Gov 2000 - 6. Hypothesis Tests

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

Where are we? Where are we going?

- Last few weeks = how to produce a best estimate of some population parameter, drawing on our knowledge of probability.
- Also learned how to derive an estimated range of plausible values of the parameter in the confidence interval.
- Now: how to use our estimates to test a particular hypothesis about the data.
- We'll draw heavily on our probability knowledge from earlier in the term!

HYPOTHESIS TESTING EXAMPLES

The lady tasting tea

- Remember the setup:

> Your advisor asks you to grab a tea with milk for him before your meeting and he says that he prefers tea poured before the milk. You stop by Darwin's and ask for a tea with milk. When you bring it to your advisor, he complains that it was prepared milk-first.

- You are skeptical that he can really tell the difference, so you devise a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups to advisor in a **random** order
  - Ask advisor to pick which 4 of the 8 were milk-first.

Assuming we know the truth

- Advisor picks out all 4 milk-first cups correctly!
- Statistical thought experiment: how often would she get all 4 correct *if she were guessing randomly*?
    - Only one way to choose all 4 correct cups.
    - But 70 ways of choosing 4 cups among 8.
    - Choosing at random ≈ picking each of these 70 with equal probability.
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%.
- ⇝ the guessing at random hypothesis might be implausible.

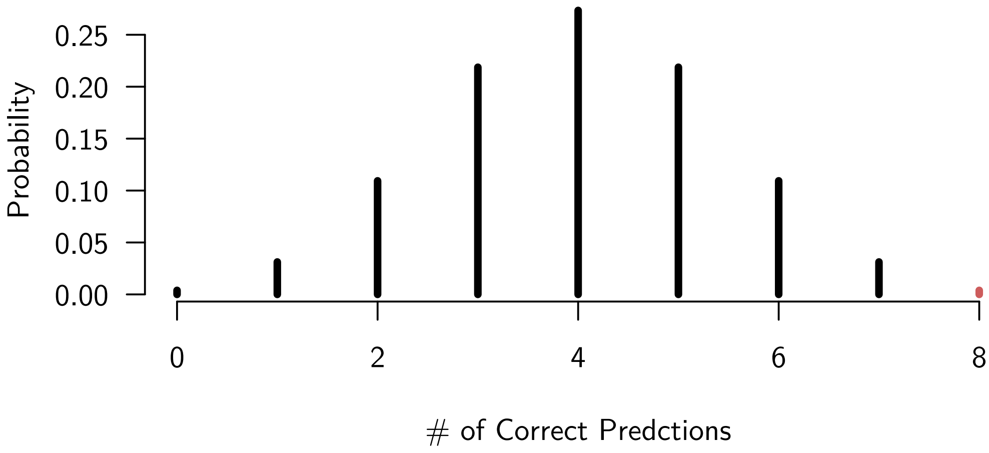
Election prediction

- Alan Lichtman (History at American U.) has predicted the winner of every presidential election all 8 elections since 1984.
    - Doesn't use any polls, just 13 true/false questions.
    - Ex: "Challenger charisma"
    - This year he's trolling liberals: predicts Trump win.
- Does he have predictive value? Does he do better than random guessing?
    - If he randomly choosing between the two candidates in each election, he'd flipping 8 coins with probability 0.5.
    - ⇝ number of correct predictions is Binomial(8, 0.5)
- What's the probability that he would do this well if he guessing at random?

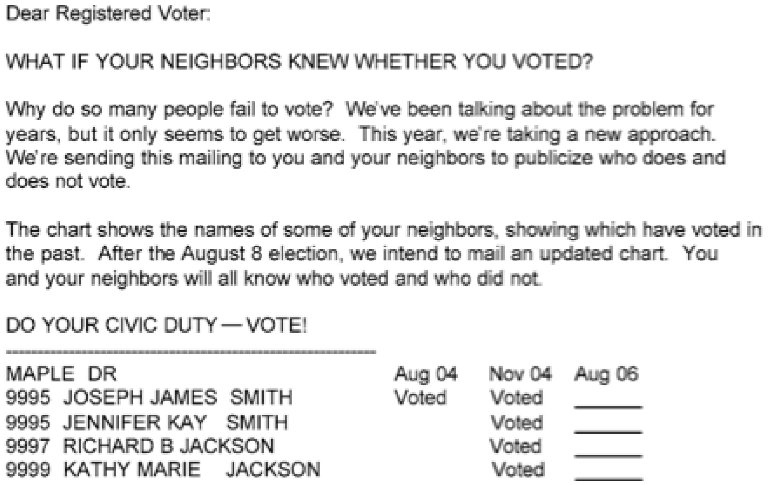
```
dbinom(x = 8, size = 8, prob = 0.5)
```

```
## [1] 0.00390625
```

```
plot(x = 0:8, y = dbinom(0:8, size = 8, prob = 0.5), type = ”h”, lwd = 4, las = 1, xlab = ”# of Correct Predctio
mtext(”Probability”, side = 2, line = 3)
```



