Estimation

Marc Meredith*

*Introduction to Statistical Methods

Week 4

- Last week:
  - Developed the concept of a sampling distribution
  - Discussed how the sampling distribution of the sample mean is approximately normal when the sample is large
- Unclear thus far why it is so useful to know that a sampling distribution is approximately normal
  - Last week I assumed that I knew some population parameters, and then backed out the expected properties of a sample
    - But I am usually collecting data because I want to learn about the population parameters
  - This week I am going to learn some properties of a sample, and then back out the expected properties of a population parameters (i.e. estimation)

The New York Times

Live Polls What is live polling?

Ky. 6 • Mich. 8 • Iowa 4 • Wash. 8 • Ill. 14 • Calif. 48 • Tex. 32 ›

Polling in Real Time: The 2018 Midterm Elections

The Upshot has partnered with Siena College to conduct polls of dozens of the most competitive House and Senate races across the country. Our poll results are updated in real time, after every phone call. We hope to help you understand how polling works, and why it sometimes doesn't. **Related article »**

2,822,889
calls made so far

We've completed 96 polls so far:

Candidates shown in **bold** have a clear lead in our poll, beyond sampling error.

New York 22 Central New York A Trumpian Republican vs. a centrist Democrat **Polled Nov. 1 to Nov. 4** 17,444 calls; 506 responses; margin of error ±4.7	46% Tenney	45% Brindisi	9% Undecided
New York 19 Catskills, Hudson Valley Will this rural swing district elect a black Rhodes scholar? **Polled Nov. 1 to Nov. 4** 16,640 calls; 505 responses; margin of error ±4.8	43% Delgado	42% Faso	8% Undecided
Kentucky 6 Lexington area Can a Democratic fighter pilot win in deeply conservative Kentucky? **Polled Nov. 1 to Nov. 4** 22,825 calls; 438 responses; margin of error ±4.9	44% Barr	44% McGrath	10% Undecided

See: `https://nyti.ms/2oOvOL2`

Agenda for week:

- Introduce the concept of estimation and estimation error
- Define the bias and variance of an estimator
- Demonstrate how to use the LLN and CLT to estimate and put a confidence interval on a population mean
- Explain why the t-distribution often is used to calculate the amount of error that you are likely to observe in an estimate of the mean in a given sample
- Show how to use R to implement an estimator of the population mean and the margin of error with polling data

Agenda for week (continued):

- Introduce the concept of difference in means, covariance, and correlation
- Define the criterion of least squares and derive the bivariate regression formula that is obtained when applying this criterion to fit a line between an independent and a dependent variable
- Introduce all of the concepts that appear in R output after using the lm() function and interpret the result of a bivariate regression

Key takeaways:

- An estimator with less bias and lower variance generally will generate less estimation error than an estimator with more bias and higher variance, but that usually isn't guaranteed in any given sample of data
- Increasing the sample size often is the best way to reduce estimation error
- When the conditions of the LLN and CLT are met, it is relatively straightforward to make probabilistic statements about the likelihood that a population mean or the difference in two population means is contained in some range
- Linear regressions often are the best way to answer (the many) empirical questions we encounter about conditional expectations or differences in conditional expectations

- We often want to estimate population parameters - numerical descriptive measures of the population at-large - based on the properties of a sample
  - E.g., what percentage of Americans approve of the president based on responses in a poll about whether individual Americans support the president
- Statistical inference is the process through which we make judgments about population parameters based on what we observe in a sample
- Two elements of statistical inference
  1. Make an estimate of the value of a population parameter
  2. Make a judgment about the accuracy of this estimate

- An estimator is a rule that tells us how to make a judgment about the value of a population parameter based on the realizations of r.v.'s in a sample
  - I.e., a function that maps a set of realization of a r.v. into a number that represents a guess of value of a population parameter
- Desirable properties of an estimator:
  - Average value of the estimator equals the value of the population parameter
  - Average deviation between the estimator and the population parameter is small
  - Rarely is the deviation between the estimator and the population parameter large

- Let $\hat{\theta}(X_1, X_2, ..., X_n)$ represent our estimator of a population parameter θ based on the realizations of $X_1, X_2, ..., X_n$
- We usually use the shorthand $\hat{\theta}$ to denote that $\hat{\theta}(X_1, X_2, ..., X_n)$ is an estimate of population parameter θ
  - E.g., $\hat{\mu} = \hat{\mu}(X_1, X_2, ..., X_n)$ tells us how to aggregate responses by respondents $1, 2, \ldots, n$ about whether (s)he approves of the president to generate our estimate of μ, the share of Americans who approve of the president
- We can rewrite the desirable properties of an estimator using notation that we developed over the past two weeks
  1. $E[\hat{\theta}] = \theta$
  2. $E[\mid \hat{\theta} - \theta \mid]$ is close to zero
  3. $p(\mid \hat{\theta} - \theta \mid > k)$ is close to zero when k becomes relatively large

- There often are many feasible estimators for the same population parameter
- To illustrate, suppose that $X_1, X_2, ..., X_n$ is a sample of n iid r.v.'s such that $E[X_i] = \mu_x$ and $var(X_i) = {\sigma_x}^2$
- Three (of many) potential estimators for μ_x:
  1. $\hat{\mu}_{x1} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$
  2. $\hat{\mu}_{x2} = \frac{\sum_{i=1}^{n} X_i}{n}$
     - I.e., Case 1 with $w_i = c$ for all i
  3. $\hat{\mu}_{x3} = X_1$
     - I.e., Case 1 with $w_1 = 1$ and $w_i = 0$ for all $i \geqslant 2$

- Given that we often have multiple estimators for the same population parameter, we need some properties that can be used to choose between them
- Three useful properties to know about:
  - The bias, $B(\hat{\theta})$, of an estimator is $E[\hat{\theta} - \theta]$
  - The variance, $var(\hat{\theta})$, of an estimator is $E[(\hat{\theta} - E[\hat{\theta}])^2]$
  - The mean-squared error, $MSE(\hat{\theta})$, of an estimator is $E[(\hat{\theta} - \theta)^2]$

- One criterion that we often impose on an estimator is that it is unbiased
  - An unbiased estimator is an estimator such that $E[\hat{\theta} - \theta] = 0$
  - Which can be rewritten $E[\hat{\theta}] = \theta$
- There often are multiple unbiased estimators of an population parameter
  - E.g., The next two slides show that $\hat{\mu}_{x1}$, $\hat{\mu}_{x2}$, $\hat{\mu}_{x3}$ are unbiased estimators of μ_x

- $E[\hat{\mu}_3] = \mu_x$

  $E[\hat{\mu}_3] =$

  $E[X_1] = \mu_x$

- $E[\hat{\mu}_2] = \mu_x$

  $E[\hat{\mu}_2] =$

  $E[\frac{\sum_{i=1}^{n} X_i}{n}] =$

  $\frac{\sum_{i=1}^{n} E[X_i]}{n} =$

  $\frac{\sum_{i=1}^{n} \mu_x}{n} =$

  $\frac{\mu_x \sum_{i=1}^{n} 1}{n} =$

  $\frac{n\mu_x}{n} = \mu_x$

- $E[\hat{\mu}_1] = \mu_x$

  $E[\hat{\mu}_1] =$

  $E[\frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}] =$

  $E_w[\frac{\sum_{i=1}^{n} w_i E[X_i|w_i]}{\sum_{i=1}^{n} w_i}] =$

  - Applying the Law of Iterated Expectations

  $E_w[\frac{\sum_{i=1}^{n} w_i \mu_x}{\sum_{i=1}^{n} w_i}] =$

  $\mu_x E_w[\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i}] = \mu_x$

- One rule for selecting among estimators is to select the unbiased estimator with the lowest variance
- The next two slides shows that $var(\hat{\mu}_2) \leqslant var(\hat{\mu}_1) < var(\hat{\mu}_3)$
  - Implying that we prefer $\hat{\mu}_2$ to $\hat{\mu}_1$ or $\hat{\mu}_3$ based on their respective variances
- In English, this means that our estimator has the lowest variance when we weight all observations equally when constructing the sample mean
  - This is an artifact of $var(X_i) = {\sigma_x}^2$ for all i (i.e., a common variance)
  - If $var(X_j) \neq var(X_k)$, then it is the case that there are $w_j \neq w_k$ st $var(\hat{\mu}_1) < var(\hat{\mu}_2)$
    - General principal that when $var(X_j) < var(X_k)$, then $w_j > w_k$

- $var(\hat{\mu}_3) = {\sigma_x}^2$

  $var(\hat{\mu}_3) = var(X_1) = {\sigma_x}^2$

- $var(\hat{\mu}_2) = \frac{{\sigma_x}^2}{n}$

  $var(\hat{\mu}_2) =$

  $var(\frac{\sum_{i=1}^{n} X_i}{n}) =$

  $\frac{1}{n}^2 var(\sum_{i=1}^{n} X_i) =$

  $\frac{1}{n}^2 \sum_{i=1}^{n} var(X_i) =$

  $\frac{1}{n}^2 \sum_{i=1}^{n} {\sigma_x}^2 = \frac{{\sigma_x}^2}{n}$

- $var(\hat{\mu}_1) = {\sigma_x}^2 E_w[\frac{\sum_{i=1}^n {w_i}^2}{(\sum_{i=1}^n w_i)^2}]$

  $var(\hat{\mu}_1) =$

  $E[\hat{\mu}_1^2] - E[\hat{\mu}_1]^2 =$

  $E[\frac{\sum_{i=1}^n w_i {X_i}^2}{\sum_{i=1}^n w_i}] - {\mu_x}^2 =$

  $E[\frac{\sum_{i=1}^n {w_i}^2 {X_i}^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^n w_i w_j X_i X_j}{(\sum_{i=1}^n w_i)^2}] - {\mu_x}^2 =$

  $E_w[\frac{\sum_{i=1}^n w_i^2 E[{X_i}^2|w] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^n w_i w_j E[X_i X_j|w]}{(\sum_{i=1}^n w_i)^2}] - {\mu_x}^2 =$

  - Applying the Law of Iterated Expectations

  $E_w[\frac{\sum_{i=1}^n w_i^2({\sigma_x}^2 + {\mu_x}^2) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^n w_i w_j ({\mu_x}^2)}{(\sum_{i=1}^n w_i)^2}] - {\mu_x}^2 =$

  - $E[{X_i}^2 \mid w] - E[{X_i}^2 \mid w]^2 = var(X_i \mid w) = {\sigma_x}^2$
  - $E[X_i X_j \mid w] - E[X_i \mid w]E[X_j \mid w] = cov(X_i, X_j) = 0$

  $E_w[\frac{{\sigma_x}^2 \sum_{i=1}^n w_i^2 + {\mu_x}^2 (\sum_{i=1}^n w_i)^2}{(\sum_{i=1}^n w_i)^2}] - {\mu_x}^2 = {\sigma_x}^2 E_w[\frac{\sum_{i=1}^n {w_i}^2}{(\sum_{i=1}^n w_i)^2}]$

  - $(\sum_{i=1}^n w_i)^2 = \sum_{i=1}^n {w_i}^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^n w_i w_j$

- Another criterion used to select an estimator is to pick the estimator with the lowest MSE
- I show on the next slide that $MSE(\hat{\theta}) = var(\hat{\theta}) + B(\hat{\theta})^2$
- Implications:
  - MSE provides a way to formalize the tradeoff between more biased estimators and estimators with more variance
  - Because an unbiased estimator will have $B(\hat{\theta}) = 0$, selecting the unbiased estimator with the lowest MSE is equivalent to selecting the unbiased estimator with the lowest variance

