CSCI 3022

intro to data science with probability & statistics

Lecture 14
March 5, 2018

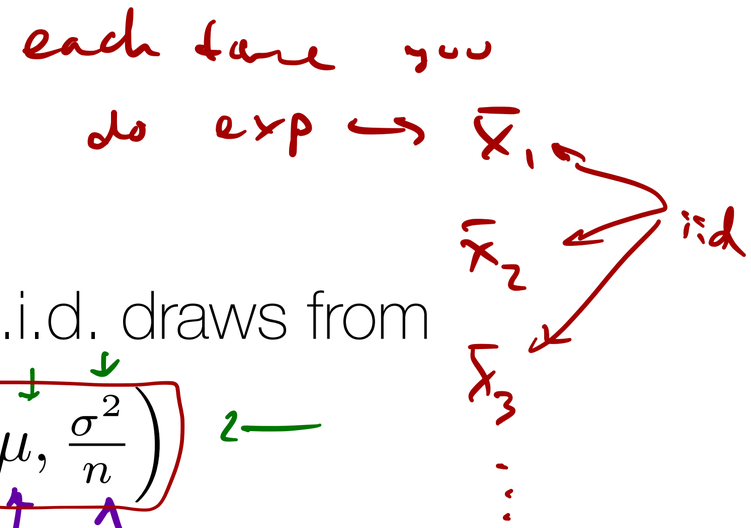
Introduction to Statistical Inference & Confidence Intervals

TONY Wed OH are cancelled

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Last time on CSCI 3022



- **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

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

Any old distribution

μ

Var $\sim \frac{1}{n}$

Examples:

Population? $p = 0.5$

Sample? $n = 50$

- **Example 2**: Suppose you have a jar of lemon and banana jelly beans where it is known that the true proportion of lemon jelly beans is 0.5. You try to estimate the proportion of lemon beans by reaching in and drawing 50 jelly beans and testing them (by eating them). What is the probability that your sample is 75% or more lemon jelly beans?
- Note: this is a little different because we're estimating a *proportion*. What changes?

Population: $p = 0.5$
Sample: $n = 50$

$P(\bar{X} \geq 0.75) = 1 - P(X \leq 0.75)$

$\hat{p} = \frac{X}{n}$

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

$= \frac{1}{n^2} n p(1-p)$

Proportions are different. $\bar{X} = \hat{p} = \frac{Bin(n,p)}{n}$

What is variance of $\hat{p}$? $Var(\hat{p}) = Var\left(\frac{Bin(n,p)}{n}\right) = \frac{1}{n^2} Var(Bin(n,p))$

Last time, on CSCI 3022…

- **Example 2**: Suppose you have a jar of lemon and banana jelly beans where it is known that the true proportion of lemon jelly beans is 0.5. You try to estimate the proportion of lemon beans by reaching in and drawing 50 jelly beans and testing them (by eating them). What is the probability that your sample is 75% or more lemon jelly beans?

$P(p \geq 0.75)$

Statistical Inference

- **Goal**: we want to learn the properties of an underlying population by analyzing sampled data.

Questions:

- Is sample mean $\bar{x}$ a good approximation of the population mean μ ?
- Is sample proportion $\hat{p}$ a good approximation of the population proportion p?
- Is there a statistically significant difference between the mean of two samples?
- If the answer is **yes**, *how sure are we*?
- How much data do we need in order to be **confident** in our conclusion?

next few weeks!

Confidence Intervals

- The Central Limit Theorem tells us that as the sample size n increases, the sample mean of X is close to *normally* distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$ of our bag of $\bar{X}$
- "Standardizing" the sample mean by first subtracting the expected value and dividing by the standard deviation yields a standard normal random variable.

$\frac{X-\mu}{\sigma}$ for CLT $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ Box-Muller Z

Question: how big does our sample need to be

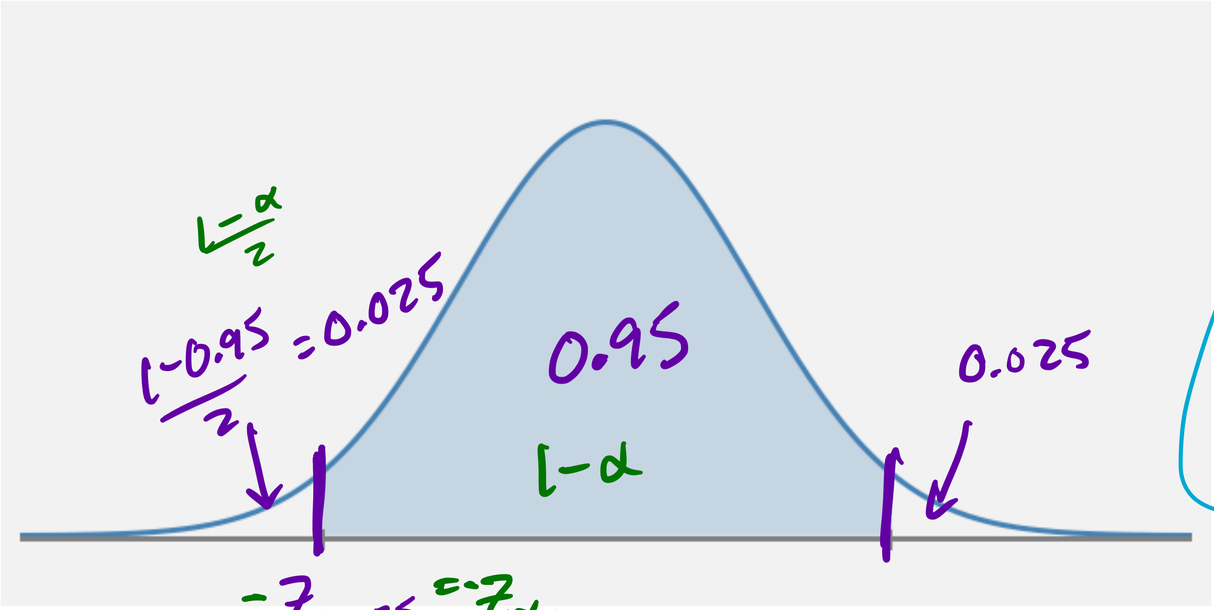
- …if the variable of interest is normally distributed?
- …if the variable of interest is not normally distributed?

Confidence Intervals

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$\alpha = 0.05$

- We saw a while ago that the **95% of the area** under the standard normal curve falls **between –1.96 and +1.96**, so we know that



we don't know μ

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

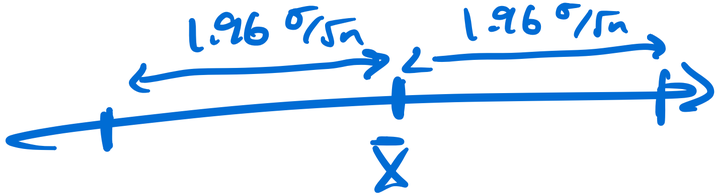
Random variable

- This is equivalent to:

$$0.95 = P\left(-1.96 \le Z \le 1.96\right) = P\left(-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.96\right)$$

$z_{0.025}$

Confidence Intervals



- The **95% confidence interval** for the mean is then given by:

$$P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

write as an interval:

middle?

LaTeX: \pm

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \;,\; \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] \rightarrow \bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

- Question: which things in this expression are random variables and which are fixed??

RANDOM: $\bar{X}$

FIXED: n, σ, $\alpha \rightarrow z_{\alpha/2}$

Confidence Intervals

- The 95% CI is centered at $\bar{X}$ and extends $1.96\frac{\sigma}{\sqrt{n}}$ to each side of $\bar{X}$

- The 95% CI's width is $2 \times 1.96\frac{\sigma}{\sqrt{n}}$ which is **not** random; only the location of the interval's midpoint $\bar{X}$ is random.

- We often write the CI $\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ as $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$.

$1.96 = z_{0.05/2}$

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

Interpreting the Confidence Interval

95%

$\bar{x} \rightarrow CI$

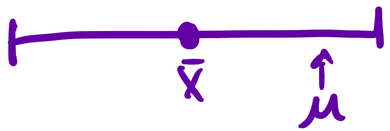
- **Statement**: We are 95% confident that the true population mean is in this interval.

- **Correct Interpretation**: In repeated sampling, 95% of all CIs obtained from sampling will actually contain the true population mean. The other 5% of CIs will not.

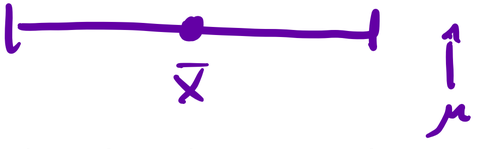
20 experiments:

IN THE LONG RUN:

$\frac{19}{20}$:



$\frac{1}{20}$:

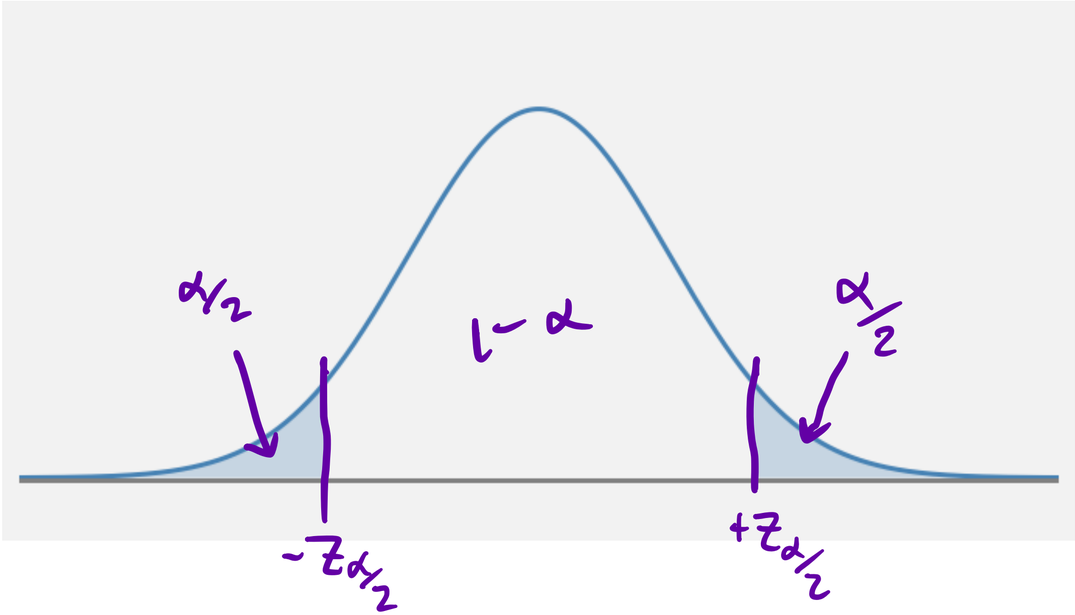


- The confidence level is not a statement about any one particular interval. Instead it describes what would happen if a very large number of CIs were computed using the same CI formula.

Other Levels of Confidence

- A probability of $1-\alpha$ is achieved by using $z_{\alpha/2}$ in place of $z_{0.05/2} = z_{0.025} = 1.96$

$\alpha = 0.05$ (for a 95% CI)



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

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

#'s we measured; α handed to you; know; data

CI Example

$z_{0.05} = 1.645$

$[3.496, 3.704]$

- The General Social Survey is a sociological survey used to collect data on demographic characteristics and attitudes of the residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours per day. Suppose further that the known standard deviation of the characteristic is 2 hours per day. **Find a 90% confidence interval for the amount of relaxation hours per day.**

$\bar{X}$ → 3.6 hours per day

σ → 2 hours per day

n = 1000

CI: $\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$\alpha = 0.1$ b/c 90% CI

$\alpha = 1 - .90$

what is $z_{\frac{0.1}{2}} = z_{0.05} = 1.645$

90% CI is $3.6 \pm 1.645 \cdot \frac{2}{\sqrt{1000}}$

90% CI = $[3.496, 3.704]$

CI Example

HERE

$z_{0.025} = 1.96$

[3.48, 3.72]

- The General Social Survey is a sociological survey used to collect data on demographic characteristics and attitudes of the residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours per day. Suppose further that the known standard deviation of the characteristic is 2 hours per day. **Find a 95% confidence interval for the amount of relaxation hours per day.**

$\bar{x} \pm z_{0.05/2} \cdot \frac{\sigma}{\sqrt{1000}}$ → ... → [3.48, 3.72] = 95% CI

[3.50, 3.70] = 90% CI

$z_{0.05/2} = 1.96$

- **Q**: what are the advantages/disadvantages of a wider confidence interval?

balance between True & useful info.

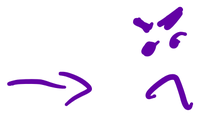
Test your understanding!

[3.48, 3.72] 95% CI

- **Concept Check**: In the previous example we found a 95% CI for relaxation time to be [3.48, 3.72]. Which of the following statements are true?

A. 95% of Americans spend 3.48 to 3.72 hours per day relaxing after work.

B. 95% of random samples of 1000 residents will yield CIs that contain the true average number of hours that Americans spend relaxing after work each day.



C. 95% of the time the true average number of hours an American spends relaxing after work is between 3.48 and 3.72 hours per day.

Nope

D. We are 95% sure that Americans in this sample spend 3.48 to 3.72 hours per day relaxing after work.

Computing required sample size

- **Example**: For the GSS data, how large would n have to be to get a 95% CI with width at most 0.1?

$z_{\alpha/2} = 1.96$

CI: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\longrightarrow$ width $= 2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq 0.1$

$\downarrow$ rearrange for n

$$\sqrt{n}^2 \geq \left(2 \cdot \frac{1.96 \cdot 2}{0.1} \cdot 10\right)^2$$

$$\Rightarrow n \geq 4 \cdot 1.96^2 \cdot 4 \cdot 100$$

estimating

$$\approx 4 \cdot 4 \cdot 4 \cdot 100$$

$$\Rightarrow \boxed{n \geq 6400}$$

Confidence IRL…?

- In the previous example we assumed that we knew the population standard deviation.
- **Question**: how often does this happen in real life?

Confidence IRL…?

- In the previous example we assumed that we knew the population standard deviation.
- **Question**: how often does this happen in real life? **never**
- **Solution**: If *n* is large we use the sample variance instead

$\sigma \rightarrow s = \sqrt{\text{var}[x]}$

$= \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$

$$CI_\alpha = \bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Just like we use data to est. μ using $\bar{x}$, use data to est. σ using s

- **Solution**: If *n* is small we have to do something else (more on this later)

Confidence intervals for proportions

- Let p denote the proportion of "successes" in a population (e.g. individuals who graduated from college, compute nodes that didn't fail on a given day)
- A random sample of n individuals is selected, and X is the number of successes in the sample

Bernoulli r.v. ($Var[x_i] = p(1-p)$)

- Then X can be modeled as a Binomial random variable with:

$E[X] = np$

$Var[X] = n\,p(1-p)$

n trials

Bernoulli variance

Confidence intervals for proportions

"was Newton a ___"

- The estimator for p is given by:

$$\hat{p} = \frac{X}{n}$$

- The estimator is approximately normally distributed with:

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}np = p$$

$$Var[\hat{p}] = Var\left[\frac{X}{n}\right] = \frac{1}{n^2}Var[X]$$

$$= \frac{1}{n^2}np(1-p)$$

$$Var[\hat{p}] = \frac{p(1-p)}{n}$$

- Standardizing the estimate yields:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \quad \left(\frac{\hat{p} - E[\hat{p}]}{SD[\hat{p}]}\right)$$

- This gives us a confidence interval of:

central est. ↓ ↓ SD

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

Confidence intervals for proportions

$z_{0.005} = 2.57$

$\frac{127}{200} = 0.635$

Sample: n, $\hat{p}$

Population: ?? want p

- Example: The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 of the sampled homes to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

200 = ✓

$\hat{p} = \frac{127}{200}$

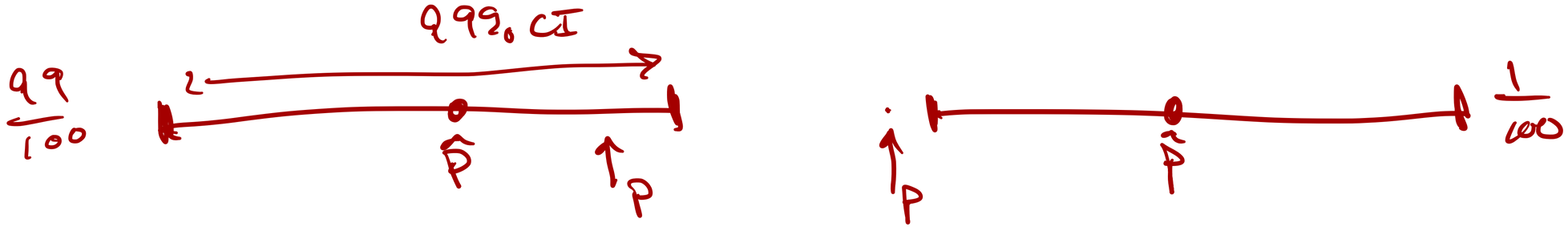
$= 0.635$

$z_{\frac{0.01}{2}}$

$$99\%\ CI = \hat{p} \pm z_{0.005}\sqrt{\frac{p(1-p)}{n}}$$

est. unknown p ($\hat{p}$ our best guess) vs.

$$= 0.635 \pm 2.57\sqrt{\frac{0.635(1-0.635)}{200}}$$



Confidence intervals for proportions

Sample: $n = 200$ #"heads" $= 127$ Population: Nada!

- Example: The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 of the sampled homes to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

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

$\alpha = .01$

$z_{\alpha/2} = z_{\frac{0.01}{2}} = 2.57$ via $\text{opt}(0.995)$

$\hat{p} = \frac{127}{200} = 0.635$

$n = 200$

$$0.635 \pm 2.57\sqrt{\frac{0.635(1-0.635)}{200}}$$

99% CI $\longrightarrow$ $[0.548, 0.722]$

Tie back to prob statement, last slide

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 (has) 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

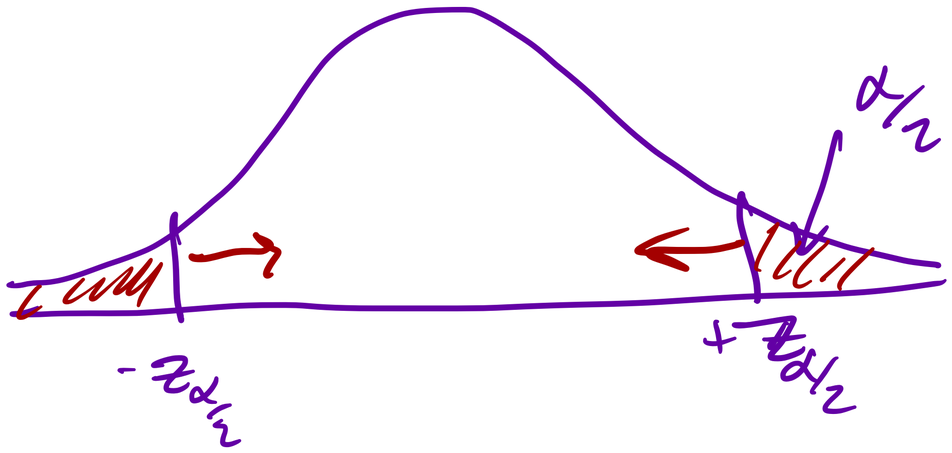
CI width: $2 \cdot z_{\alpha/2} \cdot SD$

- **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 (has) 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

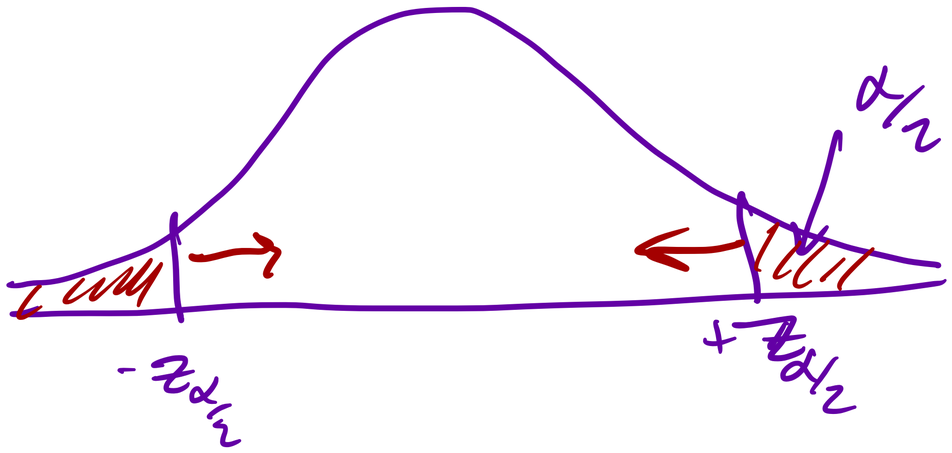
CI width: $2 \cdot z_{\alpha/2} \cdot SD$

- **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

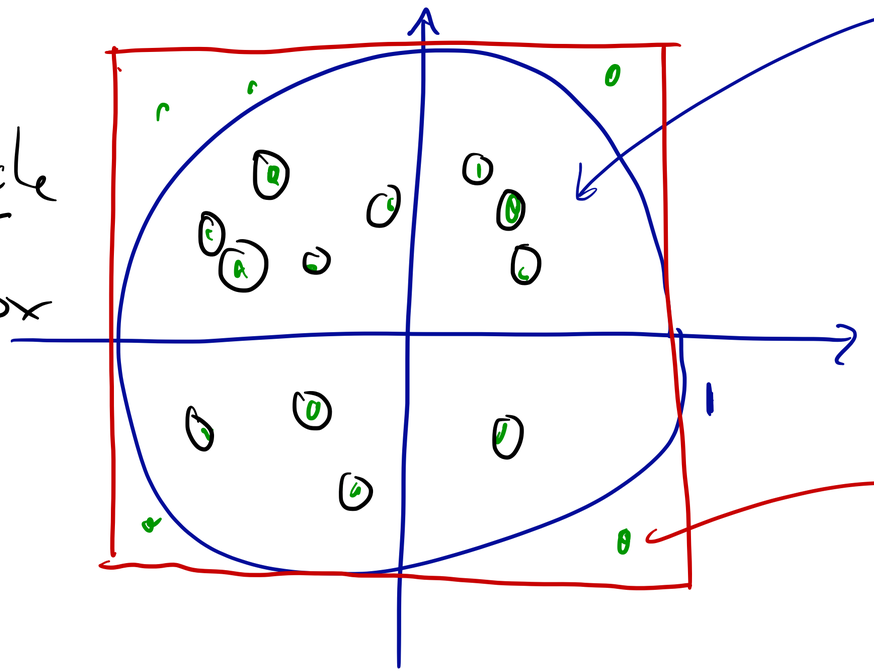
Stuff & Things

- **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 $\theta > \theta_0$ $\quad p > 0.5$