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	______

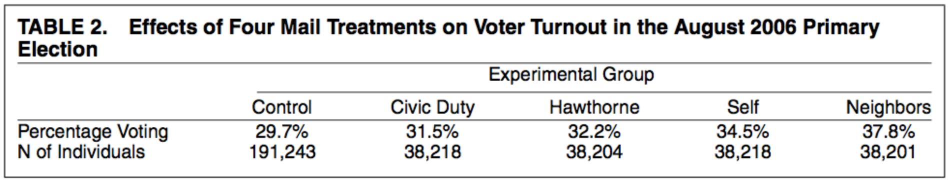


TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	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

Social pressure effect

```r
load(”../data/gerber_green_larimer.RData”)
social$voted <- 1 * (social$voted == ”Yes”)
neigh.mean <- mean(social$voted[social$treatment == ”Neighbors”])
contr.mean <- mean(social$voted[social$treatment == ”Civic Duty”])
neigh.mean - contr.mean
```

```
## [1] 0.06341057
```

- Treatment effect of 6.3410569 percentage points.
- But we know that the estimator varies from sample to sample due to random chance.
- Could this happen by random chance if there was no treatment effect at all?

Review of the difference in means

- **Treated group** $Y_1, Y_2, \ldots, Y_{n_y}$ i.i.d. with population mean μ_y and population variance σ_y^2
- **Control group** $X_1, X_2, \ldots, X_{n_x}$ i.i.d. with population mean μ_x and population variance σ_x^2
- Quantity of interest: **population differences in average turnout**: $\mathbb{E}[Y_i] - \mathbb{E}[X_i] = \mu_y - \mu_x$
- Estimator: sample difference in means: $\widehat{D}_n = \overline{Y}_{n_y} - \overline{X}_{n_x}$
- We estimated the standard error of $\widehat{D}_n$ with:

$$\widehat{\text{se}}[\widehat{D}_n] = \sqrt{\frac{S_y^2}{n_y} + \frac{S_x^2}{n_x}}$$

HYPOTHESIS TEST NOMENCLATURE

What is a hypothesis test?

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?

A **hypothesis test** is an evaluation of a particular hypothesis about the population distribution. It is a **statistical thought experiments** with a couple of steps. First, we assume that we know the true DGP or part of the true DGP. Then, we use tools of probability to see what types of data we should see under this assumption. Finally, we compare our observed data to this thought experiment. We will "reject" the assumed DGP if the data is too unusual under it. Thus, hypothesis testing is like a statistical proof by contradiction.

Null and alternative hypotheses

To perform a hypothesis test, we need to state two precise and mutually exclusive 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 $H_0 : \theta = \theta_0$.
  - H_0: Social pressure doesn't affect turnout ($H_0 : \mu_y - \mu_x = 0$)
- **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 $H_a : \theta \neq \theta_0$.
  - H_a: Social pressure affects turnout ($H_a : \mu_y - \mu_x \neq 0$)

General framework

- A **hypothesis test** chooses whether or not to reject the null hypothesis based on the data we observe.

- Rejection based on a **test statistic,** $T_n = T(Y_1, \dots, Y_n)$. This statistic will help us adjudicate between the null and the alternative. Typically, it will be the case that larger values of T_n imply that the null hypothesis is less plausible. Note that the

- The **null/reference distribution** is the distribution of T under the null. This is the key part of the statistical thought experiment. Once we assume the DGP (that is, assume that the null hypothesis is true), we will be able to figure out this null distribution. And we'll use this to assess how likely different values of T_n are under the null. We'll write its probabilities as $\mathbb{P}_0(T_n \leq t)$.

- By the CLT, we know that the standardized difference in means has a standard normal distribution in large samples:

$$T_n = \frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{\text{se}}[\widehat{D}]} \xrightarrow{d} N(0,1)$$

- Under the null hypothesis of $H_0 : \mu_y - \mu_x = 0$, then we have

$$T_n = \frac{\widehat{D}_n}{\widehat{\text{se}}[\widehat{D}_n]} \xrightarrow{d} N(0,1)$$

- If T_n is very far from 0 $\rightsquigarrow$ large sample diff-in-means $\rightsquigarrow$ no population diff-in-means is not plausible.

Rejection regions

- **Definition** The **rejection region**, R, contains the values of T_n for which we reject the null. These are the areas that indicate that there is evidence against the null.

- With a two-sided alternative ($H_0 : \mu_y - \mu_x = 0$ vs $H_a : \mu_y - \mu_x \neq 0$), the rejection region will intuitively be when T_n is much bigger than 0 or much smaller than 0. Both of these are unlikely under the null of no effect and so are evidence against the null. Thus, the rejection regions for two-sided alternatives will be $|T_n| > c$ for some value c.