Proof that the $MSE(\hat{\theta}) = var(\hat{\theta}) + B(\hat{\theta})^2$:

- $E[(\hat{\theta} - \theta)^2] =$
  $E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] =$
  $E[(\hat{\theta} - E[\hat{\theta}])^2 + 2(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta) + (E[\hat{\theta}] - \theta)^2] =$
  $E[(\hat{\theta} - E[\hat{\theta}])^2] + E[2(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] + E[(E[\hat{\theta}] - \theta)^2] =$
  $var(\hat{\theta}) + B(\hat{\theta})^2$
  - $E[(\hat{\theta} - E[\hat{\theta}])^2] = var(\hat{\theta})$
  - $E[2(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] =$
    $2(E[\hat{\theta}] - \theta)(E[\hat{\theta} - E[\hat{\theta}]]) =$
    $2(E[\hat{\theta}] - \theta)(E[\hat{\theta}] - E[E[\hat{\theta}]]) =$
    $2(E[\hat{\theta}] - \theta)(E[\hat{\theta}] - E[\hat{\theta}]) = 0$
  - $E[(E[\hat{\theta}] - \theta)^2] = B(\hat{\theta})^2$

- Later in this class we are going to need an estimator for ${\sigma_x}^2$
- Following the logic of the sample mean, someone might incorrectly assume that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{Y}_n)^2$ is unbiased estimator of ${\sigma_x}^2$
  - The average squared deviation between the value of an observation and average value over all of the observations in the sample
- But the next two slides shows that $\hat{\sigma}_x^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{Y}_n)^2$ is an unbiased estimator of ${\sigma_x}^2$
- More broadly highlights the importance of using mathematics, rather than just relying on intuition, when solving for the properties of statistics

Proof that $E[S^2] = \sigma_x^2$:

- $E[S^2] =$
  $E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{Y}_n)^2] =$
  $\frac{1}{n-1}E[\sum_{i=1}^{n} X_i^2 - 2X_i\bar{Y}_n + \bar{Y}_n{}^2] =$
  $\frac{1}{n-1}(E[\sum_{i=1}^{n} X_i^2] - 2E[\sum_{i=1}^{n} X_i\bar{Y}_n] + E[\sum_{i=1}^{n} \bar{Y}_n{}^2]) =$
  $\frac{1}{n-1}(\sum_{i=1}^{n} E[X_i^2] - 2E[\bar{Y}_n\sum_{i=1}^{n} X_i] + \sum_{i=1}^{n} E[\bar{Y}_n{}^2]) =$
  $\frac{1}{n-1}(\sum_{i=1}^{n} E[X_i^2] - 2E[\bar{Y}_n(\bar{Y}_n n)] + nE[\bar{Y}_n{}^2]) =$
  $\frac{1}{n-1}(\sum_{i=1}^{n} E[X_i^2] - nE[\bar{Y}_n{}^2])$

Proof that $E[S^2] = \sigma_x^2$ (continued):

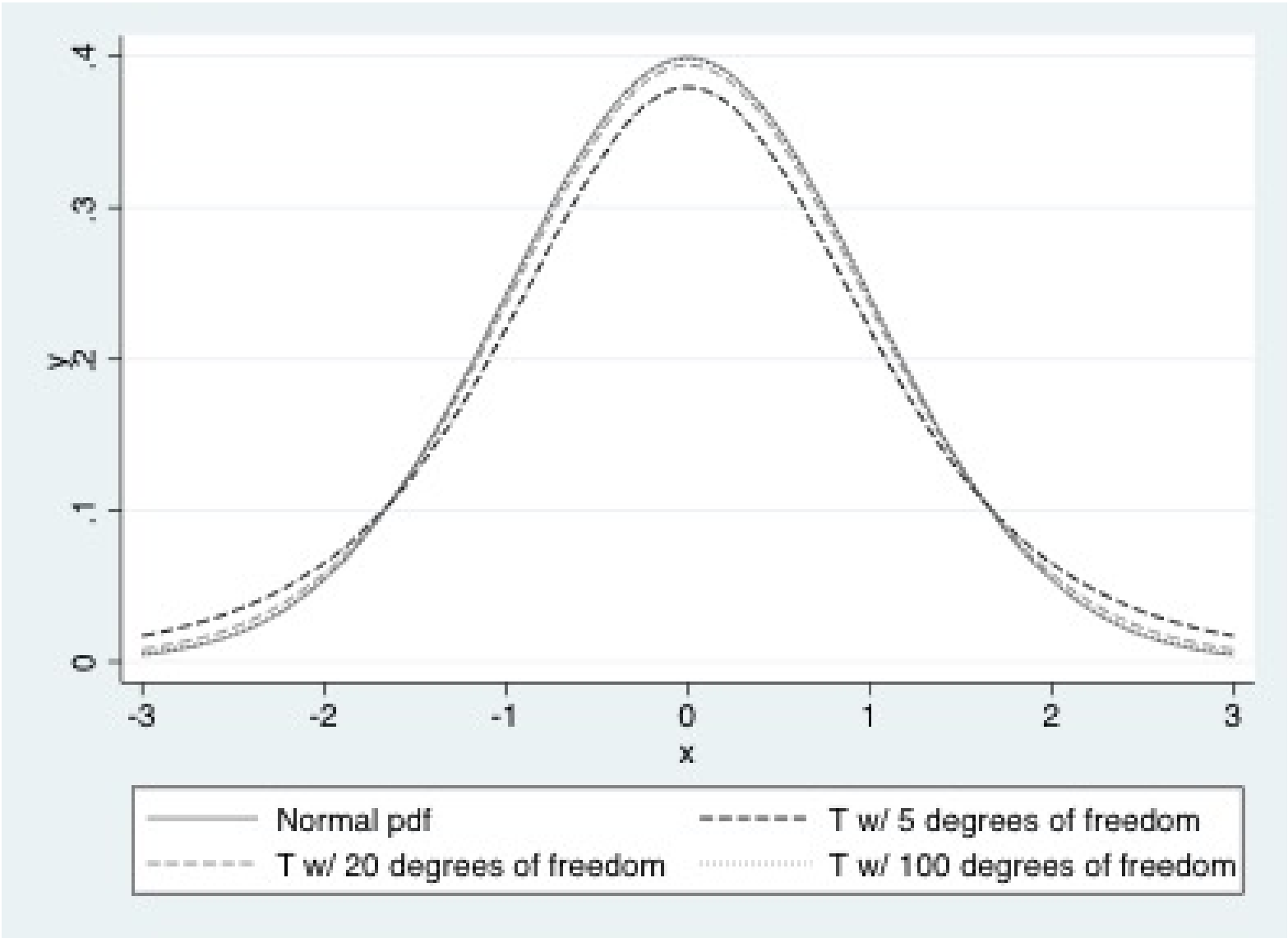
- $\frac{1}{n-1}(\sum_{i=1}^{n} E[X_i^2] - nE[\bar{Y}_n{}^2]) =$
  $\frac{1}{n-1}(n(\sigma_x^2 + \mu_x^2) - n(\frac{\sigma_x^2}{n} + \mu_x^2)) =$
  - $var(X_i) = \sigma_x{}^2 = E[X_i{}^2] - E[X_i]^2 = E[X_i{}^2] - \mu_x{}^2 \implies$
    $E[X_i{}^2] = \sigma_x{}^2 + \mu_x{}^2$
  - $var(\bar{Y}_n) = \frac{\sigma_x{}^2}{n} = E[\bar{Y}_n{}^2] - E[\bar{Y}_n]^2 = E[\bar{Y}_n{}^2] - \mu_x{}^2$
    $E[\bar{Y}_n{}^2] = \frac{\sigma_x{}^2}{n} + \mu_x{}^2$

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

- Even unbiased estimators generally have some estimation error in finite samples
  - Estimation Error: $\epsilon = \hat{\theta} - \theta$
- Because we generally don't know the value of the parameter that we are estimating, we don't know how much estimation error is present for any given estimate
- But if we know the sampling distribution of the estimator, we can assess the probability that our estimate is within a certain number of units of the population parameter
  - i.e., $p(A < \epsilon < B)$

- Last week we proved that $\sqrt{n}\frac{\bar{Y}_n-\mu_X}{\sigma_X} \sim N(0,1)$
  - Where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and for r.v.'s $X_1, X_2, \ldots, X_n$ from any distribution such that $E[X_i] = \mu_X$ and $var(X_i) = {\sigma_X}^2$
- We can leverage this to calculate the probability that the amount of estimation error is contained in (A, B):
  - $\sqrt{n}\frac{\bar{Y}_n-\mu_X}{\sigma_X} \sim N(0,1) \implies$
    $p(A < \sqrt{n}\frac{\bar{Y}_n-\mu_X}{\sigma_X} < B) = \Phi(B) - \Phi(A) \implies$
    $p(\frac{A\sigma_X}{\sqrt{n}} < \bar{Y}_n - \mu_X < \frac{B\sigma_X}{\sqrt{n}}) = \Phi(\frac{B\sigma_X}{\sqrt{n}}) - \Phi(\frac{A\sigma_X}{\sqrt{n}})$

- Using $\Phi(\frac{B\sigma_x}{\sqrt{n}}) - \Phi(\frac{A\sigma_x}{\sqrt{n}})$ to calculate the probability of a given amount of estimation error for a sample mean requires that we know σ_x
  - Unlikely that we'll know σ_x, but not μ_x
- Given that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{Y}_n)^2$ is an unbiased estimate of ${\sigma_x}^2$, could we instead estimate $\Phi(\frac{B\sqrt{S^2}}{\sqrt{n}}) - \Phi(\frac{A\sqrt{S^2}}{\sqrt{n}})$?
- No, it is not the case that $\sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sqrt{S^2}} \sim N(0,1)$ just because
  - $\sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sigma_x} \sim N(0,1)$
  - S^2 is an unbiased estimate of ${\sigma_x}^2$
- Instead $\sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sqrt{S^2}} \sim t_{n-1}$ as n get big
  - Although it does require a bit of circular logic since we are assuming that we have close to n infinite sample in order to apply the Central Limit Theorem in the first place


.4
.3
.2
.1
0
y
-3
-2
-1
0
1
2
3
x
Normal pdf
T w/ 5 degrees of freedom
T w/ 20 degrees of freedom
T w/ 100 degrees of freedom

- Using the T distribution, we can calculate the probability that $\bar{Y}_n - \mu_x$ is contained in (A, B) entirely on things that are observed data
- $\sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sqrt{S^2}} \sim t_{n-1} \implies$

  $p(A < \sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sqrt{S^2}} < B) = T_{n-1}(B) - T_{n-1}(A) \implies$

  $p(\frac{A\sqrt{S^2}}{\sqrt{n}} < \bar{Y}_n - \mu_x < \frac{B\sqrt{S^2}}{\sqrt{n}}) = T_{n-1}(\frac{B\sqrt{S^2}}{\sqrt{n}}) - T_{n-1}(\frac{A\sqrt{S^2}}{\sqrt{n}})$

  - $T_{n-1}()$ is the cdf of a t distribution with $n-1$ degrees of freedom
  - To calibrate note that $T_{10}(1.96) = 0.9608$, $T_{100}(1.96) = 0.9736$, and $\Phi(1.96) = 0.975$

