Gov 2000 - 5. Univariate Inference II: Interval Estimation and Testing

Matthew Blackwell

October 13, 2015

Where are we? Where are we going?

- Last few weeks = how to produce a single, best estimate of some population parameter, drawing on our knowledge of probability.
- Now: what if we want to express uncertainty about estimates? **Standard error** gives us some indication of how accurate our estimate is, but it's limited.
- What if we want to give a range of values instead of a single value? (Interval estimation)
- Also, what if we want to assess the plausibility of a particular hypothesis about our data (Hypothesis testing).
- This last goal is basically what we did in the Lady Tasting Tea example.

Midterm notes

- Midterm will go out 10/22 (next Thursday).
- Practice problems will go out soon.
- 3 questions, multiple parts per question.
- Available until the follow Thursday at 5pm, you have 5 hours to complete it once you check it out.
- Rules: no working or corresponding with any human being while taking the exam.
- If unsure about wording, make a clarifying assumption and note it in your answer.
- No HW that week, obvi.

Social pressure effect

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Motivating Example

```
load("gerber_green_larimer.RData")
## turn turnout variable into a numeric
social$voted <- 1 * (social$voted == "Yes")
neigh.mean <- mean(social$voted[social$treatment ==
    "Neighbors"])
neigh.mean
```

```
## [1] 0.378
```

```
contr.mean <- mean(social$voted[social$treatment ==
    "Civic Duty"])
contr.mean
```

```
## [1] 0.315
```

```
neigh.mean - contr.mean
```

```
## [1] 0.0634
```

- What is a range of plausible values? Could this happen by random chance?

1/ Large Sample Confidence Intervals

Interval estimation - what and why?

- $\overline{Y}_n - \overline{X}_n$ is our best guess about $\mu_y - \mu_x$
- But $\mathbb{P}(\overline{Y}_n - \overline{X}_n = \mu_y - \mu_x) = 0$!
- Alternative: produce a range of values that will contain the truth with some fixed probability
- An **interval estimate** of the population difference in means, $\mu_y - \mu_x$, consists of two bounds within which we expect $\mu_y - \mu_x$ to reside:

$$LB \leq \mu_y - \mu_x \leq UB$$

- How can we possibly figure out such an interval? We'll rely on the distributional properties of estimators. Ideas extend to all estimators, including regression.

Review of the difference in means

- **Treated group** $Y_1, Y_2, \ldots, Y_n$ i.i.d. with population mean μ_y and population variance σ_y^2
- **Control group** $X_1, X_2, \ldots, X_n$ i.i.d. with population mean μ_x and population variance σ_x^2
- Look the at the difference in means:

$$\widehat{D}_n = \overline{Y}_n - \overline{X}_n$$

Sampling distribution of the difference

- From last week, we know in large samples, by the CLT:

$$\widehat{D}_n \sim N\left(\mu_y - \mu_x, \mathbb{V}[\widehat{D}_n]\right)$$

- Remember the variance:

$$\mathbb{V}[\widehat{D}_n] = \mathbb{V}[\overline{Y}_n] + \mathbb{V}[\overline{X}_n] = \frac{\sigma_y^2}{n} + \frac{\sigma_x^2}{n}$$

- Today, it'll be easier to work with the standard error, $SE[\widehat{D}_n] = \sqrt{\mathbb{V}[\widehat{D}_n]}$
- From last week, remember that we can replace the SE with an estimate of the SE:

$$\widehat{SE}[\widehat{D}_n] = \sqrt{\frac{S_{yn}^2}{n} + \frac{S_{xn}^2}{n}}$$

Deriving a probabilistic bound

- When n is large, then we have

$$\widehat{D}_n \sim N((\mu_y - \mu_x), \widehat{SE}[\widehat{D}_n]^2)$$

$$\widehat{D}_n - (\mu_y - \mu_x) \sim N(0, \widehat{SE}[\widehat{D}_n]^2)$$

$$\frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{SE}[\widehat{D}_n]} \sim N(0,1)$$

- Let's use this to calculate a lower bound for $(\mu_y - \mu_x)$:

$$\mathbb{P}\left(LB < (\mu_y - \mu_x)\right) = 0.95$$

- We want to find a value so that in 95% of random samples, it will be lower than the true difference in means.
- Use the following fact:

$$\mathbb{P}\left(\frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{SE}[\widehat{D}_n]} \leq 1.64\right) = 0.95$$

Getting a lower bound for the effect

$$P\left(\frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{SE}[\widehat{D}_n]} \leq 1.64\right) = 0.95$$

$$\mathbb{P}\left(\widehat{D}_n - (\mu_y - \mu_x) \leq 1.64 \times \widehat{SE}[\widehat{D}_n]\right) = 0.95$$

$$\mathbb{P}\left(-(\mu_y - \mu_x) \leq -\widehat{D}_n + 1.64 \times \widehat{SE}[\widehat{D}_n]\right) = 0.95$$

$$\mathbb{P}\left((\mu_y - \mu_x) \geq \widehat{D}_n - 1.64 \times \widehat{SE}[\widehat{D}_n]\right) = 0.95$$

- Lower bound: $\widehat{D}_n - 1.64 \times \widehat{SE}[\widehat{D}_n]$.
- In 95% of random samples, this lower bound will be below $(\mu_y - \mu_x)$.

Getting an upper bound for the effect

$$P\left(-1.64 \leq \frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{SE}[\widehat{D}_n]}\right) = 0.95$$

$$\mathbb{P}\left(-1.64 \times \widehat{SE}[\widehat{D}_n] \leq \widehat{D}_n - (\mu_y - \mu_x)\right) = 0.95$$

$$\mathbb{P}\left(-\widehat{D}_n - 1.64 \times \widehat{SE}[\widehat{D}_n] \leq -(\mu_y - \mu_x)\right) = 0.95$$

$$\mathbb{P}\left(\widehat{D}_n + 1.64 \times \widehat{SE}[\widehat{D}_n] \geq (\mu_y - \mu_x)\right) = 0.95$$

- Upper bound: $\widehat{D}_n + 1.64 \times \widehat{SE}[\widehat{D}_n]$.
- In 95% of random samples, this upper bound will be above $(\mu_y - \mu_x)$.

Putting the bounds together

- Let $\widehat{LB} = \widehat{D}_n - 1.64 \times \widehat{SE}[\widehat{D}_n]$
- Let $\widehat{UB} = \widehat{D}_n + 1.64 \times \widehat{SE}[\widehat{D}_n]$
- The probability of each bound is 0.95:

$$\mathbb{P}\left((\mu_y - \mu_x) \leq \widehat{UB}\right) = 0.95$$

$$\mathbb{P}\left(\widehat{LB} \leq (\mu_y - \mu_x)\right) = 0.95$$

- But one of these might hold and the other might not, so:

$$\mathbb{P}\left(\left\{\widehat{LB} \leq (\mu_y - \mu_x)\right\} \cap \left\{(\mu_y - \mu_x) \leq \widehat{UB}\right\}\right) < 0.95$$

What is a confidence interval?

- **Definition** A $100(1-\alpha)\%$ **confidence interval** with is an interval estimator for a population parameter θ that will contain/cover the true value, θ, $100(1-\alpha)\%$ of the time.

- This rule will be an estimator just like the sample mean or the sample variance, but it will produce two values instead of one: the upper and lower values of the intervals.

Confidence intervals

- Remember, by the CLT, we have the following:

$$\frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{SE}[\widehat{D}_n]} \sim N(0,1)$$

- We know that the probability of being between -1.64 and 1.64 (from `pnorm()`):

$$\mathbb{P}\left(-1.64 \leq \frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{SE}[\widehat{D}_n]} \leq 1.64\right) = 0.90$$

- Implies that:

$$\mathbb{P}\left(\left(\widehat{D}_n - 1.64 \times \widehat{SE}[\widehat{D}_n]\right) \leq (\mu_y - \mu_x) \leq \left(\widehat{D}_n + 1.64 \times \widehat{SE}[\widehat{D}_n]\right)\right) = 0.90$$

- 90% Confidence Interval: $\widehat{D}_n \pm 1.64 \times \widehat{SE}[\widehat{D}_n]$
- Bounds are random! Not $(\mu_y - \mu_x)$!

