# Gov 50: 21. Hypothesis testing: Two-sample tests

Matthew Blackwell

Harvard University

Fall 2018

# 1/ Today's agenda

# Where are we? Where are we going?

- Trying to learn about (unknown) population parameters from sample data.
- Quantifying uncertainty: confidence intervals and hypothesis tests.
- Logistics:
  - Preliminary analyses due by Tuesday.
  - Final report due 12/10.

# 2/ Hypothesis testing review

# Statistical hypothesis testing

- Statistical hypothesis testing is a **thought experiment**.
- What would the world look like **if we knew the truth**?
- Example:
  - We've learned how to estimate a causal effect from an experiment or observational study.
  - But how can we tell if the difference we estimate is real or just due to chance?
  - Hypothesis test: assume there is no effect and determine what the data would look like in that world.

# Hypothesis testing procedure

Conducted with several steps:

1. Generate your **null** and **alternative hypotheses**
2. Collect sample of data
3. Calculate appropriate **test statistic**
4. Use that value to calculate a probability called a **p-value**
5. Use p-value to decide whether to reject the null hypothesis or not

# Last time

- We looked at hypothesis tests for population proportions.
  - Tested null that true population proportion was some value: $H_0 : p = p_0$
- Under the null hypothesis, we can determine the (approximate) distribution of the test statistic:

$$Z = \frac{\overline{X} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- Calculated p-values of this test statistic
- Today: generalizing to differences in means.

# 3/ Two-sample tests

# Social pressure example

- Back to the Social Pressure Mailer GOTV example.
  - Treatment group: postcards showing their own and their neighbors' voting records.
  - Control group: received nothing.
- Samples are **independent**
  - Example of dependent comparisons: **paired comparisons**

# Two-sample hypotheses

- Parameter: **population ATE** $\mu_T - \mu_C$
  - $\mu_T$: Turnout rate in the population if everyone received treatment.
  - $\mu_C$: Turnout rate in the population if everyone received control.
- Goal: learn about the population difference in means
- Usual null hypothesis: no population difference in means (no causal effect)
  - Null: $H_0 : \mu_T - \mu_C = 0$
  - Two-sided alternative: $H_1 : \mu_T - \mu_C \neq 0$
- In words: does the treatment and control group have the same distribution?

# Difference-in-means review

- Sample turnout rates: $\overline{X}_T = 0.37, \overline{X}_C = 0.30$
- Sample sizes: $n_T = 360, n_C = 1890$
- Estimator is the **sample difference-in-means**:

$$\widehat{\text{ATE}} = \overline{X}_T - \overline{X}_C = 0.07$$

- Standard error:

$$\widehat{\text{SE}}_{\widehat{\text{ATE}}} = \sqrt{\frac{\overline{X}_T(1-\overline{X}_T)}{n_T} + \frac{\overline{X}_C(1-\overline{X}_C)}{n_C}} = 0.028$$

- 95% confidence interval:

$$\begin{aligned} CI_{95} =& \widehat{\text{ATE}} \pm 1.96 \times \widehat{\text{SE}}_{\widehat{\text{ATE}}} \\ =& [0.016, 0.124] \end{aligned}$$

# CLT again and again

- $\overline{X}_T$ is a sample mean and so tends toward normal as $n_T \to \infty$
- $\overline{X}_C$ is a sample mean and so tends toward normal as $n_C \to \infty$
- $\rightsquigarrow \overline{X}_T - \overline{X}_C$ is a random variable that will tend toward normal as sample sizes get big.
- In particular, this will approximately true in large samples:

$$\overline{X}_T - \overline{X}_C \sim N\left(\mu_T - \mu_C, \frac{\mu_T(1-\mu_T)}{n_T} + \frac{\mu_C(1-\mu_C)}{n_C}\right)$$

- Using the z-transformation/standardization:

$$\frac{(\overline{X}_T - \overline{X}_C) - (\mu_T - \mu_C)}{\sqrt{\frac{\mu_T(1-\mu_T)}{n_T} + \frac{\mu_C(1-\mu_C)}{n_C}}} \sim N(0,1)$$

# Test statistic

- Null hypothesis: $H_0 : \mu_T - \mu_C = 0$
- Test statistic:

$$Z = \frac{(\overline{X}_T - \overline{X}_C) - (\mu_T - \mu_C)}{SE} = \frac{(\overline{X}_T - \overline{X}_C) - 0}{SE}$$