- We can rearrange the formula on the previous slide to construct a confidence interval (CI) on a population parameter based entirely on things that are observed data
- $p(\frac{A\sqrt{S^2}}{\sqrt{n}} < \bar{Y}_n - \mu_x < \frac{B\sqrt{S^2}}{\sqrt{n}}) =$

  $p(\bar{Y}_n + \frac{B\sqrt{S^2}}{\sqrt{n}} < \mu_x < \bar{Y}_n + \frac{A\sqrt{S^2}}{\sqrt{n}}) = T_{n-1}(\bar{Y}_n + \frac{A\sqrt{S^2}}{\sqrt{n}}) - T_{n-1}(\bar{Y}_n + \frac{B\sqrt{S^2}}{\sqrt{n}})$
- To calculate a symmetric $1-\alpha$ CI, we can set $B = {T_{n-1}}^{-1}(\frac{\alpha}{2})$ and $A = {T_{n-1}}^{-1}(1-\frac{\alpha}{2})$
  - Where ${T_{n-1}}^{-1}()$ is the inverse cdf of the T distribution with $n-1$ degrees of freedom (analog to $\Phi^{-1}()$ for the normal)
- This gives us that:

  $p(\bar{Y}_n + \frac{{T_{n-1}}^{-1}(\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}} < \mu_x < \bar{Y}_n + \frac{{T_{n-1}}^{-1}(1-\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}}) =$

  $T_{n-1}(\bar{Y}_n + \frac{{T_{n-1}}^{-1}(1-\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}}) - T_{n-1}(\bar{Y}_n + \frac{{T_{n-1}}^{-1}(\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}}) = 1-\alpha$

- I am now going to use an example to illustrate what we can learn from the formulas being presented in this section
  - To learn something about the average weight of army recruits:
    - I weigh a random sample of 40 recruits
    - Find that the sample average is 170 lbs. and the sample variance is 200 lbs.
  - What can I conclude about average weight of army recruits in the population based on what I observe in this sample?
    - Beyond that 170 lbs. is an unbiased estimate of the population mean

Example 1:

- I randomly sample the weight of 40 army recruits and find that the sample average is 170 lbs. and the sample variance is 200 lbs.
- What is the probability that I find that the population average is between 168 and 172 lbs. (i.e., the absolute value of estimation error is less than 2)?
  - $p(168 < \mu_x < 172) =$
    $p(168 - \bar{Y}_n < \mu_x - \bar{Y}_n < 172 - \bar{Y}_n) =$
    $p(\frac{\sqrt{n}}{\sqrt{S^2}}(168 - \bar{Y}_n) < \sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sqrt{S^2}} < \frac{\sqrt{n}}{\sqrt{S^2}}(172 - \bar{Y}_n)) =$
    $p(\frac{\sqrt{40}}{\sqrt{200}}(168 - 170) < \sqrt{40}\frac{\mu_x - 170}{\sqrt{200}} < \frac{\sqrt{40}}{\sqrt{200}}(172 - 170)) =$
    $p(-0.894 < \sqrt{40}\frac{\mu_x - 170}{\sqrt{200}} < 0.894) =$
    $T_{39}(0.894) - T_{39}(-0.894) \approx 0.623$
    - Because $\sqrt{40}\frac{\mu_x - 170}{\sqrt{200}} \sim t_{39}$
    - Solved in R using "pt(0.894, 39) - pt(-0.894, 39)"

Example 1:

- Suppose that I incorrectly assumed that $\sqrt{40}\frac{\mu_x - 170}{\sqrt{200}} \sim N(0,1)$
- Then I would have calculated $p(168 < \mu_x < 172) = \Phi(0.894) - \Phi(-0.894) \approx 0.629 > 0.623$
- Illustrates the general principal that applying the T distribution, rather than the normal distribution, generates more conservative estimates about the likelihood that the population mean is proximate to the sample mean
  - Difference is less consequential as the size of the sample increases

Example 2:

- I randomly sample the weight of 40 army recruits and find that the sample average is 170 lbs. and the sample variance is 200 lbs.
- What is a symmetric 95% CI?
  - $p(\bar{Y}_n + \frac{{T_{n-1}}^{-1}(\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}} < \mu_x < \bar{Y}_n + \frac{{T_{n-1}}^{-1}(1-\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}}) =$
    $T_{n-1}(\bar{Y}_n + \frac{{T_{n-1}}^{-1}(1-\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}}) - T_{n-1}(\bar{Y}_n + \frac{{T_{n-1}}^{-1}(\frac{\alpha}{2})\sqrt{S^2}}{\sqrt{n}}) \Longrightarrow$
    $p(170 + \frac{{T_{39}}^{-1}(.025)\sqrt{200}}{\sqrt{40}} < \mu_x < 170 + \frac{{T_{39}}^{-1}(.975)\sqrt{200}}{\sqrt{40}}) = 0.95$
    $p(170 + \frac{-2.023\sqrt{200}}{\sqrt{40}} < \mu_x < 170 + \frac{2.023\sqrt{200}}{\sqrt{40}}) = 0.95$
    - ${T_{39}}^{-1}(.025) = -2.023$ and ${T_{39}}^{-1}(.975) = 2.023$ are solved in R by using "qt(0.025, 39)" and "qt(0.975, 39)", respectively
    - Compared to $\Phi^{-1}(.025) = -1.96$ and $\Phi^{-1}(.975) = 1.96$
    $p(165.48 < \mu_x < 174.52) = 0.95$

Example 3:

- In my initial sample, I estimated the sample variance is 200 lbs.
- How many total recruits should I sample if I want there to be no more than a 5 percent chance of more than 2 lbs. of estimation error?
  - $p(-2 < \bar{Y}_n - \mu_x < 2) = 0.95 \implies$
    $p(-2\frac{\sqrt{n}}{\sqrt{200}} < \sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sqrt{200}} < 2\frac{\sqrt{n}}{\sqrt{200}}) = 0.95 \implies$
    $T_{n-1}(2\frac{\sqrt{n}}{\sqrt{200}}) - T_{n-1}(-2\frac{\sqrt{n}}{\sqrt{200}}) = 0.95 \implies$
    - Because $\sqrt{n}\frac{\bar{Y}_n - \mu_x}{\sqrt{200}} \sim t_n$

    $T_{n-1}(2\frac{\sqrt{n}}{\sqrt{200}}) = 0.975 \implies 2\frac{\sqrt{n}}{\sqrt{200}} = T_{n-1}^{-1}(0.975) \implies$
    $n = (\frac{\sqrt{200}}{2T_{n-1}^{-1}(0.975)})^2 \approx 195$
    - Solved in R using "(sqrt(200)*qt(0.975, 195)/2)*(sqrt(200)*qt(0.975, 195)/2)"
    - May need to be solved iteratively because $T_{n-1}^{-1}()$ is a function of n

1. What is the probability that I find that the population average is between 168 and 172 lbs. (i.e., the absolute value of estimation error is less than 2)?
   - 0.623
     - Solved in R using "pt(0.894, 39) - pt(-0.894, 39)"
     - Would have answered 0.629 if you incorrectly applied the normal distribution instead of the t_{39} distribution
2. What is a symmetric 95% CI?
   - $p(165.48 < \mu_x < 174.52) = 0.95$
     - Solved in R using "170 + qt(0.025, 39)*sqrt(200/40)" and "170 + qt(0.975, 39)*sqrt(200/40)"
3. How many total recruits should I sample if I want there to be no more than a 5 percent chance of more than 2 lbs. of estimation error?

   $$n = \left(\frac{\sqrt{200}}{2T_{n-1}^{-1}(0.975)}\right)^2 \approx 195$$

   - Solved in R using "(sqrt(200)*qt(0.975, 195)/2)*(sqrt(200)*qt(0.975, 195)/2)"
   - May need to be solved iteratively because $T_{n-1}^{-1}()$ is a function of n

- Lets use R to apply these formulas to the NY22 poll highlighted at the start of the lecture
- This is why they reported 46% Rep, 45% Dem, and 9% Und:

```
> library(survey)
> ny22 <- read.csv("RawData/NYTimes/elections-poll-ny22-3.csv")
> ny22w <- svydesign(ids = ~1,
+                    data = ny22, weights = ny22$final_weight)
> ny22output <- svymean(~response, ny22w)
> print(ny22output)
                 mean     SE
responseDem 0.449499 0.0236
responseRep 0.458816 0.0238
responseUnd 0.091684 0.0133
```

- The confint() function makes it easier to calculate the margin of error of 4.7%

```
> numobs <- nrow(ny22)
> confint(ny22output, level = 0.95, df = numobs - 1)
                  2.5 %    97.5 %
responseDem 0.40305918 0.4959390
responseRep 0.41215018 0.5054828
responseUnd 0.06555933 0.1178096
```

- What is happening underneath the hood of confint()

```
> print(coef(ny22output))
responseDem responseRep responseUnd
 0.44949909  0.45881647  0.09168444
> print(vcov(ny22output))
              responseDem   responseRep   responseUnd
responseDem  5.587306e-04 -4.730498e-04 -8.568078e-05
responseRep -4.730498e-04  5.641909e-04 -9.114105e-05
responseUnd -8.568078e-05 -9.114105e-05  1.768218e-04
> print(coef(ny22output)[2] +
+         vcov(ny22output)[2, 2]^(1/2)*qt(0.025, numobs - 1))
responseRep
  0.4121502
> print(coef(ny22output)[2] +
+         vcov(ny22output)[2, 2]^(1/2)*qt(0.975, numobs - 1))
responseRep
  0.5054828
```

- The difference-in-means estimator is one of the most common techniques summarizing the relationship between a dependent variable and a binary independent variable
- Few examples:
  - Support for a political candidate among females and males
  - Health outcomes for people with and without insurance
  - Revenue generated when people view version A/B of a web page
  - Outcomes for people who receive a treatment in an experiment / people who receive a control in an experiment

- Let $X_{11}, X_{12}, ..., X_{1q}$ be a series of q r.v.'s observed from population 1 (e.g., random variables realized from the subsample of the population for which the independent variable is equal to 1)
  - $E[X_{1i}] = \mu_1$
  - $var(X_{1i}) = {\sigma_1}^2$
- Let $X_{21}, X_{22}, ..., X_{2m}$ be a series of m r.v.'s observed from population 2 (e.g., random variables realized from the subsample of the population for which the independent variable is equal to 0)
  - $E[X_{2i}] = \mu_2$
  - $var(X_{2i}) = {\sigma_2}^2$
- Our interest when doing difference-in-mean analysis is to estimate and make inferences about $\mu_1 - \mu_2$

- Our difference-in-means estimator is $\hat{\mu}_1 - \hat{\mu}_2$
  - $\hat{\mu}_1 = \frac{1}{q} \sum_{i=1}^{q} X_{1i}$
  - $\hat{\mu}_2 = \frac{1}{m} \sum_{i=1}^{m} X_{2i}$
- Appealing features of our difference-in-means estimator:
  1. An unbiased estimator of $\mu_1 - \mu_2$
  2. In large enough samples the estimator is distributed normally
  3. We have an estimator for the variance of the estimator that can be calculated based on data readily observed in a sample

- Next few slides prove two useful facts about the difference-in-means estimator:
  1. $E[\hat{\mu}_1 - \hat{\mu}_2] = \mu_1 - \mu_2$
  2. When population 1 and population 2 are sampled independently, then $var(\hat{\mu}_1 - \hat{\mu}_2) = \frac{{\sigma_1}^2}{q} + \frac{{\sigma_2}^2}{m}$
     - Where q is the number of observations sampled from population 1 and m is the number of observations sampled from population 2
     - Sampling independently (e.g., no observations in both population 1 and population 2) guarantees that $cov(\hat{\mu}_1, \hat{\mu}_2) = 0$