② $\theta < \theta_0$ $\quad p < 0.5$

③ $\theta \neq \theta_0$ $\quad p \neq 0.5$

} 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: $p \neq 0.5$

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

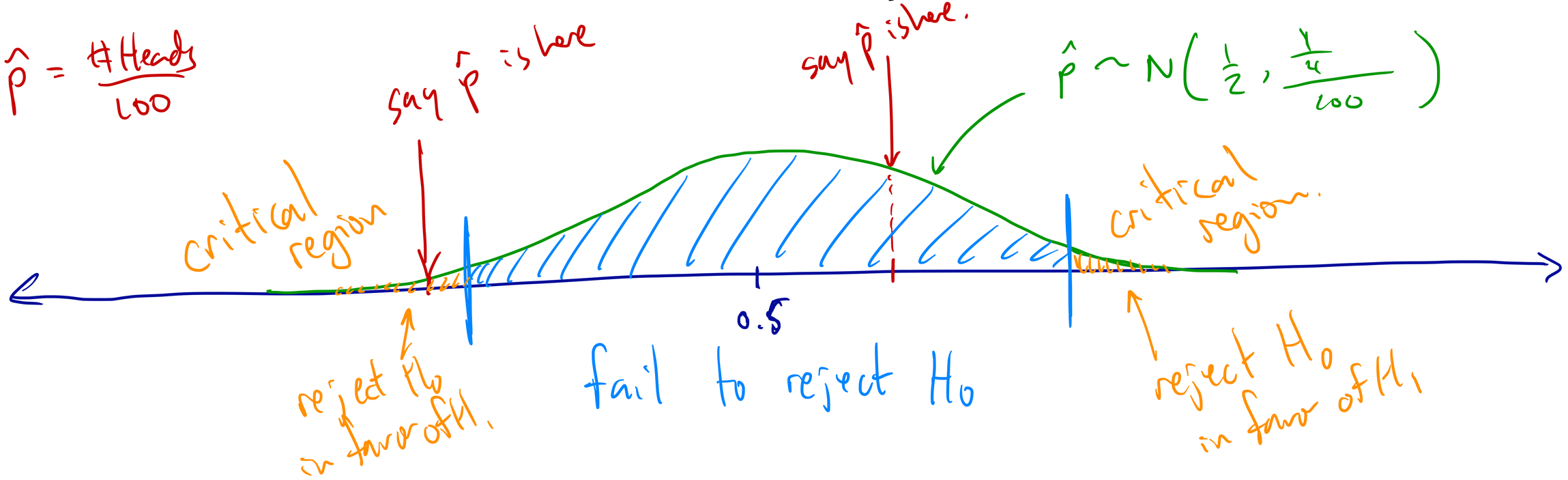
Under the null, $\hat{p} \sim N\left(0.5, \frac{0.5(1-0.5)}{100}\right)$

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?

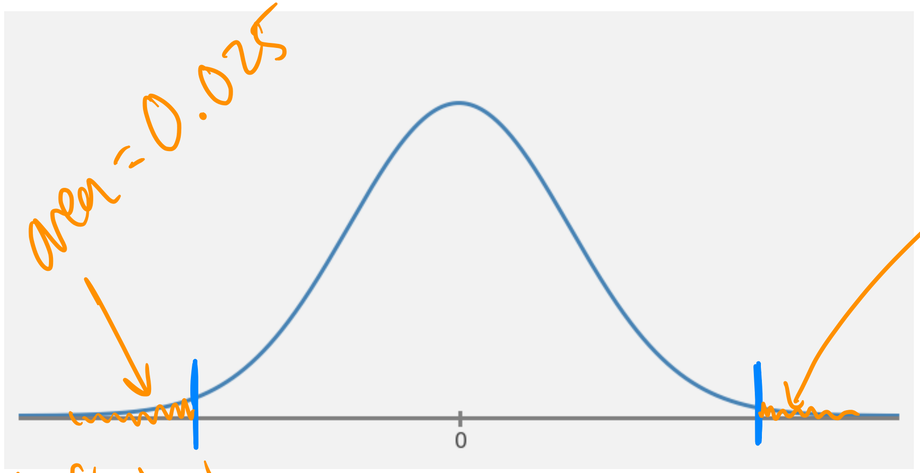


Test statistics and evidence

- **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

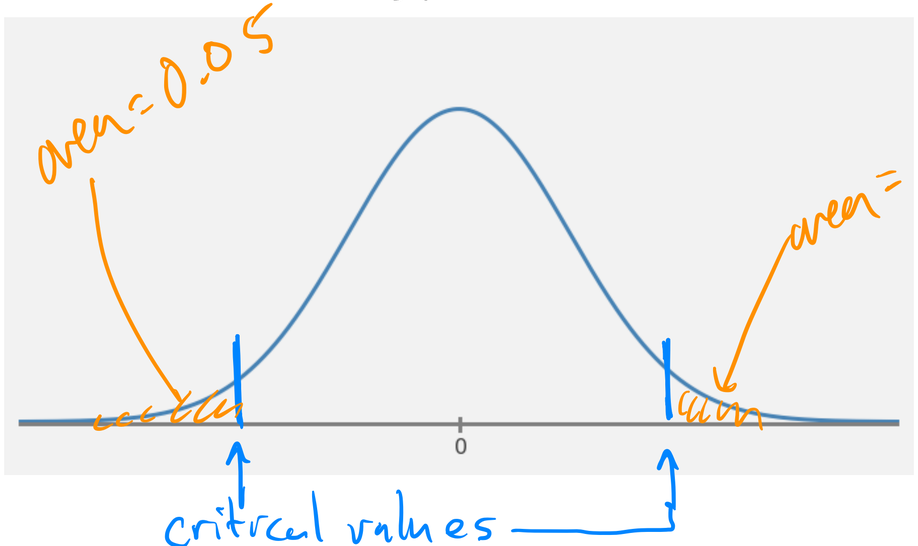


Rejection regions and significance level

- **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?

$H_0: p = 0.5$

$H_1: p \neq 0.5$

$\alpha = 0.05$

$z_{\alpha/2} = 1.96$

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

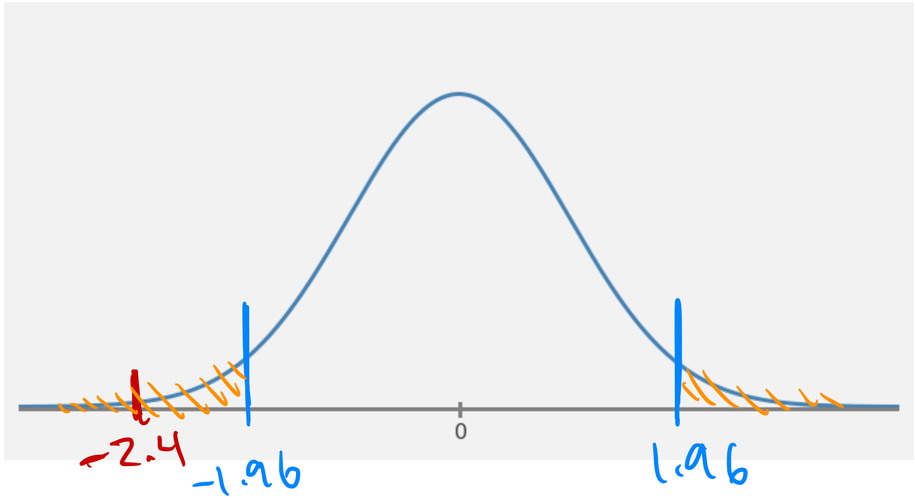
$\hat{p} \sim N\left(\frac{1}{2}, \frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{100}\right)$ (CLT)

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

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

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

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



Detecting Biased Coins

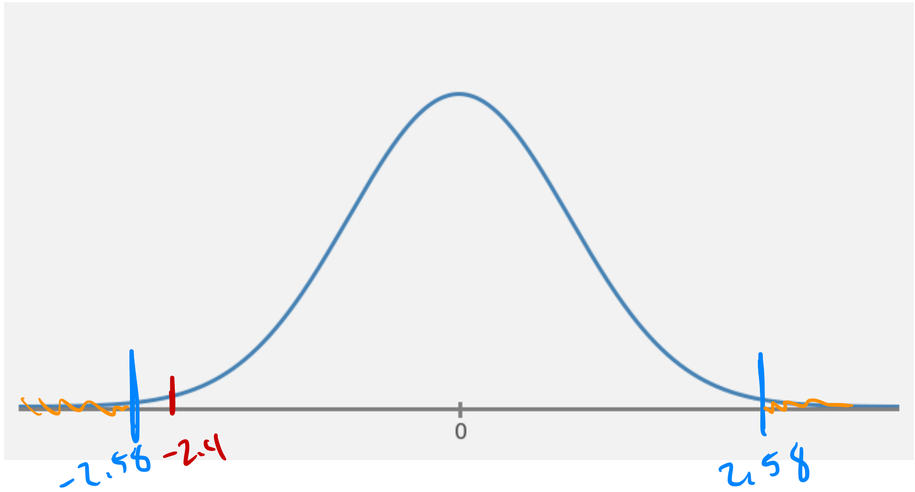
- **Example**: To test if the Belgian 1 Euro coin is fair you flip it 100 times and get 38 Heads. Do you reject the Null at the .01 significance level or not?

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

$-z_{0.005} = -2.58$

prev slide:

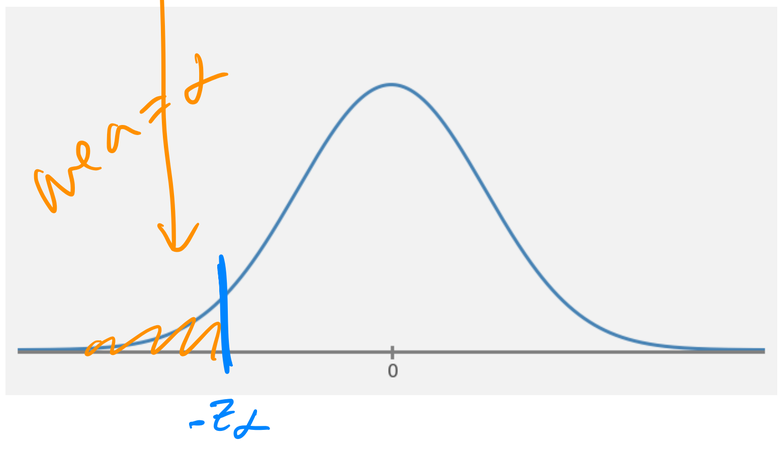
test statistic = -2.4

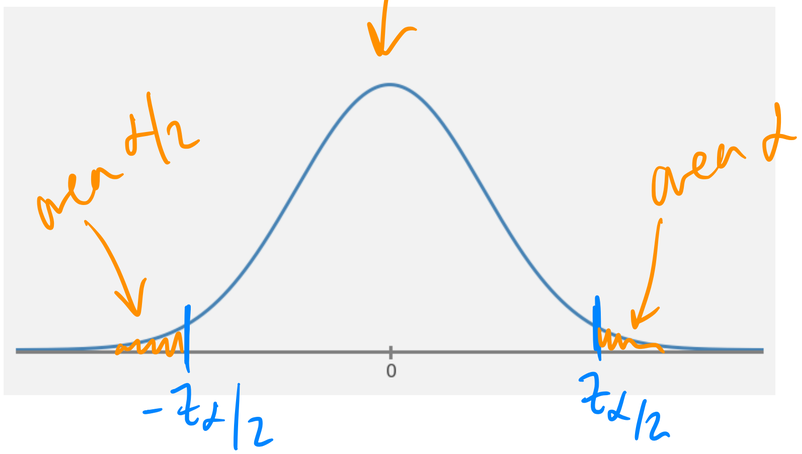


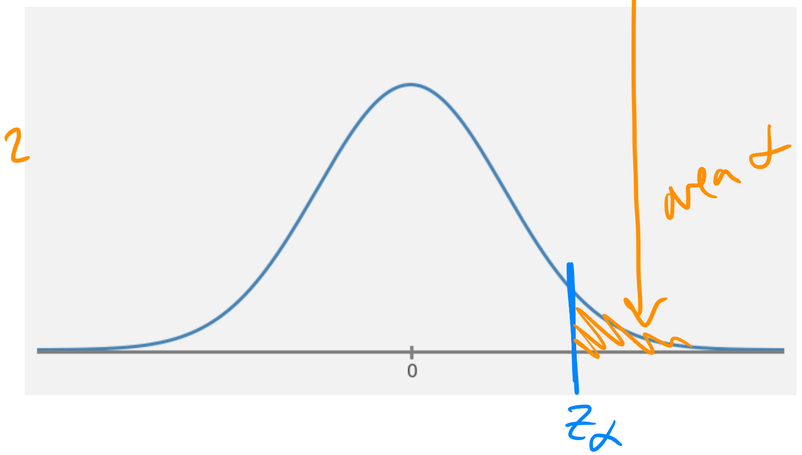
Different tests for different hypotheses

- 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.

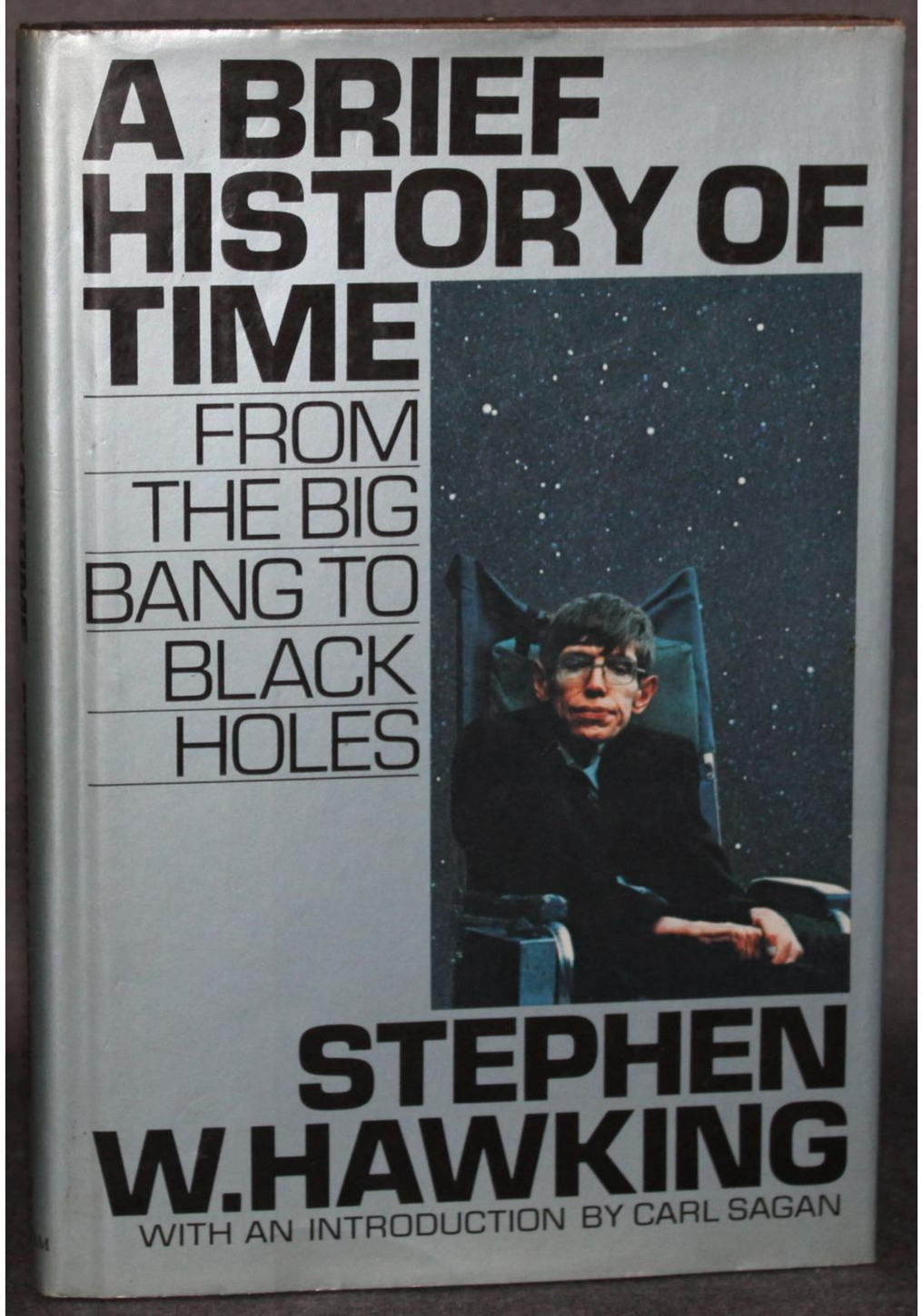
Space is limited, so act fast! 🏃‍♀️🏃‍♂️

"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

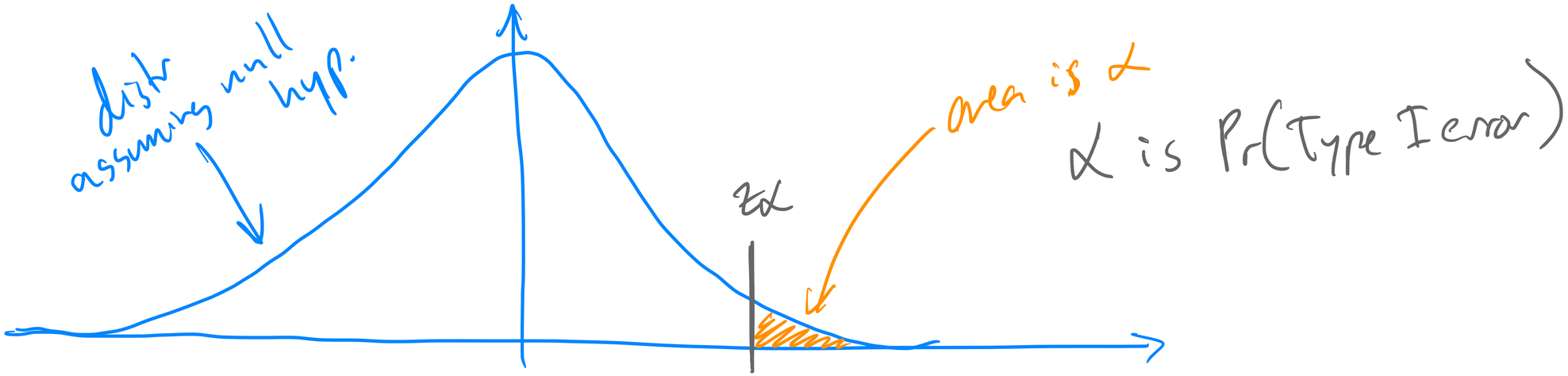
- **Definitions**:
- A **Type I Error** occurs when the Null hypothesis is rejected, but the Null hypothesis is in fact true ("**False Positive**")

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.

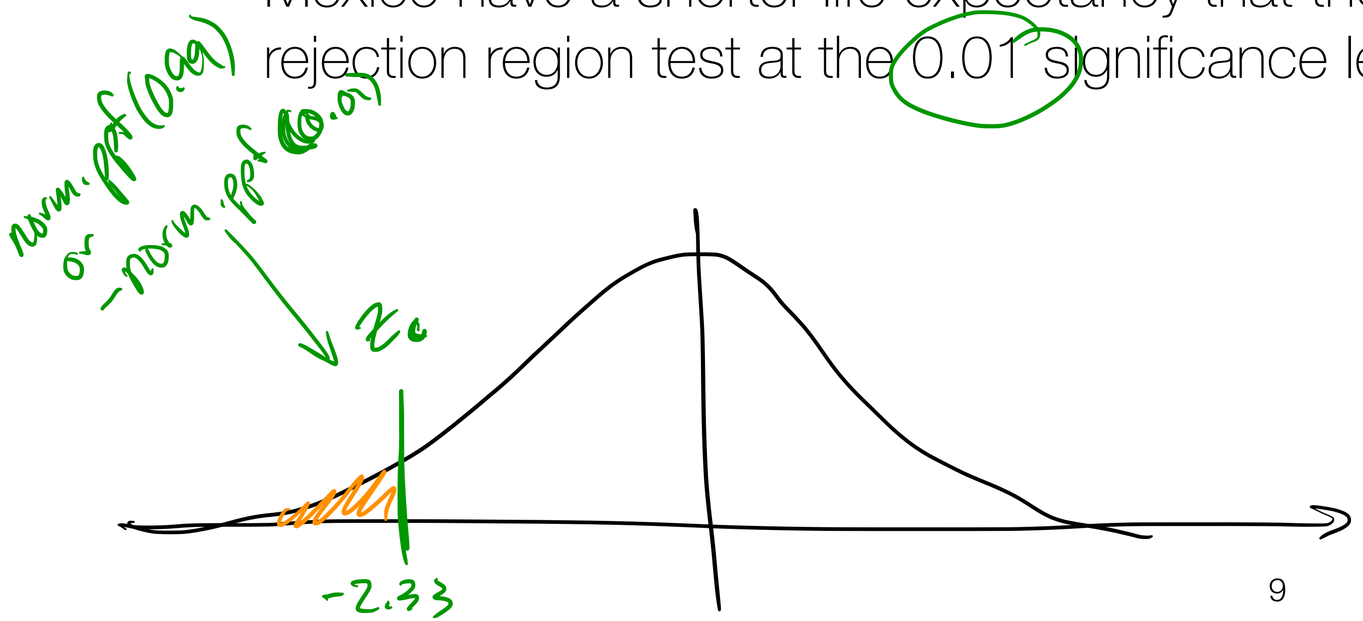
- **Question 2**: Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out a rejection region test at the 0.01 significance level.