- Here, the SE is:

$$SE = \sqrt{\frac{\mu_T(1-\mu_T)}{n_T} + \frac{\mu_C(1-\mu_C)}{n_C}}$$

- In large samples, we can replace true SE with an estimate:

$$\widehat{SE} = \sqrt{\frac{\overline{X}_T(1-\overline{X}_T)}{n_T} + \frac{\overline{X}_C(1-\overline{X}_C)}{n_C}}$$

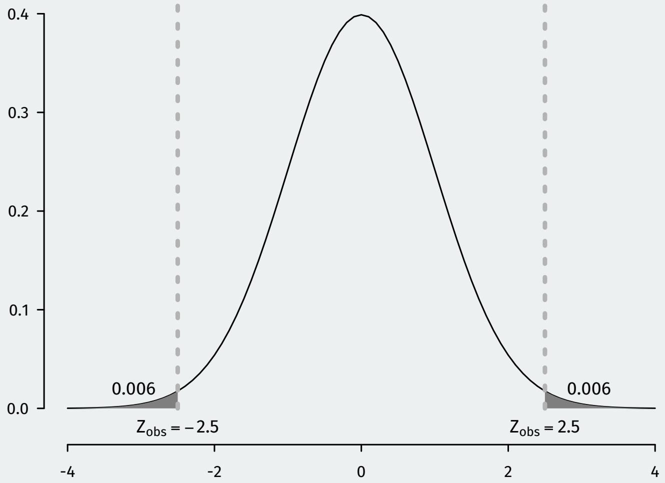
- By CLT, $Z \sim N(0,1)$

# Calculating p-values

- Finally! Our test statistic in this sample:

$$Z = \frac{\overline{X}_T - \overline{X}_C}{\widehat{\text{SE}}} = \frac{0.07}{0.028} = 2.5$$

- p-value based on a two-sided test: probability of getting a difference in means this big (or bigger) if the null hypothesis were true
    - Lower p-values $\rightsquigarrow$ stronger evidence against the null.




```r
2 * pnorm(2.5, lower.tail = FALSE)
```

```
## [1] 0.0124
```

# Tests and confidence intervals

- There is a deep connection between confidence intervals and tests.
- Any value outside of a $100 \times (1 - \alpha)\%$ confidence interval would have a p-value less than $\alpha$ if we tested it as the null hypothesis.
  - 95% CI for social pressure experiment: $[0.016, 0.124]$
  - $\rightsquigarrow$ p-value for $H_0 : \mu_T - \mu_C = 0$ less than 0.05.
- Confidence intervals are all of the null hypotheses we **can't reject** with a test.

# 4/ Example: checking randomization

# Checking randomization

- Load the social pressure experiment data:

```
social <- read.csv("data/social.csv")
social <- subset(social, hhsize == 2)
treated <- subset(social, messages == "Neighbors")
control <- subset(social, messages == "Control")
head(treated[,1:4])
```

```
##        sex yearofbirth primary2004  messages
## 28    male        1946           0 Neighbors
## 29  female        1932           0 Neighbors
## 80  female        1946           0 Neighbors
## 81    male        1941           0 Neighbors
## 116   male        1970           1 Neighbors
## 117 female        1971           1 Neighbors
```

# Checking randomization

- If randomization was successful, there should be no differences between the treated and control group on pretreatment variables.
- One variable: **year of birth**

```
mean(treated$yearofbirth) - mean(control$yearofbirth)
```

```
## [1] -0.115
```

- Treatment group is older than control group!!
- Did randomization fail?!
  - Or...could this just be due to random chance?

# More general difference in means

- Null hypothesis: $H_0 : \mu_T - \mu_C = 0$
- Estimator is still sample difference in means: $\overline{X}_T - \overline{X}_C$
- Year of birth isn't binary $\rightsquigarrow$ more general standard error:

$$\widehat{\text{SE}} = \sqrt{\widehat{\text{SE}}_T^2 + \widehat{\text{SE}}_C^2} = \sqrt{\frac{\widehat{\sigma}_T^2}{n_T} + \frac{\widehat{\sigma}_C^2}{n_C}}$$

  - $\widehat{\sigma}_T^2$ is the sample variance of year of birth in the treated group.
  - $\widehat{\sigma}_C^2$ is the sample variance of year of birth in the control group.
- Test statistic is the same: $(\overline{X}_T - \overline{X}_C)/\widehat{\text{SE}}$

# R can do the work

```r
t.test(treated$yearofbirth, control$yearofbirth)
```

```
##
##  Welch Two Sample t-test
##
## data:  treated$yearofbirth and control$yearofbirth
## t = -1.26, df = 33600, p-value = 0.21
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -0.292963  0.063707
## sample estimates:
## mean of x mean of y
##    1954.6    1954.7
```