Different confidence intervals

- What about confidence intervals other than 90%? Say we want an $100(1-\alpha)\%$ confidence interval:

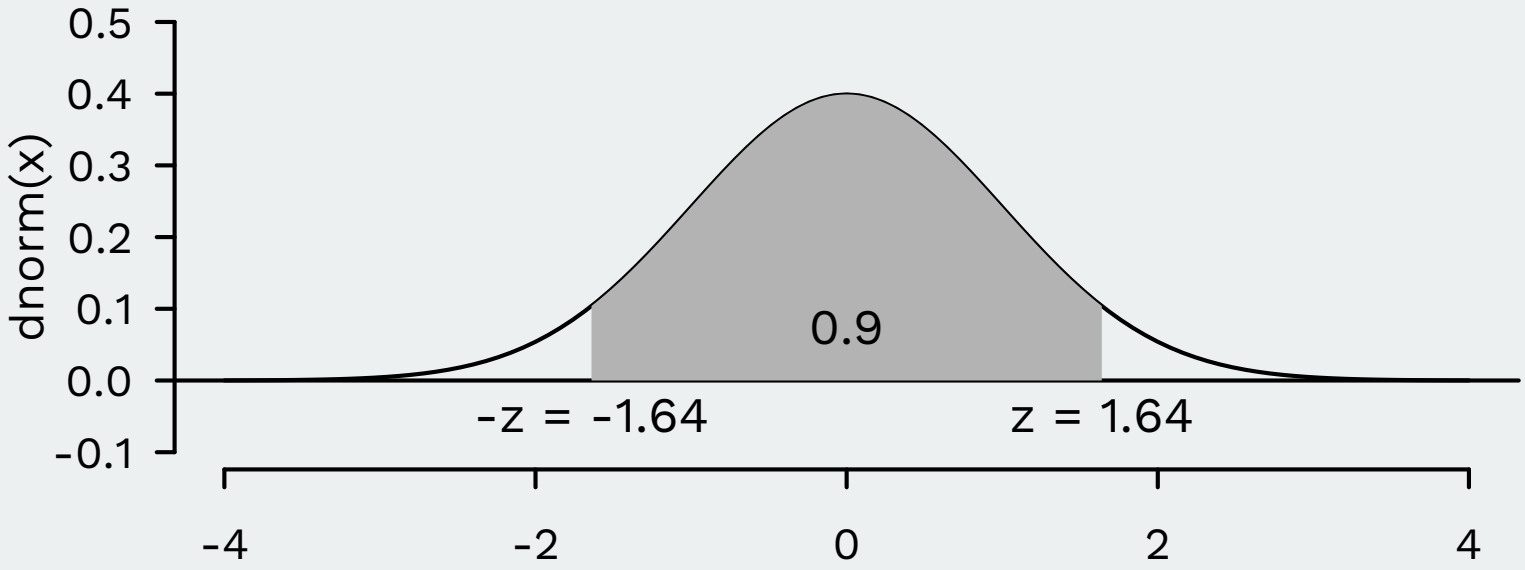
$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{SE}[\widehat{D}_n]} \leq z_{\alpha/2}\right) = (1-\alpha)$$

- General formula for a $100(1-\alpha)\%$ confidence interval:

$$\widehat{D}_n \pm z_{\alpha/2} \times \widehat{SE}[\widehat{D}_n]$$

- Here we call the $z_{\alpha/2}$ values the z-values for the particular confidence intervals.

Finding the z values



- How do we figure out what $z_{\alpha/2}$ will be? Need to find the values such that for $Z \sim N(0, 1)$:

$$\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

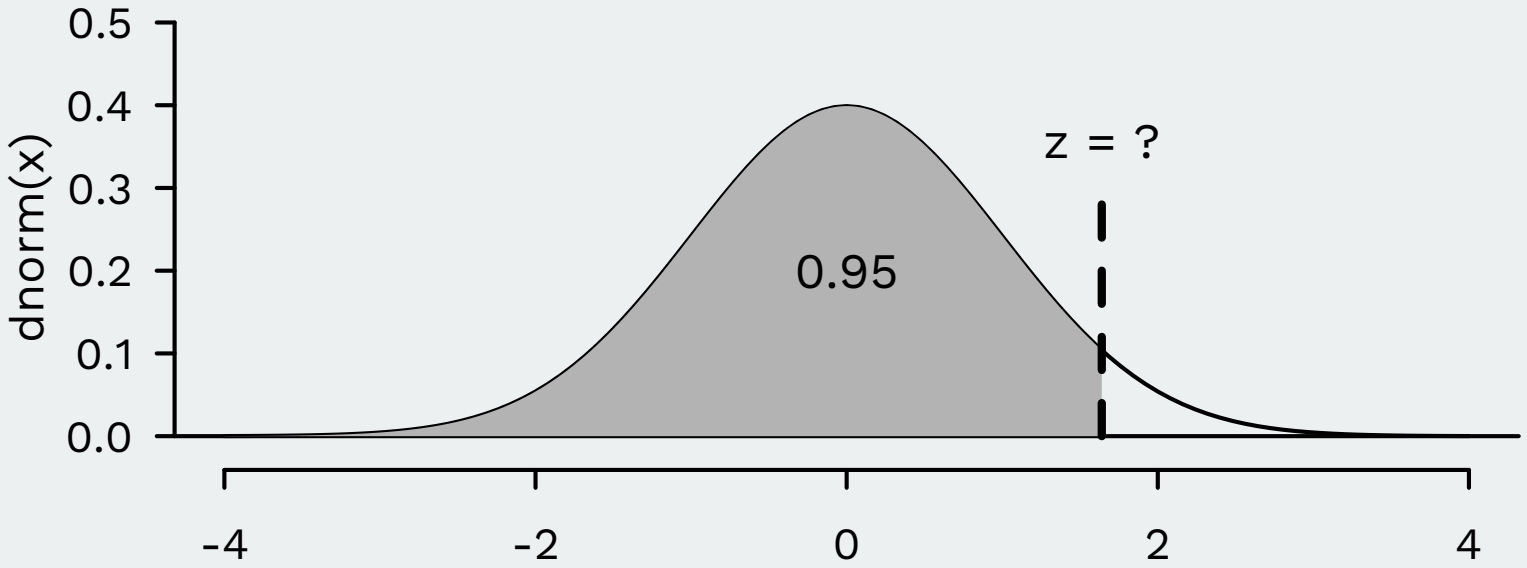
- Intuitively, we want the z values that puts $\alpha/2$ in each of the tails.
- For example, with $\alpha = 0.1$ for a 90% confidence interval, we want the z values that put 0.05 (5%) in each of the tails.