- We determine these rejection regions by attempting to control the probability of making mistakes in our tests. There are two types of mistakes we might make.

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

Definition 1. Type I errors 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 (very bad, thus type I).

Definition 2. Type II errors A **Type II** error is when we fail to reject the null hypothesis when it is false.

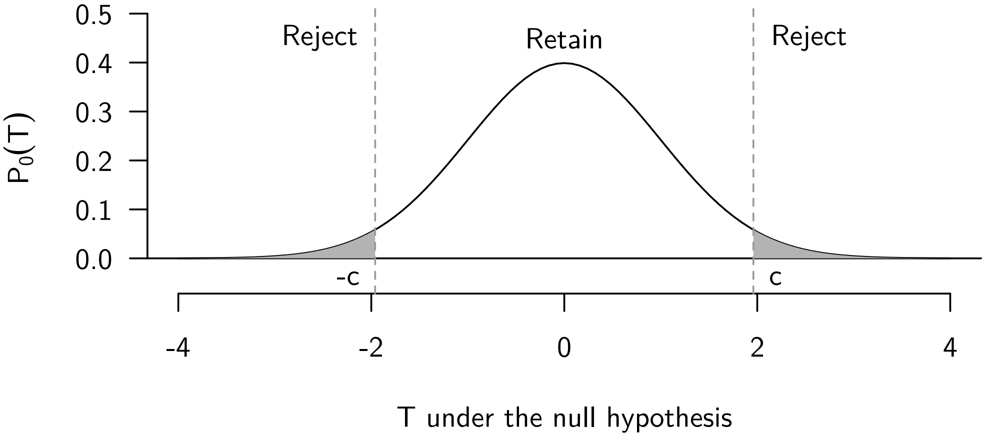
- We say that the Lady is just guessing when she is truly discerning. An undetected finding (not as bad, thus type II).
- **Defintion** The **level/size of the test**, or α, is the probability of a Type I error. With two-sided alternative, we reject when $|T_n| > c$, which implies that the size of test then is: $\mathbb{P}_0(|T_n| > c) = \alpha$
- Choose a level α based on aversion to false discovery. The convention in social sciences is $\alpha = 0.05$, but nothing magical there. For instance, particle physicists at CERN use $\alpha \approx \frac{1}{1{,}750{,}000}$. The key tradeoff here is that lower values of α guard against "flukes" but increase barriers to discovery.

CONDUCTING HYPOTHESIS TESTS

Hypothesis testing procedure

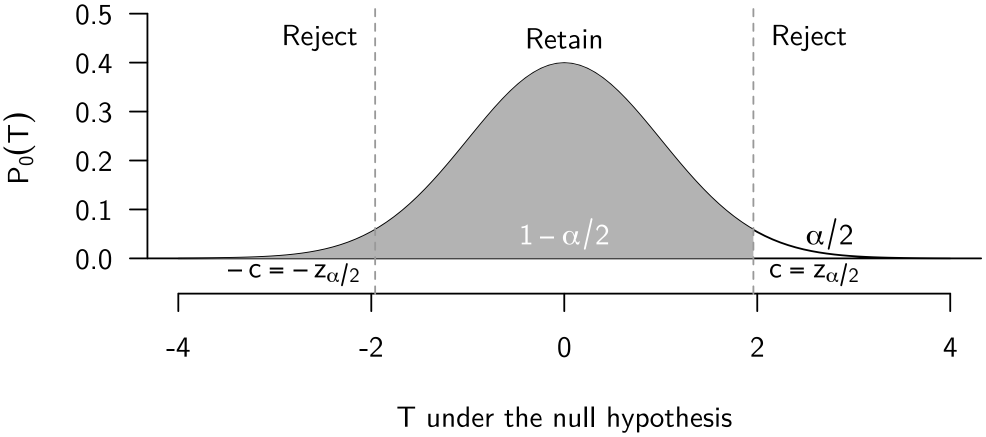
1. Choose null and alternative hypotheses
2. Choose a test statistic, T_n
3. Choose a level, α
4. Determine rejection region
5. Reject if T_n in rejection region, fail to reject otherwise

Rejection region



- What's the rejection region $|T_n| > c$ if $\alpha = 0.05$?
- *Under the null hypothesis of no effect*, we want T_n to be in the rejection region only 5% of the time.
  - $\rightsquigarrow$ false rejection of the null only 5% of the time.
  - Can find c based on the null distribution being $\approx$ standard normal!

