Week 4: Hypothesis Testing and Causal Inference

Brandon Stewart[1]

Princeton

September 24–26, 2024

[1]Many of the slides are adapted from Matt Blackwell, Adam Glynn, Jens Hainmueller, and Erin Hartman

Where We've Been and Where We're Going...

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing
  - potential outcomes approach to causal inference
- Next Week
  - directed acyclic graphs / nonparametric structural equation models
  - causal inference with measured confounding.
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1 Hypothesis Testing

2 Power

3 p-values and their Problems
- Fun With Salmon
- The Significance of Significance

4 Potential Outcomes

We Secretly Already Covered This!

American Political Science Review Vol. 102, No. 1 February 2008

DOI: 10.1017/S000305540808009X

Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER *Yale University*
DONALD P. GREEN *Yale University*
CHRISTOPHER W. LARIMER *University of Northern Iowa*

We Secretly Already Covered This!

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!

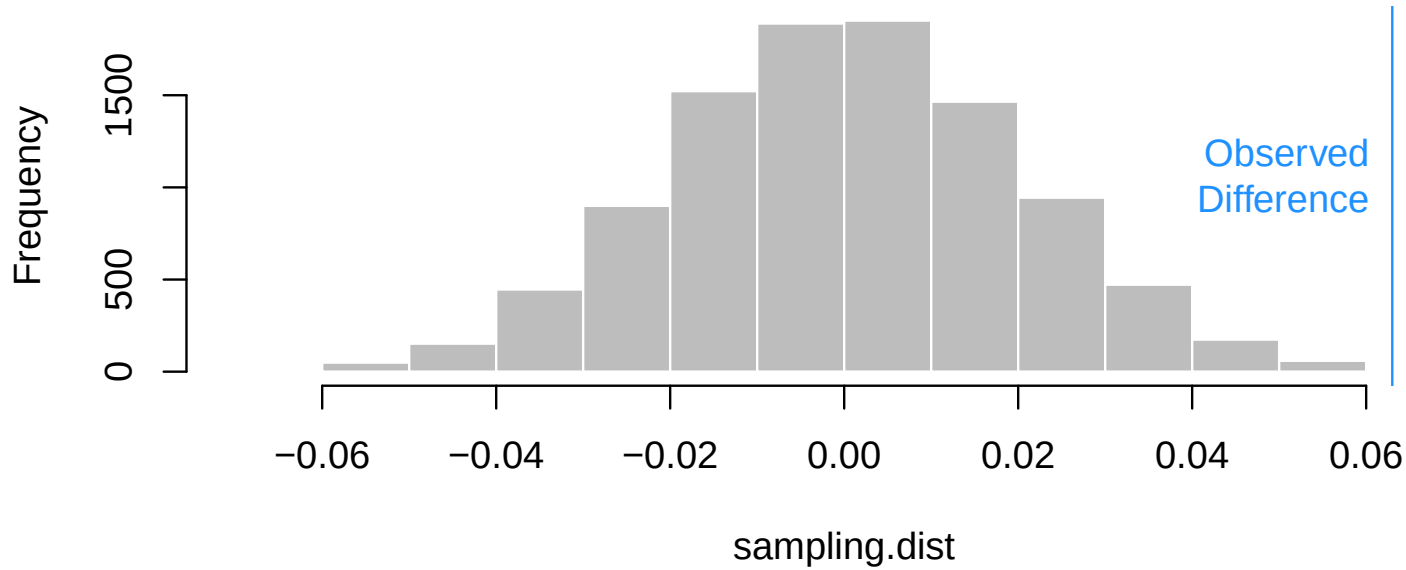
MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____