Proof that the difference-in-means estimator is an unbiased estimator of $\mu_1 - \mu_2$:

- $E[\hat{\mu}_1 - \hat{\mu}_2] =$
  $E[\frac{1}{q}\sum_{i=1}^{q} X_{1i} - \frac{1}{m}\sum_{i=1}^{m} X_{2i}] =$
  $E[\frac{1}{q}\sum_{i=1}^{q} X_{1i}] - E[\frac{1}{m}\sum_{i=1}^{m} X_{2i}] =$
  $\frac{1}{q}\sum_{i=1}^{q} E[X_{1i}] - \frac{1}{m}\sum_{i=1}^{m} E[X_{2i}] =$
  $\frac{1}{q}(q\mu_1) - \frac{1}{m}(m\mu_2) = \mu_1 - \mu_2$

- When population 1 and population 2 are sampled independently, then $var(\hat{\mu}_1 - \hat{\mu}_2) = \frac{{\sigma_1}^2}{q} + \frac{{\sigma_2}^2}{m}$
  - $var(\hat{\mu}_1 - \hat{\mu}_2) =$
    $var(\hat{\mu}_1) - 2cov(\hat{\mu}_1, \hat{\mu}_2) + var(\hat{\mu}_2) =$
    $var(\hat{\mu}_1) + var(\hat{\mu}_2)$
    - We'll come back shortly to what it means for $cov(\hat{\mu}_1, \hat{\mu}_2) = 0$, but for now just think about it meaning that whether or not observation i is sampled from population 1 has no bearing on whether observation j is sampled from population 2
    - Which means that no observations can both be in population 1 and population 2
- On the next slide, I show that $var(\hat{\mu}_1) = \frac{{\sigma_1}^2}{q}$ and $var(\hat{\mu}_2) = \frac{{\sigma_2}^2}{m}$

- Proof that $var(\hat{\mu_1}) = \frac{{\sigma_1}^2}{q}$
  - $var(\hat{\mu_1}) =$
    $var(\frac{1}{q}\sum_{i=1}^{q} X_{1i}) =$
    $\frac{1}{q}^2 var(\sum_{i=1}^{q} X_{1i}) =$
    $\frac{1}{q}^2 \sum_{i=1}^{q} var(X_{1i}) =$
    $\frac{1}{q}^2 \sum_{i=1}^{q} {\sigma_1}^2 =$
    $\frac{1}{q}^2 q{\sigma_1}^2 = \frac{{\sigma_1}^2}{q}$
- We solve that $var(\hat{\mu_2}) = \frac{{\sigma_2}^2}{m}$ in an analogous manner

- We know that as long as both q and m are sufficiently large then:
  1. $\hat{\mu}_1 \sim N(\mu_1, \frac{{\sigma_1}^2}{q})$
  2. $\hat{\mu}_2 \sim N(\mu_2, \frac{{\sigma_2}^2}{m})$
- Therefore, by the convolution property of normal r.v.'s
  1. $\hat{\mu}_1 - \hat{\mu}_2 \sim N(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{q} + \frac{{\sigma_2}^2}{m}) \implies$
  2. $\frac{(\mu_1 - \mu_2) - (\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\frac{{\sigma_1}^2}{q} + \frac{{\sigma_2}^2}{m}}} \sim N(0, 1)$

- We usually cannot apply the formula on the previous slide, because we don't know σ_1 or σ_2
- But we can estimate them with:
  - $\hat{\sigma_1}^2 = {S_1}^2 = \frac{1}{q-1} \sum_{i=1}^{q} (X_{1i} - \hat{\mu_1})^2$
  - $\hat{\sigma_2}^2 = {S_2}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_{2i} - \hat{\mu_2})^2$
- So $\frac{(\mu_1 - \mu_2) - (\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\frac{{S_1}^2}{q} + \frac{{S_2}^2}{m}}} \sim t_\nu$
  - $\nu = \frac{(\frac{{S_1}^2}{q} + \frac{{S_2}^2}{m})^2}{\frac{\frac{{S_1}^2}{q}^2}{q-1} + \frac{\frac{{S_2}^2}{m}^2}{m-1}}$
  - Which always is greater than $min(q-1, m-1)$

Example 1:

- Suppose we are running a clinical trial to see if a taking a sleeping drug makes people sleep better
  - 49 people took a sleeping drug. They averaged 7.5 hours of sleep and the sample variance was 1 hour.
  - 81 people took a placebo. They averaged 7 hours of sleep and the sample variance was .64 hours.
- Construct a symmetric 95% CI for the difference-in-means
- Steps:
  1. Calculate ν
  2. Construct a r.v. $T \sim t_\nu$ from μ_1, μ_2, $\hat{\mu}_1$, $\hat{\mu}_2$, S_1, S_2
  3. Calculate a symmetric 95% CI for a r.v. $T \sim t_\nu$
  4. Rearrange terms in CI to solve for bounds on $\mu_1 - \mu_2$

Example 1:

- Suppose we are running a clinical trial to see if taking a sleeping drug makes people sleep better
  - 49 people took a sleeping drug. They averaged 7.5 hours of sleep and the sample variance was 1 hour.
  - 81 people took a placebo. They averaged 7 hours of sleep and the sample variance was .64 hours.

1. Calculate ν
   - $\nu = \frac{(\frac{1}{49} + \frac{.64}{81})^2}{\frac{\frac{1}{49}^2}{48} + \frac{\frac{.64}{81}^2}{80}} \approx 84$
2. Construct a r.v. $T \sim t_{84}$ from μ_1, μ_2, $\hat{\mu}_1$, $\hat{\mu}_2$, S_1, S_2
   - $\frac{(\mu_1 - \mu_2) - (\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu \implies$

     $\frac{(\mu_1 - \mu_2) - (7.5 - 7)}{\sqrt{\frac{1}{49} + \frac{.64}{81}}} \sim t_{84}$

Example 1:

3. Calculate a symmetric 95% CI for a r.v. $T \sim t_{84}$
   - $p(T_{84}^{-1}(.025) < T < T_{84}^{-1}(.975)) = .95 \implies$
     $p(-1.989 < T < 1.989) = .95$
     - $T_{84}^{-1}(.975) = 1.989$, which I solved in R using qt(0.975, 84)
4. Rearrange terms to solve for bounds on $\mu_1 - \mu_2$
   - $p(-1.989 < T < 1.989) = .95 \implies$
     $p(-1.989 < \frac{(\mu_1 - \mu_2) - (7.5 - 7)}{\sqrt{\frac{1}{49} + \frac{.64}{81}}} < 1.989) = .95 \implies$
     $p(.5 - 1.989 * 0.168 < \mu_1 - \mu_2 < .5 + 1.989 * 0.168) = .95 \implies$
     $p(0.165 < \mu_1 - \mu_2 < 0.835) = 0.95$

Example 2:

- Suppose we are a political campaign that sends get-out-to-vote (GOTV) mailings to some members of a population that we would like to vote
  - 400 members of group A receive no political mailing; turnout in this group is 0.35
  - 500 members of group B receive a GOTV mailing; turnout in this group is 0.38
- Steps in calculating whether the mobilized population turns out to vote at a higher rate than the unmobilized population
  1. Estimate the sample variances and the standard error of the difference in means
  2. Calculate ν
  3. Construct a r.v. $T \sim t_\nu$ from μ_1, μ_2, $\hat{\mu}_1$, $\hat{\mu}_2$, S_1, S_2
  4. Rearrange terms so that $p(\mu_1 - \mu_2 < 0)$ is written as $p(T <?)$
  5. Evaluate $T_\nu(?)$

Example 2:

- Suppose we are a political campaign that sends get-out-to-vote (GOTV) mailings to some members of a population that we would like to vote
  - 400 members of group A receive no political mailing; turnout in this group is 0.35
  - 500 members of group B receive a GOTV mailing; turnout in this group is 0.38

1. Estimate the sample variances and the standard error of the difference in means
   - ${S_A}^2 = \frac{1}{399}\sum_{i=1}^{400}(Y_{ai} - .35)^2 = \frac{1}{399}(140 * (1 - .35)^2 + 260 * (0 - .35)^2) \approx 0.2275$
   - ${S_B}^2 = \frac{1}{499}\sum_{i=1}^{500}(Y_{bi} - .38)^2 = \frac{1}{499}(190 * (1 - .38)^2 + 310 * (0 - .38)^2) \approx 0.2356 \implies \sqrt{\frac{{S_A}^2}{n_A} + \frac{{S_B}^2}{n_B}} = \sqrt{\frac{0.2275}{400} + \frac{0.2356}{500}} \approx 0.032$

Example 2:

- Suppose we are a political campaign that sends get-out-to-vote (GOTV) mailings to some members of a population that we would like to vote
  - 400 members of group A receive no political mailing; turnout in this group is 0.35
  - 500 members of group B receive a GOTV mailing; turnout in this group is 0.38

2. Calculate ν
   - $\nu = \frac{\left(\frac{0.2275}{400} + \frac{0.2356}{500}\right)^2}{\frac{0.2275^2}{400} + \frac{0.2356^2}{500}} \approx 882$

Example 2:

- Suppose we are a political campaign that sends get-out-to-vote (GOTV) mailings to some members of a population that we would like to vote
  - 400 members of group A receive no political mailing; turnout in this group is 0.35
  - 500 members of group B receive a GOTV mailing; turnout in this group is 0.38

3. Construct a r.v. $T \sim t_{882}$ from μ_A, μ_B, $\hat{\mu}_A$, $\hat{\mu}_B$, S_A, S_B
   - $\frac{(\mu_A - \mu_B) - (\hat{\mu}_A - \hat{\mu}_B)}{\sqrt{\frac{{S_A}^2}{n_A} + \frac{{S_B}^2}{n_B}}} \sim t_\nu$

Example 2:

4. Rearrange terms so that $p(\mu_1 - \mu_2 < 0)$ is written as $p(T <?)$
   - $p(\mu_A - \mu_B < 0) =$
   - $p((\mu_A - \mu_B) - (\hat{\mu}_A - \hat{\mu}_B) < -(\hat{\mu}_A - \hat{\mu}_B)) =$
   - $p(\frac{(\mu_A - \mu_B) - (\hat{\mu}_A - \hat{\mu}_B)}{\sqrt{\frac{{s_A}^2}{n_A} + \frac{{s_B}^2}{n_B}}} < \frac{-(\hat{\mu}_A - \hat{\mu}_B)}{\sqrt{\frac{{s_A}^2}{n_A} + \frac{{s_B}^2}{n_B}}}) =$
   - $p(T < \frac{-(0.35 - 0.38)}{0.032}) =$
   - $p(T < 0.9375)$
5. Evaluate $T_{882}(0.9375) = 0.826$
   - Solved in R using "pt(.9375, 882)"

- Below is the R script I used to solve this problem
- Note that it produces an answer of 0.823, because it doesn't do all of the rounding that gets done when doing the problem by hand
  - Making this is a much better way to solve a question like this

```
SA <- 1/399*(140*(1 -.35)^2 + 260*(0 -.35)^2)
SB <- 1/499*(190*(1 -.38)^2 + 310*(0 -.38)^2)
SE <- (SA/400 + SB/500)^(1/2)
nu <- (SA/400 + SB/500)^2 /
  ((SA/400)^2/399 + (SA/500)^2/499)
T <- (-(0.35 - 0.38) / SE)
pt(T, nu)
```