# 5/ 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, $\mu_T - \mu_C = 0.05$
  - 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 $\mu_T - \mu_C = 0.05$.
- 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 $\sigma_T^2 = \sigma_C^2 = 0.2$
  - Implies $\mathbb{V}[\overline{X}_T - \overline{X}_C] = 0.2/250 + 0.2/250 = 0.0016$.
- Using these assumptions, we can derived the sampling distribution of the estimator under the proposed effect size:

$$\overline{X}_T - \overline{X}_C \approx N(0.05, 0.0016)$$

# Power analysis

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

$$|T| = \left| \frac{\overline{X}_T - \overline{X}_C - 0}{\widehat{\text{SE}}} \right| > 1.96 \iff |\overline{X}_T - \overline{X}_C| > 1.96 \times \widehat{\text{SE}}$$

- Can figure out the probability of this from the sampling distribution!
- Since $1.96 \times \sqrt{0.0016} = 0.078$:

$$\mathbb{P}\left(\overline{X}_T - \overline{X}_C < -0.078\right) + \mathbb{P}\left(\overline{X}_T - \overline{X}_C > 0.078\right)$$

# Power in R

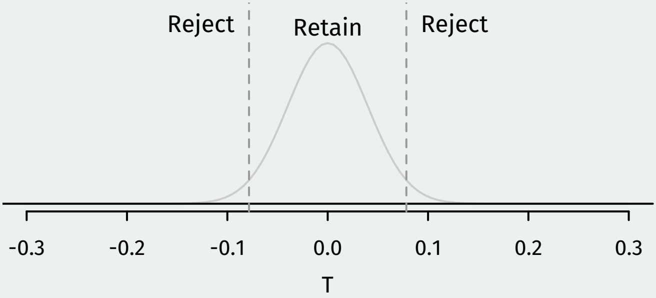
- Power of the test against $\mu_y - \mu_x = 0.05$, using the fact that $\overline{X}_T - \overline{X}_C \approx N(0.05, 0.0016)$:

```
pnorm(-0.078, mean = 0.05, sd = sqrt(0.0016)) +
  pnorm(0.078, mean = 0.05, sd = sqrt(0.0016), lower.tail = FALSE
```

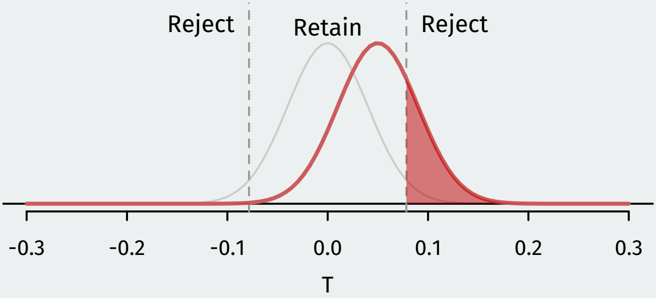
```
## [1] 0.24265
```

- 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

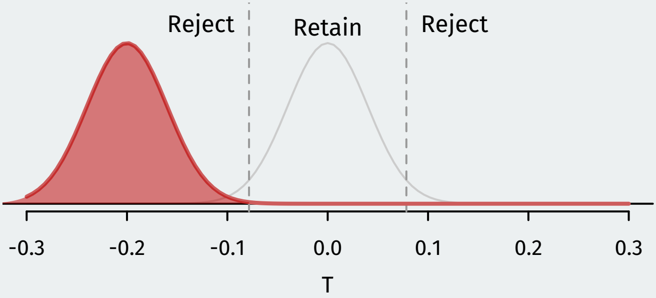


# Power graph



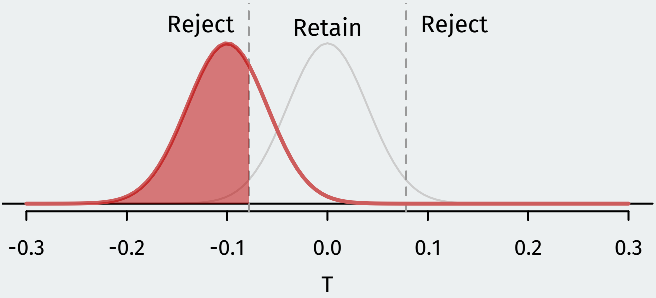

Assumed treatment effect = 0.05 and power = 0.23952.