Putting it in the tails

- How to get the z values? Put α probability in the tails:

$$\begin{aligned}
\mathbb{P}(\{Z < -z_{\alpha/2}\} \cup \{Z > z_{\alpha/2}\}) &= \alpha \\
\mathbb{P}(Z < -z_{\alpha/2}) + \mathbb{P}(Z > z_{\alpha/2}) &= \alpha && \text{(additivity)} \\
2 \times \mathbb{P}(Z > z_{\alpha/2}) &= \alpha && \text{(symmetry)} \\
\mathbb{P}(Z < z_{\alpha/2}) &= 1 - \alpha/2
\end{aligned}$$

- Find the z-value that puts probability $1 - \alpha/2$ below it:



Calculating z-values in R

- Inverse of the CDF of the standard Normal evaluated at $1 - \alpha/2$!
- We can find this using the `qnorm()` function in R.
- Procedure for a 95% confidence interval:
  1. Choose a value α (0.05 for example) for a $100(1-\alpha)\%$ confidence interval (95% in this case)
  2. Convert this to $1 - \alpha/2$ (0.975 in this case)
  3. Plug this value into `qnorm()` to find $z_{\alpha/2}$:

```
qnorm(0.975)
```

```
## [1] 1.96
```

- 95% CI: $\widehat{D}_n \pm 1.96 \times \widehat{SE}[\widehat{D}_n]$

Question

- **Question** What happens to the size of the confidence interval when we increase our confidence, from say 95% to 99%? Do confidence intervals get wider or shorter?
- **Answer** Wider!
- Decreases $\alpha \rightsquigarrow$ increases $1 - \alpha/2 \rightsquigarrow$ increases $z_{\alpha/2}$

Interpreting the confidence interval

- **Caution!** An often recited, but **incorrect** interpretation of a confidence interval is the following:
  - "I calculated a 95% confidence interval of [0.05,0.13], which means that there is a 95% chance that the true difference in means in is that interval."
  - This is WRONG.
- The true value of the population difference in means, $\mu_y - \mu_x$, is **fixed**.
  - It is either in the interval or it isn't—there's no room for probability at all.
- The randomness is in the interval: $\widehat{D}_n \pm 1.64 \times \widehat{SE}[\widehat{D}_n]$. This is what varies from sample to sample.
- Correct interpretation: **across 95% of random samples, the constructed confidence interval will contain the true value.**

Confidence interval simulation

- Draw samples of size 500 (pretty big) from $N(1, 10)$
- Calculate confidence intervals for the sample mean:

$$\overline{Y}_n \pm 1.96 \times \widehat{SE}[\overline{Y}_n] \rightsquigarrow \overline{Y}_n \pm 1.96 \times S_n/\sqrt{n}$$

```
set.seed(2143)
sims <- 10000
cover <- rep(0, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for (i in 1:sims) {
    draws <- rnorm(500, mean = 1, sd = sqrt(10))
    low.bound[i] <- mean(draws) - sd(draws)/sqrt(500) *
        1.96
    up.bound[i] <- mean(draws) + sd(draws)/sqrt(500) *
        1.96
    if (low.bound[i] < 1 & up.bound[i] > 1) {
        cover[i] <- 1
    }
}
mean(cover)
```

```
## [1] 0.95
```

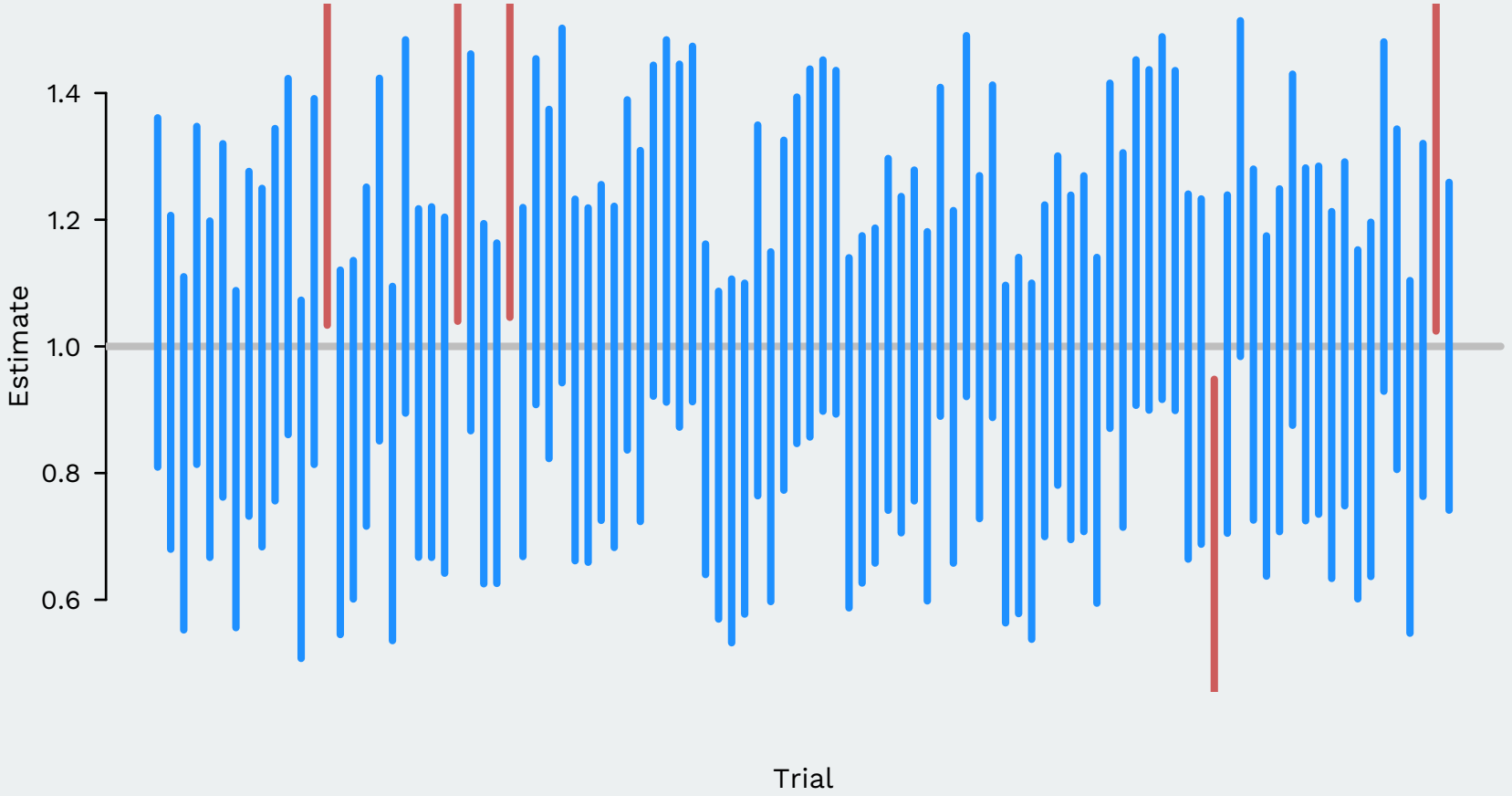
Plotting the CIs

Plotting the CIs

Plotting the CIs

Plotting the CIs

Plotting the CIs



- You can see that in these 100 samples, exactly 95 of the calculated confidence intervals contains the true value.

Understanding check

- **Question** What happens to the intervals when we increase n, the sample size? Do more or fewer of them contain the true value?

2/ Confidence Intervals Example

Gerber, Green, and Larimer experiment

- Let's go back to the Gerber, Green, and Larimer experiment from last week. Here are the results of their experiment:

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Let's use what we have learned up until now and the information in the table to calculate a 95% confidence interval for the difference in mean turnout between the Neighbors group and the Civic Duty group.

Interval estimation of the population proportion

- Note that Y_i and X_i are Bernoulli with probability μ_y and μ_x
  - Different sample sizes: n_y and n_x
- μ_y is the **population proportion** of turnout in the Neighbors group.
- $\overline{Y}_{n_y}$ is the **sample proportion** (and sample mean).
- Bernoulli party trick: $\sigma_y^2 = \mathbb{V}[Y_i] = \mu_y(1-\mu_y)$.
- $\rightsquigarrow$ sampling variance is a function of the mean:

$$\mathbb{V}[\overline{Y}_{n_y}] = \frac{\mu_y(1-\mu_y)}{n_y}$$

- $\rightsquigarrow$ easy estimate of the SE:

$$\widehat{SE}[\overline{Y}_{n_y}] = \sqrt{\frac{\overline{Y}_{n_y}(1-\overline{Y}_{n_y})}{n_y}}$$

Calculating the CI for social pressure effect

- Putting this together, we have that the estimated SE for $\widehat{D}$ is:

$$\widehat{SE}[\widehat{D}] = \sqrt{\widehat{SE}^2[\overline{Y}_{n_y}] + \widehat{SE}^2[\overline{X}_{n_x}]}$$

$$= \sqrt{\frac{\overline{Y}_{n_y}(1-\overline{Y}_{n_y})}{n_y} + \frac{\overline{X}_{n_x}(1-\overline{X}_{n_x})}{n_x}}$$

- Apply usual formula to get 95% confidence interval:

$$\widehat{D} \pm 1.96 \times \widehat{SE}[\widehat{D}]$$

Calculating the CI for social pressure effect (II)

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

		Experimental Group			
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

```
n.n <- 38201
samp.var.n <- (0.378 * (1 - 0.378))/n.n
n.c <- 38218
samp.var.c <- (0.315 * (1 - 0.315))/n.c
se.diff <- sqrt(samp.var.n + samp.var.c)
## lower bound
(0.378 - 0.315) - 1.96 * se.diff
```

```
## [1] 0.0563
```

```
## upper bound
(0.378 - 0.315) + 1.96 * se.diff
```

```
## [1] 0.0697
```

3/ Hypothesis Tests

What is a hypothesis test?