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

$\mu = 300$

$\sigma = 150$

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



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

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

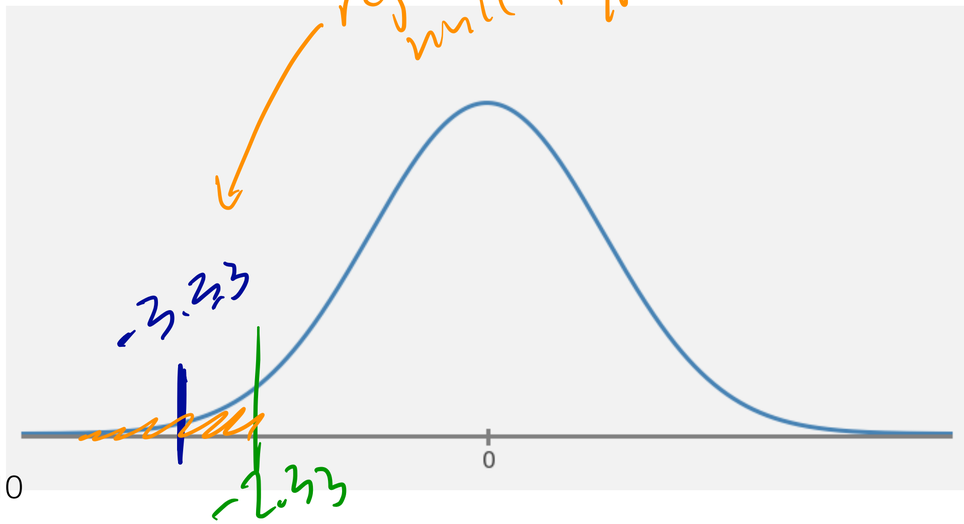
CLT

null hypothesis

sample size

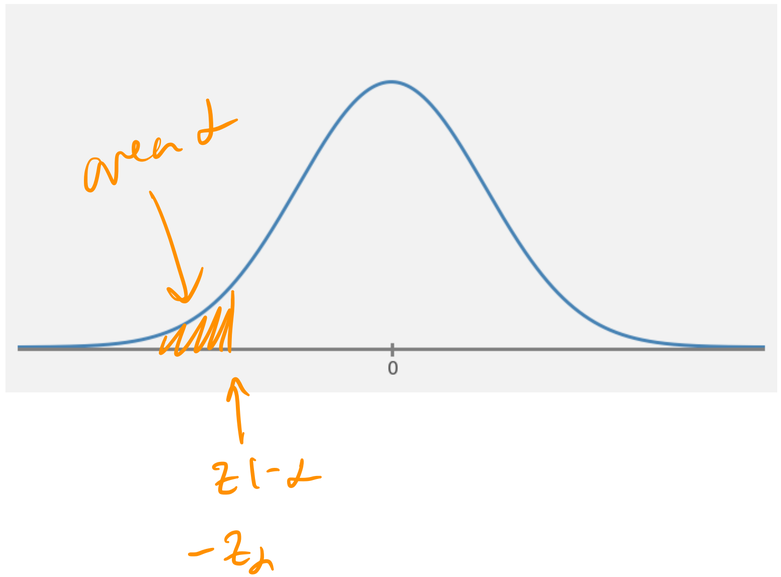
$$\frac{250 - 300}{150/\sqrt{100}} = -3.33$$

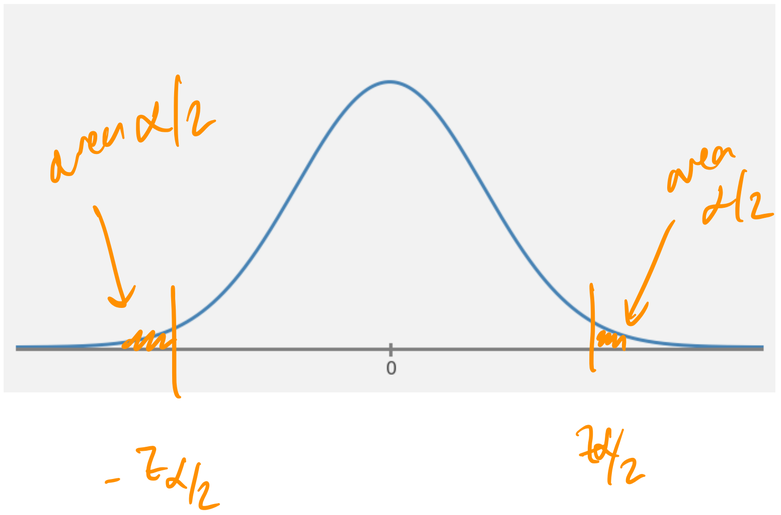
$$z_c = -2.33$$

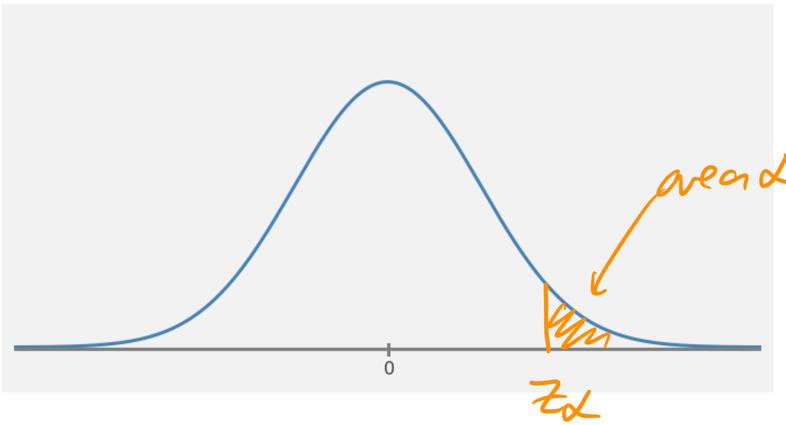


Rejection region & critical value summary

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

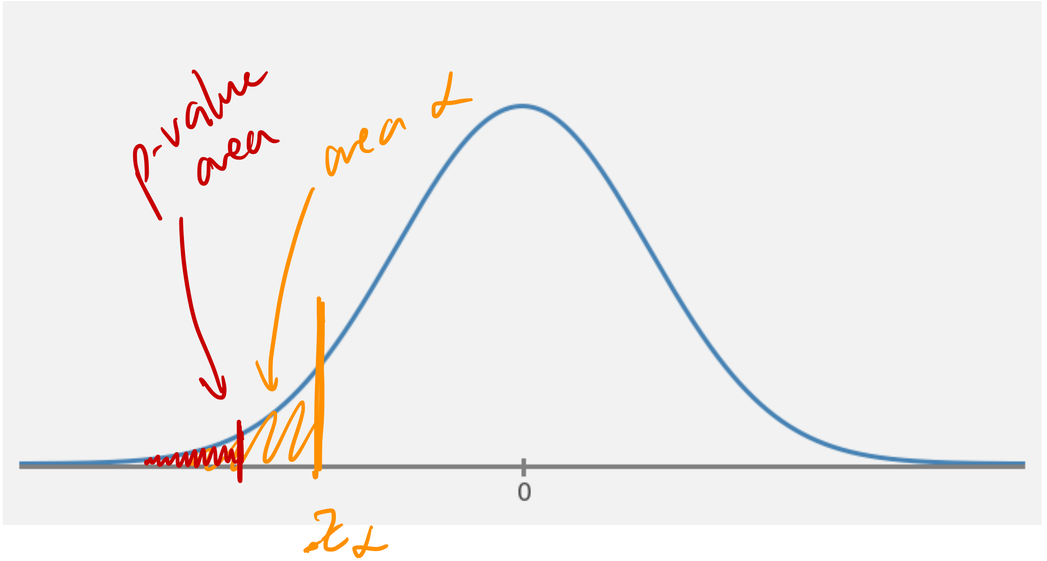
- **Critical Region** is region where test statistic has low probability under Null Hypothesis.
- Requires normally distributed data, or large enough sample for Central Limit Theorem.
- Under these assumptions we call this a Z-Test
- Rejecting the Null when the Null is true is called a Type I Error
- 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 $\dfrac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} = \dfrac{250 - 300}{150/\sqrt{100}} = -3.33$

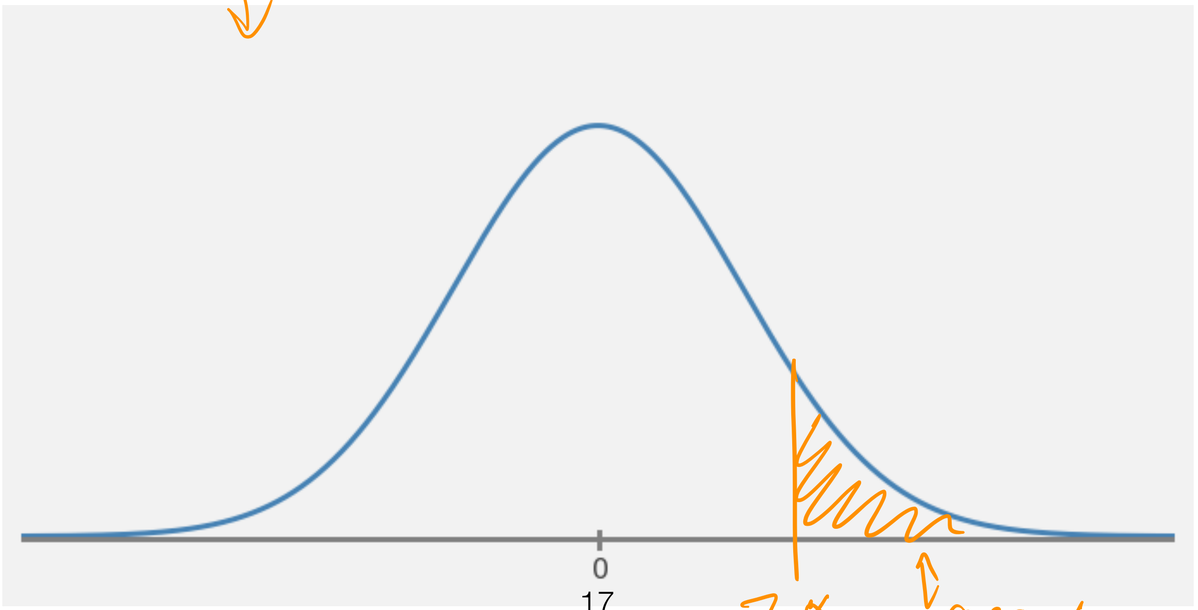
α

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

Reject Null Hypothesis

p-values for different z-tests

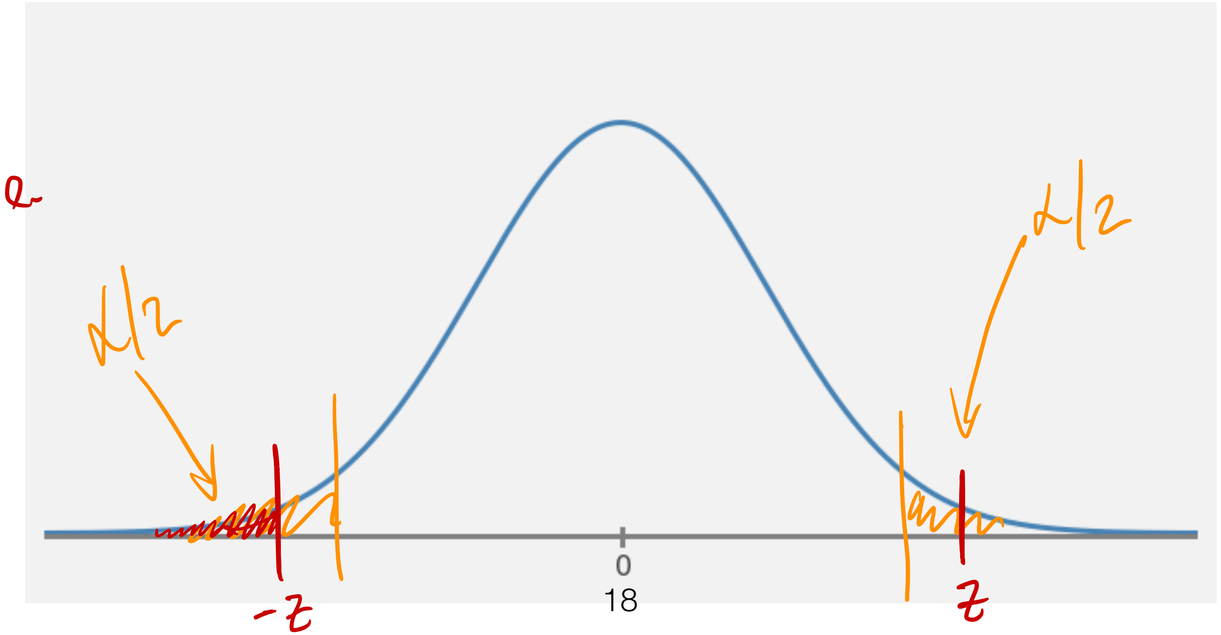
Alternative Hypothesis	Critical Region Level α Test	p-value Level α Test
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$	If $1-\Phi(z) \leq \alpha$
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$	If $\Phi(z) \leq \alpha$
$H_1 : \theta \neq \theta_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	Next slide.



p-values for different z-tests

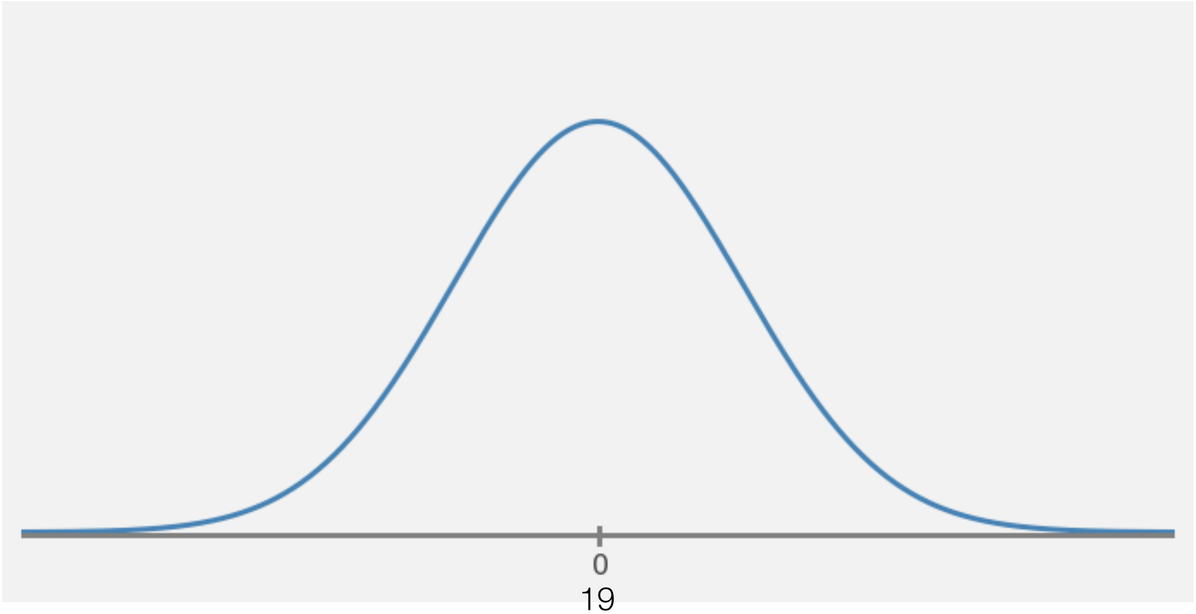
Alternative Hypothesis	Critical Region Level α Test	p-value Level α Test
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$	
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$	
$H_1 : \theta \neq \theta_0$	$z \leq -z_\alpha$ or $z \geq z_\alpha$	p-value: $2 \times \Phi(-\lvert z \rvert) \leq \alpha$

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

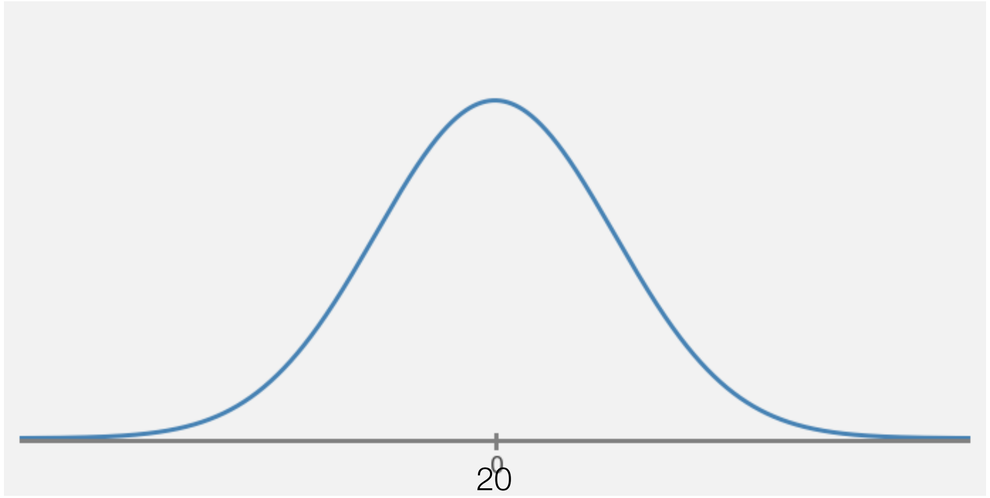
$\alpha = 0.05$ two sided

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

$= 0.0164$ 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

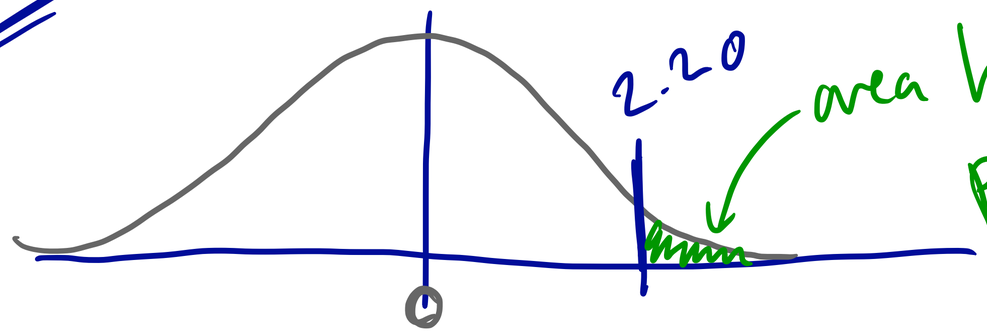
- **Example**: Data on calorie intake both for a sample of teens that reported that they do not typically eat fast food and another sample of teens who said they did usually eat fast food is as follows:

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?

$z = 2.20$



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

Nope

large effect.

Reject H_0 in favor of alt. Hyp. H_1

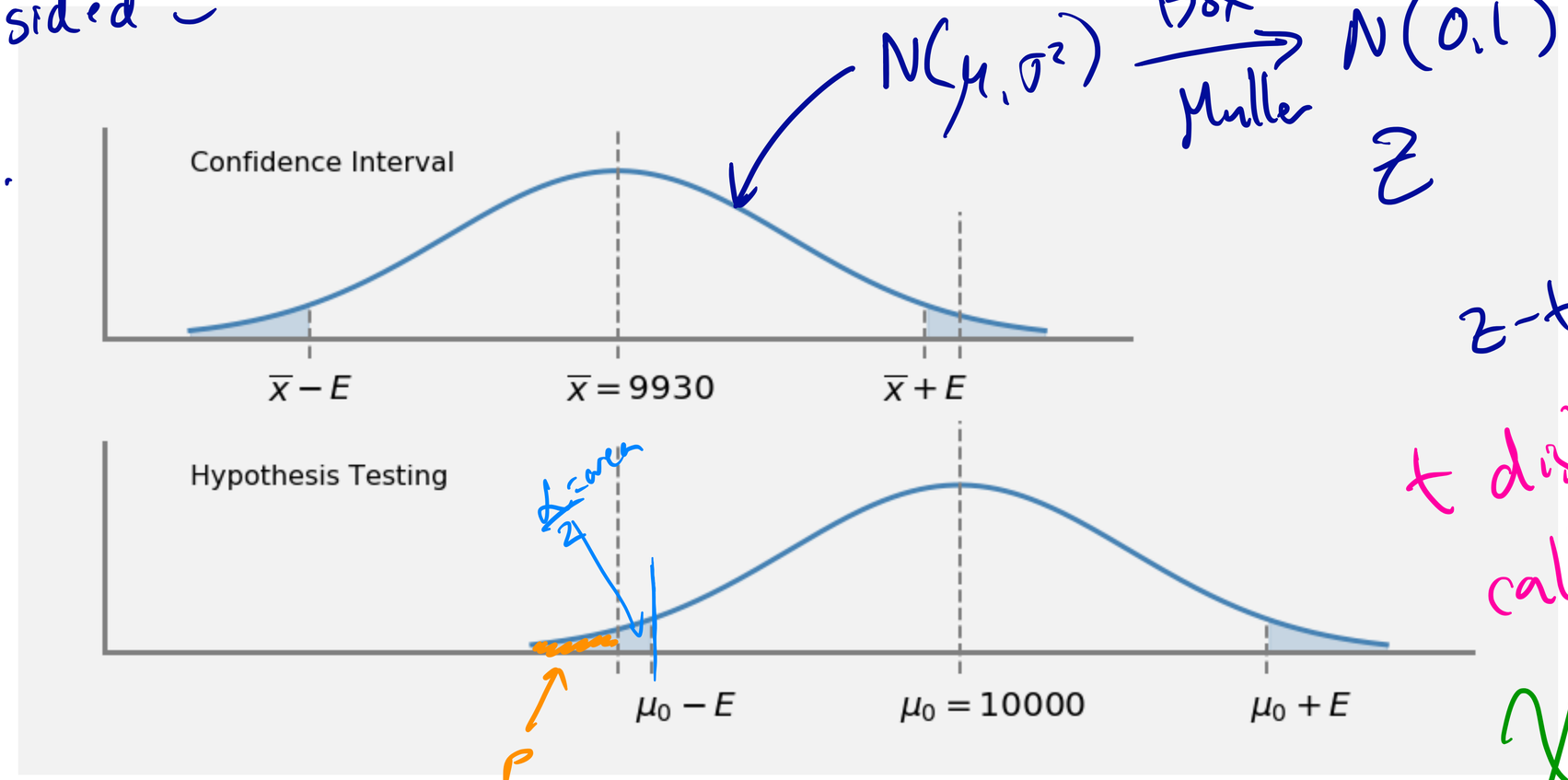
$\theta = \theta_0$

$\theta > \theta_0$

"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…

- σ is known:

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

"z tests"

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

center — window

- σ is unknown:

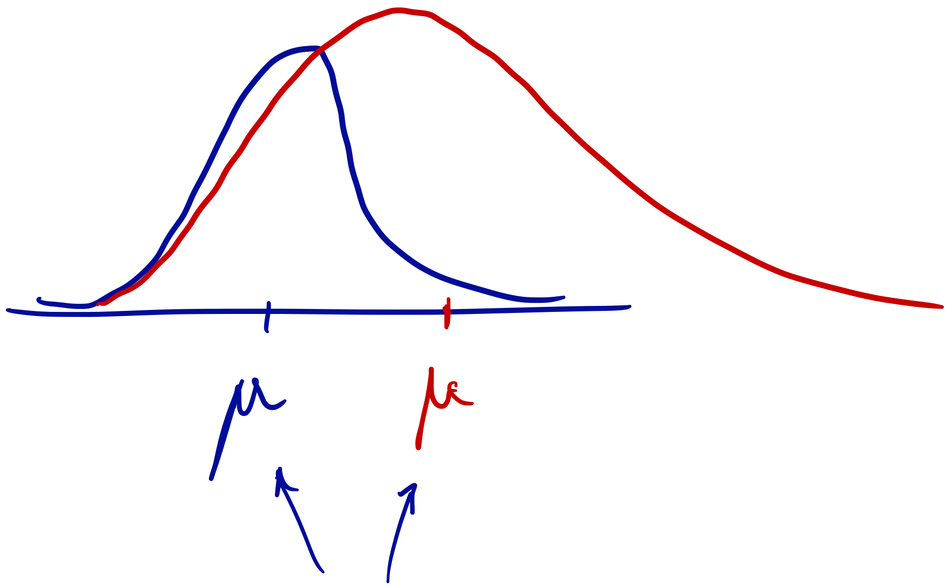
$$\left(\frac{\bar{x}-\mu}{s/\sqrt{n}}\right) \sim N(0,1)$$

"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:

	$n \geq 30$ ("n is large")	$n < 30$ ("n is small")
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

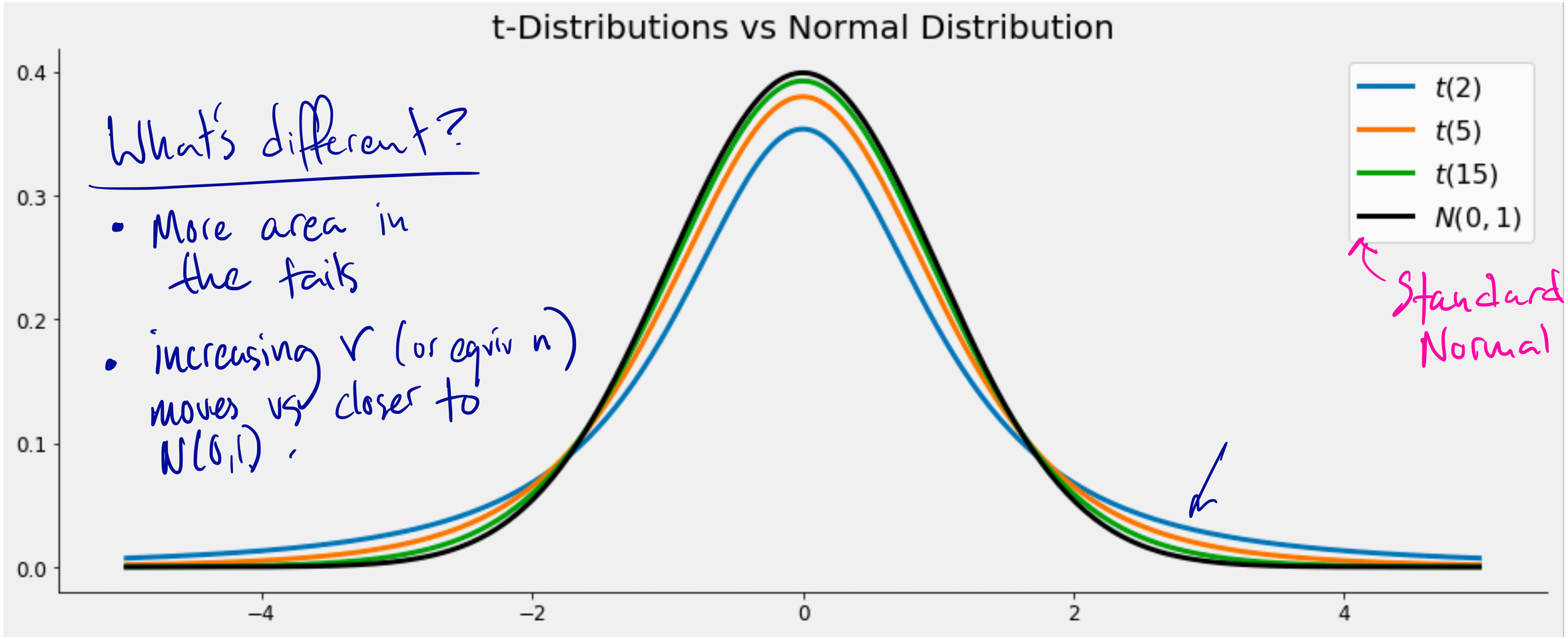
- When n is small we cannot invoke the Central Limit Theorem
- When n is small and the variance is unknown we need to do something else …
- When $\bar{X}$ is the sample mean of a random sample of size *n* from a normal distribution with mean μ, the random variable

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

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.)

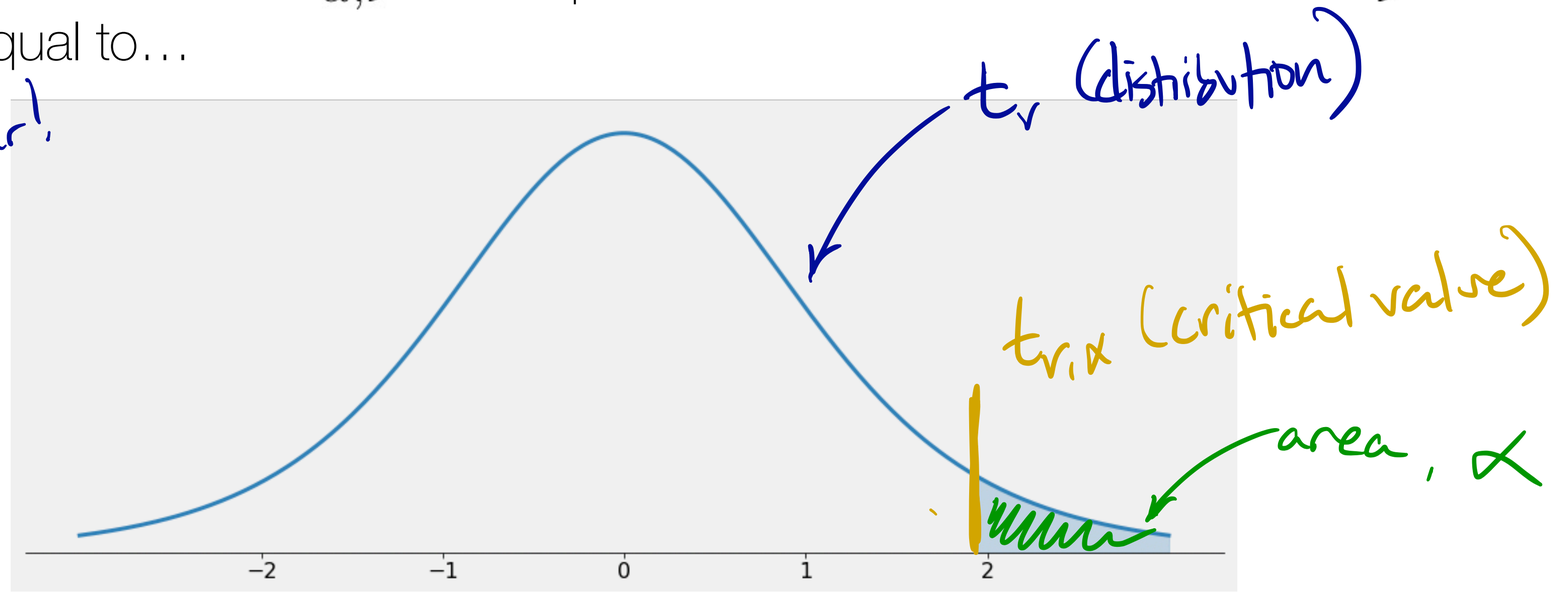
- Each t_ν -curve is bell-shaped and centered at 0
- Each t_ν-curve is more spread out than the standard normal distribution
- As ν increases, the spread of the corresponding t_ν-curve decreases
- As $\nu \to \infty$ the sequence of t_ν-curves approaches the standard normal curve

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:

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

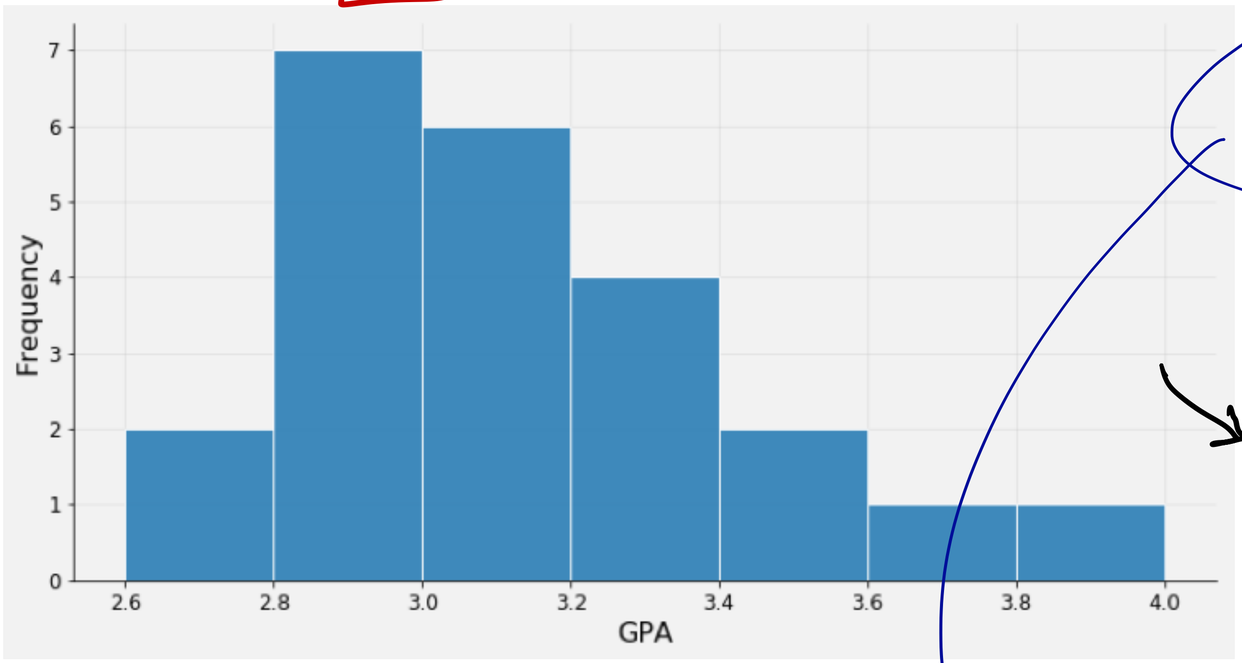
- Or more compactly:

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

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

$H_0 : \theta = \theta_0$

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

t test statistic looks just like z test statistics!

confidence

degrees of freedom

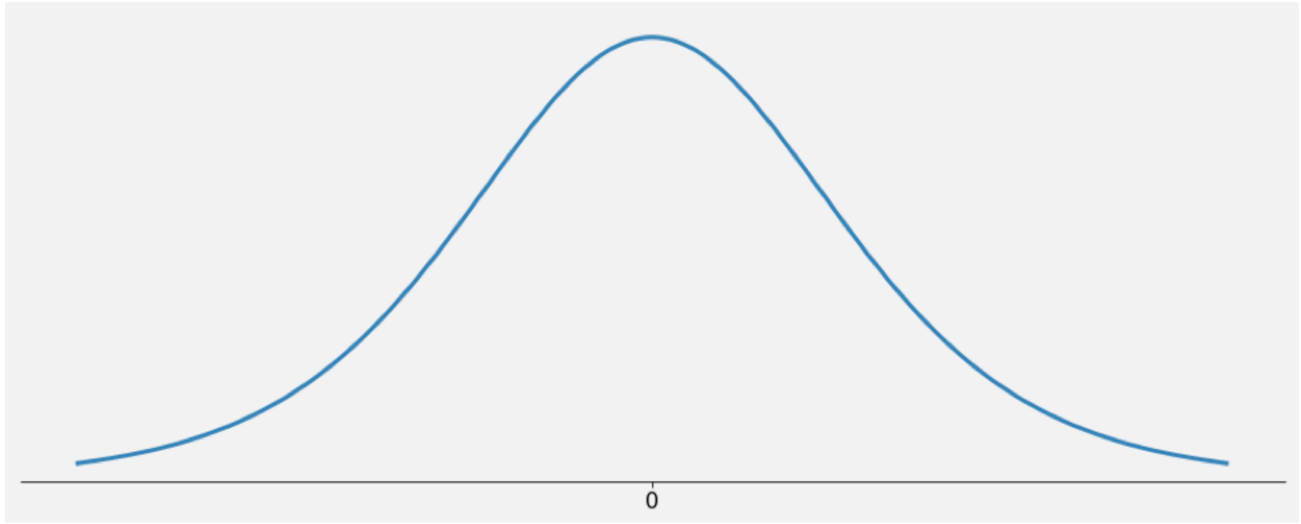
The only difference ... is n is small ($n < 30$)

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

"standardized statistic"

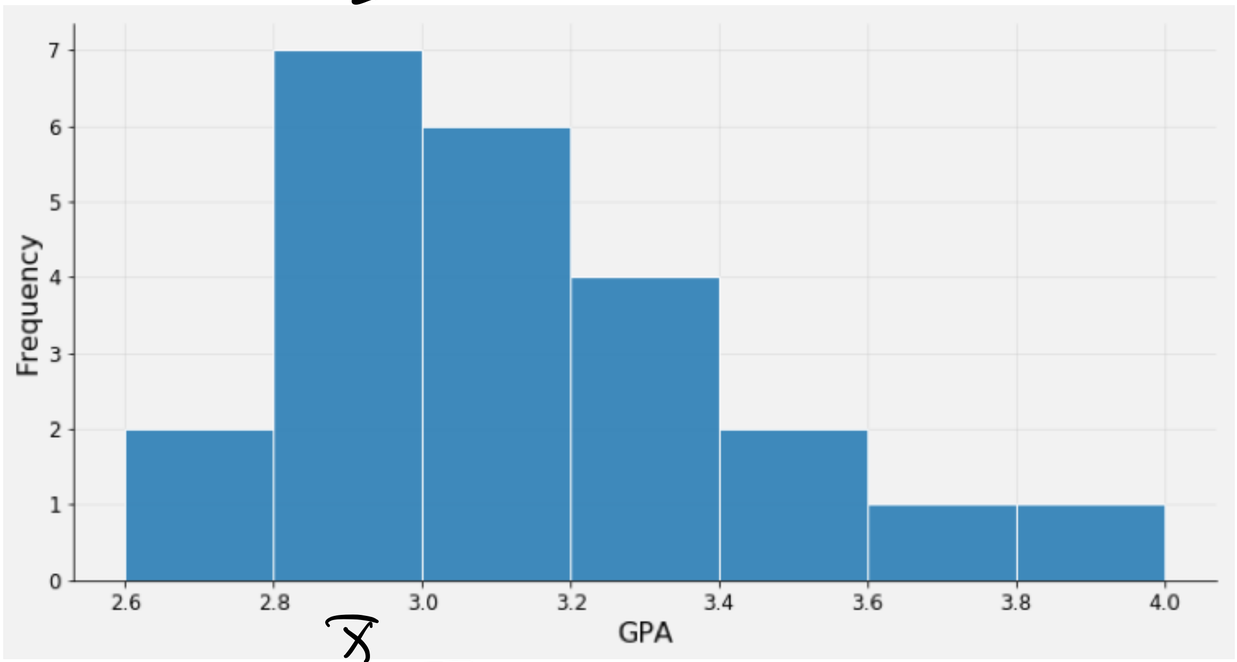
The t-Test, Critical Regions and P-Values

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:

	$n \geq 30$ ("n is large")	$n < 30$ ("n is small")
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

$H_0 : \theta = \theta_0$

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

t test statistic looks just like z test statistics!

confidence

degrees of freedom = n - 1

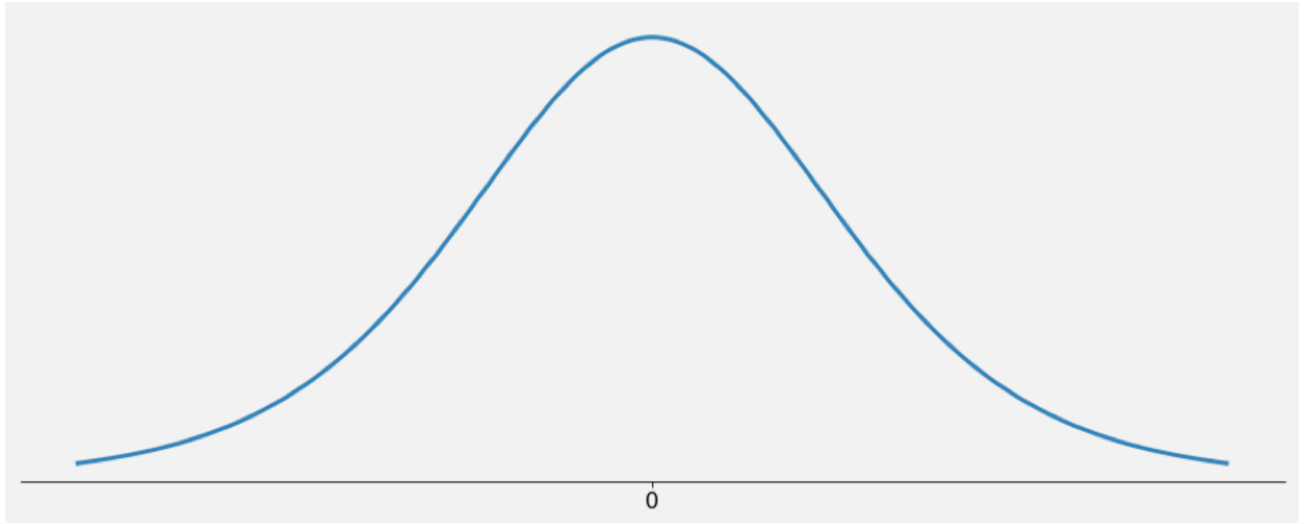
The only difference ... is n is small (n < 30)

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

"standardized statistic"

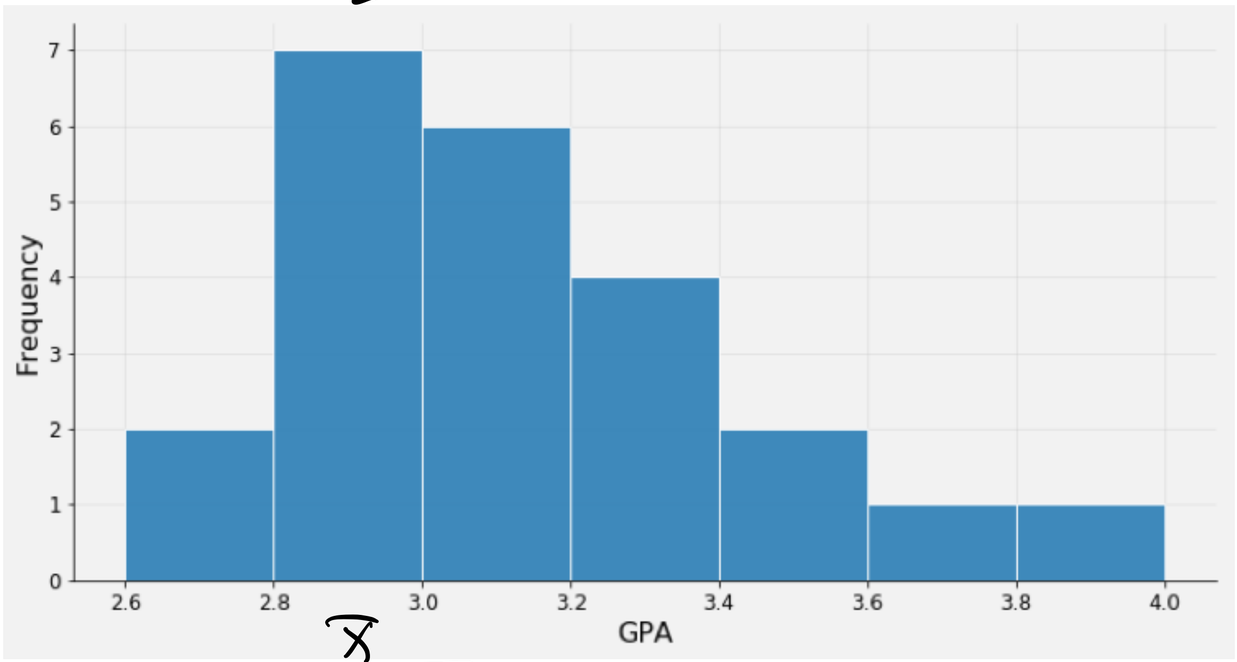
The t-Test, Critical Regions and P-Values

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

$\alpha = 0.1$

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

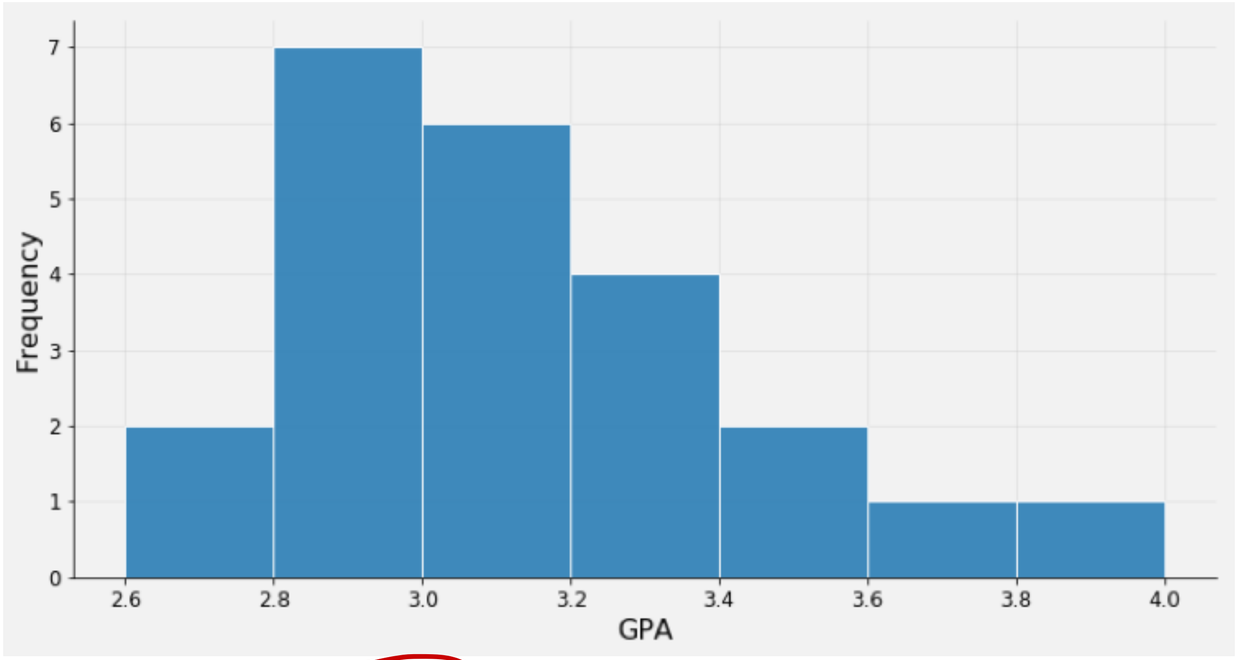
t-Test example (p-value method)

$2 \times$ 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.