- Lets apply to examine gender differences in the NY22 poll highlighted at the start of the lecture:

```
> ny22men <- svymean(~response, subset(ny22w, gender == "Male"))
> print(ny22men)
                mean     SE
responseDem 0.345412 0.0323
responseRep 0.566192 0.0339
responseUnd 0.088396 0.0188
>
> ny22female <- svymean(~response, subset(ny22w, gender == "Female"))
> print(ny22female)
                mean     SE
responseDem 0.544214 0.0327
responseRep 0.361110 0.0316
responseUnd 0.094677 0.0188
```

- Lets apply to examine gender differences in the NY22 poll highlighted at the start of the lecture:

```
> tt <- svyttest(I(response == "Rep")~gender, ny22w)
> print(tt)

        Design-based t-test

data:  I(response == "Rep") ~ gender
t = 4.4272, df = 504, p-value = 1.171e-05
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 0.1142907 0.2958736
sample estimates:
difference in mean
         0.2050821

> confint(tt, level = 0.95, df = numobs - 2)
                2.5 %    97.5 %
genderMale 0.1142907 0.2958736
attr(,"conf.level")
[1] 0.95
```

- Difference-in-means is useful for summarizing a bivariate relationship when the independent variable takes on two values
- It becomes less useful when our independent variable takes on more values
  - Three values means three possible comparisons
    - 1 vs. 2, 1 vs. 3, 2 vs. 3
  - Four values means six possible comparisons
    - 1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
  - Five values means ten possible comparisons
    - 1 vs. 2, 1 vs. 3, 1 vs. 4, 1 vs. 5, 2 vs. 3, 2 vs. 4, 2 vs. 5, 3 vs. 4, 3 vs. 5, 4 vs. 5
  - ...
- Thus, we need to develop some alternative ways to summarize a relationship between a dependent variable and independent variables that can take on a wider range of values

- The covariance between two r.v.'s ($cov(X, Y)$) is defined as $cov(X, Y) = \sigma_{xy} = E[(X - E[X])(Y - E[Y])]$
  - Also use the notation σ_{xy} to represent $cov(X, Y)$
- We can show that:
  $E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
  - $E[(X - E[X])(Y - E[Y])] =$
    $E[XY - XE[Y] - E[X]Y + E[X]E[Y]] =$
    $E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]] =$
    $E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]E[1] =$
    $E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]E[1] =$
    $E[XY] - E[Y]E[X]$

- To calculate the covariance, we need to calculate $E[XY]$
- When X and Y are discrete r.v.'s, then $E[XY] = \sum_{x \in X(S)} \sum_{y \in Y(S)} xyp(x, y)$
  - $p(x, y)$ is notation for the probability of the event $X \cap Y$ occurring
- When X and Y are continuous r.v.'s, then $E[XY] = \int_{x \in X(S)} \int_{y \in Y(S)} xyf(x, y)dxdy$
  - $f(x, y)$ is notation for the relative frequency of the event $X \cap Y$ occurring

- Example: Calculate cov(X, Y) of the r.v.'s described by the following pdf:

$$p(x,y) = \begin{array}{cc|c} x & y & p(x,y) \\ 0 & 0 & \frac{1}{12} \\ 0 & 1 & \frac{1}{6} \\ 0 & 2 & \frac{1}{4} \\ 1 & 0 & \frac{1}{4} \\ 1 & 1 & \frac{1}{6} \\ 1 & 2 & \frac{1}{12} \end{array}$$

- Calculate $E[X]$, $E[Y]$, $E[XY]$
  - $E[X] = \frac{1}{12} * 0 + \frac{1}{6} * 0 + \frac{1}{4} * 0 + \frac{1}{4} * 1 + \frac{1}{6} * 1 + \frac{1}{12} * 1 = \frac{1}{2}$
  - $E[Y] = \frac{1}{12} * 0 + \frac{1}{6} * 1 + \frac{1}{4} * 2 + \frac{1}{4} * 0 + \frac{1}{6} * 1 + \frac{1}{12} * 2 = 1$
  - $E[XY] = \frac{1}{12} * 0 * 0 + \frac{1}{6} * 0 * 1 + \frac{1}{4} * 0 * 2 + \frac{1}{4} * 1 * 0 + \frac{1}{6} * 1 * 1 + \frac{1}{12} * 1 * 2 = \frac{2}{6}$
- Use to solve for $cov(X, Y)$
  - $cov(X, Y) = E[XY] - E[Y]E[X] = \frac{2}{6} - \frac{1}{2}1 = -\frac{1}{6}$

- Let X and Y be r.v.'s and a and b be constants
- Construct $Z = aX + bY$
- $var(Z) = a^2 var(X) + 2abcov(X, Y) + b^2 var(Y)$
  - $var(Z) =$
    $E[Z^2] - E[Z]^2 =$
    $E[(aX + bY)^2] - E[aX + bY]^2 =$
    $E[a^2X^2 + 2abXY + b^2Y^2] - (aE[X] + bE[Y])^2 =$
    $(a^2E[X^2] + 2abE[XY] + b^2E[Y^2]) - (a^2E[X]^2 +$
    $2abE[X]E[Y] + b^2E[Y]^2) =$
    $a^2(E[X^2] - E[X]^2) + 2ab(E[XY] - E[X]E[Y]) +$
    $b^2(E[Y^2] - E[Y]^2) =$
    $a^2var(X) + 2abcov(X, Y) + b^2var(Y)$

- Suppose we observe n pairs of realizations of independent r.v.'s: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
- $\hat{\sigma_{xy}}$, an estimator of the covariance, equals $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
- The next slides shows the proof that $E[\hat{\sigma_{xy}}] = \sigma_{xy}$
- Leverages the fact that:
  - $E[x_i y_j] = \sigma_{xy} + E[x_i]E[y_j]$ if $i = j$
  - $E[x_i y_j] = E[x_i]E[y_j]$ if $i \neq j$
    - Because these are independent r.v.'s

Proof that $E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})]=\sigma_{xy}$:

- $E[\hat{\sigma_{xy}}]=$
  $E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})]=$
  $\frac{1}{n-1}E[\sum_{i=1}^{n}(x_iy_i-\bar{x}y_i-\bar{y}x_i+\bar{x}\bar{y})]=$
  $\frac{1}{n-1}E[\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}\bar{x}y_i-\sum_{i=1}^{n}\bar{y}x_i+\sum_{i=1}^{n}\bar{x}\bar{y})]=$
  $\frac{1}{n-1}E[\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+\bar{x}\bar{y}\sum_{i=1}^{n}1]=$
  $\frac{1}{n-1}E[\sum_{i=1}^{n}x_iy_i-\bar{x}(n\bar{y})-\bar{y}(n\bar{x})+\bar{x}\bar{y}(n)]=$
  $\frac{1}{n-1}(\sum_{i=1}^{n}E[x_iy_i]-nE[\bar{x}\bar{y}])=$
  $\frac{1}{n-1}(\sum_{i=1}^{n}E[x_iy_i]-nE[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}x_iy_j}{n^2}])=$
  $\frac{1}{n-1}(\sum_{i=1}^{n}E[x_iy_i]-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E[x_iy_j])=$
  - $E[x_iy_j]=\sigma_{xy}+E[x_i]E[y_i]$ if $i=j$
  - $E[x_iy_j]=E[x_i]E[y_j]$ if $i\neq j$

  $\frac{1}{n-1}(n(\sigma_{xy}+E[x_i]E[y_i])-\frac{1}{n}(n\sigma_{xy}+n^2E[x_i]E[y_i]))$
  $\frac{1}{n-1}((n-1)\sigma_{xy})=\sigma_{xy}$

Properties of the covariance:

- A positive (negative) covariance indicates that higher realizations of X tend to associate with higher (lower) realizations of Y
  - And that positive realizations of Y tend to associate with higher realizations of X
- When r.v.'s X and Y are independent, then $cov(X, Y) = 0$
  - Independence implies that $E[XY] = E[X]E[Y] \implies$
  - $E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0$
- Observing that $cov(X, Z) > cov(X, Y)$ does not necessarily imply that there is a stronger relationship between X and Z than X and Y
  - Let $Z = 100 * Y$
  - $cov(X, Z) =$
    $E[XZ] - E[X]E[Z] =$
    $E[X(100 * Y)] - E[X]E[100 * Y] =$
    $100(E[XY] - E[X]E[Y]) = 100 * cov(X, Y)$

- Because of the difficultly in interpreting the covariance, we often use correlation to summarize the strength of the relationship between two r.v.'s.
- The Pearson correlation (or correlation) between X and Y, ρ_{xy}, equals $\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$
- Things we can prove to be true about this measure:
  1. $-1 < \rho_{xy} < 1$
  2. If $X^{'} = qX$ then $\rho_{x^{'}y} = \rho_{xy}$
  3. $\mid \rho_{xy} \mid = 1$ if $Y = \lambda_1 X + \lambda_2$

Proof that $-1 < \rho_{xy} < 1$:

- Cauchy-Schwartz Inequality says that for any r.v.'s W and Z:
  $E[WZ]^2 \leqslant E[W^2]E[Z^2]$
- Define $W = X - E[X]$ and $Z = Y - E[Y] \implies$
  $(E[(X - E[X])(Y - E[Y])])^2 \leqslant$
  $E[(X - E[X])^2]E[(Y - E[Y])^2] \implies$
  $cov(X, Y)^2 \leqslant var(X)var(Y) \implies$
  $\frac{cov(X,Y)^2}{var(X)var(Y)} \leqslant 1 \implies$
  $-1 \leqslant \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \leqslant 1$

Proof that if $X' = qX$ then $\rho_{x'y} = \rho_{xy}$:

- $\rho_{x'y} =$

$$\frac{cov(X',Y)}{\sqrt{var(X')var(Y)}} =$$

$$\frac{cov(qX,Y)}{\sqrt{var(qX)var(Y)}} =$$

$$\frac{qcov(X,Y)}{\sqrt{q^2 var(X)var(Y)}} =$$

$$\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \rho_{xy}$$

Proof that $\mid \rho_{xy} \mid = 1$ if $Y = \lambda_1 X + \lambda_2$:

- Meaning that correlation is a measure of the linear association between two r.v.'s
- $var(Y) =$
  $var(\lambda_1 X + \lambda_2) =$
  ${\lambda_1}^2 var(X)$
- $cov(X, Y) = cov(X, \lambda_1 X + \lambda_2) =$
  $cov(X, \lambda_1 X) + cov(X, \lambda_2) =$
  $\lambda_1 var(X)$
- $\rho_{xy} =$
  $\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} =$
  $\frac{\lambda_1 var(X)}{\sqrt{var(X){\lambda_1}^2 var(X)}} = 1$