- A hypothesis test is just an evaluation of a particular hypothesis about the population distribution.
- These are just **statistical thought experiments**:
  - Assume we know the true DGP
  - See how likely the observed data is under that assumed DGP
- Key question: how unusual would our data be in this statistical thought experiment?
- We will "reject" the assumed DGP if the data is too unusual under it.
- This is exactly like the Lady Tasting Tea.

What is a hypothesis?

- **Definition** A **hypothesis** is just a statement about population parameters.
- We might have hypotheses about causal inferences
  - Does social pressure induce higher voter turnout? (mean turnout higher in social pressure group compared to Civic Duty group?)
  - Do daughters cause politicians to be more liberal on women's issues? (voting behavior different among members of Congress with daughters?)
  - Do treaties constrain countries? (behavior different among treaty signers?)
- We might also have hypotheses about other parameters:
  - Is the share of Hillary Clinton supporters more than 50%?
  - Are traits of treatment and control groups different?
  - Is there evidence of electoral manipulation (Iran HW question)?

Null and alternative hypotheses

- **Defintion** The **null hypothesis** is a proposed, conservative value for a population parameter.
  - This is usually "no effect/difference/relationship."
  - We denote this hypothesis as $H_0 : \mu = a$.
- **Definition** The **alternative hypothesis** for a given null hypothesis is the research claim we are interested in supporting.
  - Usually, "there is a relationship/difference/effect."
  - We denote this as $H_a : \mu \neq a$.
- Always mutually exclusive

Null and alternative examples

- Causal inference: Does social pressure (Neighbors) mailer affect turnout?
  - H_0: Social pressure doesn't affect turnout ($\mu_y - \mu_x = 0$)
  - H_a: Social pressure affects turnout ($\mu_y - \mu_x \neq 0$)
- Parameters: presence of electoral manipulation?
  - H_0: no electoral manipulation (7s in the last digit occur with probability $1/10$)
  - H_a: electoral manipulation (proportion of 7s higher than $1/10$)

General framework

- A **hypothesis test** chooses whether or not to reject the null hypothesis based on the data we observe.
- **Definition** We base these rules on a **test statistic**, $T(Y) = T(Y_1, \dots, Y_n)$.
  - Will help us adjudicate between the null and the alternative.
  - Typically: larger values of $T(Y) \rightsquigarrow$ null less plausible.
  - A test statistic is a r.v.
- **Definition** The **null/reference distribution** is the distribution of $T(Y)$ under the null.
  - We'll write as $\mathbb{P}_0(T(Y) > t)$.
- **Definition** The **critical value** is the value that determines our rejection of the null:
  - $T(Y) > c \rightsquigarrow$ we reject the null hypothesis
  - $T(Y) < c \rightsquigarrow$ we retain (or fail to reject) the null hypothesis.

Lady Tasting Tea redux

- Your advisor tastes n cups of tea and either guesses the ith cup correctly $Y_i = 1$ or incorrectly, $Y_i = 0$.
- Parameter of interest: π = probability of a correct guess.
  - $\pi = 0.5$ guessing at random
  - $\pi > 0.5$ able to discern tea-first vs milk-first
  - $\pi < 0.5$ worse than random
- Null and alternative hypotheses:
  - $H_0 : \pi = 0.5$
  - $H_a : \pi > 0.5$
- Assume she tastes LOTS of cups, so n is large, so that we have

$$\overline{Y}_n \sim N(\pi, S_n^2/n)$$

Type I and Type II errors

- **Definition** A **Type I** error is when we reject the null hypothesis when it is in fact true.
  - We say that the Lady is discerning when she is just guessing.
  - A false discovery.
- **Definition** A **Type II** error is when we fail to reject the null hypothesis when it is false.
  - We say that the Lady is just guessing when she is truly discerning.
  - An undetected finding.

Size, power, and errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Defintion** The **level/size of the test**, or α, is the probability of a Type I error.
- Choose a level α based on aversion to false discovery:
  - Convention in social sciences is $\alpha = 0.05$, but nothing magical there
  - Particle physicists at CERN use $\alpha \approx \frac{1}{1{,}750{,}000}$
  - Lower values of α guard against "flukes" but increase barriers to discovery
- **Definition** The **power** of a test is the probability that a test correctly rejects the null.
  - Power $= 1 - \mathbb{P}(\text{Type II error})$
  - Better tests $=$ higher power.

Hypothesis testing procedure

1. Choose null and alternative hypotheses
2. Choose a test statistic, $T(Y)$
3. Choose a level, α
4. Find the critical value, c, which is $\mathbb{P}_0(T(Y) > c) = \alpha$
5. Reject if $T(Y) > c$, fail to reject otherwise

Deriving the test statistic

- Null hypothesis: $H_0 : \pi = 0.5$
- Alternative hypothesis: $H_a : \pi > 0.5$
- Test statistic should be big/unusual under the null when the null is false:

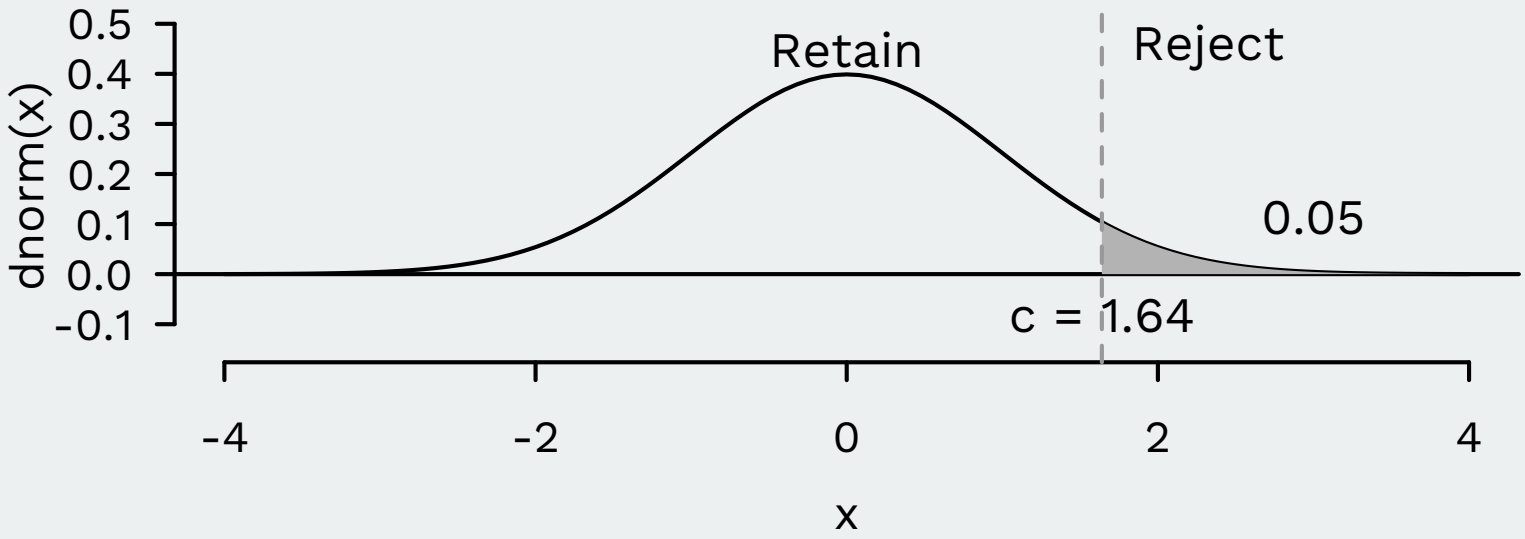
$$T(Y) = \frac{\overline{Y}_n - 0.5}{S_n/\sqrt{n}}$$