$n = 23$

$s = 0.308$

$\bar{x} = 3.146$

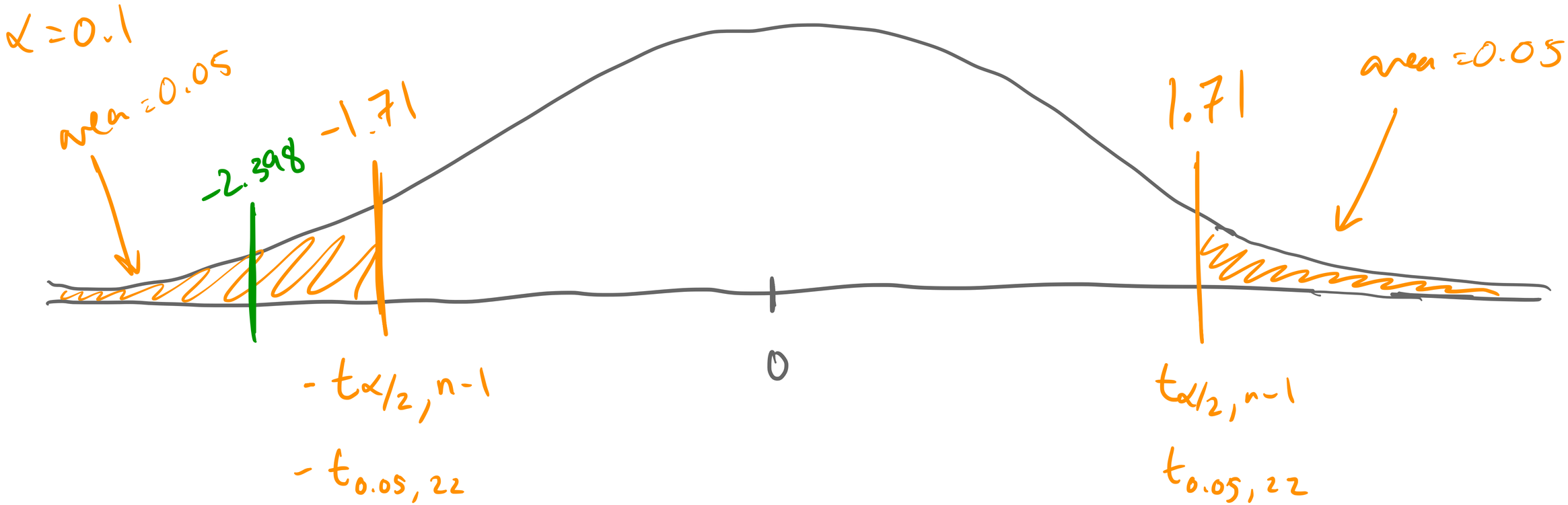
$\alpha = 0.1$

$\mu = 3.30$

$H_0: \mu = 3.30$

$H_1: \mu \neq 3.30$

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*

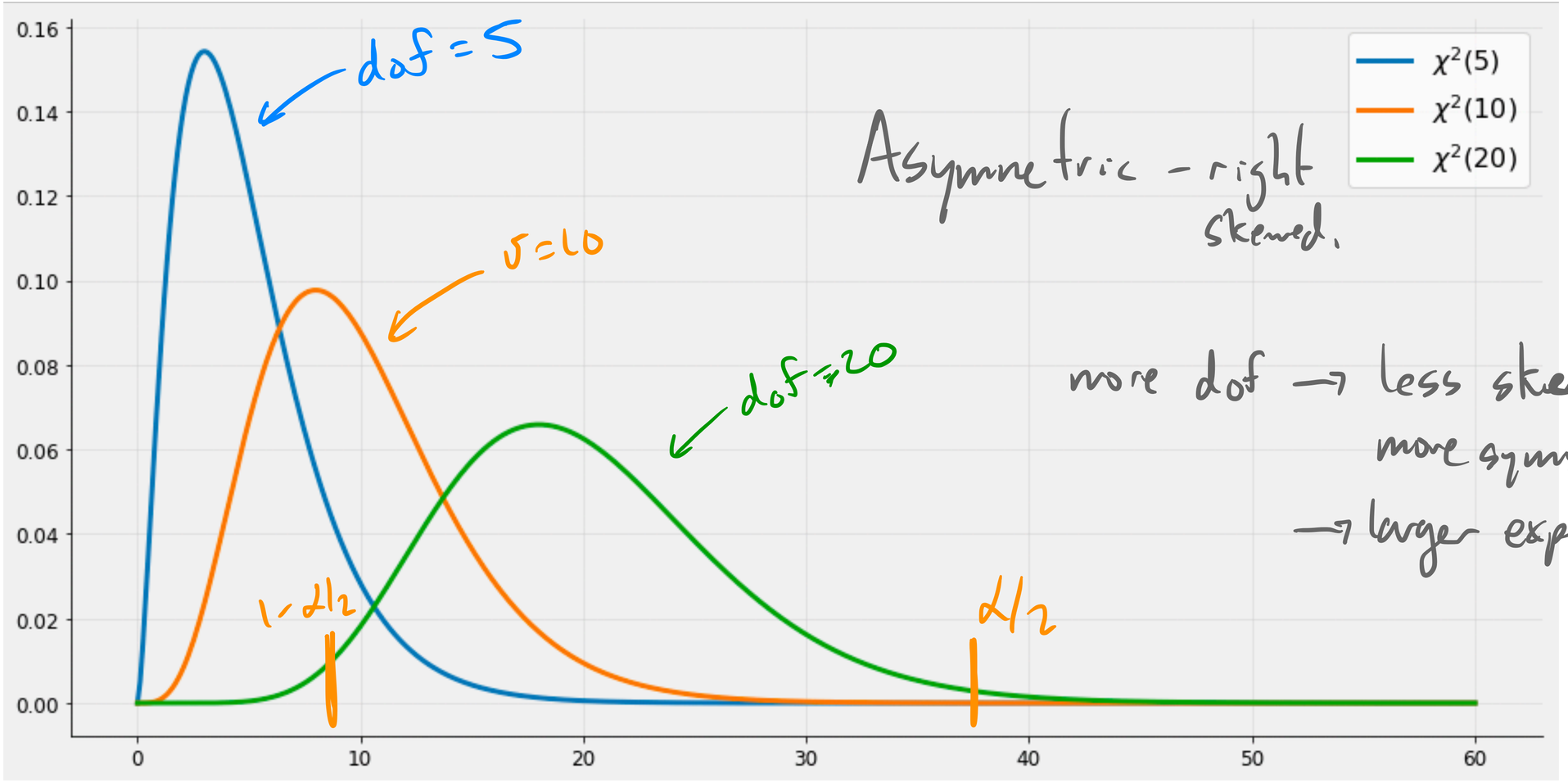
- ~~After Spring Break~~ Today, we'll talk about estimating confidence intervals for the variance of a population using something [wonderful] called **The Bootstrap**.

- 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

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

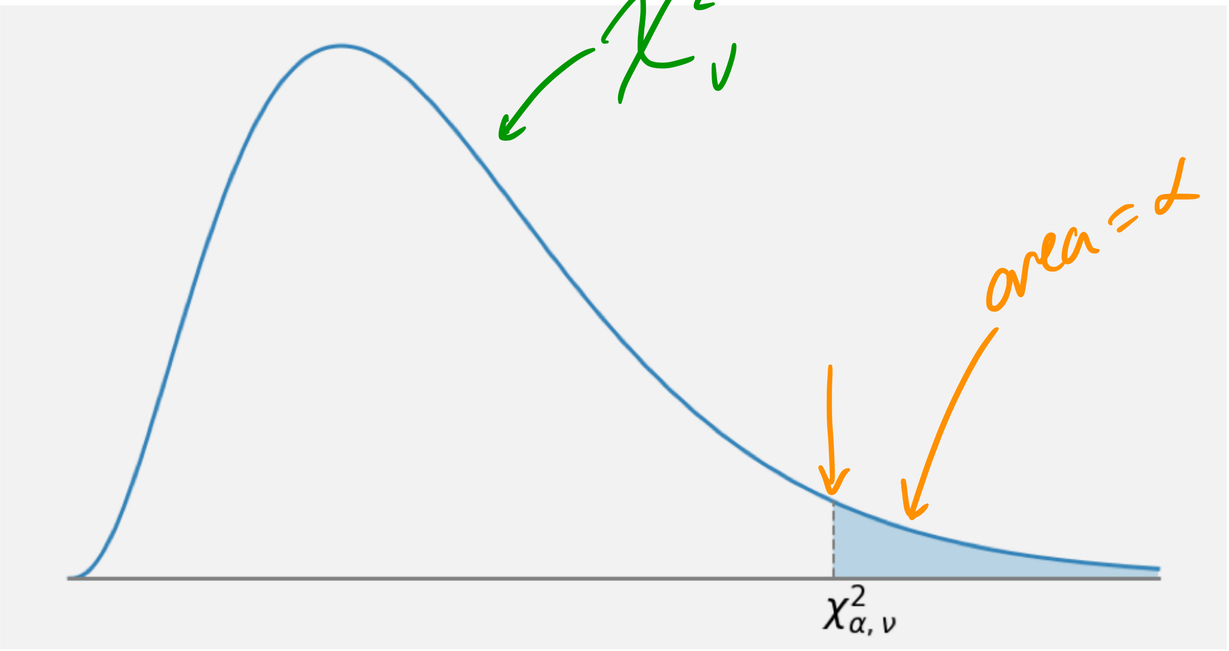
d.o.f.

Left bound

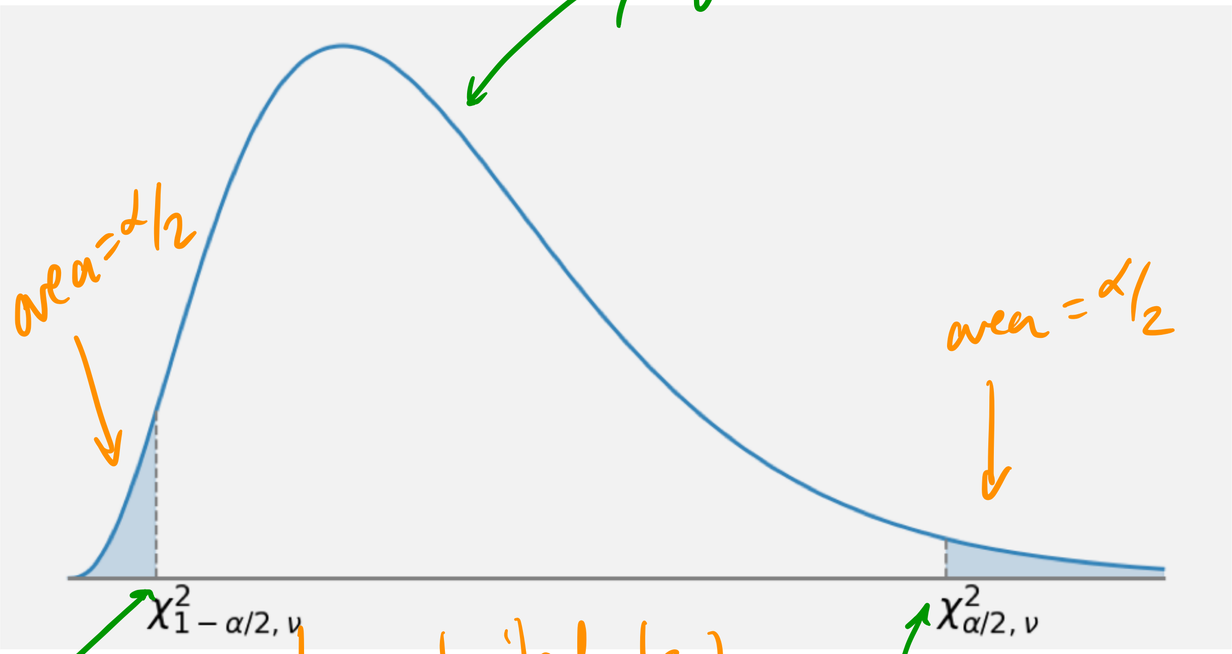
Right bound

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.

$s^2 = 4.2$

$n = 10$

$\alpha = 0.05$

$\alpha/2 = 0.025$

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

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

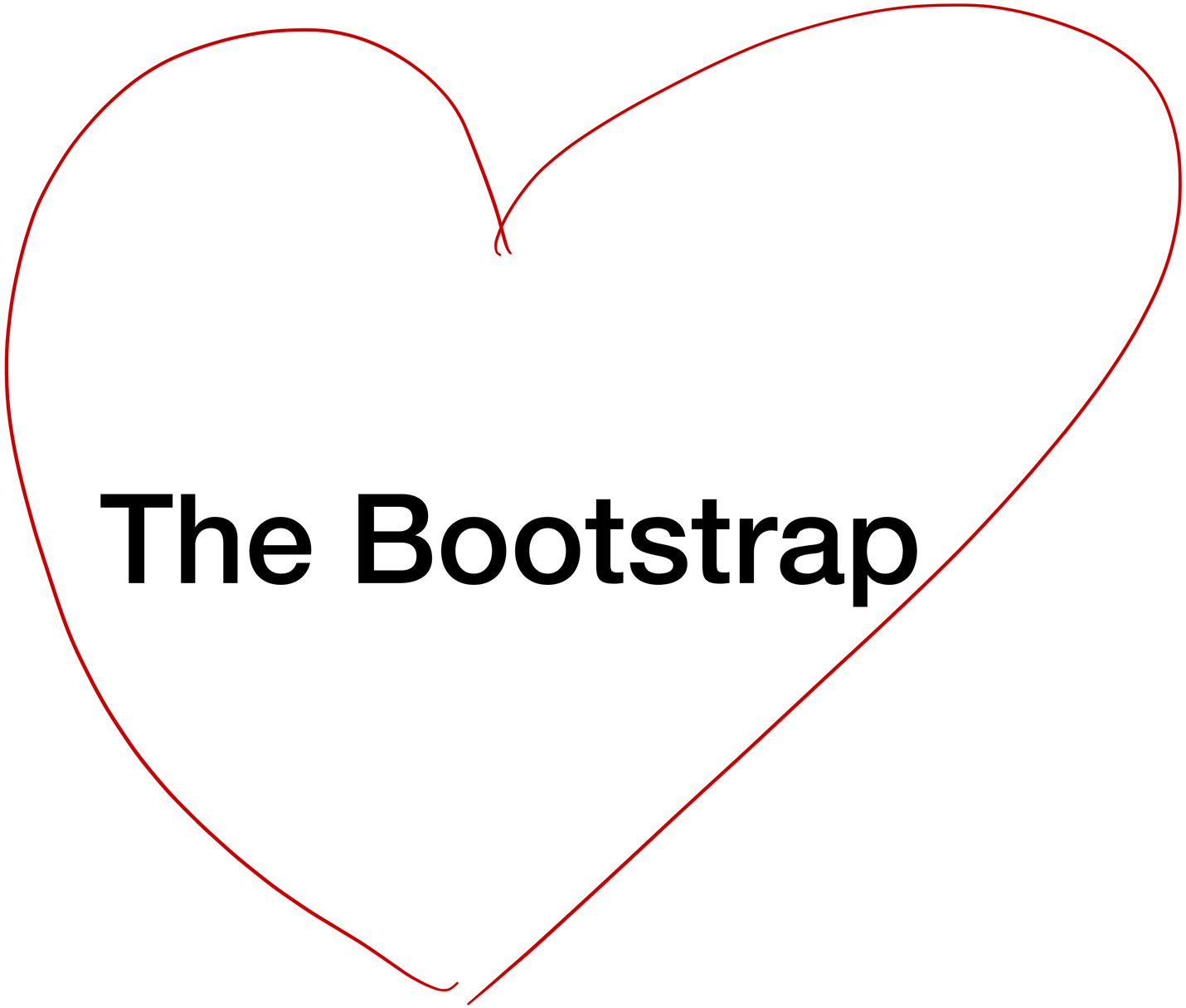
$\frac{(n-1)s^2}{\chi^2}$

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

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

95% CI for σ^2: $[1.99, 14.0]$

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!

CSCI 3022

intro to data science with probability & statistics

Lecture 21
April 4, 2018

The Bootstrap wrapup
Intro to Regression (maybe)

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

Dan Larremore

Stuff & Things

- HW5 due **this Friday**.

OH today 11-1

Fr 8-~~10~~ 9:45

Previously on CSCI 3022

- **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.

Bootstrap: why we like it

- The bootstrap for a confidence interval around the mean is convenient, particularly when there are **not enough samples** to use the CLT.
- Of course, if we *can* use the CLT, we should. So why is the bootstrap so exciting?

Bootstrap: why we like it

- The bootstrap for a confidence interval around the mean is convenient, particularly when there are **not enough samples** to use the CLT.
- Of course, if we *can* use the CLT, we should. So why is the bootstrap so exciting?

We can bootstrap CIs for things other than the mean!

- Median. ✓
- Standard Deviation. ✓
- Other statistical measures that we don't have a theory for.

Bootstrap for the median

90%

- Let's write **pseudocode** for how we would bootstrap a CI for the median:

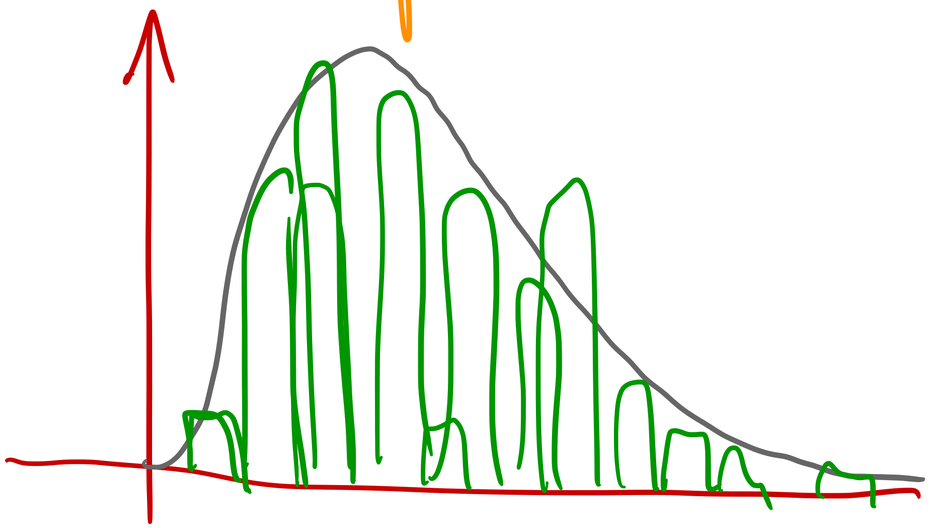
1. Resample. Create M resampled datasets (with replacement)
2. For each resampled dataset, compute the median
3. Take that distr. of medians, and compute
   5th percentile and 95th percentile

CI: $\left[\frac{100-\alpha}{2}\%, 100 - \frac{100-\alpha}{2}\%\right]$

Bootstrap for the variance

- Let's write **pseudocode** for how we would bootstrap a CI for the variance:

No. See prev. slide.



The Non-Parametric Bootstrap

- In the literature—your book, the Wikipedia, etc—you may read about a "non-parametric bootstrap." What is this?

The Non-Parametric Bootstrap

- In the literature—your book, the Wikipedia, etc—you may read about a "non-parametric bootstrap." What is this?
- Let's decode this word, "non-parametric"
- **Definition**: *parametric statistics* assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters.
- Can you name some **examples** of distributions with parameters?

Urchin Eating $Pois(\lambda)$ $N(\mu, \sigma^2)$ $Bin(n, p)$

- Can you name a *non*-parametric distribution we've talked about in class?