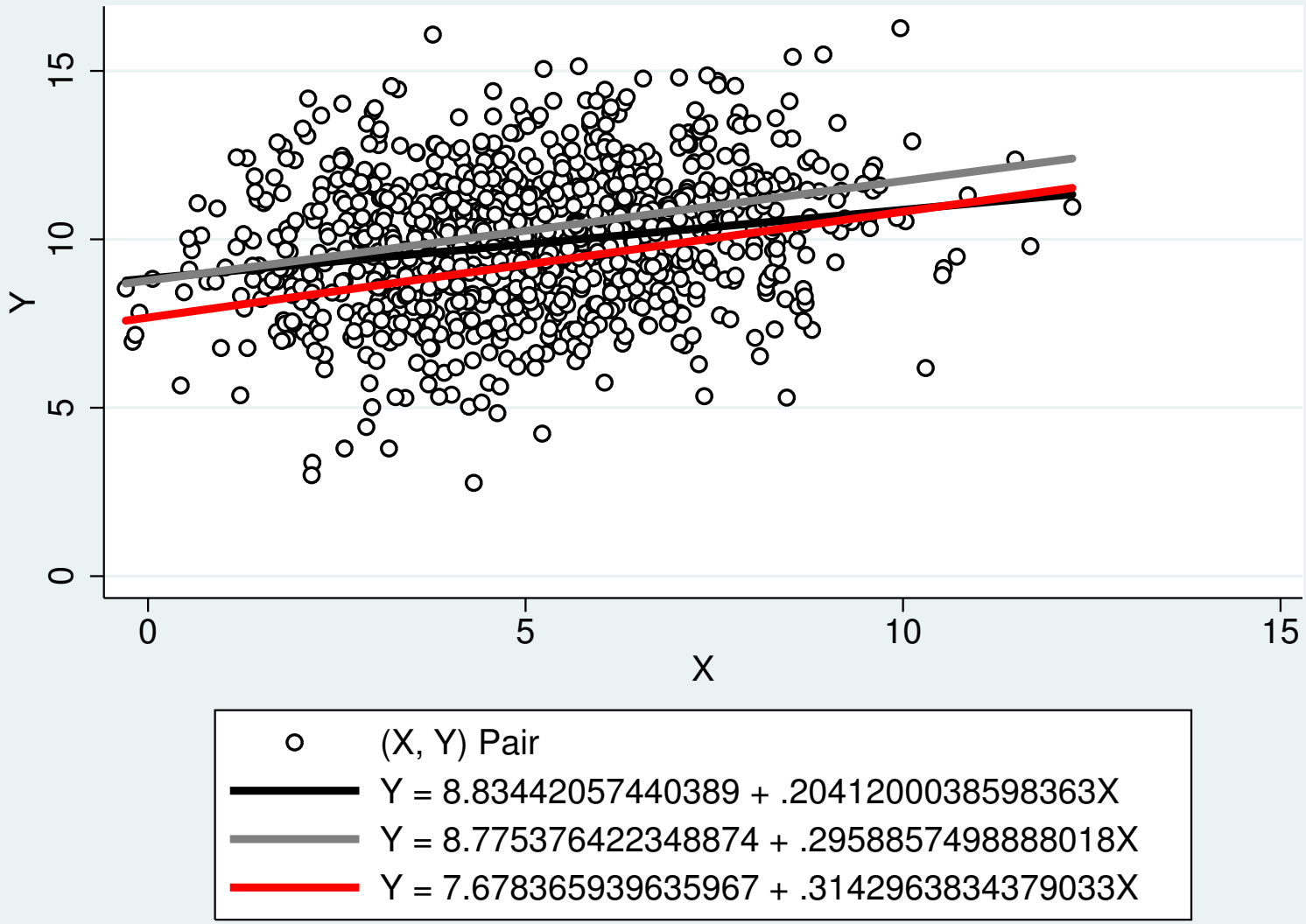
- The previous slide reveals a weakness of correlation, which is that all linear associations produce the same correlation
- All of following relationships would produce a correlation of 1:
  1. $Y = X$
  2. $Y = 100X$
  3. $Y = \frac{1}{100}X$
  4. $Y = \frac{1}{100}X + 400$
- Thus, it is somewhat problematic to interpret a high correlation as a measure of the similarity of X and Y

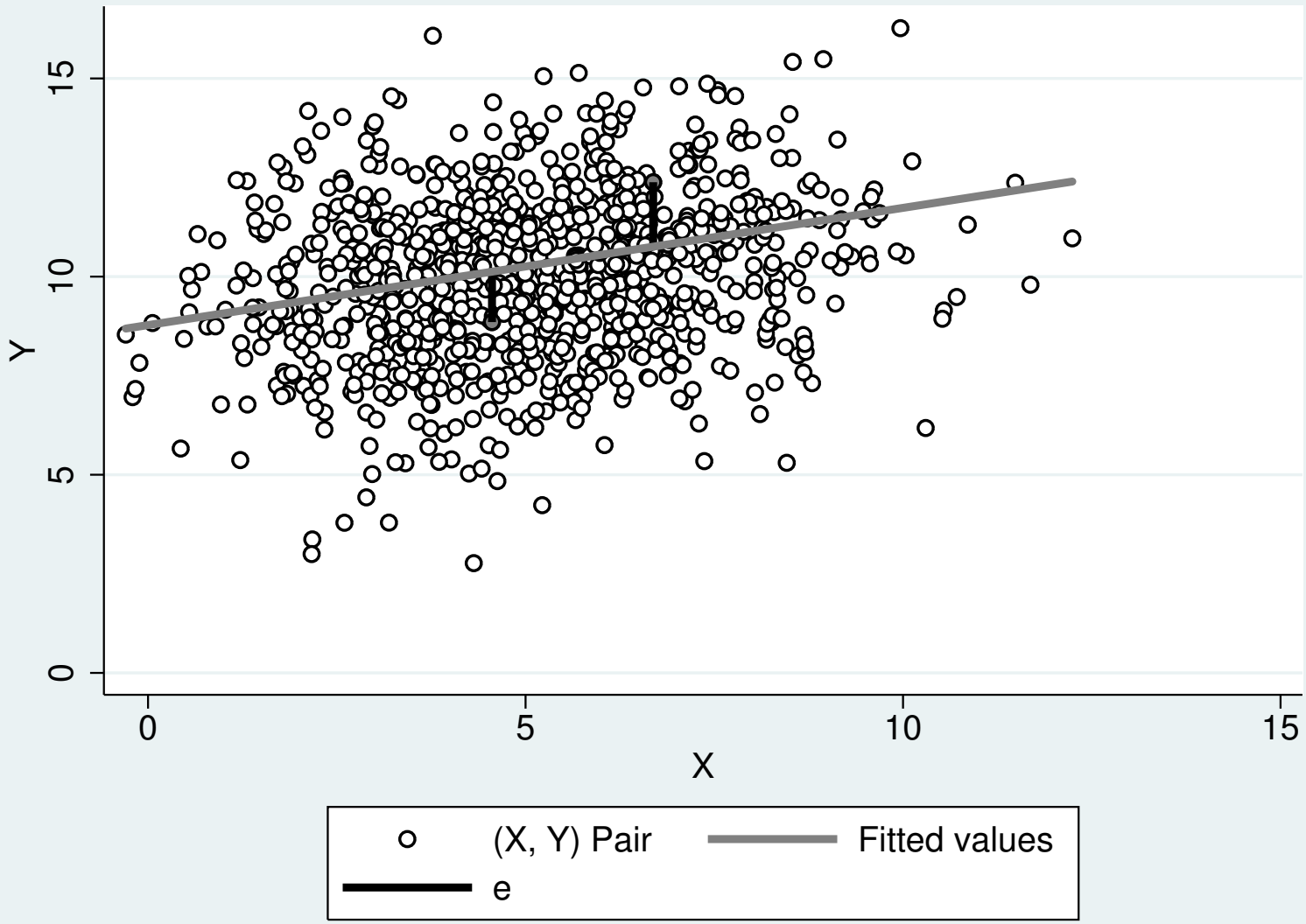
- We estimate ρ_{xy} with $\hat{\rho}_{xy} = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$
- It can be demonstrated that $E[\hat{\rho}_{xy}] \neq \rho_{xy}$
  - E.g., $\hat{\rho}_{xy}$ is a biased estimator of r_{xy}
  - When X and Y are distributed bivariate normal, it is biased towards zero
- But it is a consistent estimator, meaning that the bias goes to zero as the sample size gets large

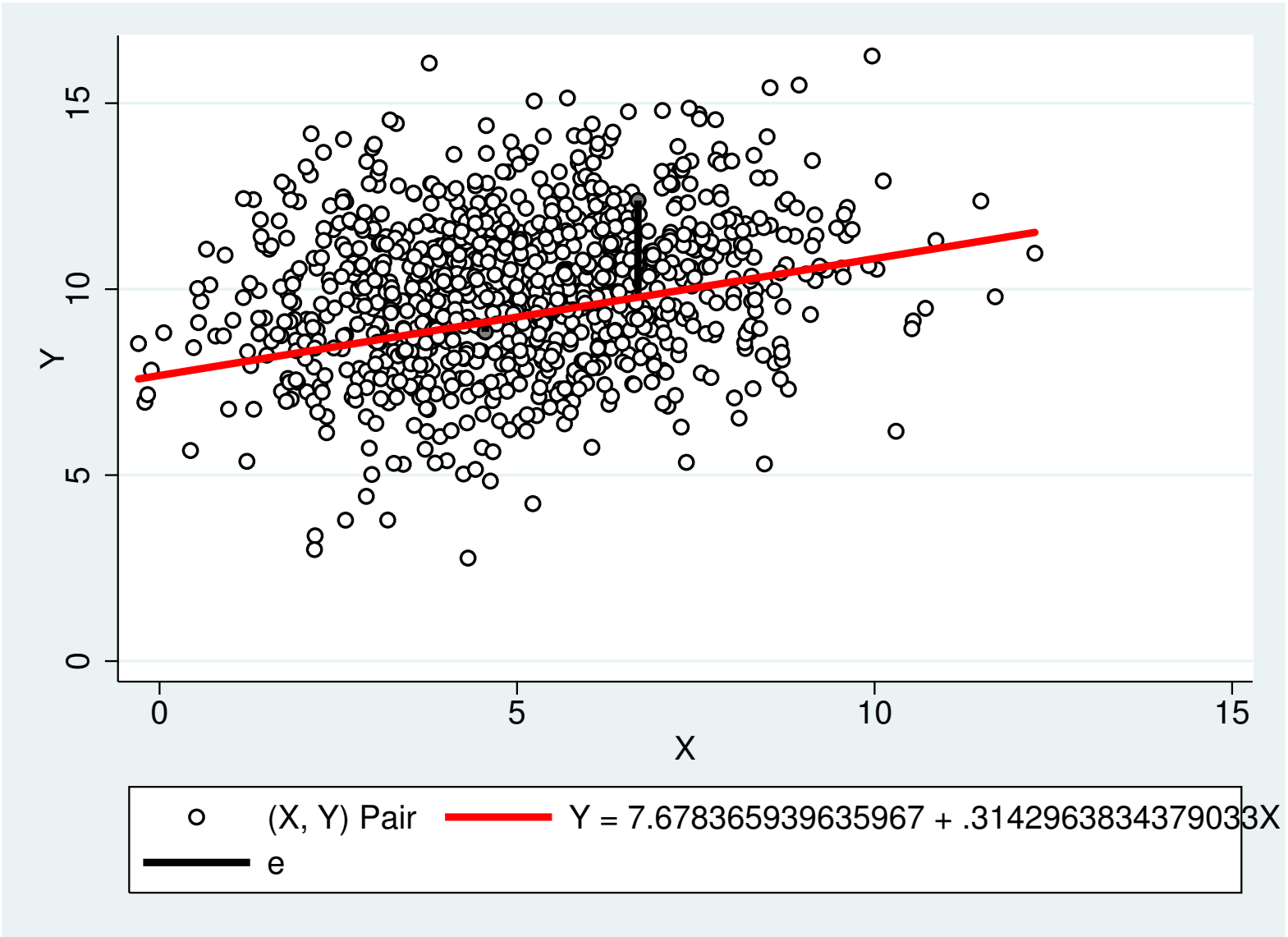
- Spearman's rank correlation coefficient is an alternate correlation that is based on the similarity of the ranking of X_i and Y_i in a sample
  - Addresses the issue that we'll see in a few slides that the Pearson correlation can be quite sensitive to outliers
- How to compute Spearman's rank correlation:
  1. Order the X's from lowest to highest. Define $\tilde{x}_i$ as the rank in that ordering of x_i
     - If there are ties, give every observation with the same value the average ranking
  2. Do the same for the Y's
     - Let $(x_1, y_1) = (3, 12)$, $(x_2, y_2) = (7, 8)$, $(x_3, y_3) = (-2, -5)$, $(x_4, y_4) = (2, 10)$
     - Then $(\tilde{x}_1, \tilde{y}_1) = (3, 4)$, $(\tilde{x}_2, \tilde{y}_2) = (4, 2)$, $(\tilde{x}_3, \tilde{y}_3) = (1, 1)$, $(\tilde{x}_4, \tilde{y}_4) = (2, 4)$
  3. Construct $d_i = \tilde{x}_i - \tilde{y}_i$
  4. $\rho_{spearman} = 1 - \frac{6 \sum_{i=1}^{N} {d_i}^2}{n(n^2-1)}$

