data [2,4,6,7,9]
  - Resample 1 might be: [4, 6, 7, 4, 9]
  - Resample 2 might be: [2, 6, 7, 9, 2]
  - Resample 3 might be: [9, 7, 7, 7, 4]

- Given the example above, what does "*sample with replacement*" mean?

A confidence interval for the mean

- **Definition**: a bootstrapped resample is a set of *n* draws from the original dataset (drawn IID from *X*), sampled *with replacement*.
- **Proposition**: a suitable estimate of the 95% confidence interval for the mean of the distribution *X* is given by [*a*,*b*], where *a* and *b* are the 2.5 percentile and 97.5 percentile of the means of a large number of bootstrapped resamples.

- **In plain English**: resample your original data many times. Compute the mean for each resample. Compute the 2.5 and 97.5 percentiles of those means.

Magic!