Determining the rejection region



- Find $z_{\alpha/2}$ such that

$$\mathbb{P}_0(T_n < -z_{\alpha/2}) = \mathbb{P}_0(T_n > z_{\alpha/2}) = \alpha/2$$

- $\rightsquigarrow$ find quantile $\mathbb{P}_0(T_n < z_{\alpha/2}) = 1 - \alpha/2$

  - if $\alpha = 0.05 \rightsquigarrow z_{\alpha/2} =$ `qnorm(1-0.05/2) = ` 1.959964

Final hypothesis test

1. Hypotheses: $H_0 : \mu_y - \mu_x = 0$ vs. $H_a : \mu_y - \mu_x \neq 0$
2. Test statistic: $T_n = \widehat{D}_n / \widehat{\text{se}}[\widehat{D}_n]$
3. Use $\alpha = 0.05$
4. Rejection region is $|T_n| > 1.96$.

Social pressure test

- Calculate test statistic for social pressure mailers:

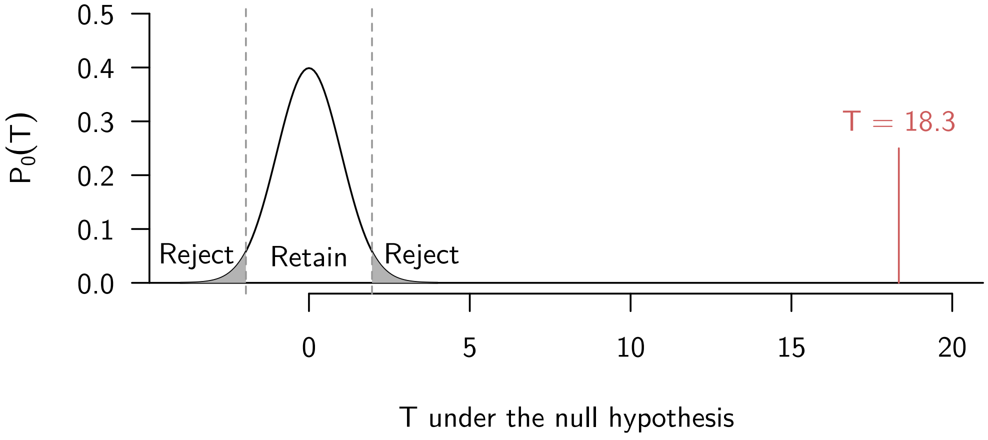
```
neigh_var <- var(social$voted[social$treatment == ”Neighbors”])
neigh_n <- 38201
civic_var <- var(social$voted[social$treatment == ”Civic Duty”])
civic_n <- 38218
se_diff <- sqrt(neigh_var/neigh_n + civic_var/civic_n)

## Calcuate test statistic
(0.378-0.315)/se_diff
```

```
## [1] 18.34304
```

- $|T_n| = 18.3430374 > 1.96 \rightsquigarrow$ REJECT!

Perform the test



t-test

- These ideas extend to *any* asymptotically normal estimator, $\widehat{\theta}$ for parameter θ. Consider testing $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_a$. A **size-$\alpha$ t-test** (or **Wald test**) rejects H_0 when $|T_n| > z_{\alpha/2}$ where

$$T_n = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathrm{se}}[\widehat{\theta}]}$$

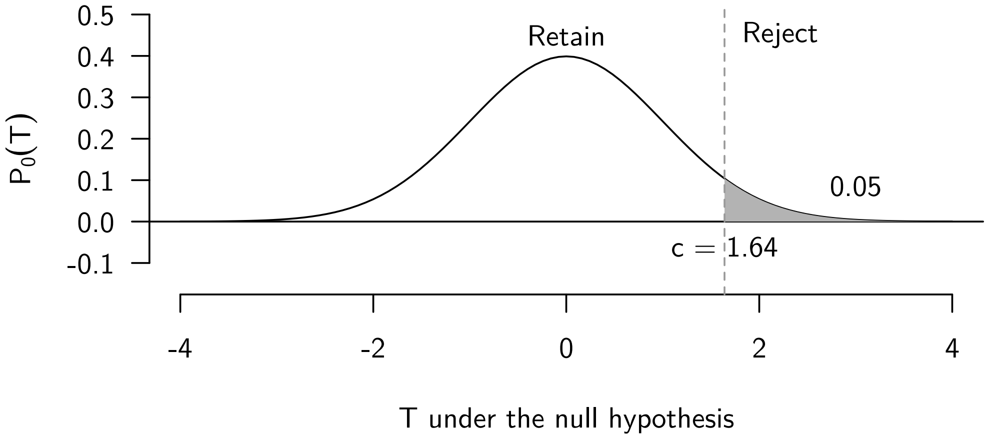
- Critical value $z_{\alpha/2}$ calculated in the exact same way as above. For standard normal Z, find $z_{\alpha/2}$ such that $\mathbb{P}(Z \leq z_{\alpha/2}) = 1 - \alpha/2$.
- Size of the test converges to the *nominal size* as n gets big $\mathbb{P}_0(|T_n| > z_{\alpha/2}) \xrightarrow{p} \alpha$.