- Because of the issues the previous slides identified with correlation, bivariate regressions are the most common way to describe the relationship between X and Y
- Let $Y_i = \alpha + \beta X_i + \epsilon_i$
  - Where ϵ_i is a r.v.
- The goal of bivariate regression is to identify the line (i.e., $\alpha + \beta X_i$) that "best" describes the relationship between X_i and Y_i when we apply the same α and β to all observations i in the sample
  - Which requires a definition of best
- One common definition of best is the line that minimizes the average of $(Y_i - \alpha - \beta X_i)^2$ in the population

- We do something analogous to define the line that best describes the relationship between X and Y in a sample
- Let $\hat{\alpha}$ and $\hat{\beta}$ be estimates of α and β from data contained in a sample
- For any $\hat{\alpha}$ and $\hat{\beta}$, we can define:
  - $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, which is the our guess of Y_i based on $\hat{\alpha}$ and $\hat{\beta}$
    - Which is the called the fitted value
- $e_i = Y_i - \hat{Y}_i = (Y_i - \hat{\alpha} - \hat{\beta} X_i)$
  - Which is the called the residual
- The criterion of least squares defines $L(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{N} {e_i}^2$
- A least squares regression estimates $\hat{\alpha}$ and $\hat{\beta}$ by selecting the values that minimize $L(\hat{\alpha}, \hat{\beta})$


Y
0
5
10
15
0
5
10
15
X
(X, Y) Pair
Y = 8.83442057440389 + .2041200038598363X
Y = 8.775376422348874 + .2958857498888018X
Y = 7.678365939635967 + .3142963834379033X


Y
0
5
10
15
X
0
5
10
15
(X, Y) Pair
Fitted values
e

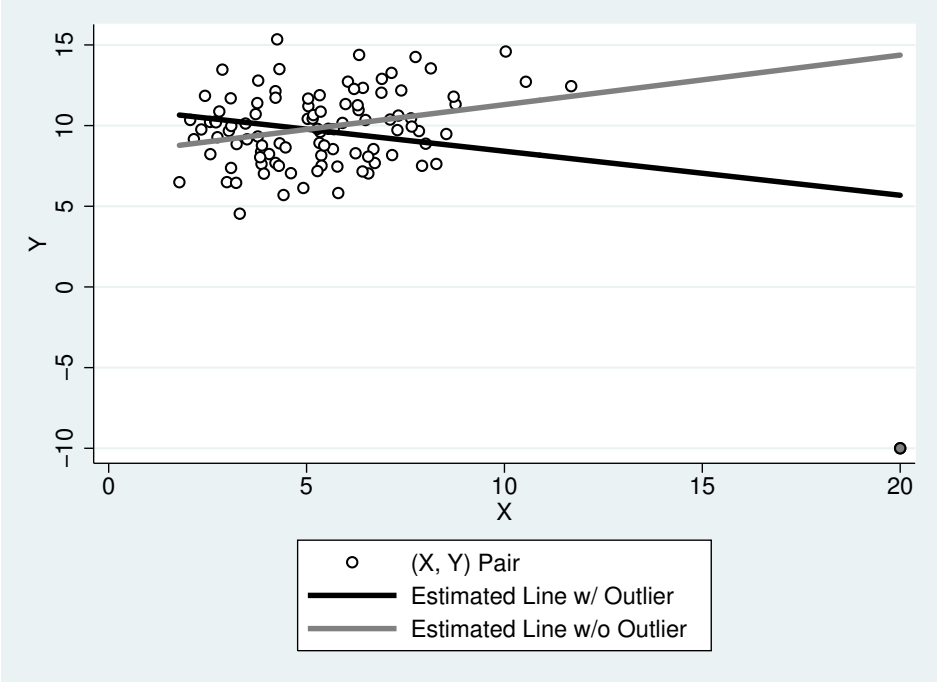

Y
0
5
10
15
X
0
5
10
15
(X, Y) Pair
Y = 7.678365939635967 + .3142963834379033X
e

- To minimize functions, we generally take the derivatives w.r.t. to the parameters that we wish to minimize function w.r.t
  - $L(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{N}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$
  - So $\frac{dL(\hat{\alpha},\hat{\beta})}{d\hat{\alpha}} = \sum_{i=1}^{N} -2(Y_i - \hat{\alpha} - \hat{\beta}X_i)$
  - And $\frac{dL(\hat{\alpha},\hat{\beta})}{d\hat{\beta}} = \sum_{i=1}^{N} -2X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i)$
- So we solve for the value of $\hat{\alpha}$ that sets the first derivative equal to zero
  - $\sum_{i=1}^{N} -(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \implies$
  - $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$
- And plug into the second equation in order to solve for $\hat{\beta}$
  - $\sum_{i=1}^{N} -X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \implies$
  - $\sum_{i=1}^{N} -X_i(Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i) = 0 \implies$
  - $\hat{\beta}(\sum_{i=1}^{N} {X_i}^2 - \bar{X}^2) = \sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y} \implies$
  - $\hat{\beta} = \frac{\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}}{\sum_{i=1}^{N} {X_i}^2 - \bar{X}^2}$

- An implication of the formula for $\hat{\beta}$ is that $\hat{\beta} = \frac{cov(\hat{X},Y)}{var(\hat{X})}$
  - $\hat{\beta} = \frac{\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y}}{\sum_{i=1}^{N} X_i^2 - \bar{X}^2} = \frac{\frac{1}{n-1}(\sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y})}{\frac{1}{n-1}(\sum_{i=1}^{N} X_i^2 - \bar{X}^2)} = \frac{cov(\hat{X},Y)}{var(\hat{X})}$
- Thus, $\hat{\beta} = \hat{\rho}_{xy}\sqrt{\frac{var(\hat{Y})}{var(\hat{X})}}$
  - Because $\hat{\rho}_{xy} = \frac{cov(\hat{X},Y)}{\sqrt{var(\hat{X})var(\hat{Y})}} = \hat{\beta}\sqrt{\frac{var(\hat{X})}{var(\hat{Y})}}$

- We established on the previous slide that $\hat{\beta} = \hat{\rho}_{xy}\sqrt{\frac{\hat{var}(Y)}{\hat{var}(X)}}$ when X is the independent variable and Y is the dependent variable
- So $\hat{\beta} = \hat{\rho}_{xy}\sqrt{\frac{\hat{var}(X)}{\hat{var}(Y)}}$ if we made X the dependent variable and Y the independent variable
- Both $\sqrt{\frac{\hat{var}(Y)}{\hat{var}(X)}}$ and $\sqrt{\frac{\hat{var}(X)}{\hat{var}(Y)}}$ can never be negative
  - Any variance is positive as long as there is some variability in the value of the variable
- Thus, the sign of $\hat{\beta}$ is determined by the sign of $\hat{\rho}_{xy}$
  - Highlights that we cannot necessarily interpret $\hat{\beta}$ causally

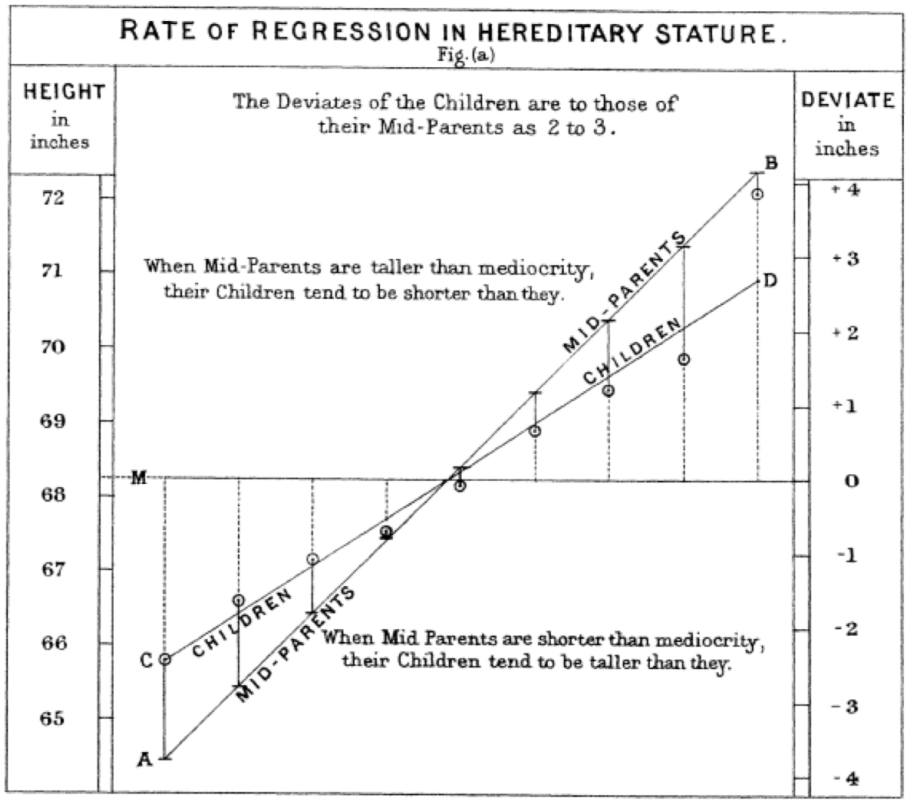
- Because covariance is sensitive to outliers, both correlations and least squares regressions also are sensitive to outliers



- Other types of regressions (e.g., quantile regressions) exist, often in the spirit of the Spearman's rank correlation, that are designed to be less sensitive to outliers

- We can show that $E[\hat{Y_i} \mid X_i] = \hat{\alpha} + \hat{\beta} X_i$
  - $E[\hat{Y_i} \mid X_i] =$
    $E[\hat{\alpha} + \hat{\beta} X_i + e_i \mid X_i] =$
    $E[\hat{\alpha} \mid X_i] + E[\hat{\beta} X_i \mid X_i] + E[e_i \mid X_i] = \hat{\alpha} + \hat{\beta} X_i$
    - Residuals are constructed so that $E[e_i \mid X_i] = 0$
- Implication is that $E[Y_i \mid \hat{X}_i = c] = \hat{\alpha} + \hat{\beta} c$
  - See the Appendix of Ch. 2 of Angrist and Pischke for further discussion about how to interpret regression coefficients
- Interpretation:
  - $\hat{\alpha}$ is the expected value of Y_i when $X_i = 0$
  - $\hat{\beta}$ is the expected change in Y_i from a one-unit change in X_i