- If she can discern, then $\overline{Y}_n$ will be high and thus $T(Y)$ will be high.
  - $\rightsquigarrow$ null hypothesis less plausible
- Note that in "null world" (null is true), then we know the distribution of $T(Y)$:

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

The critical value

- **Definition** The **critical value**, c is the value that determines rejection of the null.
  - How large should $T(Y)$ before we reject?
- Let's say that we were happy with $\alpha = 0.05$ in this case.
- We would need to find the value, c: $\mathbb{P}_0(T(Y) > c) = 0.05$. Since we know this is normal, we can plot it:



The critical value and the rejection region (II)

- If we want $\mathbb{P}_0(T(Y) > c) = \alpha$, need to find c so that $\mathbb{P}_0(T(Y) \leq c) = 1 - \alpha$.
- As with CIs, use the inverse of the CDF:

```
qnorm(0.95)
```

```
## [1] 1.64
```

p-values

- **Definition** The **p-value** for an observed test statistic is the probability of seeing a test statistic at least as extreme as observed statistic, **given that the null hypothesis is correct**.
  - If we have $T(Y) = t_{\text{obs}}$ in our sample, then p-value is $\mathbb{P}_0(T(Y) > t_{\text{obs}})$.
- The smallest value of α that we could still reject the null.
- What p-values are **not**:
  - An indication of a large substantive effect
  - The probability that the null hypothesis is false
  - The probability that the alternative hypothesis is true
- Can use `pnorm()` in many cases with $T(Y) \sim N(0, 1)$ under the null.

One-sided tests

- **Definition** A **one-sided test** is a test of an alternative hypothesis that only goes in one direction.
  - In the Lady Tasting Tea (redux) example, we are interested in a one-sided test.
- Only deviations from the null hypothesis in one direction cast doubt on the null hypothesis.
  - $T(Y)$ only rejected when it was big and so $\pi > 0.5$
- A **two-sided test** allows for evidence against the null in either direction.
  - Null: $H_0 : \pi = 0.5$, Alternative: $H_a : \pi \neq 0.5$
  - Test statistic needs to be large when null is false $\rightsquigarrow$ use absolute value: $|T(Y)|$.
- Two-sided tests are almost always the right thing to do, since one-sided easier to reject at the same α

Connection between confidence intervals and hypothesis tests

- **CI/Test duality:** A $100(1-\alpha)\%$ confidence interval represents all null hypotheses that we would not reject with a α-level test.
- Example:
  - Construct a 95% CI (a, b) for $\mu_y - \mu_x$.
  - If $0 \in (a, b) \rightsquigarrow$ cannot reject $H_0 : \mu_y - \mu_x = 0$ at $\alpha = 0.05$
  - If $0 \notin (a, b) \rightsquigarrow$ reject $H_0 : \mu_y - \mu_x = 0$ at $\alpha = 0.05$
- Two-sided CIs $\iff$ two-sided tests.

4/ Hypothesis Testing Example

Example of a two-sided test: Social pressure experiment

- **Step 1**: Write down null and alternative hypotheses:
  - $H_0 : \mu_y - \mu_x = 0$ (no effect/difference in turnout)
  - $H_a : \mu_y - \mu_x \neq 0$ (some effect/difference)
- **Step 2**: choose test statistic, using the difference in means $\widehat{D} = \overline{Y}_{n_y} - \overline{X}_{n_x}$:

$$T = \frac{\widehat{D} - 0}{\widehat{SE}[\widehat{D}]}$$

- We want a two-sided alternative, so we'll use $T^* = |T|$ as our test statistic.
- **Step 3**: We'll set $\alpha = 0.05$ to be standard.

Social pressure test, continued

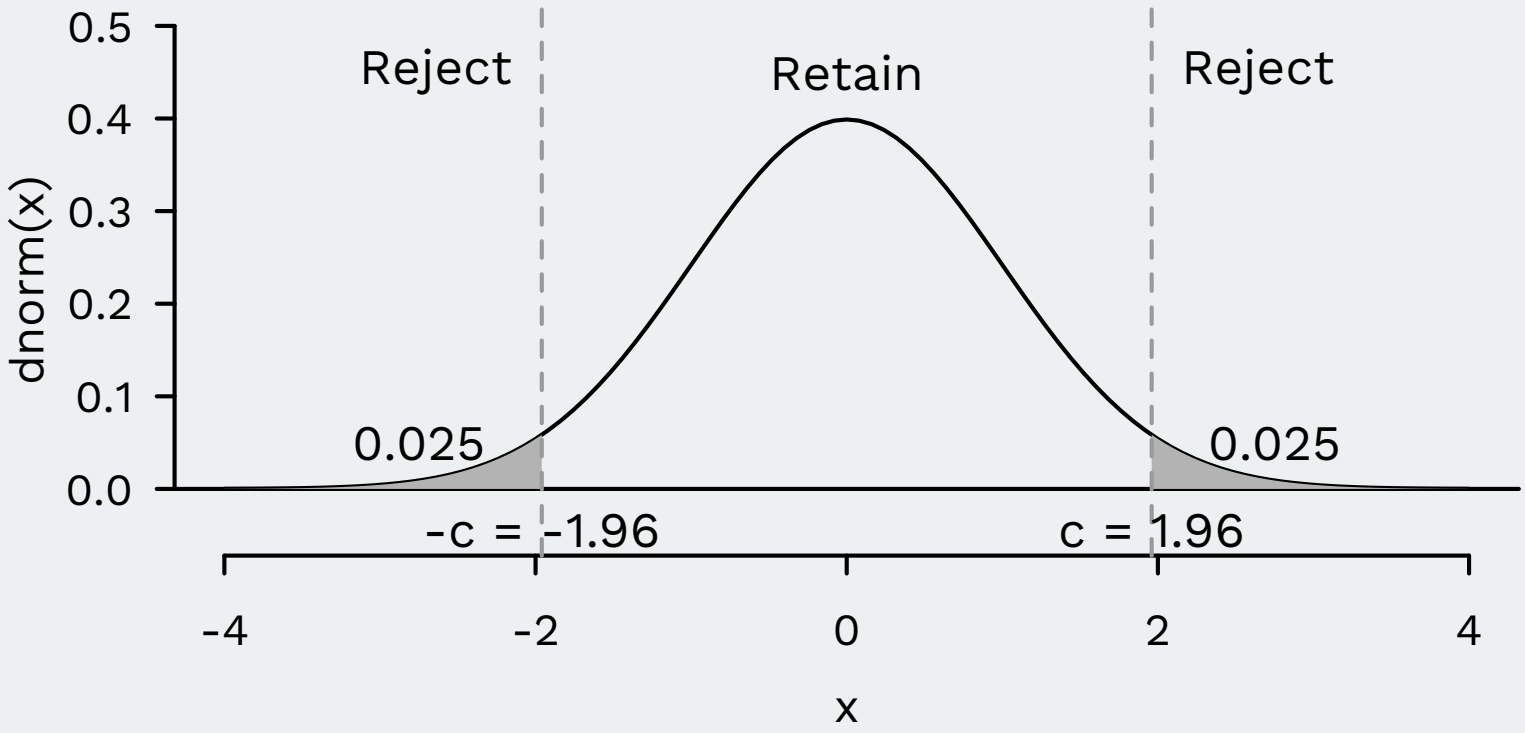
- **Step 4**: Calculate the critical value.
- Note that under the null, $T \sim N(0,1)$, and we can use this to figure out distribution of T^*
- What's the critical value? We need c such that $\mathbb{P}_0(T^* > c) = \alpha$.

$$\begin{aligned}
\alpha &= \mathbb{P}_0(T^* > c) \\
&= \mathbb{P}_0(|T| > c) \\
&= \mathbb{P}_0(\{T > c\} \cup \{T < -c\}) \\
&= \mathbb{P}_0(T > c) + \mathbb{P}_0(T < -c) \\
&= 2 \times \mathbb{P}_0(T > c)
\end{aligned}$$

- Implies we need to find the value such that $\mathbb{P}_0(T > c) = \alpha/2$

The two-sided test critical values

- Implies we need to find the value such that puts $\alpha/2$ in each of the tails.