Review of the Gerber, Green and Larimer Result

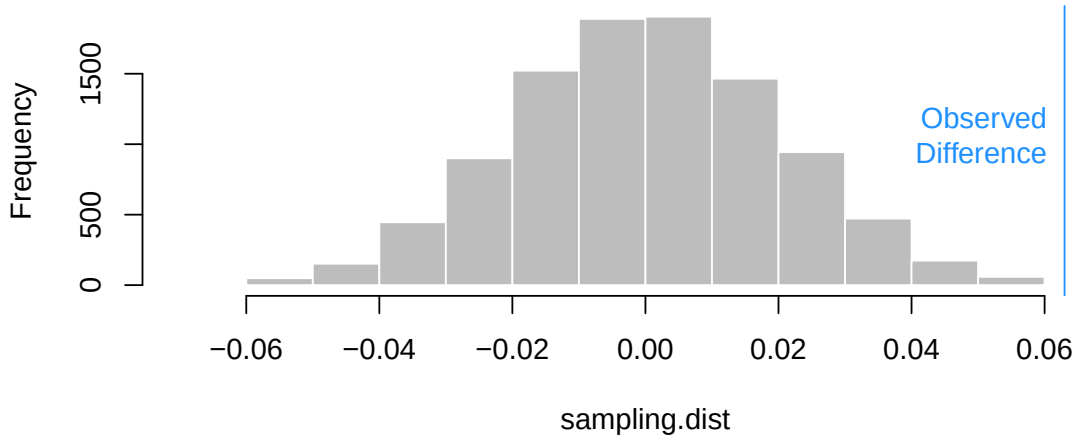
- Our estimand is the difference in population means: $\theta = \mu_y - \mu_x$. Where μ_y is those receiving the social pressure mailer and μ_x is those receiving the civic duty mailer.
- We used difference in means to get an estimate of 0.063
- We estimated the estimator's standard error $\widehat{\text{SE}}(\hat{\theta})$.
- What if there was no difference in means in the population $(\mu_y - \mu_x = 0)$?
- By asymptotic Normality $(\hat{\theta} - 0)/\text{SE}(\hat{\theta}) \sim N(0, 1)$
- This implies that $\hat{\theta} \sim \mathcal{N}(0, \text{SE}(\hat{\theta})^2)$
- Which we then approximate with $\hat{\theta} \sim \mathcal{N}(0, \widehat{\text{SE}}(\hat{\theta})^2)$

We Secretly Already Covered This!

Example from Gerber, Green and Larimer (2008).



Overall Idea



- We assumed we knew the difference in means was zero.
- We derived the sampling distribution under that value.
- We asked 'if the population difference was zero (i.e. this was the sampling distribution), how likely would it be to see an estimate as extreme (or more extreme) than our observed estimate?'
- Our observed difference was so implausible we concluded it was unlikely the population difference was really zero.

Basically, we did a hypothesis test.

What is a Hypothesis Test?

- A hypothesis test is a way of assessing evidence against a particular hypothesis about the world—it is a statistical thought experiment.
- By using probability, we define what the data would look like under a given state of the world, and then assess whether or not that's consistent with our observed data.
- Hypothesis testing is an alternative way to think about inference than confidence intervals, but using much of the same infrastructure.
- Hypothesis tests lead to discrete decisions.

Statistical hypothesis tests

General procedure:

1) Choose a null hypothesis (H_0) and a complementary alternative hypothesis (H_A)
2) Choose a test statistic $T_{(n)}$
3) Derive the sampling distribution (or reference distribution) of $T_{(n)}$ under H_0
4) Is the observed value (or anything more extreme) of $T_{(n)}$ likely to occur under H_0 with some pre-specified probability?
   - Yes: Fail to reject H_0 (note: NOT ACCEPT H_0)
   - No: Reject H_0

Test Statistics and Null Distributions

Test Statistic: which we denote T_n is a function of the sample, the estimator and the null hypothesis.

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}[\hat{\theta}]}$$

This is a random variable because it is a function of the sample.

Null Distribution: the sampling distribution of the statistic/test statistic assuming that the null is true.

Test Statistics and Null Distribution

Returning to the voting experiment. We know from the Central Limit Theorem that the standardized difference in means has a standard normal distribution asymptotically,

$$T_n = \frac{\hat{\theta} - (\mu_y - \mu_x)}{\widehat{\mathsf{SE}}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0,1)$$

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

$$T_n = \frac{\hat{\theta}}{\widehat{\mathsf{SE}}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0,1)$$

If T_n is very far from zero—in the sense that it has low probability under $\mathcal{N}(0,1)$—then we reject the null hypothesis as not plausible.