Let X be a r.v. such that $P(X=-1)=0.2$, $P(X=0)=0.5$, $P(X=19)=0.3$

The Parametric Bootstrap

Replace with your preferred distr. and statistic

- We call the bootstrap discussed in class today the non-parametric bootstrap because it doesn't assume any parametric distribution. What you resample is what you get.
- **Definition**: the parametric bootstrap estimates a CI for a desired property in two steps: (1) repeatedly estimate the parameter(s) of the known distribution, and then (2) compute a CI for the desired property by sampling from the known known distribution using the parameters that you inferred.

1. Create M bootstrapped datasets.
2. Assume that the data came from Pois(λ), and estimate $\hat{\lambda}$ for each of the M datasets.
3. Use those parameters, and for each one, compute the median.
4. Compute the CI from those medians.

The Parametric Bootstrap

- We call the bootstrap discussed in class today the non-parametric bootstrap because it doesn't assume any parametric distribution. What you resample is what you get.
- **Definition**: the parametric bootstrap estimates a CI for a desired property in two steps: (1) repeatedly estimate the parameter(s) of the known distribution, and then (2) compute a CI for the desired property by sampling from the known known distribution using the parameters that you inferred.
- **Why**? The parametric bootstrap can be shown to do a better job than the non-parametric bootstrap in various scenarios.
- Why not use the parametric bootstrap all the time?
  1. Might not be getting data from a parametric distrib.
  2. Might not know what the parametric distrib. is!

Let’s notebook it up!

CSCI 3022

intro to data science with probability & statistics

Lecture 22
April 6, 2018

Introduction to statistical regression

CI for std. dev of Normally distr. draws is indeed

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

slides are correct :)

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- **HW5** due today. Giddyup!

Thx Google Image

Today: linear regression.

- **Examples**:
  - given a person's age and gender, predict their height.
  - given the square footage and number of bathrooms in a house, predict its sale price.
  - given unemployment, inflation, number of wars, and economic growth, predict the president's approval rating.
  - given a user's browsing history, predict how long they will stay on a product page.
  - given the advertising budget expenditures in various media markets, predict the number of products sold.

Today, we start in the notebook

- Pull that in-class notebook, and let's get started!

Simple Linear Regression Model

aka. features predictors

- **Definitions and Assumptions** of the simple [one independent variable] linear regression model:

one x

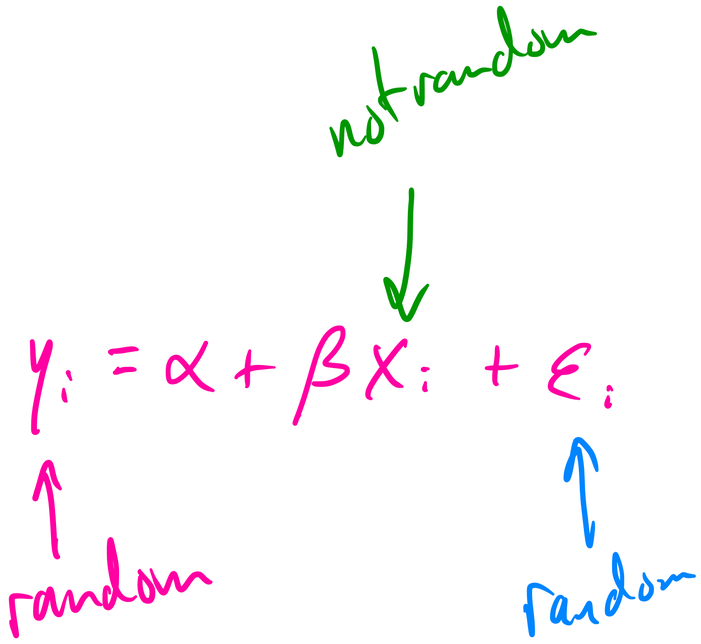
1. $y_i = \underbrace{\alpha + \beta x_i}_{\text{linear}} + \underbrace{\varepsilon_i}_{\text{noise}}$ true underlying relationship is $y = \alpha + \beta x$
2. Each ε_i is drawn independently from same distr. IID
3. $\varepsilon_i \sim N(0, \sigma^2)$

key! mean is 0

SLR Model

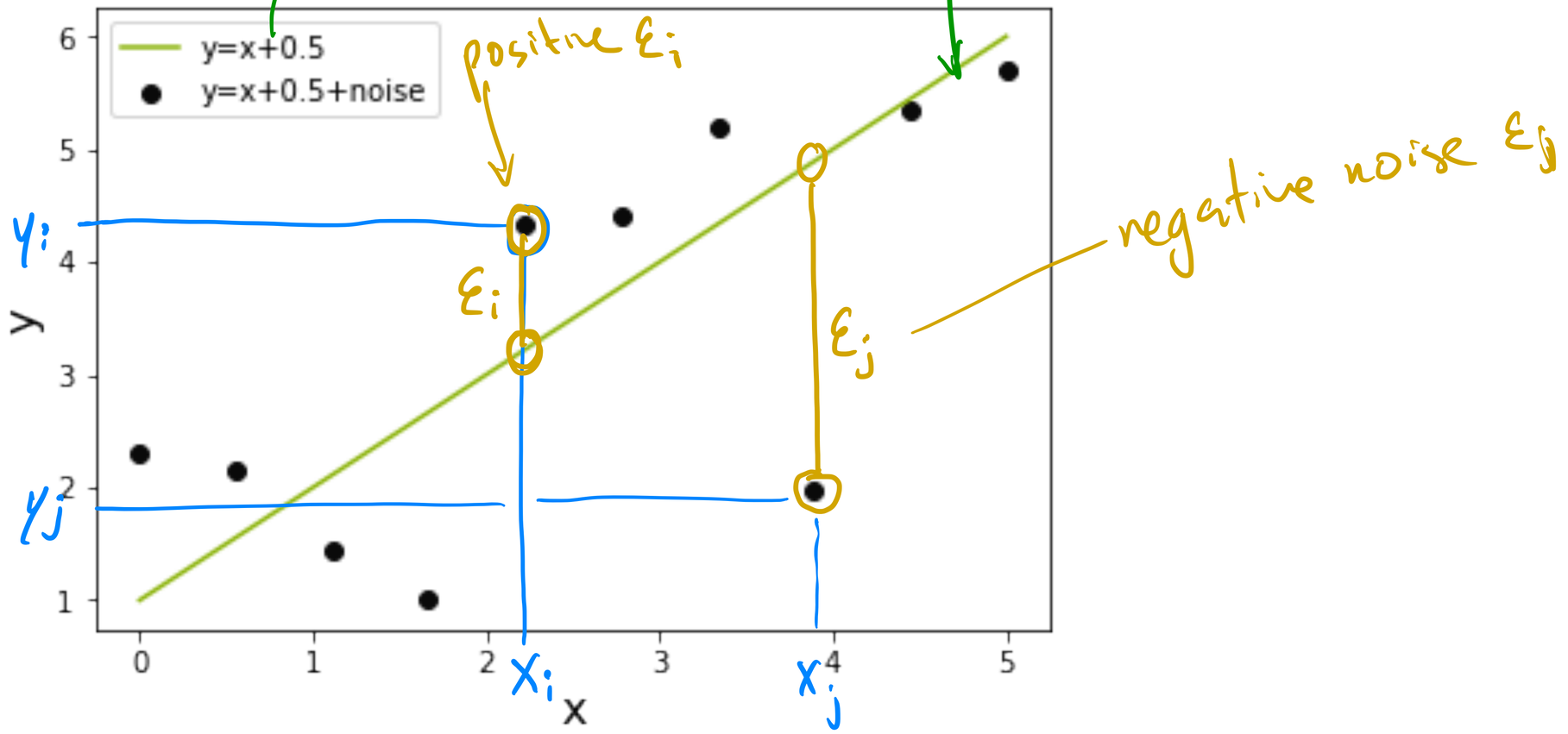
- **Vocabulary** for the SLR model:
- ***X****:* the **independent variable**, the **predictor**, the **explanatory variable**, the **feature**.
  - *X is not random!*
- ***Y***: the **dependent variable**, the **response variable**.
  - For a fixed *x*, Y is *random.*
- **ε**: the **random deviation** or **random error** term.
  - For a fixed *x*, ε *is random*.

What exactly is ε doing?



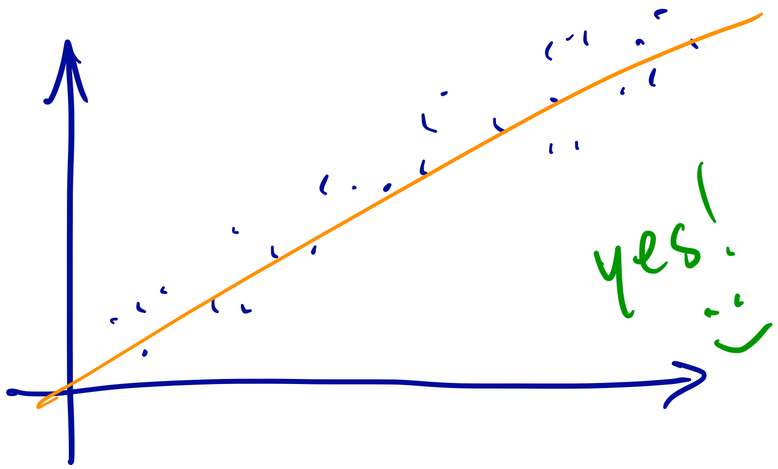
SLR Model

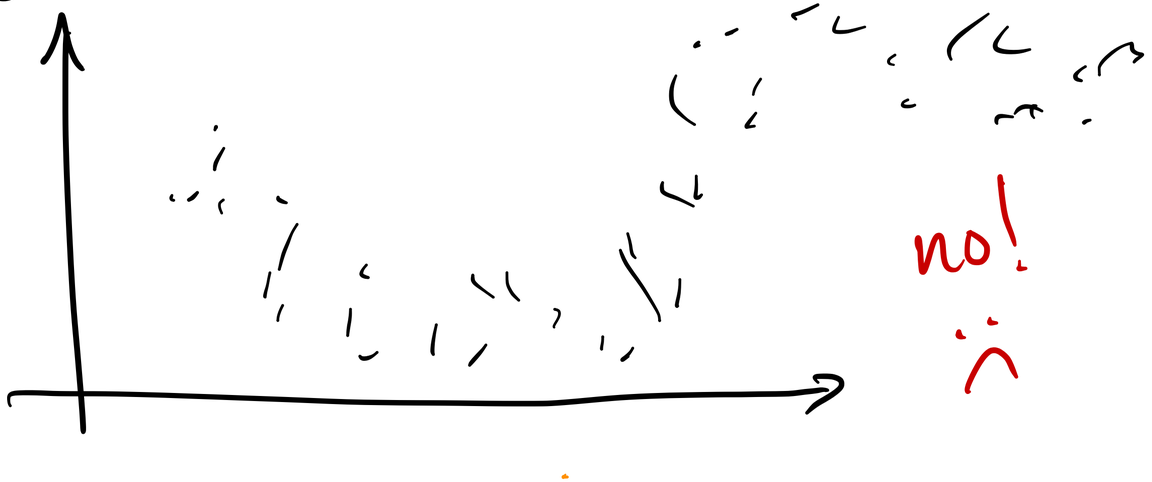
- The points $(x_1,y_1),\ldots,(x_n,y_n)$ resulting from n independent observations will then be scattered about the true regression line:



SLR: theory

- How do we know that a simple linear regression is appropriate?
- Theoretical considerations
  - ① Belief or knowledge about where the noise comes from in my real application.
  - ② relationship between X, Y.
- Scatterplots
- Knowledge of the process generating the data.





SLR Model

$$Y = \alpha + \beta X + \varepsilon$$

$$E[Y] = E[\alpha + \beta X + \varepsilon]$$

$$= E[\alpha] + E[\beta X] + E[\varepsilon]$$

$$= \alpha + \beta X + 0$$

recall $\varepsilon \sim N(0, \sigma^2)$

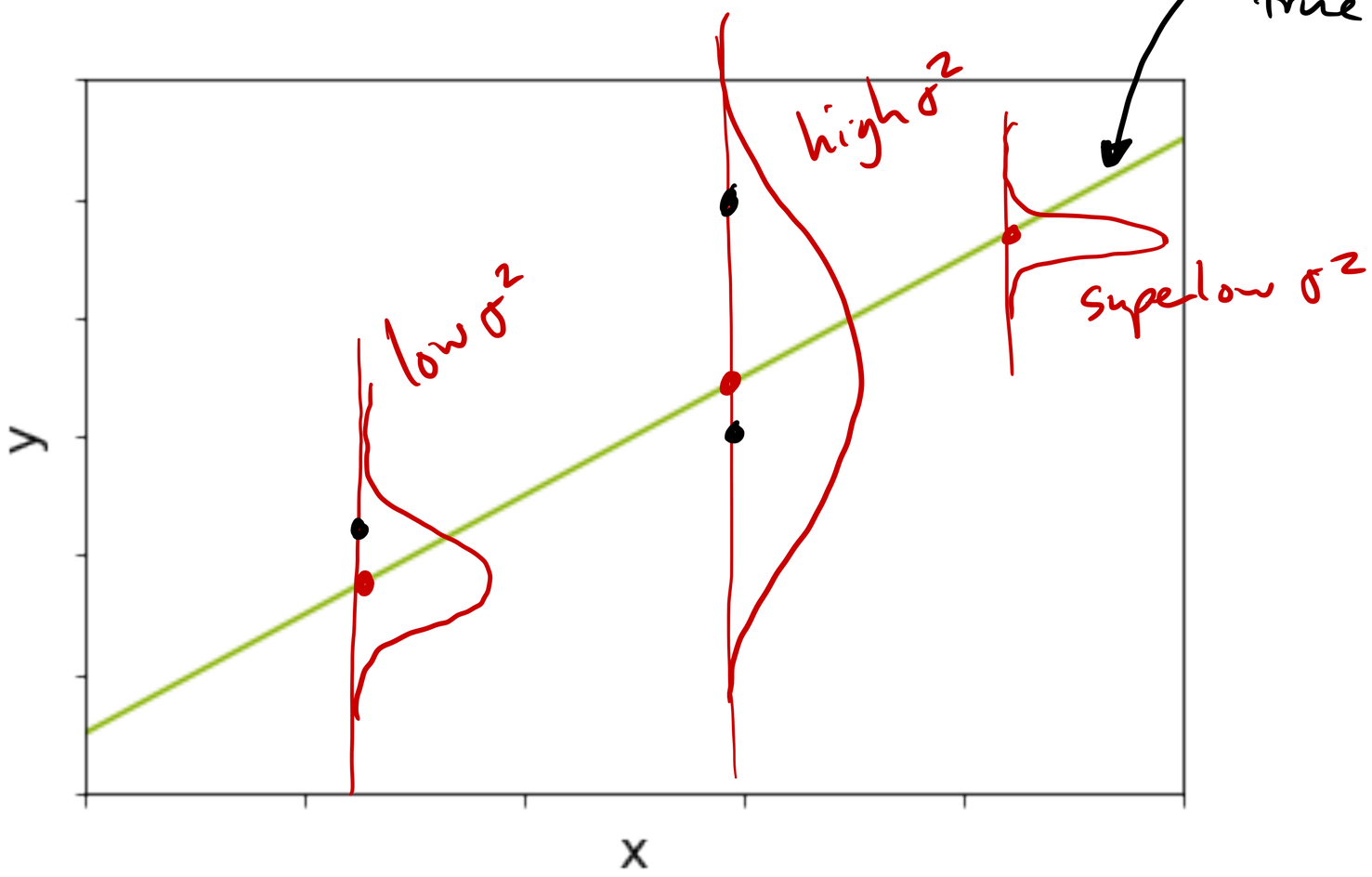
- **Interpreting parameters**:

- Y is a random variable. What is its expectation, E[Y]? $= \alpha + \beta X$
- α (the intercept of the true regression line):
  - The average value of Y when x is zero. This is sometimes called the **baseline average**.
- β (the slope of the true regression line):
  - The average change in Y associated with a 1-unit increase in the value of x.

The Error Term

$\varepsilon_i \sim N(0, \sigma^2)$

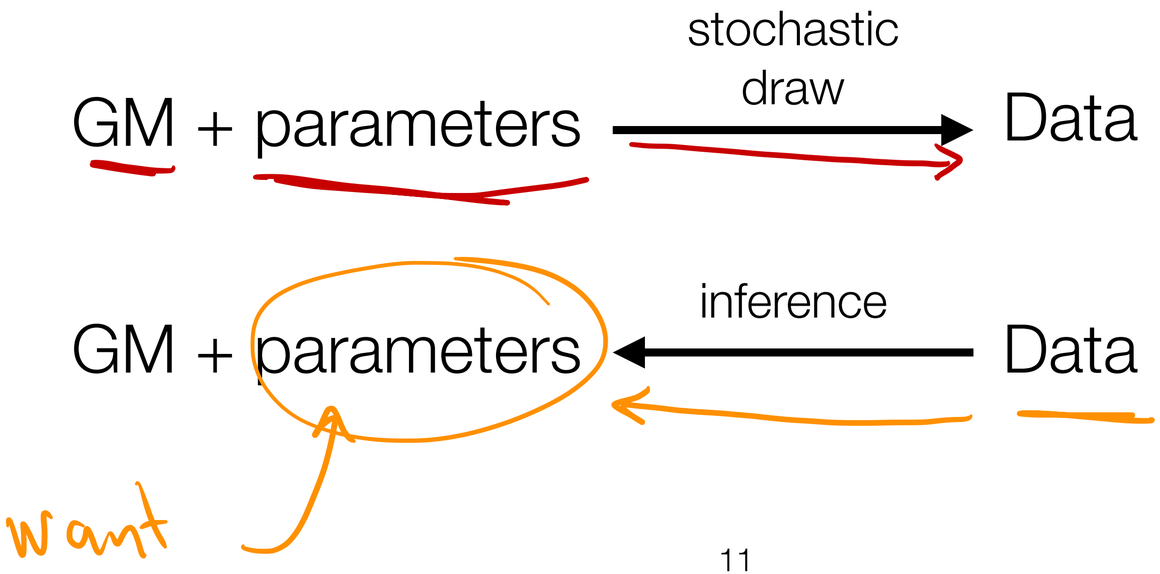
$y_i = \alpha + \beta x_i$ true regression line



- The variance parameter σ^2 determines the extent to which each normal curve spreads out about the regression line.

Generative model vs regression

- So far, we've written down a **generative model** where we choose **parameters** and then **generate data stochastically**.
- But really, we want to run this process in reverse. We have data, and we want to **find/learn/estimate the parameters** that explain the data.



How can we estimate model parameters?

- Plan of attack: the variance of our model σ^2 will be smallest if the differences between the estimate of the true line and each point is the smallest. **This is our goal: minimize** σ^2

- We use our sample data, which consists of *n* observed (x,y) pairs to *estimate* the regression line. $(x_1,y_1), \ldots, (x_n,y_n)$ ingredients

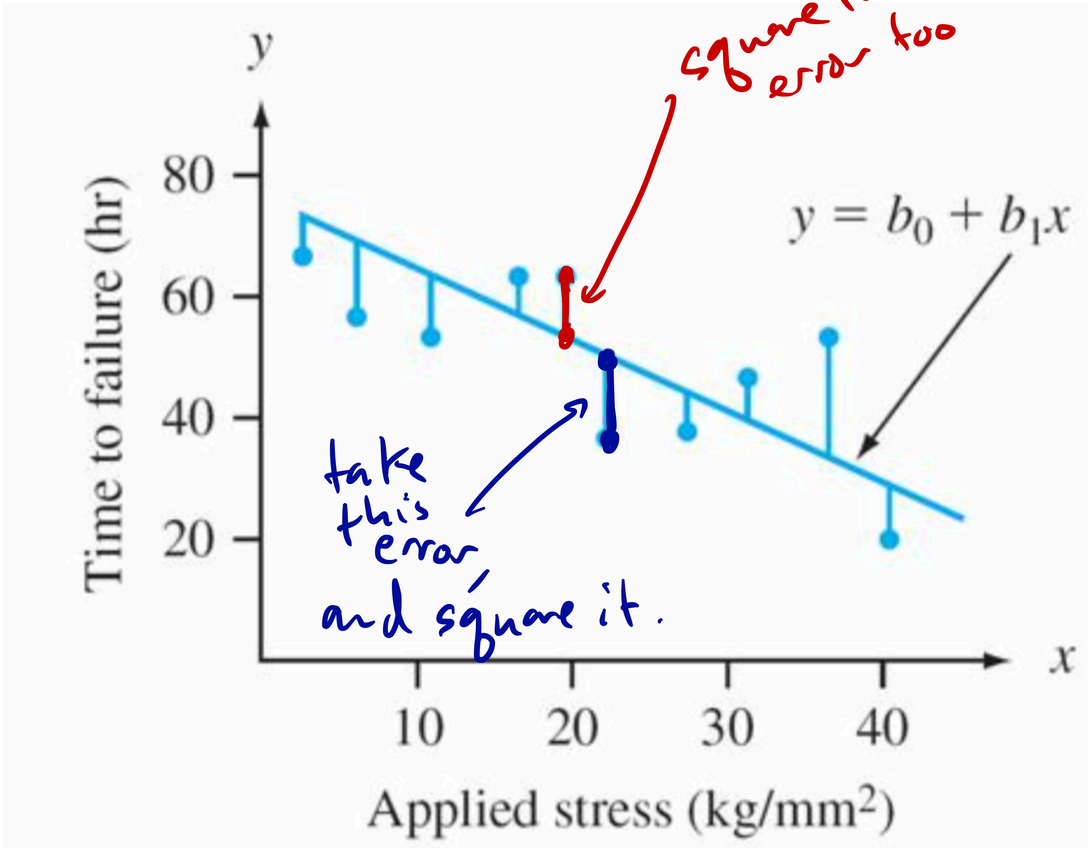
goal: cook up α, β

- What are we assuming about each of the data pairs?

Independence of errors ε_1, ε_1 has no bearing on $\varepsilon_2, \varepsilon_3 \ldots$

Estimating model parameters

- The **best fit line** is motivated by the principle of least squares, which can be traced back to the German mathematician **Gauss** (1777– 1855):.
- A line provides the best fit to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.



in fact, square all the errors.

add 'em up!