- Again, CDF evaluated at $1 - \alpha/2$

Social pressure test, continued

- $1 - \alpha/2 = 0.975 \rightsquigarrow c =$ `qnorm(0.975)` $= 1.96$
- Earlier we got an estimate of the standard error:

```
n.n <- 38201
samp.var.n <- (0.378 * (1 - 0.378))/n.n
n.c <- 38218
samp.var.c <- (0.315 * (1 - 0.315))/n.c
se.diff <- sqrt(samp.var.n + samp.var.c)
se.diff
```

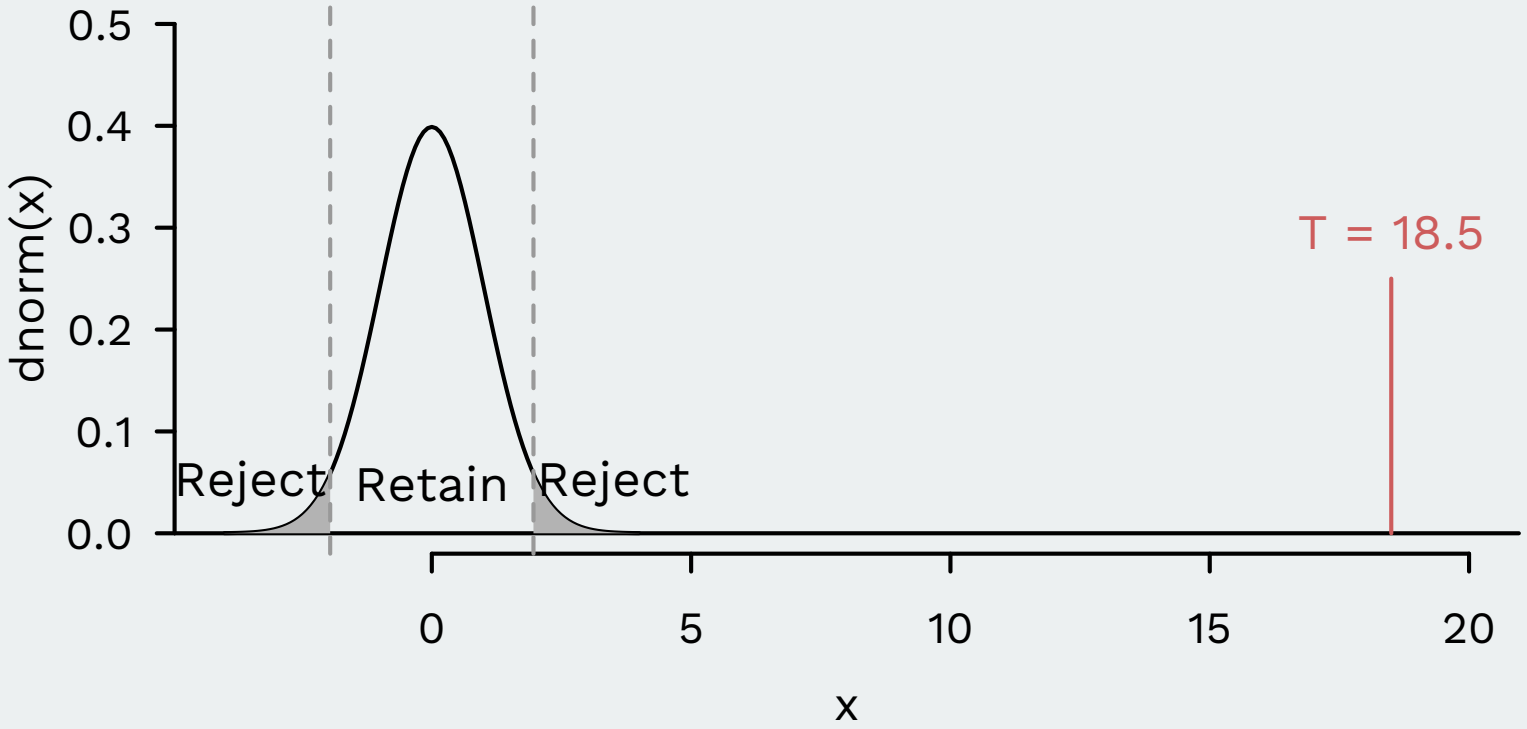
```
## [1] 0.00344
```

```
## Calcuate test statistic
(0.378 - 0.315)/se.diff
```

```
## [1] 18.3
```

Perform the test

- **Step 5**: Is 18.5 large relative to the null distribution? Is it bigger than c?



- Obviously an easy reject!

Calculate the p-value

- How unlikely would it be to get a difference this extreme or more extreme (either above or below 0)? That is, what is the p-value?
- Remember two-tailed alternative:

$$\begin{aligned}\mathbb{P}_0(T^* > 18.5) &= \mathbb{P}_0(|T| > 18.5) \\ &= \mathbb{P}_0(T > 18.5) + \mathbb{P}_0(T < -18.5) \\ &= 2 \times \mathbb{P}_0(T < -18.5)\end{aligned}$$

- Use the `pnorm()` function:

```
2 * pnorm(-18.5)
```

```
## [1] 2.06e-76
```

- Likely?

5/ Small-Sample Problems

Small sample complications

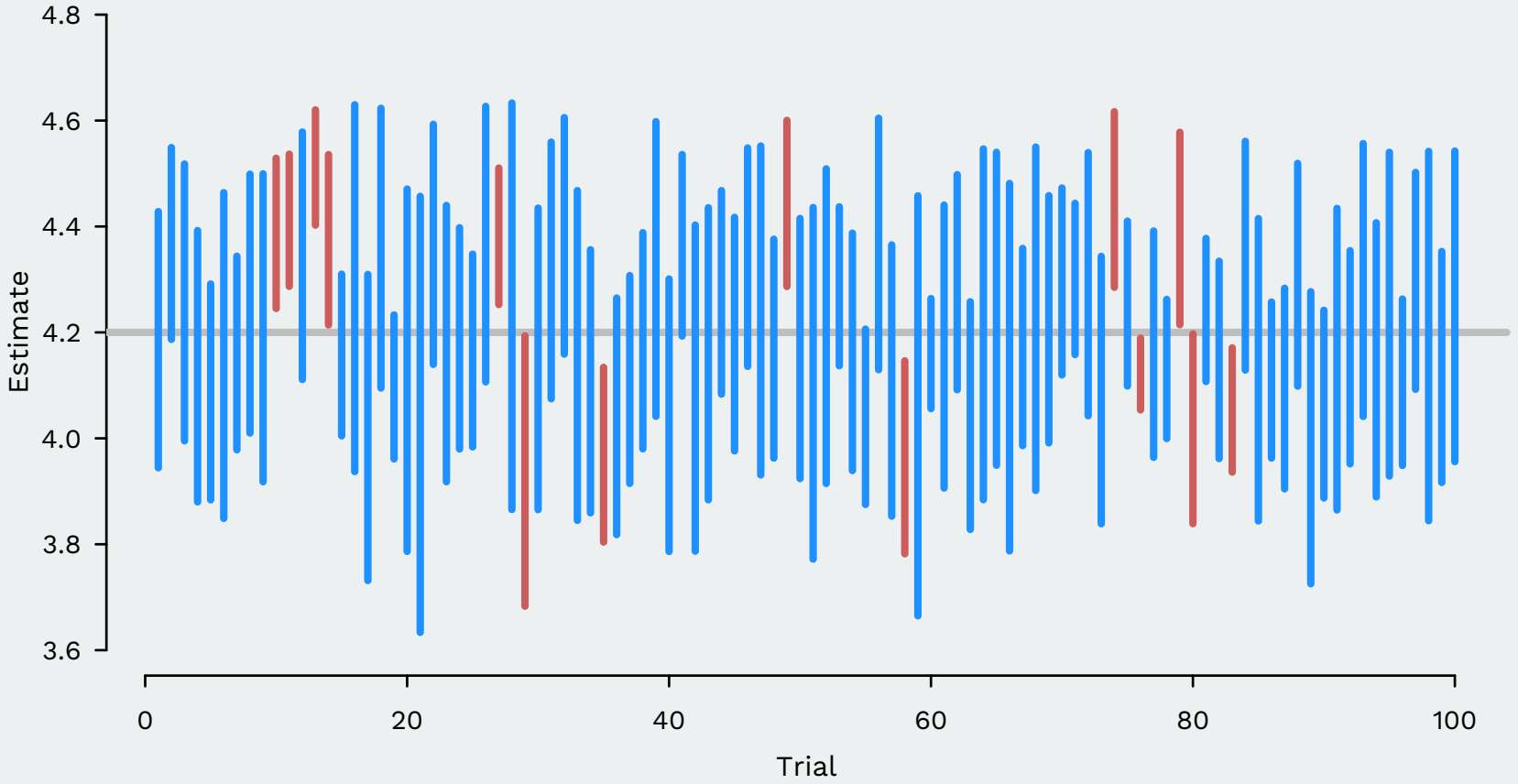
- What if n is not large?
  - CLT doesn't hold
  - $\widehat{D}_n$ not approximately normal.
- To take a somewhat canonical example, let's say that we are trying to estimate the percent alcohol by volume of Guinness beer
- A can of Guinness has alcohol distributed $N(4.2, 0.09)$.
- Imagine that we could only take samples of six packs, so we can only use $n = 6$.
- What happens when we apply the usual, normal-based CI formula to that data?