Confidence intervals and hypothesis tests

- 95% confidence interval: $\widehat{D}_n \pm 1.96 \times \widehat{\mathrm{se}}$
- **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 $\alpha = 0.05$
- CIs are a range of plausible values in the sense we cannot reject them as null hypotheses.

One-sided tests

- **Definition** A **one-sided test** is a test of an alternative hypothesis that only goes in one direction.
  - The social pressure effect is positive ($H_a : \mu_y - \mu_x > 0$)
- Only deviations from the null hypothesis in one direction cast doubt on the null hypothesis.
  - Rejection region is only in one tail: $T_n > c$, with c adjusted downward relative to two-sided test with the same level.
- Really only valid when one side is a priori not possible.



p-values

Just rejecting or not rejecting the null hypothesis is not too informative. We rejected null of no population diff-in-means ($H_0 : \mu_y - \mu_x = 0$) at $\alpha = 0.05$. What about all the other levels like $\alpha = 0.01$? **p-values** are a useful way to summarize all possible levels at once.

Definition 3. p-value The **p-value** is the smallest value α such that an α-level test would reject the null hypothesis.

- If p-value is less than α, then we often say it is *statistically significant* at level α. For example, if p-value is 0.03, then we can reject at $\alpha = 0.05$ of $\alpha = 0.1$.
- **Theorem** For a two-sided test with observed test statistic $T_n = t_{\text{obs}}$, the p-value is the probability (under H_0) of observing a value of the test statistict at least as extreme as the one observed:

$$\mathbb{P}_0(|T_n| > t_{\text{obs}})$$

- Low p-value $\rightsquigarrow$ data was unlikely given the null $\rightsquigarrow$ evidence against the null.

Calculating the p-value

- Social pressure test statistic, $t_{\text{obs}} = 18.5$. How likely would it be to get a test statistic this extreme or more extreme if there were no treatment effect?

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

- Use the `pnorm()` function:

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

```
## [1] 2.06474e-76
```

Be careful with p-values

- 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

- Using a p-value cutoff ($p < 0.05$) can be very misleading. Leads to a clustering of p-values at 0.049. False discovery rates actually quite high (p-value fallacy).

- As difficult as they are to interpret, confidence intervals actually make more sense. CIs allow easy assessment substantive and statistical significance.

POWER ANALYSES

Effect sizes

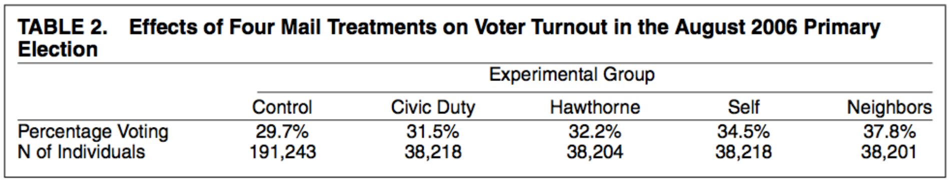


TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	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

- Why did Gerber, Green, and Larimer use sample sizes of 38,000 for each treatment condition?

- Choose the sample size to ensure that you can *detect* what you think might be the true treatment effect:
  - Small effect sizes (half percentage point) will require huge n
  - Large effect sizes (10 percentage points) will require smaller n

- *Detect* here means "reject the null of no effect"

Power of a test

- *Definition* The *power* of a test is the probability that a test rejects the null.
  - Probability that we reject given some specific value of the parameter $\mathbb{P}_\theta(|T| > c)$
  - Power = $1 - \mathbb{P}(\text{Type II error})$
  - Better tests = higher power.
- If we fail to reject a null hypothesis, two possible states of the world:
  - Null is true (no treatment effect)
  - Null is false (there is a treatment effect), but test had low power.

Why care about power?

- Imagine you are a company being sued for racial discrimination in hiring.
- Judge forces you to conduct hypothesis test:
  - Null hypothesis is that hiring rates for white and black people are equal, $H_0 : \mu_w - \mu_b = 0$
  - You sample 10 hiring records of each race, conduct hypothesis test and fail to reject null.
- Say to judge, "look we don't have any racial discrimination"! What's the problem?

Power analysis procedure

- Power can help guide the choice of sample size through a *power analysis*.
  - Calculate how likely we are to reject different possible treatment effects at different sample sizes.
  - *Can be done before the experiment*: which effects will I be able to detect with high probability at my n?
- Steps to a power analysis:
  - Pick some hypothetical effect size, $\mu_y - \mu_x = 0.05$
  - Calculate the distribution of T under that effect size.
  - Calculate the probability of rejecting the null under that distribution.
  - Repeat for different effect sizes.