Rejection Region

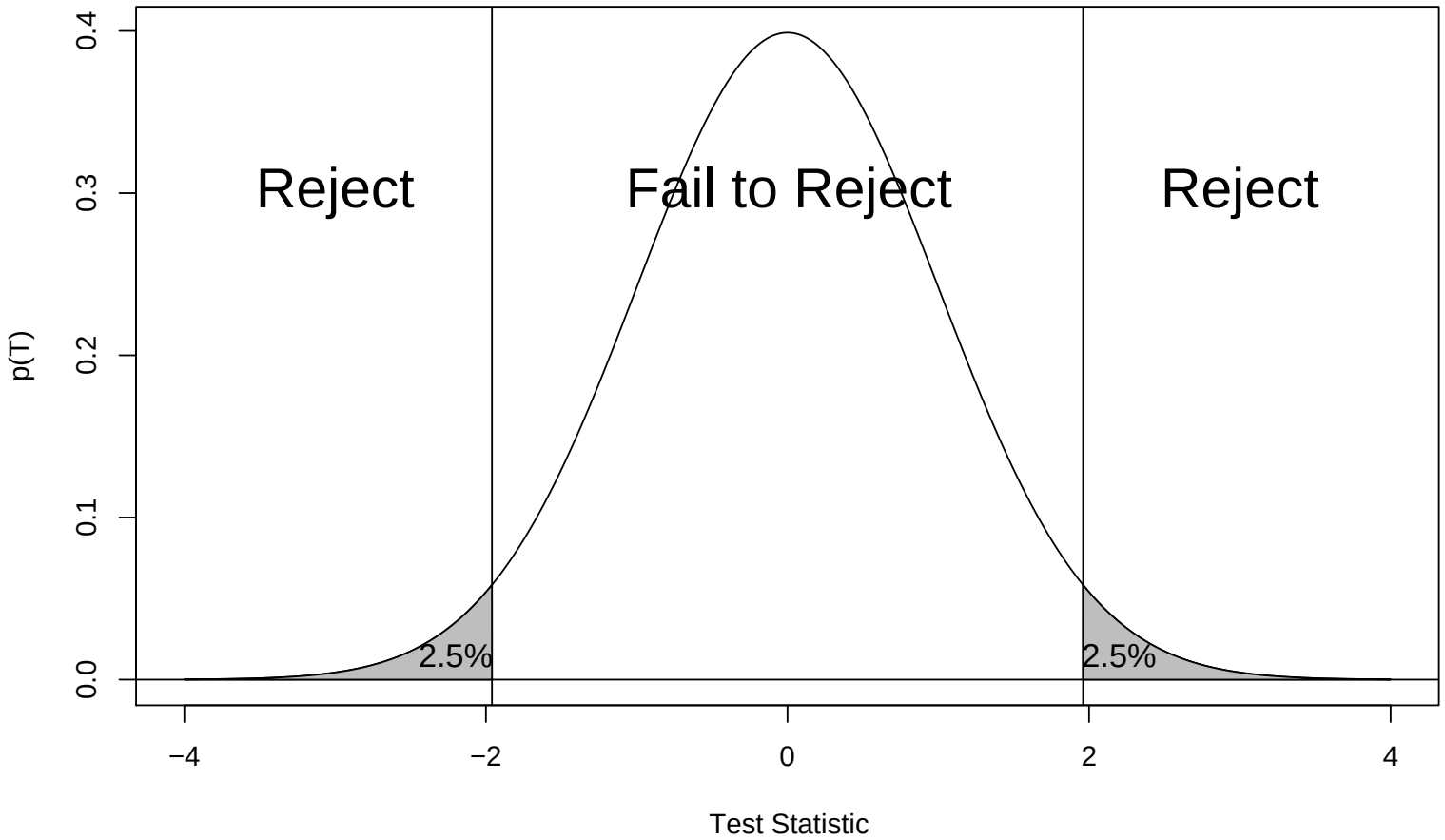
- The rejection region R contains the values of the test statistics, T_n for which we reject the null. This is defined by the null distribution and our chosen tolerance for Type I error.
- Denote the probability of rejecting the null when it is true (called a Type I error in this literature) as α.
- If we have a two-sided test (i.e. our null hypothesis equals a given value and thus extreme values on either side could reject the null), we reject when $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.
- We call c the critical value.
- Much like the 95% confidence interval, we pick $\alpha = .05$ by convention.