If you're having small-sample problems, I feel bad for you son...

```
set.seed(2143)
sims <- 10000
cover <- rep(NA, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for (i in 1:sims) {
    draws <- rnorm(6, mean = 4.2, sd = sqrt(0.09))
    low.bound[i] <- mean(draws) - sd(draws)/sqrt(6) *
        1.96
    up.bound[i] <- mean(draws) + sd(draws)/sqrt(6) *
        1.96
    if (low.bound[i] < 4.2 & up.bound[i] > 4.2) {
        cover[i] <- 1
    } else {
        cover[i] <- 0
    }
}
mean(cover)
```

```
## [1] 0.89
```

Plotting the CIs



- Obviously, we can see here that far fewer than 95 of the 100 confidence intervals contains the true value. We call this **undercoverage**.

Small samples break us out of asymptopia

- But what is the distribution of $\overline{Y}_n$ in small samples?

$$\overline{Y}_n \sim ?(\mu, \sigma^2/n)$$

- If we're given no other information we actually don't know.
- Can make progress if $Y_1, \ldots, Y_n$ are i.i.d. samples from $N(\mu, \sigma^2)$, then:

$$\frac{\overline{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- Have to estimate σ, which changes the sampling distribution in small samples:

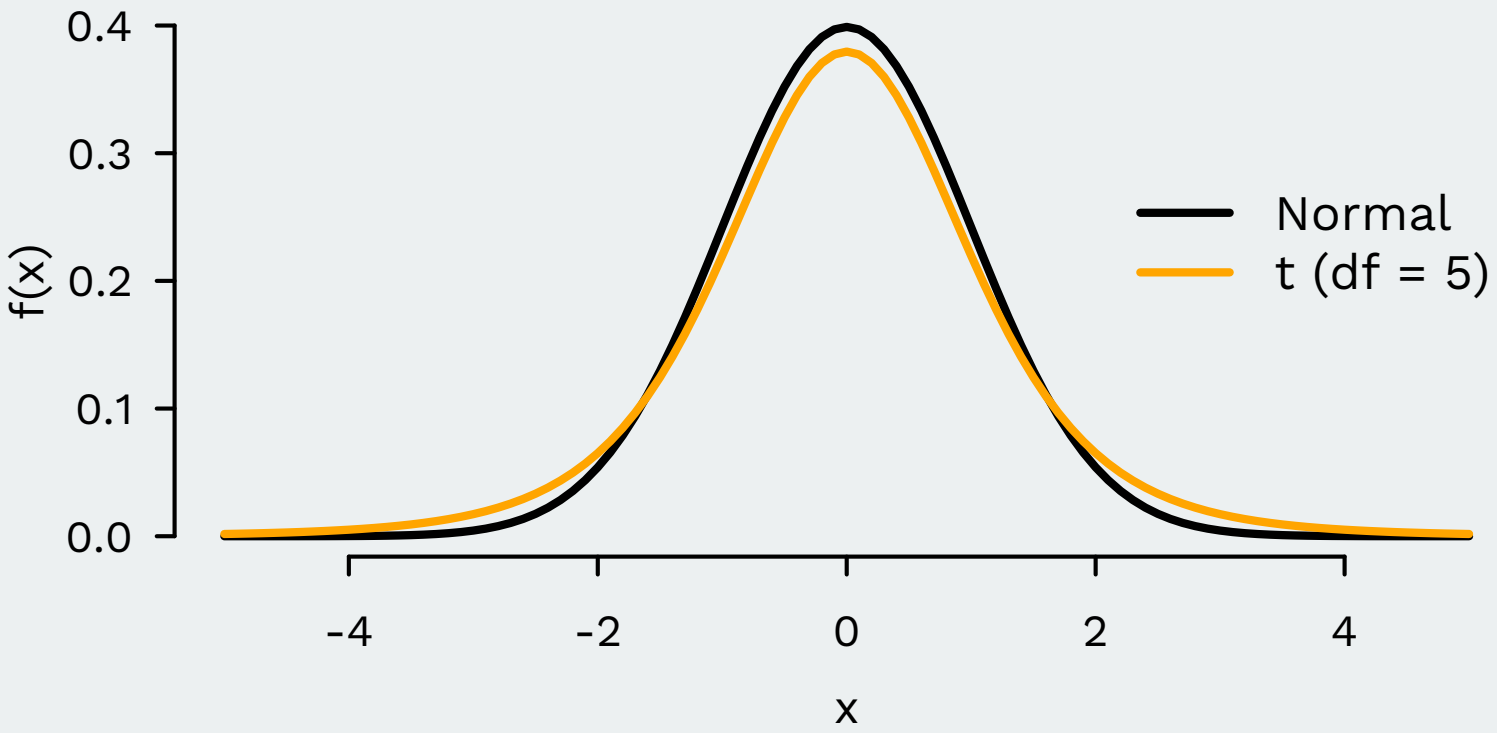
$$\frac{\overline{Y}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t_{n-1}$$

Student's t distribution

- Here, t_{n-1} is the **Student's t-distribution** (usually just called the t distribution) with $n-1$ degrees of freedom (df).
  - Family of distributions with parameter df.
- Named after **William Sealy Gossett** who published under the pen name, Student, while he was an employee at Guinness Brewery in Ireland. He developed the distribution while working on quality control problems in the brewery, where sample sizes could be quite small.

The shape of the t

- The t distribution is completely summarized by its degrees of freedom, which here is dictated by the sample size.
- As sample sizes increase, tends toward the $N(0, 1)$
- Similar shape to the Normal, but with fatter tails.
- You can think of this extra variance as coming from the extra variance of estimating SE.