The first regression:



Source: Galton 1886 Plate IX, fig. a), as reprinted in Han, Ma, and Zhu (2015)

- Translating Galton's picture into an equation: $\hat{Child}_i = 21.328 + 0.688 * Parents_i$
  - E.g., the expected value of a child's height (in inches) given their parent's height (in inches)
- Implications:
  - If a child's parents average height is 65 inches, then we would predict the child would be 21.328 + 0.688*65 = 66.016 inches
  - If a child's parents average height is 68.25 inches, then we would predict the child would be 21.328 + 0.688*68.25 = 68.284 inches
  - If a child's parents average height is 72 inches, then we would predict the child would be 21.328 + 0.688*68.25 = 70.828 inches

- In Galton, both the independent (e.g., parents height) and dependent (e.g., kids height) variables take on many values
- However, all that is necessary to be able to run a bivariate regression is that the independent variable takes on at least two values
- A binary independent variable (i.e., $X_i = \{0, 1\}$) can be used to take generate a difference in means analysis
  - $\hat{\alpha}$ equals the average value of Y when $X_i = 0$
  - $\hat{\alpha} + \hat{\beta}$ equals in the average value of Y when $X_i = 1$
  - Meaning that $\hat{\beta}$ equals the difference in the average value of Y when $X_i = 1$ relative to when $X_i = 0$

```
reg1 <- lm(Voted ~ Mobilized, data = LectureDifferenceMeans)
reg1summary <- summary(reg1)
print(reg1summary)

Call:
lm(formula = Voted ~ Mobilized, data = LectureDifferenceMeans)

Residuals:
   Min     1Q Median     3Q    Max
 -0.38  -0.38  -0.35   0.62   0.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35000    0.02411  14.517   <2e-16 ***
Mobilized    0.03000    0.03235   0.927    0.354
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4822 on 898 degrees of freedom
Multiple R-squared:  0.0009569,        Adjusted R-squared:  -0.0001556
F-statistic: 0.8602 on 1 and 898 DF,  p-value: 0.3539
```

How do we interpret these results?

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.3800 -0.3800 -0.3500  0.6200  0.6500
```

- This summarizes the distribution of $e_i = Y_i - \hat{Y}_i$:
  - The smallest and 25th percentile value is -0.38
    - A non-voter ($Y_i = 0$) who was mobilized ($\hat{Y}_i = 0.38$)
  - The 50th percentile value is -0.35
    - A non-voter ($Y_i = 0$) who was not mobilized ($\hat{Y}_i = 0.35$)
  - The 75th percentile value is 0.62
    - A voter ($Y_i = 1$) who was mobilized ($\hat{Y}_i = 0.38$)
  - The maximum value is 0.65
    - A voter ($Y_i = 1$) who was not mobilized ($\hat{Y}_i = 0.35$)

How do we interpret these results?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35000    0.02411  14.517   <2e-16 ***
Mobilized    0.03000    0.03235   0.927    0.354
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- The first column tells us about the regression coefficients:
  - $\hat{\alpha} = 0.35$ (i.e., average turnout is 35 percent in the non-mobilized group)
  - $\hat{\beta} = 0.03$ (i.e., average turnout is 3 percent higher in the mobilized group than non-mobilized group)
  - An implication of $\hat{\alpha} + \hat{\beta} = 0.38$ is that average turnout is 38 percent in the mobilized group

How do we interpret these results?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35000    0.02411  14.517   <2e-16 ***
Mobilized    0.03000    0.03235   0.927    0.354
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- The second column tells us about the degree of uncertainty in the regression coefficients:
  - $\sigma_{\hat{\alpha}} = 0.024$ (i.e., the standard error on turnout in the non-mobilized group is 2.4 percent)
  - $\sigma_{\hat{\beta}} = 0.032$ (i.e., the standard error on the difference in turnout between the mobilized and non-mobilized group is 3.2 percent)
  - We cannot directly access the standard error on turnout in the mobilized group from this output

How do we interpret these results?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35000    0.02411  14.517   <2e-16 ***
Mobilized    0.03000    0.03235   0.927    0.354
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- The third column tells us the value of $\frac{\hat{c}-c}{\sigma_{\hat{c}}}$ if $c = 0$ (more next week about why)
  - $\frac{\hat{\alpha}}{\sigma_{\hat{\alpha}}} = 14.517$
  - $\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} = 0.927$

How do we interpret these results?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.35000    0.02411  14.517   <2e-16 ***
Mobilized    0.03000    0.03235   0.927    0.354
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- The fourth column tells us $p(\mid T \mid > \frac{\hat{c}-c}{\sigma_{\hat{c}}}$ if $c = 0$ and $T \sim t_{n-2}$ (more next week about why)
  - $p(\mid T \mid > 14.517) < 0.0000000000000002$ when $T \sim t_{898}$
    - 898 is the number of observations minus the number of coefficients estimated in the regression (i.e., the degrees of freedom)
  - $p(\mid T \mid > 0.927) < 0.354$ when $T \sim t_{898}$
- The "***", "**", "*", or "." denote whether the value of $p(\mid T \mid > \frac{\hat{c}-c}{\sigma_{\hat{c}}})$ is within a certain range
  - E.g., ** denotes the value is greater than 0.001, but less than 0.01

How do we interpret these results?

```
Residual standard error: 0.4822 on 898 degrees of freedom
Multiple R-squared:  0.0009569,Adjusted R-squared:  -0.0001556
F-statistic: 0.8602 on 1 and 898 DF,  p-value: 0.3539
```

- These are all measures, in some form, of the goodness-of-fit of the model
- All things equal, the model fits the data better when the residual standard error is lower, the multiple R-squared and Adjusted R-squared is higher, and the F-statistic is higher
- We'll define these concept more precisely in a couple of weeks

- One thing that is not directly available in R output are confidence intervals on the coefficients
- However, the next slide shows that they easily can be constructed either using
  - The confint() function
    - Although limited to symmetric CI
  - The data contained when running the summary() function on a data.frame with regression output
- Requires either a large enough sample to apply the CLT or an assumption that $\epsilon_i \sim N(0, \sigma^2)$

- A rule of thumb is that we roughly are 95% certain that the true relationship between a dependent variable and independent variables is somewhere within two standard errors of the regression coefficient

```
> confint(reg1, level = 0.95)
                 2.5 %     97.5 %
(Intercept)  0.30268151 0.39731849
Mobilized   -0.03348442 0.09348442
> print(reg1summary$coefficients)
            Estimate Std. Error    t value     Pr(>|t|)
(Intercept)     0.35 0.02410999 14.5168023 4.749587e-43
Mobilized       0.03 0.03234695  0.9274445 3.539450e-01
> print(reg1summary$df)
[1]   2 898   2
> print(reg1summary$coefficients[1, 1]  +
+         reg1summary$coefficients[1, 2]*
+         qt(0.025, df = reg1summary$df[2]))
[1] 0.3026815
> print(reg1summary$coefficients[1, 1]  +
+         reg1summary$coefficients[1, 2]*
+         qt(0.975, df = reg1summary$df[2]))
[1] 0.3973185
```

- Sometimes we may want to construct a confidence interval on a combination of two or more coefficients
  - E.g., We need to construct a confidence interval of $\alpha + \beta$ to put a confidence interval on the rate of turnout in the mobilized population
- Unfortunately, this is not easily done using the confint() function so we need to follow the steps shown on the previous slide to manually construct a confidence interval on regression coefficients
- But before we can do this we need to construct the standard error on the combination of two or more coefficients
  - E.g.,

$$\sigma_{\hat{\alpha}+\hat{\beta}} = \sqrt{var(\hat{\alpha} + \hat{\beta})} = \sqrt{var(\hat{\alpha}) + 2cov(\hat{\alpha}, \hat{\beta}) + var(\hat{\beta})}$$

Constructing $\sigma_{\hat{\alpha}+\hat{\beta}}$

```
> vcov = vcov(reg1)
> print(vcov)
             (Intercept)     Mobilized
(Intercept)  0.0005812918 -0.0005812918
Mobilized   -0.0005812918  0.0010463252
> print(vcov^(1/2))
            (Intercept)  Mobilized
(Intercept)  0.02410999        NaN
Mobilized           NaN 0.03234695
> alphabeta_se <- (vcov[1, 1] + 2*vcov[2, 1] +
+                   vcov[2, 2])^(1/2)
```

Using $\sigma_{\hat{\alpha}+\hat{\beta}}$ to construct confidence interval on $\alpha + \beta$

```
> print(reg1summary$coefficients[1, 1]  +
+         reg1summary$coefficients[2, 1] +
+         alphabeta_se*
+         qt(0.025, df = reg1summary$df[2]))
[1] 0.3376771
> print(reg1summary$coefficients[1, 1]  +
+         reg1summary$coefficients[2, 1] +
+         alphabeta_se*
+         qt(0.975, df = reg1summary$df[2]))
[1] 0.4223229
```

- How to "replicate" the NY22 poll analysis using lm()

```
> meanreg <- lm(I(response == "Rep")~1, weight = final_weight, data = ny22)
> summary(meanreg)

Call:
lm(formula = I(response == "Rep") ~ 1, data = ny22, weights = final_weight)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-0.7088 -0.4569 -0.3462  0.5329  0.9040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45882    0.02217   20.69   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4988 on 505 degrees of freedom

> confint(meanreg, level = 0.95, df = meanreg$df.residual)
                2.5 %    97.5 %
(Intercept) 0.4152516 0.5023813
```

- How to "replicate" the NY22 poll analysis using lm()

```
> diffreg <- lm(I(response == "Rep")~gender, weight = final_weight, data = ny22)
> summary(diffreg)

Call:
lm(formula = I(response == "Rep") ~ gender, data = ny22, weights = final_weight)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-0.8648 -0.4227 -0.2803  0.4763  0.8906

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.36111    0.03002  12.029  < 2e-16 ***
genderMale   0.20508    0.04349   4.715 3.13e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4886 on 504 degrees of freedom
Multiple R-squared:  0.04225,        Adjusted R-squared:  0.04035
F-statistic: 22.23 on 1 and 504 DF,  p-value: 3.128e-06

> confint(diffreg, level = 0.95, df = meanreg$df.residual)
                2.5 %    97.5 %
(Intercept) 0.3021296 0.4200899
genderMale  0.1196330 0.2905313
```

Key takeaways:

- Estimation is the process through which we use the data contained in a sample to make judgements about the value of population parameters, as well as our certainty about these judgments
- An estimator with less bias and lower variance generally will generate less estimation error than an estimator with more bias and higher variance, but that usually isn't guaranteed in any given sample of data.
- Increasing the sample size often is the best way to reduce estimation error
    - Although thinking back to last week we might be better off with a smaller random sample of data than a larger non-random sample

Key takeaways:

- When the conditions of the LLN and CLT are met, we can construct probabilistic statements about the likelihood that a population mean or the difference in two population means is contained in some range
- Least squares regression is one tool that can be used to produce this analysis
- This is just one example of how least squares regressions are useful in addressing empirical questions we encounter about conditional expectations or differences in conditional expectations
- Moving forward, we'll see that the real value of a least squares regression is that we'll be able to simultaneously account for the effect of many independent variables