The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

- Ronald Fisher, *Design of Experiments* (1922)

Two-sided Rejection Region

Rejection region with $\alpha = .05$, $H_0 : \theta = 0$, $H_A : \theta \neq 0$:



The Gerber, Green and Larimer Example

```
diff <- mean(treated) - mean(control)
se_diff <- sqrt(var(treated)/length(treated) +
                var(control)/length(control))
test_statistic <- diff/se_diff
```

This yields 18.3 which is much larger in absolute value than our .05 critical value of 1.96.

We reject the null.

One-Sided Test: An Example from the FDA

In Phase II clinical trials the goal is to assess Efficacy (is there any evidence it helps?) before proceeding to Phase III trials which compare to existing treatments.

	Drug works	Drug doesn't work
FDA approves	Good!	Bad!
FDA doesn't approve	Bad!	Good!

Drug trials are expensive and *ex ante* we can specify that we only care about one direction in particular. Consider a drug combination which claims to improve blood pressure.

Data Example: Sowers et al. (2006) Phase II Trial

Subject	SBP_{before}	SBP_{after}	Improvement
1	147	135	12
2	153	122	31
3	142	119	23
4	141	134	7
⋮	⋮	⋮	⋮
345	155	115	40

On average blood pressure improved by 21 points and sample standard deviation was 14.3. Should the FDA approve?

Blood Pressure Example

We can define our hypotheses for a one-sided test.

$$H_0 : \mu_{\text{improves BP}} \leq 0 \tag{1}$$
$$H_a : \mu_{\text{improves BP}} > 0 \tag{2}$$

We can calculate the test statistic:

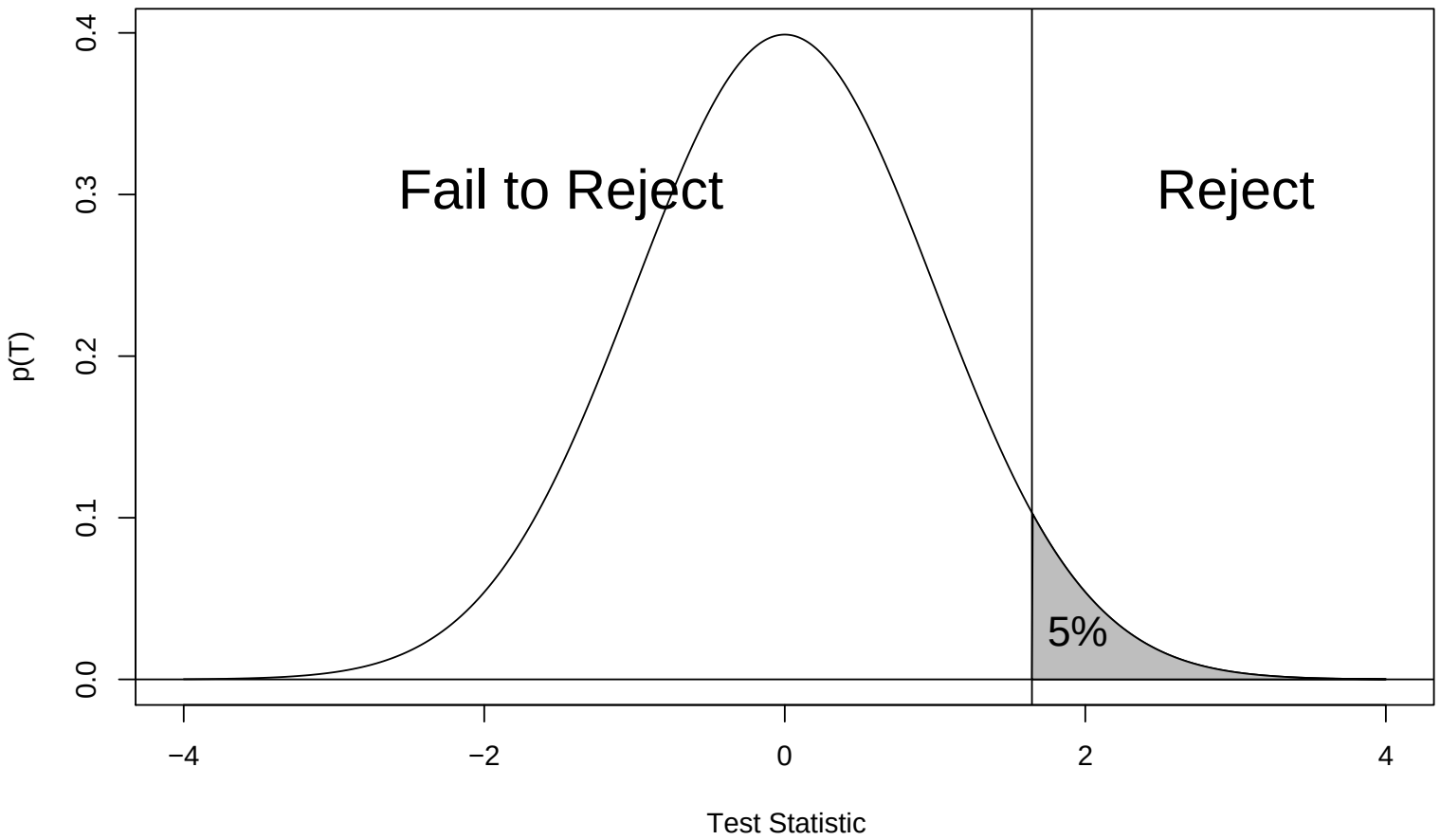
- $\overline{x} = 21.0$
- $\hat{\sigma} = 14.3$
- $n = 345$