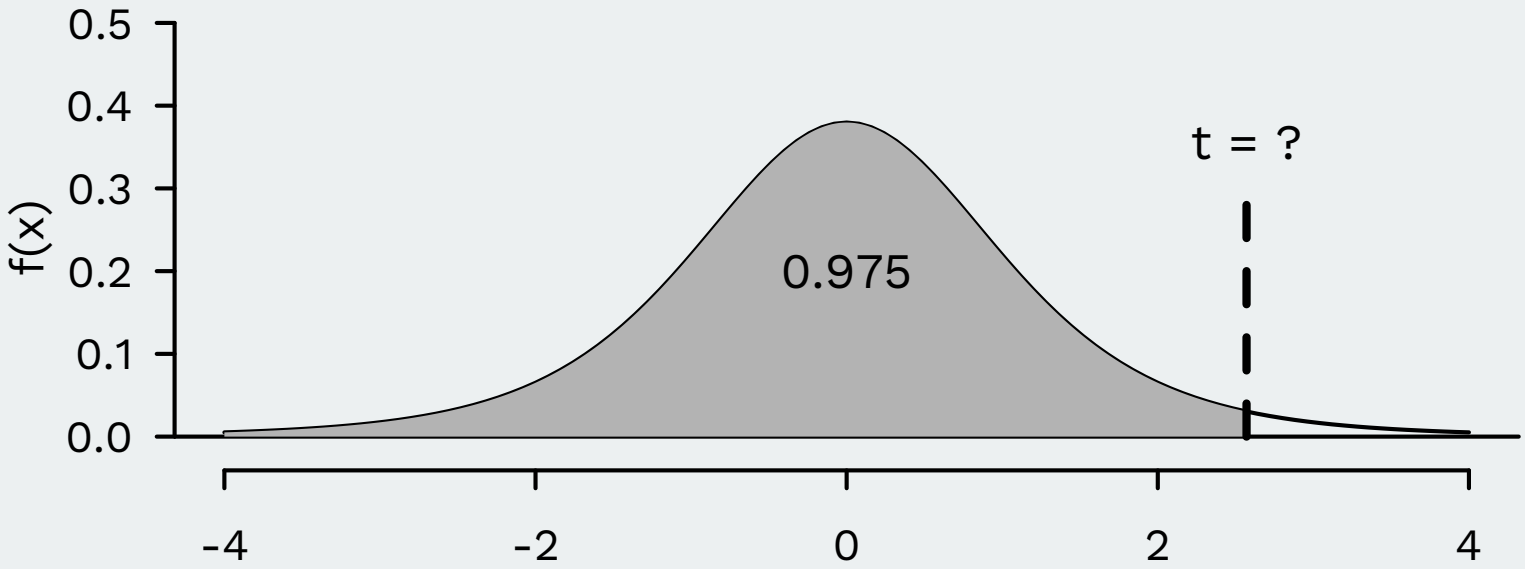
Using the t for small sample confidence intervals

- We need to figure this out:

$$\mathbb{P}\left(-? \leq \frac{\overline{Y}_n - \mu}{\widehat{SE}[\overline{Y}_n]} \leq ?\right) = (1-\alpha)$$

- Using same logic as above, find $t_{n-1,\alpha/2}$ such that

$$\mathbb{P}(T_{n-1} \leq t_{n-1,\alpha/2}) = 1 - \alpha/2$$



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

$$\mathbb{P}(T_{n-1} \leq t_{n-1,\alpha/2}) = 1 - \alpha/2$$

- Here, T_{n-1} has a t distribution with $n-1$ d.f.s
- Use the `qt()` function in R:

```
qt(0.975, df = 6 - 1)
```

```
## [1] 2.57
```

- Note that this is different the value we would get with a Normal:

```
qnorm(0.975)
```

```
## [1] 1.96
```

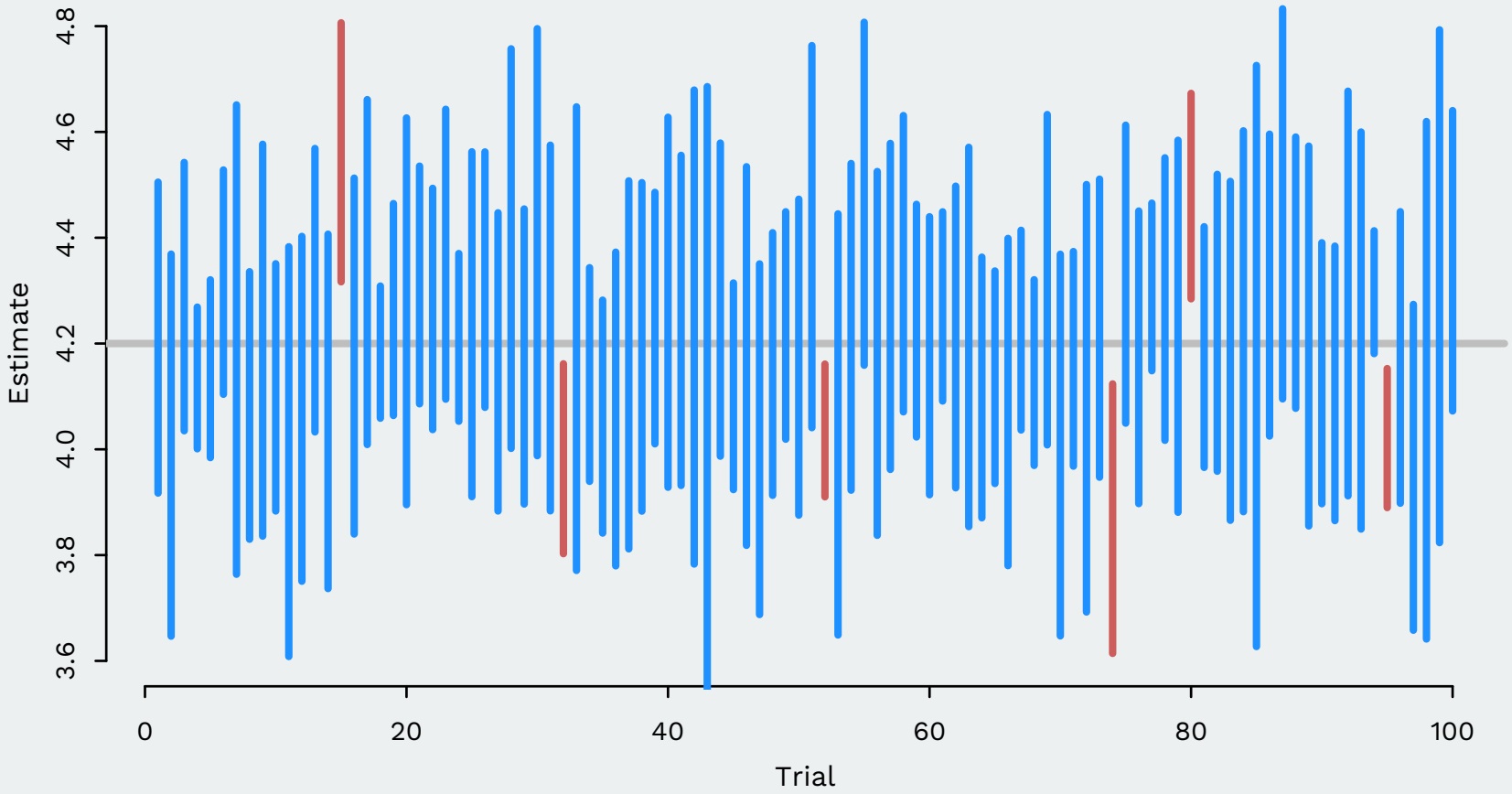
Small sample simulation redux

- Calculate the CIs using the t distribution: $\overline{Y}_n \pm t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}$

```
sims <- 10000
cover <- rep(NA, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for (i in 1:sims) {
    draws <- rnorm(6, mean = 4.2, sd = sqrt(0.09))
    tval <- qt(0.975, df = 6 - 1)
    low.bound[i] <- mean(draws) - sd(draws)/sqrt(6) *
        tval
    up.bound[i] <- mean(draws) + sd(draws)/sqrt(6) *
        tval
    if (low.bound[i] < 4.2 & up.bound[i] > 4.2) {
        cover[i] <- 1
    } else {
        cover[i] <- 0
    }
}
mean(cover)
```

```
## [1] 0.951
```

Plotting the (correct) CIs



- Here we can see that our coverage is back on track even though we have really small samples!
- Same ideas apply to calculating critical values for tests with small samples.