Power analysis

- You want to run another turnout experiment want to make sure you have a high probability of rejecting the null if the true effect is $\mu_y - \mu_x = 0.05$.
- Unfortunately, your grant $$ are minimal so you can only send 500 mailers (250 for each type).
- Need to assume values for unknown variances:
  - Assume we know that $\sigma_y^2 = \sigma_x^2 = 0.2$
  - Implies $\mathbb{V}[\widehat{D}_n] = 0.2/250 + 0.2/250 = 0.0016$.
- Using these assumptions, we can derived the sampling distribution of the estimator under the proposed effect size:

$$\widehat{D}_n \approx N(0.05, 0.0016)$$

Power analysis

- What is the probability of rejecting the null if $\mu_y - \mu_x = 0.05$?
- We reject when

$$|T| = \left| \frac{\widehat{D}_n - 0}{\widehat{\text{se}}[\widehat{D}_n]} \right| > 1.96 \iff |\widehat{D}_n| > 1.96 \times \widehat{\text{se}}[\widehat{D}_n]$$

- Since we assumed that $\mathbb{V}[\widehat{D}_n] = 0.0016$ then we reject when:

$$\left\{ \widehat{D}_n < -1.96 \times \sqrt{0.0016} \right\} \cup \left\{ \widehat{D}_n > 1.96 \times \sqrt{0.0016} \right\}$$

- Can figure out the probability of this from the sampling distribution we just derived!

$$\mathbb{P}\left(\widehat{D}_n < -1.96 \times \sqrt{0.0016} \right) + \mathbb{P}\left(\widehat{D}_n > 1.96 \times \sqrt{0.0016} \right)$$

Power in R

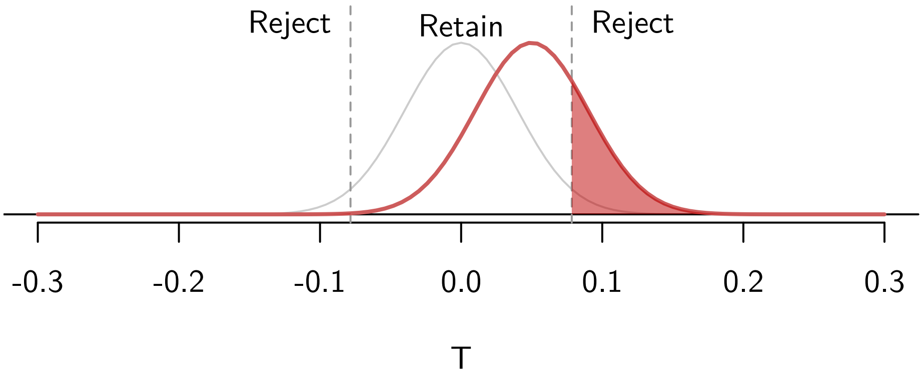
- Power of the test against $\mu_y - \mu_x = 0.05$, using the fact that $\widehat{D}_n \approx N(0.05, 0.0016)$:

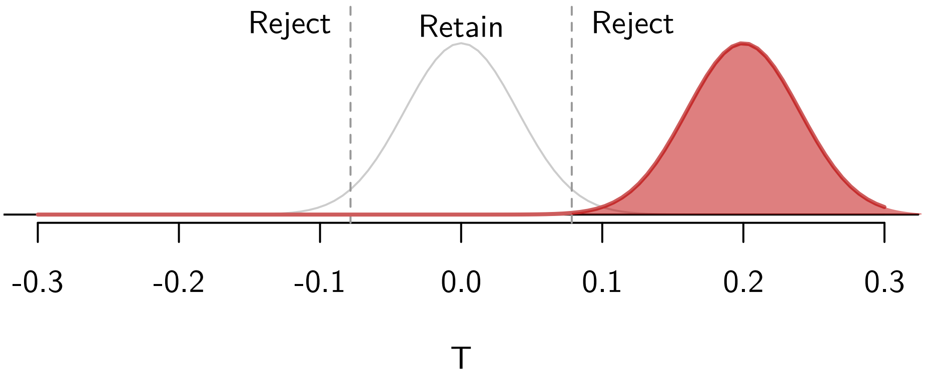
```r
se <- sqrt(0.2/250 + 0.2/250)
pnorm(-1.96 * se, mean = 0.05, sd = se) + pnorm(1.96 * se, mean = 0.05, sd = se, lower.tail = FALSE)
```

```
## [1] 0.2395157
```

- Interpretation: if the true effect was a 5 percentage point increase in voter turnout, then we would be able to reject the null of no effect about *a quarter of the time*.