Therefore,

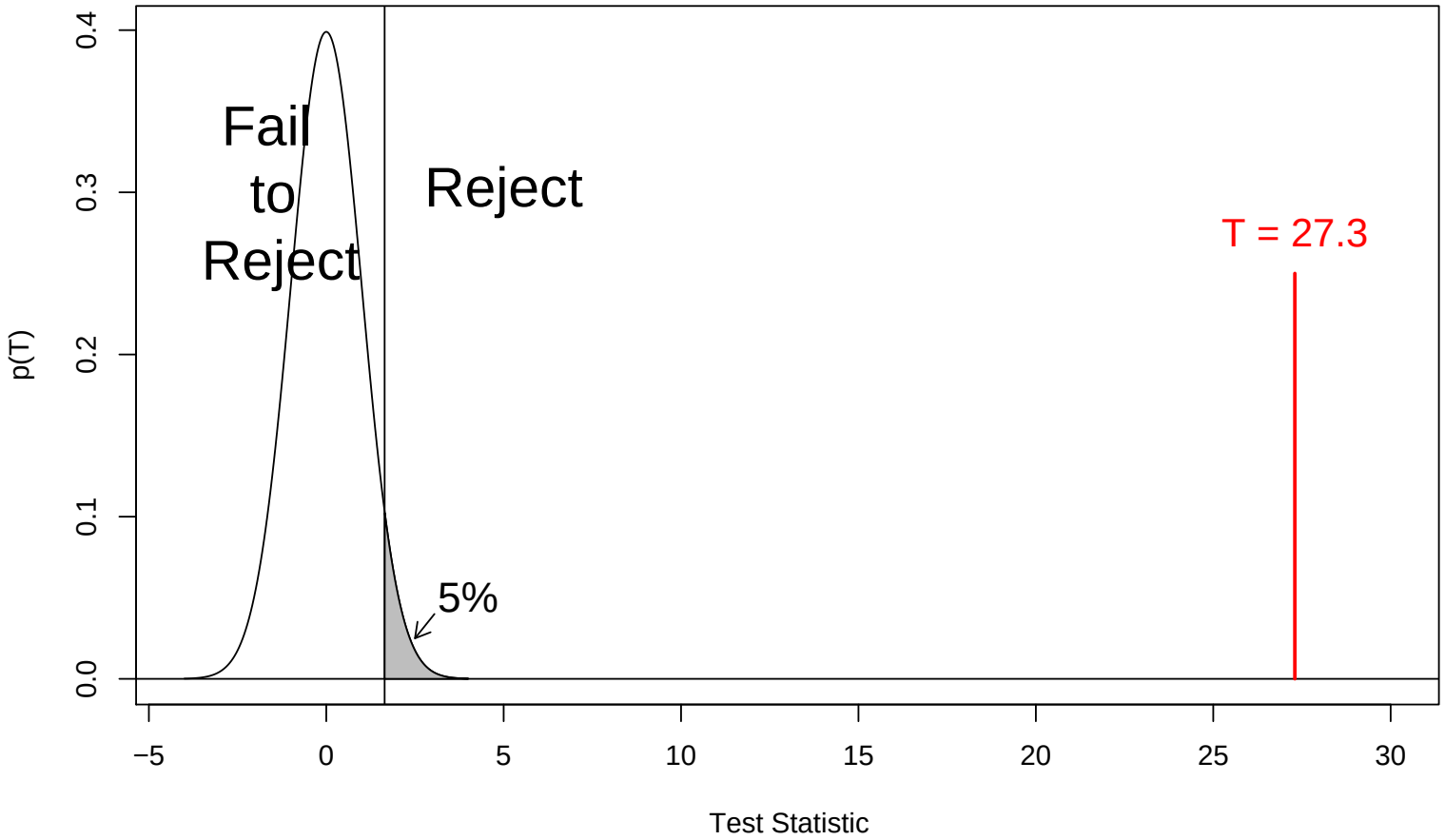
$$t_n = \frac{21.0 - 0}{\frac{14.3}{\sqrt{345}}} = 27.3$$

We construct our rejection region with $c = \texttt{qnorm(.95)} = 1.644$.

Rejection Region with $\alpha = .05$



Rejection Region with $\alpha = .05$



Zooming Out: t-test

- This strategy works for any asymptotically normal estimator.
- A size-α t-test (also called a Wald test) rejects H_0 when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}[\hat{\theta}]}$$

- We get the critical value c by using the standard normal to get the probability such that $P_{H_0}(T_n \leq c) = 1 - \alpha/2$
- This will guarantee the nominal probability of Type I error as n gets large.

Connections to Confidence Intervals

You may have noticed that there is a 1:1 mapping between CIs and hypothesis tests (same math, different story).

- If the confidence interval includes the null hypothesis, fail to reject your null.
- If the confidence interval does not include the null hypothesis, reject your null.

A $100(1-\alpha)\%$ confidence interval represents all null hypotheses that we would not reject with a α level test.

We can think of confidence intervals as a range of plausible values in the sense that we would not have rejected them had they been our null hypotheses.

1 Hypothesis Testing

2 Power

3 p-values and their Problems
- Fun With Salmon
- The Significance of Significance

4 Potential Outcomes

Power

We designed our tests to minimize Type I error (the false positive). However we might also worry we are failing to detect important findings.

Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

Definition (Power)

The power of a test is the probability that a tests rejects the null given some assumed population distribution $P_\theta(|T_n| > c)$.

$$\text{Power} = 1 - P(\text{Type II error})$$

If we fail to reject the null hypothesis, there are basically three possible states of the world:

1. Null is true (no difference between the mailer populations).
2. Null is false (there is a difference between the mailer populations), but test had low power.
3. Null is false, the test is well-powered and we got incredibly unlucky.