# Power graph



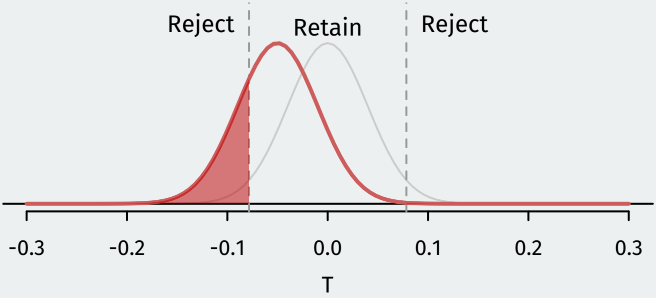

Assumed treatment effect = -0.2 and power = 0.99882.

# Power graph



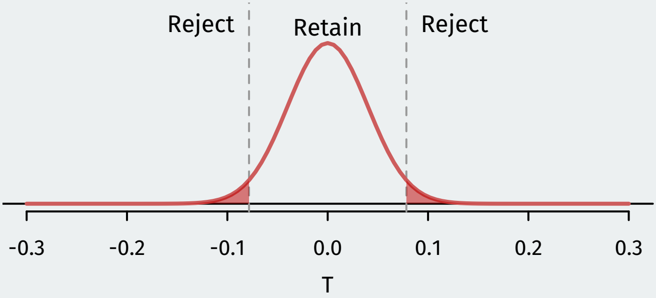

Assumed treatment effect = -0.1 and power = 0.70541.

# Power graph



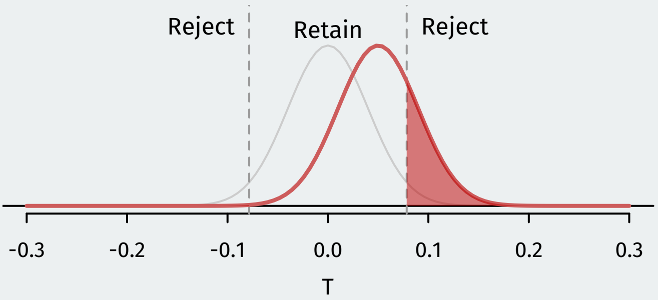

Assumed treatment effect = -0.05 and power = 0.23952.

# Power graph



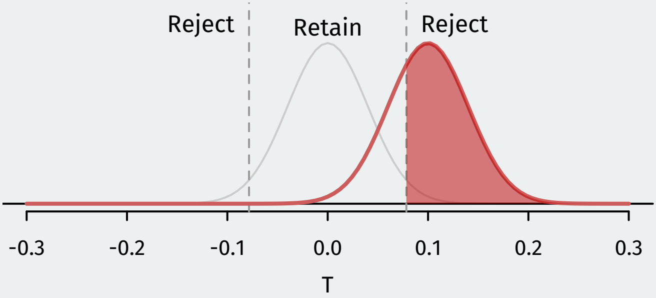

Assumed treatment effect = 0 and power = 0.05.

# Power graph



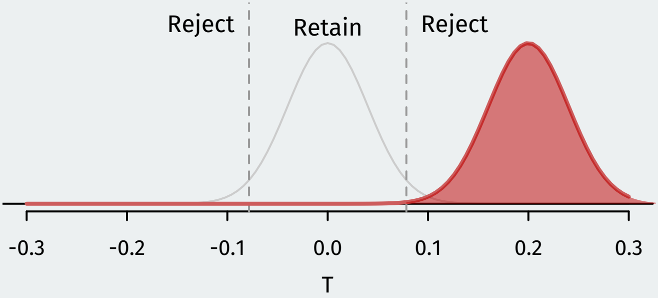

Assumed treatment effect = 0.05 and power = 0.23952.

# Power graph




Assumed treatment effect = 0.1 and power = 0.70541.

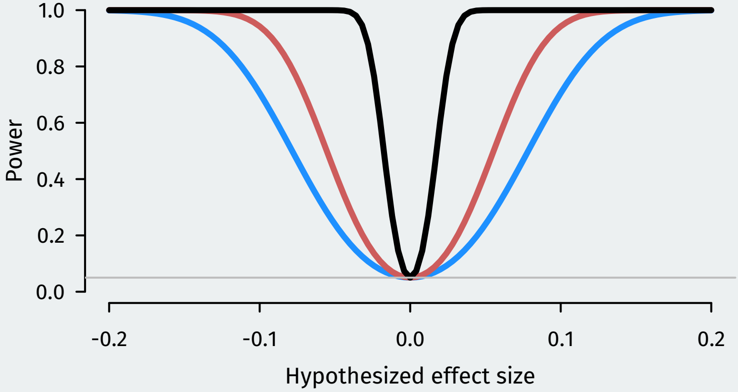
# Power graph




Assumed treatment effect $= 0.2$ and power $= 0.99882$.

# 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)



# Next time

- How to conduct inference on regression coefficients.