Power graph



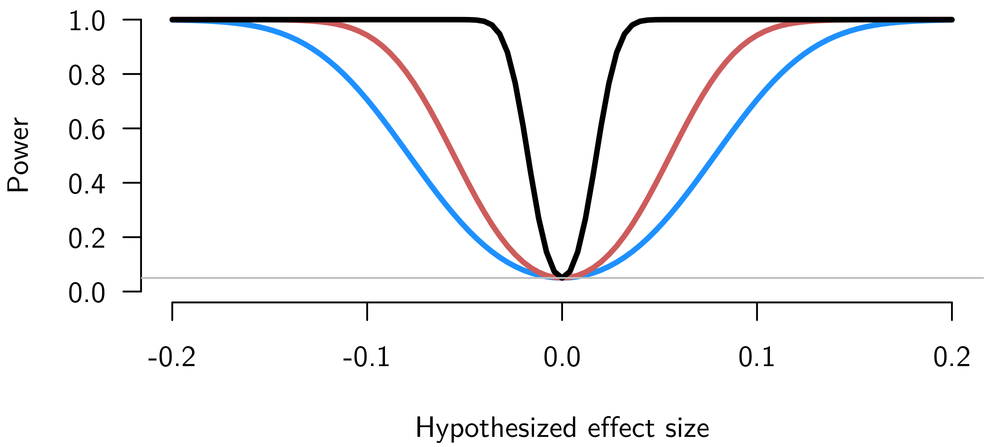


A power analysis

- We can calculate the power for every possible effect size and plot the resulting *power curve*:
    - $n =$ 500 (blue), 1000 (red), 10000 (black)

```
hold <- seq(-0.2, 0.2, by = 0.005)
power500 <- function(x) pnorm(-1.96 * se, x, se) + pnorm(1.96*se, x, se, lower.tail = FALSE)
se1k <- sqrt(0.2/500 + 0.2/500)
power1k <- function(x) pnorm(-1.96 * se1k, x, se1k) + pnorm(1.96*se1k, x, se1k, lower.tail = FALSE)
```

```
se10k <- sqrt(0.2/5000 + 0.2/5000)
power10k <- function(x) pnorm(-1.96 * se10k, x, se10k) + pnorm(1.96*se10k, x, se10k, lower.tail = FALSE)
curve(power500, -0.2, 0.2, col = ”dodgerblue”, lwd = 3, ylim = c(0,1), ylab = ”Power”, las = 1, xlab = ”Hypothes
curve(power1k, -0.2, 0.2, col = ”indianred”, lwd = 3, ylim = c(0,1), add = TRUE)
curve(power10k, -0.2, 0.2, col = ”black”, lwd = 3, ylim = c(0,1), add = TRUE)
abline(h=0.05, col = ”grey”)
```



EXACT INFERENCE*

Small sample complications

- Asymptotics are approximations. Can we ever get *exact* inferences at any sample size?
    - *Exact* means that we know or can figure out the distribution of a statistic without relying on an approximation.
- Remember: we are using a nonparametric model
    - Y_i are i.i.d. with $\mathbb{E}[Y_i] = \mu < \infty$ and $\mathbb{V}[Y_i] = \sigma^2 < \infty$
    - Relied on large n to get distribution of $\overline{Y}_n$ (CLT)
- Alternative: use a *parametric model* and assume $Y_1, \ldots, Y_n$ are i.i.d. samples from $N(\mu, \sigma^2)$
    - Stronger assumptions $\rightsquigarrow$ learn more with lower n
    - *Model dependence*: If the model is wrong (Y_i are not normal), inferences will be wrong!

Exact inference for the normal distribution

- Remember that the CLT gives us the following approximation:

$$T_n = \frac{\overline{Y}_n - \mu}{\frac{S_n}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$

- If we additionally know that $Y_i \sim N(\mu, \sigma^2)$, then we know the following for *any* sample size:

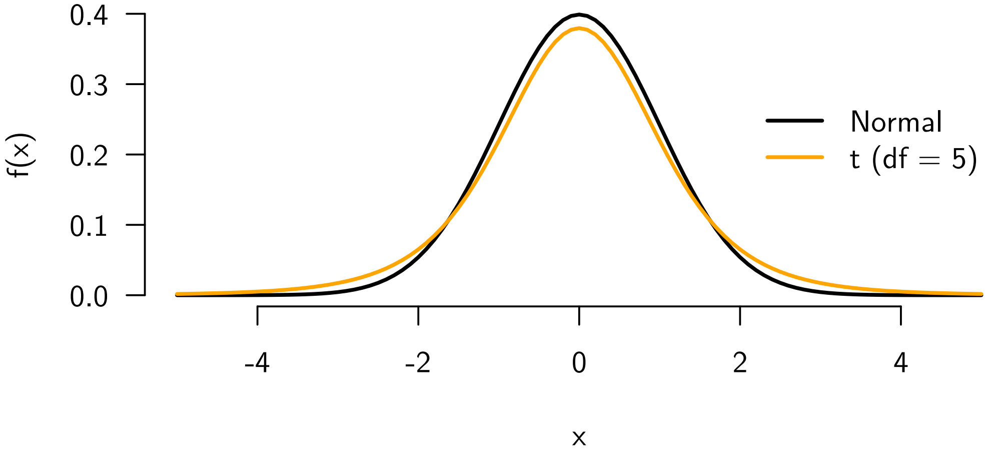
$$T_n = \frac{\overline{Y}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t_{n-1}$$

- 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.