Estimating model parameters

- The sum of the squared deviations (also called errors) from the points $(x_1,y_1), \ldots, (x_n,y_n)$ to the line is then

$$SSE(\alpha,\beta) = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i) \right)^2$$

sum of squared errors

y_i: actual data

$\alpha + \beta x_i$: my prediction

- The "point estimates" of the slope and intercept parameters are called the **least squares estimates**, and are defined to be the values that minimize the SSE.

Find α, β to minimize $SSE(\alpha, \beta)$

Estimating model parameters

- The **fitted regression line** or **least squares line** is then the line whose equation is:

$$y = \hat{\alpha} + \hat{\beta} x$$

\hat{\alpha}

hat means "best fit" or "point estimates"

- The minimizing values of α and β are found by taking [partial] derivates of SSE with respect to α and β , setting each equal to zero, and solving.
- [Take a derivative and set=0? Sounds like calculus!]

Calc III

Estimating model parameters

$$SSE(\alpha,\beta) = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i) \right)^2$$

$$\frac{\partial SSE(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} \left(y_i - (\alpha + \beta x_i) \right)^2$$

set to zero to minimize

$$= \sum_{i=1}^{n} 2 \left(y_i - (\alpha + \beta x_i) \right)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \alpha - \beta x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \underbrace{\sum_{i=1}^{n} \alpha}_{-n\alpha} - \beta \sum_{i=1}^{n} x_i = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} n\alpha - \beta \frac{1}{n} \sum_{i=1}^{n} x_i = 0$$

$$\bar{y} - \alpha - \beta \bar{x} = 0$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \quad (2)$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (1)$$

CSCI 3022

intro to data science with probability & statistics

Lecture 23
April 9, 2018

Statistical regression
and
Inference in Regression

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

Tony
~~Dan Larremore~~

Stuff & Things

- **HW6** posted tonight!. Giddyup!

Last time on CSCI3022: SLR

- Given data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ fit a simple linear regression of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

- Compute estimates of the intercept and slope parameters by minimizing:

$$SSE = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

- The least-squares estimates of the parameters are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Does it work?

- Let’s dig into problem 2 in the in-class notebook to see how this works.

Residuals

- The **fitted** or **predicted** values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ are obtained by substituting $x_1, \ldots x_n$ into the equation of the estimated regression line.

- The **residuals** are the differences between the observed and fitted *y* values:

$$r_i = y_i - \hat{y}_i = y_i - [\hat{\alpha} + \hat{\beta} x_i]$$

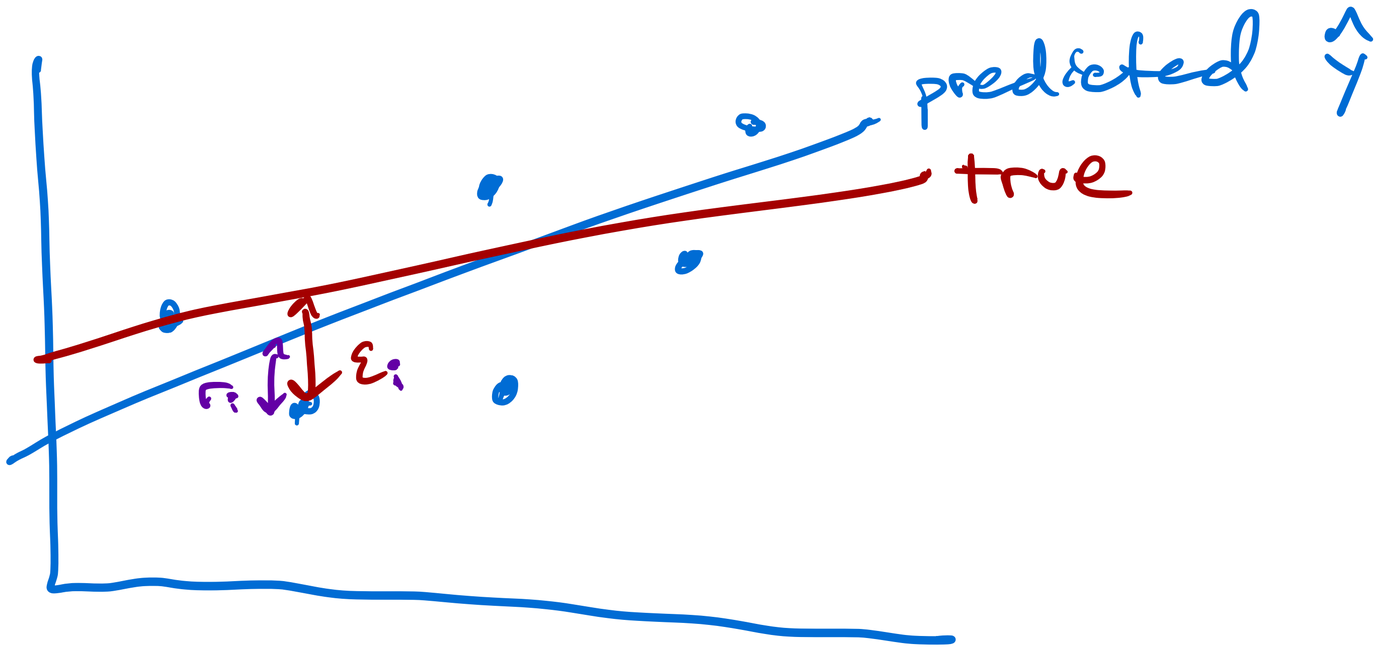
Residuals

truth → $y = \alpha + \beta x$

$N(0, \sigma^2)$

measure: $y_i = \alpha + \beta x_i + \varepsilon_i$

- Why are the residuals estimates of the error?



Want to estimate true line as well as possible

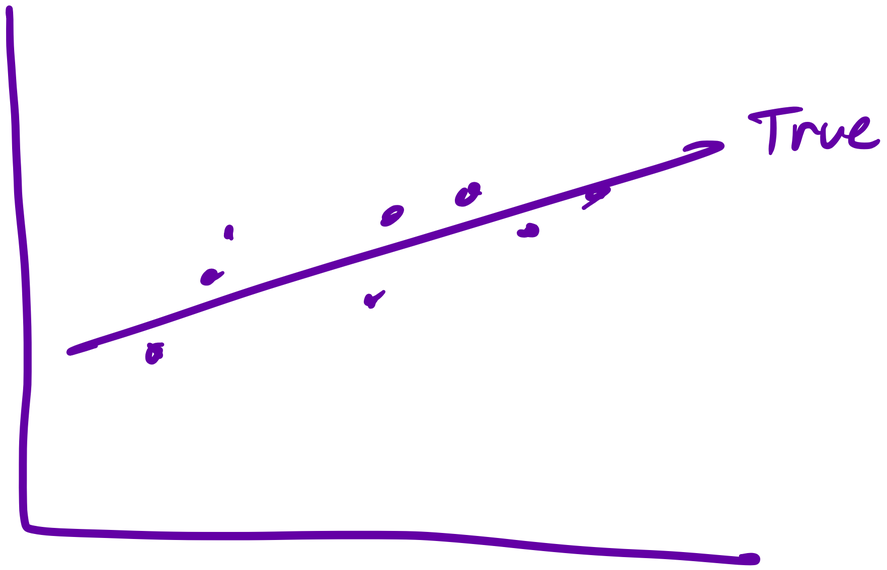
→ minimize sum of squared r_i (SSE)

For the rest of today:

- **How can we**:
  - Estimate the variance in the population of estimates?
  - Quantify the goodness-of-fit in our simple linear regression model?
  - Perform inference on the regression parameters?

Estimating the variance

- The parameter σ^2 determines the spread of the data about the true regression line. [We experimented with this in the notebooks!]



σ low

σ large

Generally, we don't know what σ is!

Estimating the variance

$Y = \alpha + \beta x + \varepsilon$

$\varepsilon \sim N(0, \sigma^2)$

$\hat{\sigma}^2 = \frac{SSE}{n-2}$

- The divisor (n-2) in the estimate of σ^2 is the number of *degrees of freedom* (abbreviated df) associated with the estimate of SSE.

- This is because to obtain $\hat{\sigma}^2$, the two parameters $\hat{\alpha}$ and $\hat{\beta}$ must first be estimated, which results in a loss of 2 degrees of freedom.

Mean: $\bar{x} = \frac{1}{n} \sum x_i$

Var: $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

The coefficient of determination

- The coefficient of determination, R^2 quantifies how well the model explains the data.

$$SSE = \sum (y_i - \hat{y}_i)^2$$ ← how much uncertainty/variance in y_i can be explained by model $\hat{y}_i$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$ ← regression sum of squares

$$SST = \sum (y_i - \bar{y})^2$$

$$SST = SSR + SSE$$ = what can be explained by regression + what can't be explained by regression

- R^2 is a value between 0 and 1.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \leq 1$$

The coefficient of determination

The **sum of squared errors** (SSE)

can be interpreted as a measure of how much variation in *y* is left unexplained by the model: how much variation *cannot* be attributed to a linear relationship?

The **regression sum of squares** is given by

SSR = defined before

A quantitative measure of the total amount of variation in observed *y* values is given by the so-called **total sum of squares**

SST

The coefficient of determination

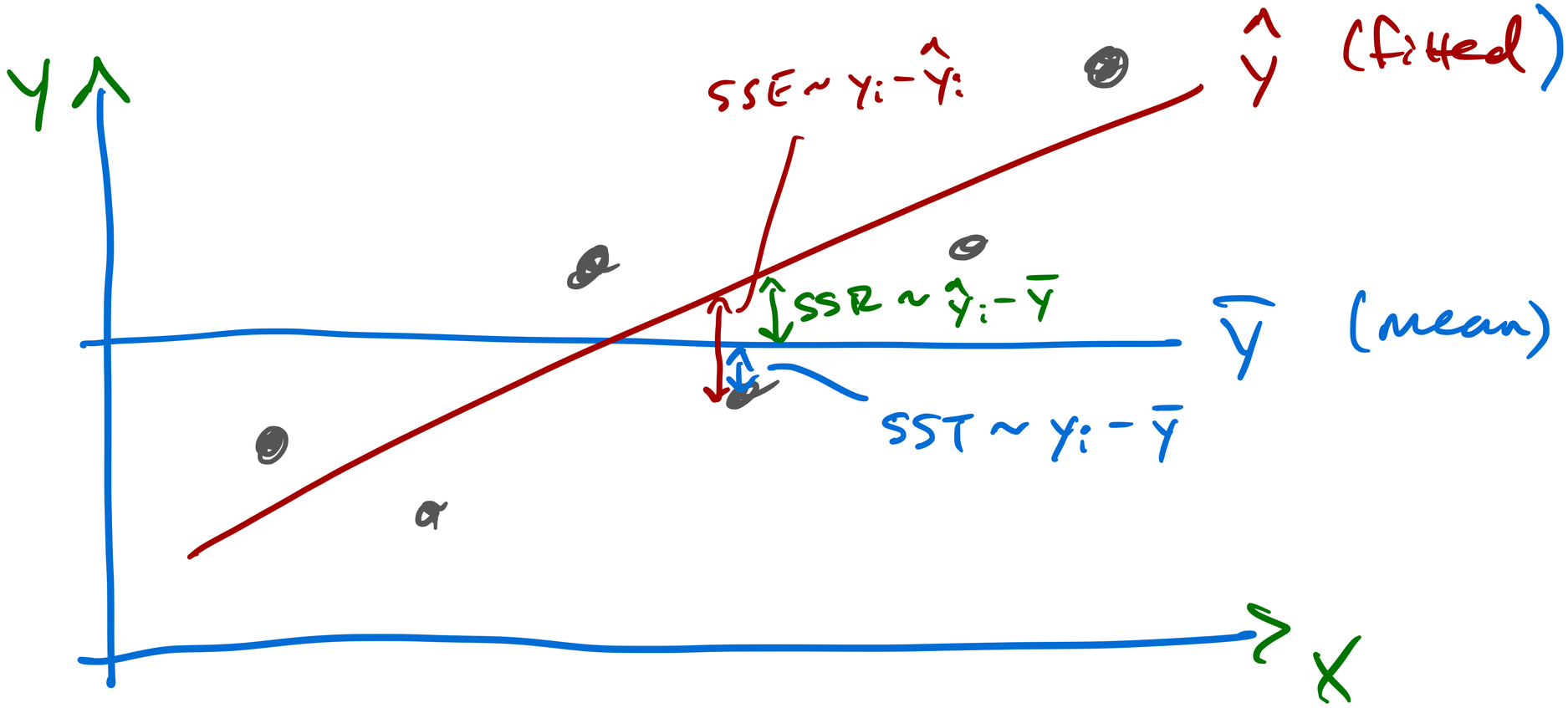
- The sum of squared deviations about the least-squares line is smaller than the sum of squared deviations about any other line, i.e. SSE < SST unless the horizontal line itself is the least-squares line
- The ratio SSE/SST is the proportion of total variation in the data that cannot be explained by the simple linear regression model, and the coefficient of determination is

$$R^2 = 1 - \frac{SSE}{SST}$$

SSE ← variability we can't explain w/ SLR

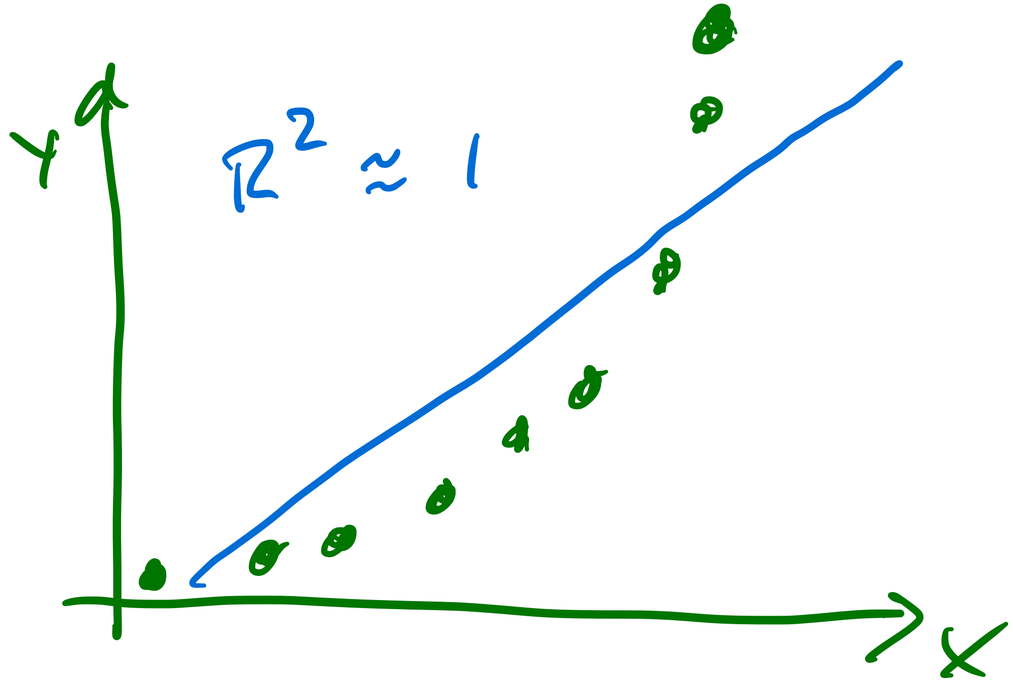
SST ← total variability

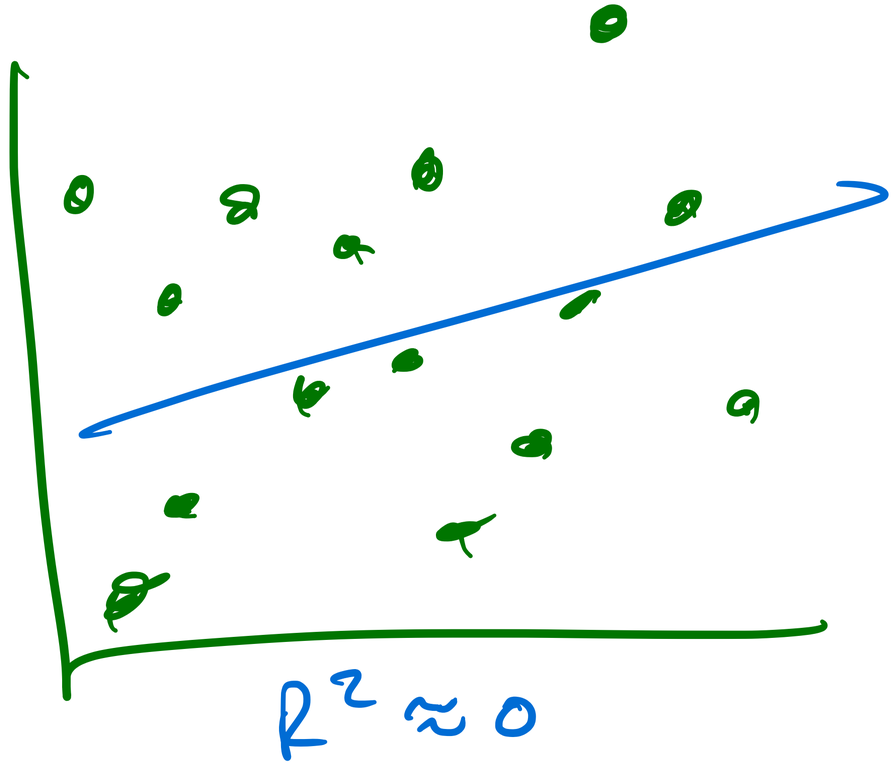
The coefficient of determination



The coefficient of determination

- Note: R^2 is the proportion of total variation in the data that is explained by the model.
- But: R^2 does *not* tell you that you necessarily have the correct model!





Inference about parameters

- The parameters in simple linear regression have distributions! We demonstrated this in the in-class notebook last time.
- From these distributions, we can conduct hypothesis tests (e.g.: H?), compute confidence intervals, etc.
- **Distributions**:

especially for β: $H_0: \beta = c$ (e.g. 0)
$H_1: \beta \neq c$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

$$SE(\hat{\beta}) = \sqrt{\frac{SSE}{n-2} \Big/ \sum_i (x_i - \bar{x})^2}$$

Inferences about the parameters

- **Confidence intervals**: $100 \times (1-\alpha)\%$ CI:

$$\hat{\beta} \pm t_{\alpha/2,\, n-2} \times SE(\hat{\beta})$$

- **Tests**:

$H_0: \beta = 0$? $\longrightarrow$ Test statistic: $t = \dfrac{\hat{\beta} - 0}{SE(\hat{\beta})}$ $\longrightarrow$ Compute p-value or C.I.

$H_1: \beta \neq 0$

CSCI 3022

intro to data science with probability & statistics

Lecture 24
April 11, 2018
and 13

Multiple Linear Regression

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- HW6 due Next Friday. Suggested milestones:
  - Problems 1-3 **this** week.
  - Problems 4-6 next week.

2 times ago on CSCI3022: SLR

- Given data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ fit a simple linear regression of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

- Compute estimates of the intercept and slope parameters by minimizing:

$$SSE = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

- The least-squares estimates of the parameters are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Last time on CSCI 3022:

- We can perform inference on slope to determine if relationship is significant.

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

$$se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

middle, signif, width

$$CI : \hat{\beta} \pm t_{\alpha/2,n-2} \times se(\hat{\beta})$$

- We can use the Coefficient of Determination to evaluate goodness-of-fit of SLR model

R^2

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

- If R^2 is close to 1 then the model fits the data relatively well.

Regression with Multiple Features

In most practical applications there are multiple features or predictors that potentially have an effect on the response.

Example: Suppose that y represents the sale price of a house. Reasonable features associated with sale price might be:

- x_1 : the interior size of the house
- x_2 : the size of the lot on which the house sits
- x_3 : the number of bedrooms in the house
- x_4 : the number of bathrooms in the house
- x_5 : the age of the house

Regression with Multiple Features

Questions we would like to answer in the next few classes:

- Is at least one of the features useful in predicting the response?
- Do all of the features help to explain the response, or is it just a subset?
- How well does the model fit the data?
- Given a set of predictor values, what response should we predict, and how accurate is our prediction?

We will look at these questions over the course of the week, but first let's do a little exploration of a multiple feature data set and remind ourselves about SLR

Advertising Budget Example

- Get in groups (pairs at least!), get out your laptops, and open the Lecture 22 In-Class Notebook
- **Example**: Data is provided about the sales of a particular product in 200 different markets, along with advertising budgets for each market for three different media types: TV, Radio, and Newspaper.
- The sales response is given in thousands of units, and each of the advertising budget features are given in thousands of dollars.
- We ~~will begin~~ began by fitting individual SLR models with the advertising budget as the feature and the sales as a response.