Power is Important

- Imagine that you are trying to prove that your approximate measurement strategy is good.
- You take a sample of 20 measurements where you can obtain the truth and do a hypothesis test.
- You fail to reject the null. Is this good evidence that there are no differences and your measurement is perfect?
- No! You haven't been able to confidently reject the possibility that there is no difference, but that's different than rejecting the possibility that there is a difference.
- This might seem obvious but conflating a lack of evidence with evidence for a zero effect is a problem that crops up all the time.
- Power analysis is a way of guiding the choice of sample size prior to an experiment to avoid this kind of mistake.

How Well Powered Are Social Science Studies?

Quantitative Political Science Research is Greatly Underpowered*

Vincent Arel-Bundock† Ryan C. Briggs‡
Hristos Doucouliagos§ Marco Mendoza Aviña¶ T.D. Stanley‖

July 05, 2022

Abstract

We analyze the statistical power of political science research by collating over 16,000 hypothesis tests from about 2,000 articles. Even with generous assumptions, the median analysis has about 10% power, and only about 1 in 10 tests have at least 80% power to detect the consensus effects reported in the literature. There is also substantial heterogeneity in tests across research areas, with some being characterized by high-power but most having very low power. To contextualize our findings, we survey political methodologists to assess their expectations about power levels. Most methodologists greatly overestimate the statistical power of political science research.

The median analysis has about 10% power ($\alpha = 0.05$), and only about 1 in 10 statistical tests have at least 80% power.

Steps for Power Analysis

1) Specify the null hypothesis to be tested at significance level α
2) Choose a true value for population parameters and derive the sampling distribution of test statistic
3) Calculate the probability of rejecting the null hypothesis under this sampling distribution.
4) Find the smallest sample size such that this rejection probability equals a pre-specified power level.
5) Possibly repeat under different assumptions about the population.

Example: Power Analysis

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 they use 38 thousand people per group? Power analysis!
- If their results were exactly true, would we be expect to be able to confirm that the effect is non-zero with a replication using only 500 mailers?
- For this example, let's ignore the household sampling (which is really important in practice, but we want to make this simple!).

What Do We Know?

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

$$
\begin{aligned}
\mu_y &= .378 \\
\sigma_y^2 &= (.378)(1-.378) = 0.235 \\
\mu_x &= .315 \\
\sigma_x^2 &= (.315)(1-.315) = 0.216 \\
\theta &= \mu_y - \mu_x = .063 \\
\theta_0 &= 0 \\
V[\hat{\theta}_n] &= .235/250 + 0.216/250 = .001804 \\
\hat{\theta}_n &\xrightarrow{d} \mathcal{N}(0.63, 0.001804)
\end{aligned}
$$

Calculating Probability of Rejecting the Null

- We reject when

$$|T| = \frac{|\hat{\theta}_n - 0|}{\hat{\text{SE}}[\hat{\theta}_n]} > 1.96$$

- Rearranging we get rejection when

$$|\hat{\theta}_n| > 1.96\hat{\text{SE}}[\hat{\theta}_n]$$

- Under our assumption of the truth we know $V[\hat{\theta}_n] = .001804$ and thus:

$$\left\{\hat{\theta}_n < -1.96\sqrt{.001804}\right\} \cup \left\{\hat{\theta}_n > 1.96\sqrt{.001804}\right\}$$

- Using the sampling distribution we derived can calculate:

$$P\left(\hat{\theta}_n < -1.96\sqrt{.001804}\right) + P\left(\hat{\theta}_n > 1.96\sqrt{.001804}\right)$$