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 the SE.

```
curve(dnorm(x), from = -5, to = 5, lwd = 2, bty = ”n”, ylab = ”f(x)”, las = 1)
curve(dt(x, df = 5), from = -5, to = 5, add = TRUE, lwd = 2, col = ”orange”)
legend(x = 2, y = 0.3, legend = c(”Normal”, ”t (df = 5)”), lwd = 2, col = c(”black”, ”orange”), bty = ”n”)
```



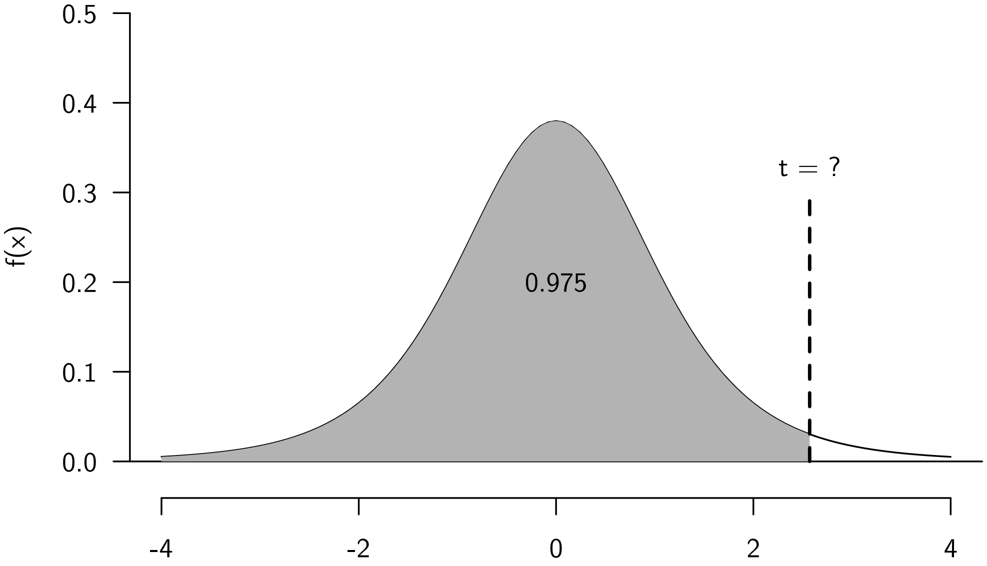
Using the t for small samples

- Use the same test statistic:
$$T_n = \frac{\overline{Y}_n - \mu_0}{S_n/\sqrt{n}}$$
- Assuming the null hypothesis, $T_n \sim t_{n-1}$, so use this distribution in place of the normal
- Use `qt()` in place of `qnorm()` for:
  - Testing: finding critical values $t_{n-1,\alpha/2}$ such that $\mathbb{P}_0(T \leq t_{n-1,\alpha/2}) = 1 - \alpha/2$
  - CIs: for $t_{n-1,\alpha/2}$ in place of z-values: $\overline{Y}_n \pm t_{n-1,\alpha/2} \times \frac{S_n}{\sqrt{n}}$
- *Conservative approach* relative to using asymptotic normality:
  - The t distribution has fatter tails $\rightsquigarrow t_{n-1,\alpha/2} > z_{\alpha/2}$
  - $\rightsquigarrow$ wider CIs, smaller rejection regions

Rejection region with the t

```r
par(mar = c(2.1, 4, 0.1, 0.1))
curve(dt(x, df = 5), from = -4, to = 4, ylim = c(-.02, 0.5), bty = ”n”, las = 1, ylab = ”f(x)”)
abline(h = 0)
thisq <- qt(0.975, df = 5)
polygon(c(-4,seq(-4,thisq,0.01),thisq), c(0,dt(seq(-4,thisq,0.01), df = 5), 0), col = ”grey70”, border = NA)
text(x = 0, y = 0.2, ”0.975”)
```

```
segments(x0 = thisq, y0 = 0, y1 = 0.3, lwd = 2, lty = 2)
text(x = thisq, y = 0.3, ”t = ?”, pos = 3)
```



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

```
## [1] 2.570582
```

WRAP UP

Key points

- Hypothesis testing:
  - Statistical thought experiments.
  - Allow us to test specific hypotheses about parameters.
- p-values:
  - Summarize evidence against the null in this data set.
  - Can be misleading, better to use confidence intervals.
- Deep connection between confidence intervals and hypothesis tests.
- Sometimes exact inference is possible, but only under strong assumptions.

- Power analyses help to guide what sample size we need.
- Next week: beginning to think about regression.