Multiple Linear Regression

- We've seen from the Advertising example that SLR analysis has indicated that there is a significant relationship between each of the media types: TV, Radio, and Newspaper on the sales of the product. for each $p < 0.01$
- But individual SLR models only show the effect of each media type in a vacuum. To get a clearer picture of what's going on, we want to consider the effect of all three advertising types on sales simultaneously
- This is where Multiple Linear Regression (MLR) comes in
- **Def**: In MLR, the data is assumed to come from a model of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

X_1 news
X_2 radio
X_3 TV

intercept

Multiple Linear Regression

$tv $news $radio outcome (sales)

- This means that for each of n data points $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ we assume

intercept

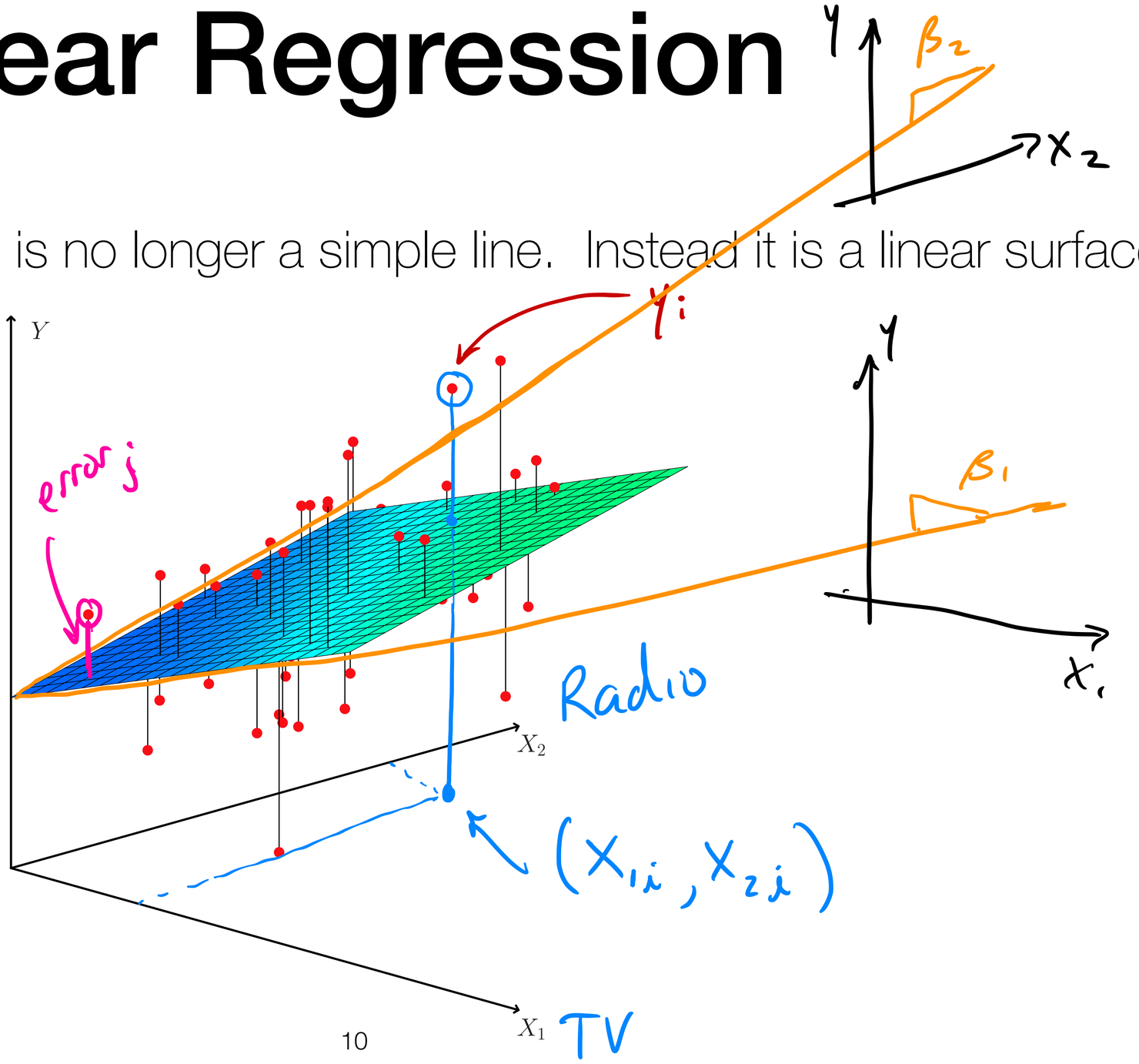
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

↖ noise!

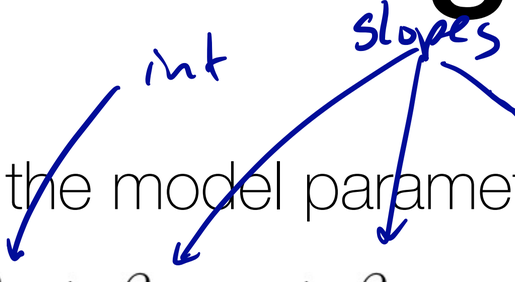
- We make similar assumptions as in the case of SLR:
  - ϵ_i are independent of each other
  - $\epsilon_i \sim N(0, \sigma^2)$

Multiple Linear Regression

- Note that our model is no longer a simple line. Instead it is a linear surface



Multiple Linear Regression



- The interpretation of the model parameters are similar to that of SLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Parameter β_k is the expected change in the response associated with a 1-unit change in the value of x_k while all other features are held fixed.

- **House Sale Price Example**:

$y = 100 + 3x_1 + 2x_2 + 2.5x_3$

If I have a house w/ same lot size (x_2) and same sq. ft. (x_1), then, if I ↑ the # of cool dogs (x_3) in nbhd by 1, then price(y) ↑ 2.5.

Estimating the MLR parameters

- Just as in the case of SLR, we have no hope of discovering the true model parameters, and so have to estimate them from the data. Our estimated model will be

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- As before, we will determine the estimated parameters by minimizing the sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi}\right]\right)^2$$

- The SSE is again interpreted as the measure of how much variation is left in the data that cannot be explained by the model.

- Note: Without linear algebra, it is difficult to write down a closed-form expression for the parameter estimates. For now we will simply see how we can find them in Python. Later we'll see how to estimate parameters using the method of **Stochastic Gradient Descent**.

Advertising Budget Example

- Group back up! Let's see how we can find an MLR model for the Advertising data…

Advertising Budget Example

- OK. We've determined that the MLR model for the advertising data is:

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

- **Question**: Why did our SLR models indicate a positive relationship between newspaper advertising and product sales, but our MLR model did not?

With multiple features, we must consider the relationship between the features themselves.

If 2 features have a strong rel'ship w/ each other, then it's possible that only 1 of them has a rel'ship w/ the response variable.

A Correlation Parable

- **Example**: A simple linear regression analysis of **shark attacks vs ice cream sales** at a Southern California beach indicates that there is a strong relationship between the two.
- **Question**: Do you think that this relationship is real?

?
=

A Correlation Parable

- **Example**: A simple linear regression analysis of **shark attacks vs ice cream sales** at a Southern California beach indicates that there is a strong relationship between the two.
- **Question**: Do you think that this relationship is real?
- **Answer**: Probably not. Higher temps cause more people to head to the beach, increasing the chance of shark attacks. And, higher temps cause more people to buy ice cream.
- If we ran a MLR analysis with shark attacks as the response and temperature and ice cream sales as features, our model would show the strong relationship between temperature and shark attacks, and an insignificant relationship between shark attacks and ice cream sales!
- In such an analysis, we say that when we **adjust** or **control** for temperature, the relationship between ice cream sales and shark attacks disappears.

Advertising Budget Example

- **Question**: Based on our rather ~~absurd~~ reasonable shark attack example, can you explain why newspaper spending became less significant in our MLR of product sales?

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

Newspaper sales $ was a surrogate for one of our other features.

When we control for radio $, rel'ship between sales and news disappears.

Covariance and Correlation of Features

- One way to discover this relationship between features is to do a **correlation analysis**. We want to know, if the value of one feature goes up is it likely that the other feature will go up as well? Similarly, we might find that if one feature goes up is it likely that the other feature will go down?

- **Def**: Let X and Y be random variables. The covariance between X and Y is given by

$$Cov(X,Y) = E\Big[(X - E[X])(Y - E[Y])\Big]$$

$Cov(X,X) = E[(X-E[X])^2] = Var(X)$

- **Def**: The correlation coefficient $\rho(X,Y)$ is a measure between -1 and 1, given by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Estimating Covariance and Correlation

- We can estimate these relationships from the data using formulas analogous to the sample variance.
- **Def**: The sample covariance is given by

$$S_{xy} = \sum_{i} (x_i - \bar{x})(y_i - \bar{y})$$

est of cov.

- **Def**: The sample correlation coefficient is then given by

$$\hat{\rho}_{xy} = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

est. of variance.

Advertising Budget Example

- Let's compute the pairwise correlation coefficients for the TV, radio, and newspaper spending features in the advertising data.

```
In [40]:  1 dfAd[["tv", "radio", "news"]].corr()
```

Out[40]:

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

- **Question**: What do you notice?

radio, news are correlated.

Looking Forward

- **Next time** we'll look performing inference on MLR parameters. We'll see how to
  - Perform HT to determine if any of the features are related to the response
  - Perform HT to determine if a subset of features is related to the response.
  - Extend SLR Goodness-of-Fit measures to the MLR setting R^2, SSE, SST etc.
  - Perform model selection to get the best lean-and-mean MLR model that we can
- For the rest of today we'll look how we can use MLR to explain nonlinear relationships between single-feature data and the response.
- Regroup & get out your laptops! Problem 3 in in-class nb.

Polynomial regression

- For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression where the additional features are powers of the original single-feature, *x*.

Using Residual Plots in Polynomial Reg.

- Recall that the assumed nature of our true model is:

CSCI 3022

intro to data science with probability & statistics

Lecture 25 (TWENTY FIVE?!?)
April 16, 2018

Inference & Model Selection in Multiple Linear Regression

"Dreams come true!"

- you, in ref. to today's ~~exciting~~ astounding material.

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Stuff & Things

- HW6 due Friday.
- No more homework! Instead: practicum, posted ~~on~~ Tonight.
  - Real data & real questions. Combining EDA, stats, pandas, and all your favorite tricks. A little Monte Carlo simulation. Other fun things to practice data science, "where the rubber meets the code."
  - 3 problems.
  - Due Weds before finals week, 5/2.
  - No collaborating. Sry. Show me what you've learned! 🎉😎🤩 (Ask in OH!)

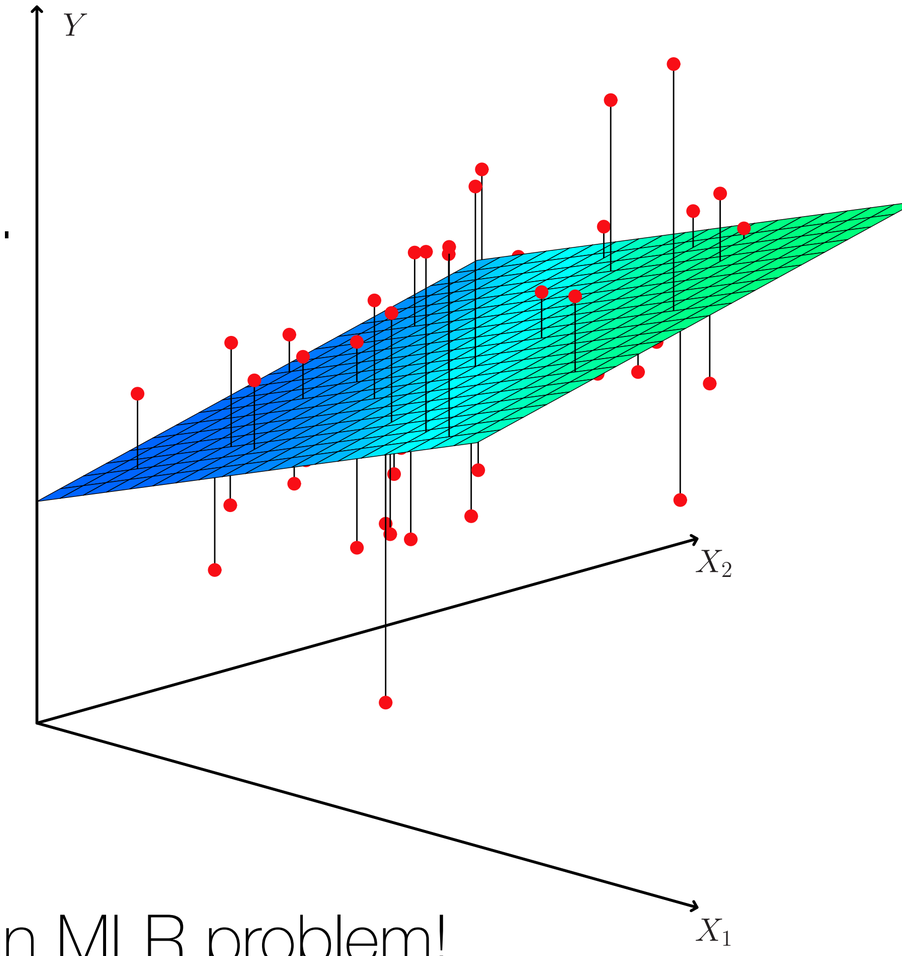
No late days.

Last time on CSCI 3022:

- Multiple Linear Regression assumes that the response *y* may be affected by multiple features.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Instead of fitting a line to the data, MLR fits a plane.
- What did we learn about MLR vs SLR?



- ~~Recall~~ that we can cast *polynomial regression* as an MLR problem!

Polynomial regression

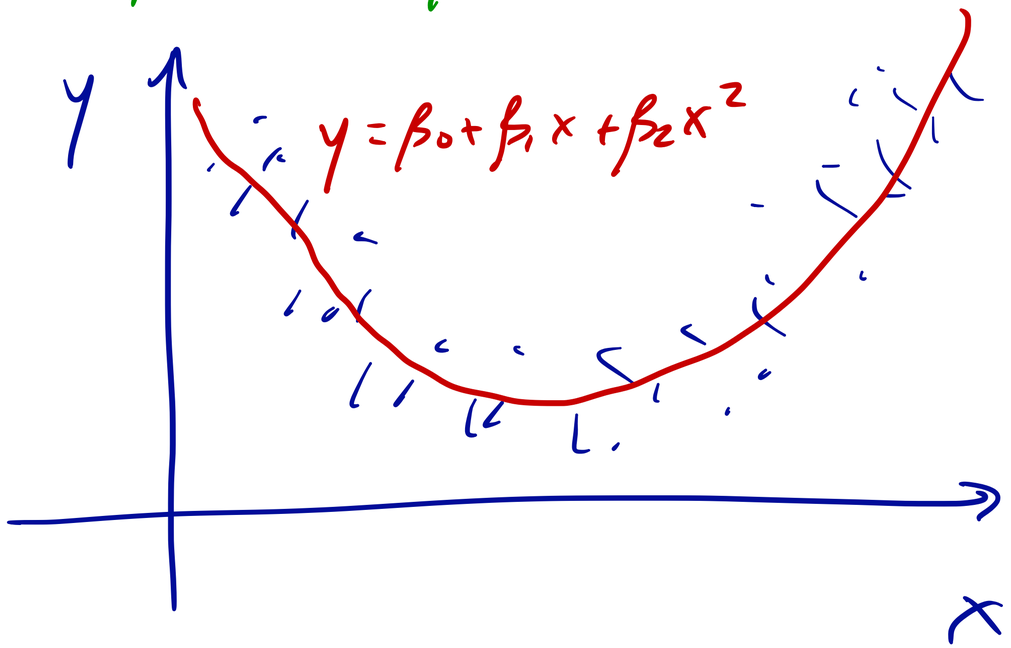
- For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression where the additional features are powers of the original single-feature, *x*.

recall polynomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots$

first feature $x_1 = x$

second feature $x_2 = x^2$ ←

third feature $x_3 = x^3$



Using Residual Plots in Polynomial Reg.

- Recall that the assumed nature of our true model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \cdots + \beta_p x_1^p + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

If true model is $y = \beta_0 + \beta_1 x + \varepsilon$

and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

then $r = y - \hat{y} \sim N(0, \sigma^2)$

If true model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

then $r = y - \hat{y} \sim N(\beta_2 x^2, \sigma^2)$

$\Rightarrow$ If I plot the residual (x_i, r_i) should be normally distr. around missing feature.

See last notebook Prob. #3

Recap: advertising budgets

SLR

```
SLR for tv vs sales
----------------------
intercept = 7.0326
slope = 0.0475
p-value = 1.4673897001945922e-42


SLR for radio vs sales
----------------------
intercept = 9.3116
slope = 0.2025
p-value = 4.354966001766913e-19


SLR for news vs sales
----------------------
intercept = 12.3514
slope = 0.0547
p-value = 0.0011481958688882112
```

MLR

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

Under SLR, each feature shows a significant slope.
Under MLR, the coefficient for newspapers disappears.

Recap: advertising budgets

SLR

```
SLR for tv vs sales
----------------------
intercept = 7.0326
slope = 0.0475
p-value = 1.4673897001945922e-42


SLR for radio vs sales
----------------------
intercept = 9.3116
slope = 0.2025
p-value = 4.354966001766913e-19


SLR for news vs sales
----------------------
intercept = 12.3514
slope = 0.0547
p-value = 0.0011481958688882112
```

MLR

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

Under SLR, each feature shows a significant slope.
Under MLR, the coefficient for newspapers disappears.

This is because *news* is a surrogate for *radio*, which we learned from the correlation matrix.

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

Inference in Multiple Linear Regression

- Questions we would like to answer:
  1. Is at least one of the features useful in predicting the response?
  2. Do all of the features help to explain the response, or is it just a subset?
  3. How well does the model fit the data?

Hypothesis Testing for MLR

- Recall our question from last time:

Is there a relationship between the response and predictors?

- In the simple linear regression setting, we can simply check whether $\beta_1 = 0$.
- In the MLR setting, with *p* features (aka predictors) we n eed to ask whether *all* of the coefficients are zero:
  - H_0 : $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p = 0$
  - H_1 : At least one is not 0, so $\beta_j \neq 0$ for at least one j

this is not $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 \neq 0$ ...

Is at Least One Feature Important?

- We test the hypothesis via the F-statistic.

$$F = \frac{\frac{(SST - SSE)}{df_{SST} - df_{SSE}}}{\frac{SSE}{df_{SSE}}} = \frac{(SST - SSE)/p}{SSE/(n-p-1)}$$

$df_{SST}: n-1$

$df_{SSE}: n-(p+1)$

$df_{SST} - df_{SSE} = p$

- Recall:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad df: n-1$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi})\right)^2 \quad df: n-(p+1) = n-p-1$$

Is at Least One Feature Important?

- We test the hypothesis via the F-statistic.

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} \qquad SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

- Suppose H_0 were true. What would F be?

F nearly 1

- Suppose that H_1 were true. What would F be?

F > 1

The F-statistic

- We test the hypothesis via the F-statistic.

$$\tilde{F} = \frac{(SST - SSE)/p}{SSE/(n-p-1)}$$

two diff d.o.f. parameters.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

- F distribution will give us a critical value so we can do a p-value test!

Is $F \geq F_{critical}$? Always one tailed.

compare this to α

$\Pr(\tilde{F} \geq F_{p,n-p-1})$ ← $\text{scipy.stats.f.cdf}(\tilde{F}, p, n-p-1)$

Is a Subset of Features Important?

- **Full Model**: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (p=4 features in full model)
- **Reduced Model**: $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4$ (k=2 features in reduced model)
  - **Question**: Are the missing features important, or are we OK going with the reduced model?
  - **Partial F-Test**: $H_0 : \beta_1 = \beta_3 = 0$
  - Since the features in the reduced model are also in the full model, we expect the full model to perform at least as well as the reduced model.
  - **Strategy**: Fit the Full and Reduced models. Determine if the difference in performance is real or due to just chance.

dropped p-k features

df_{full}: $n-(p+1)$

df_{red}: $n-(k+1)$

$n-(p+1)-[n-(k+1)]$

$n-p-1-n+k+1$

$k-p$ $\quad$ $p-k$

Is a Subset of Features Important?

p is # features in full model

k = features in reduced model

- SSE_{full} = variation unexplained by the full model

- SSE_{red} = variation unexplained by the reduced model

Intuitively, if SSE_{full} is much smaller than SSE_{red}, the full model fits the data much better than the reduced model. The appropriate test statistic should depend on the difference $SSE_{red} - SSE_{full}$ in unexplained variation.

- Test Statistic: $$F = \frac{(SSE_{\text{red}} - SSE_{\text{full}})/(p-k)}{SSE_{\text{full}}/(n-p-1)} \sim F_{p-k,n-p-1}$$

- Rejection Region: $$F \geq F_{\alpha,p-k,n-p-1}$$ stats.f.ppf(signif, dof1, dof2)

http://homepage.divms.uiowa.edu/~mbognar/applets/f.html

F… why even?

- Why compute the p-value for F-statistic when instead, we already have p-values for each of the covariates?

- Doing so would not be testing one hypothesis, but rather *p* hypotheses!

- At α=0.05, how many *p* values do we expect to be significant if the null hypothesis is, in fact, true?

p·0.05, so if 100 features, 100·0.05=5

```
In [27]: 1 model.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Tue, 28 Nov 2017	**Prob (F-statistic):**	1.58e-96
Time:	20:28:02	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
tv	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
news	-0.0010	0.006	-0.177	0.860	-0.013	0.011

The road to R^2 for MLR

- Just as with simple regression, the error sum of squares is:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \hat{\sigma}^2 = \frac{SSE}{n-(p+1)} = \frac{SSE}{n-p-1}$$

You may see SSE written as RSS: "residual sum of squares"

- It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.
- The number of df associated with SSE is n–(p+1) because p+1 df are lost in estimating the p+1 β coefficients.

The road to R^2

- Just as before, the **total sum of squares** is:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad df: n-1$$

- And the **sum of squared errors** is:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad df: n-p-1$$

- Then the coefficient of multiple determination R^2 is:

$$R^2 = \frac{SSR}{SST} = \boxed{1 - \frac{SSE}{SST} = R^2}$$

- It is interpreted in the same way as before. (Do you remember?)

$\frac{SSE}{SST}$ ~ how much variance is left unexplained after model fit.

Hacking R^2

Unfortunately, there is a problem with R^2: Its value can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous!

Hacking R^2

- For example, suppose y is the sale price of a house. Then:
- Sensible predictors include
  x_1 = the interior size of the house,
  x_2 = the size of the lot on which the house sits,
  x_3 = the number of bedrooms,
  x_4 = the number of bathrooms, and
  x_5 = the house's age.
- But now suppose we add in
  x_6 = the diameter of the doorknob on the coat closet,
  x_7 = the thickness of the cutting board in the kitchen,
  x_8 = the thickness of the patio slab.

Adjusted R^2

- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R^2 to take account of the size of the model:

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2_a = 1 - \frac{SSE/df_{SSE}}{SST/df_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

↑ adjusted

Adjusted R^2

- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R^2 to take account of the size of the model:

$$R_a^2 = 1 - \frac{SSE/df_{SSE}}{SST/df_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Adjusted R^2

```
In [27]:  1 model.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Tue, 28 Nov 2017	**Prob (F-statistic):**	1.58e-96
Time:	20:28:02	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
tv	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
news	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Deciding on important variables

- Suppose that we have 100 data points (n=100), but we have 200 different features (p=200). How can we learn which features are important and which are not?
- **Some options**:
  - Try all the possible combinations of features in models to see which gives the best fit.

Bad idea! Reason 2^p different models.

$p=3 \rightarrow 8$ models

$p=30 \rightarrow 1,073,741,824$ models.