Getting an Answer in R

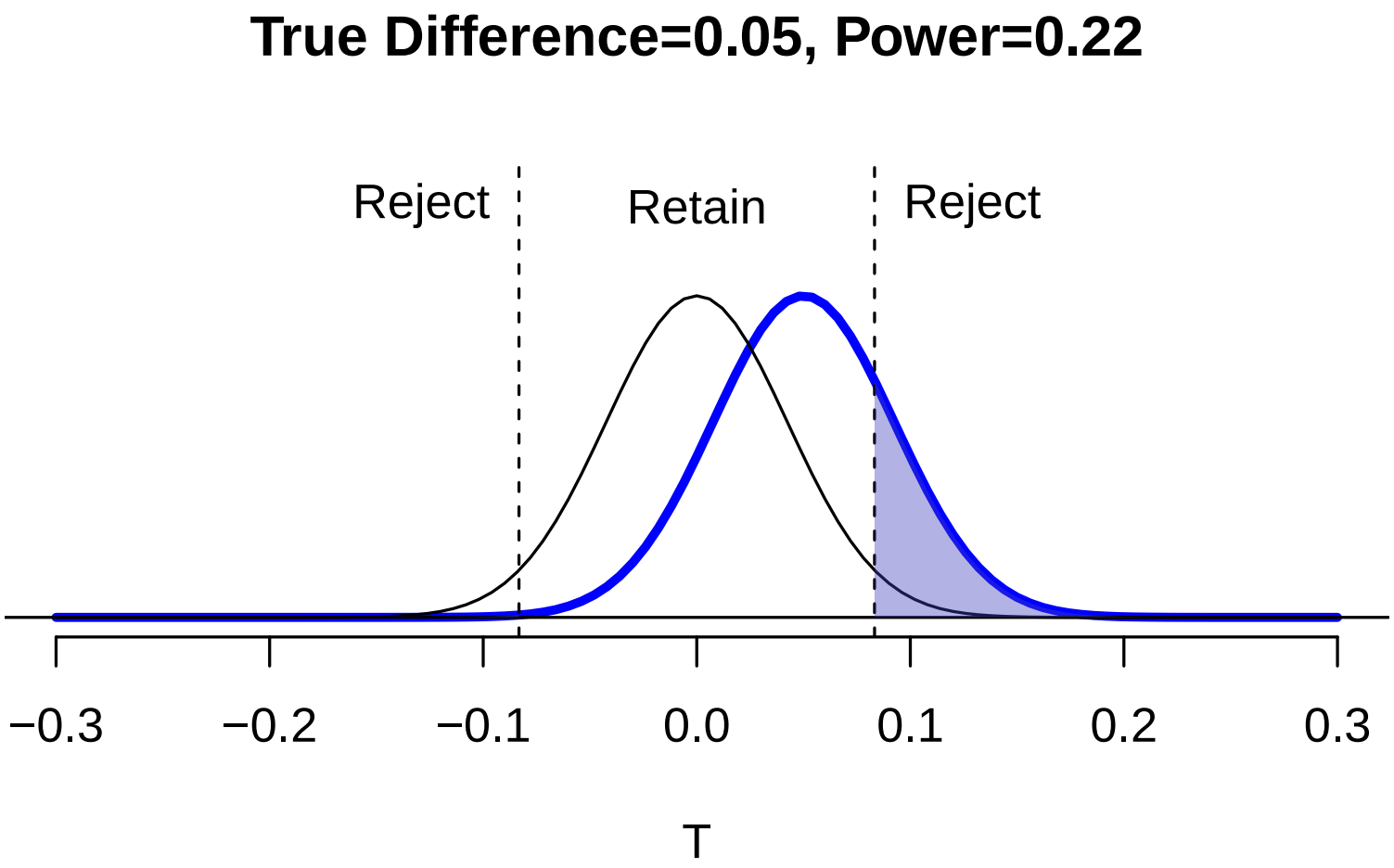
Let's calculate this using the fact that $\hat{\theta}_n \xrightarrow{d} \mathcal{N}(0.63, 0.001804)$

```
se <- sqrt(.001804)
pnorm(-1.96*se, mean=.05, sd=se) +
pnorm(1.96*se, mean=.05, sd=se, lower.tail=FALSE)
```

We get 0.2177 which tells us that if the true population means were those calculated in the experiment, we would be able to reject the null of no effect about 22% of the time using 500 mailers.

Yikes! That is not well powered.

Power Graph

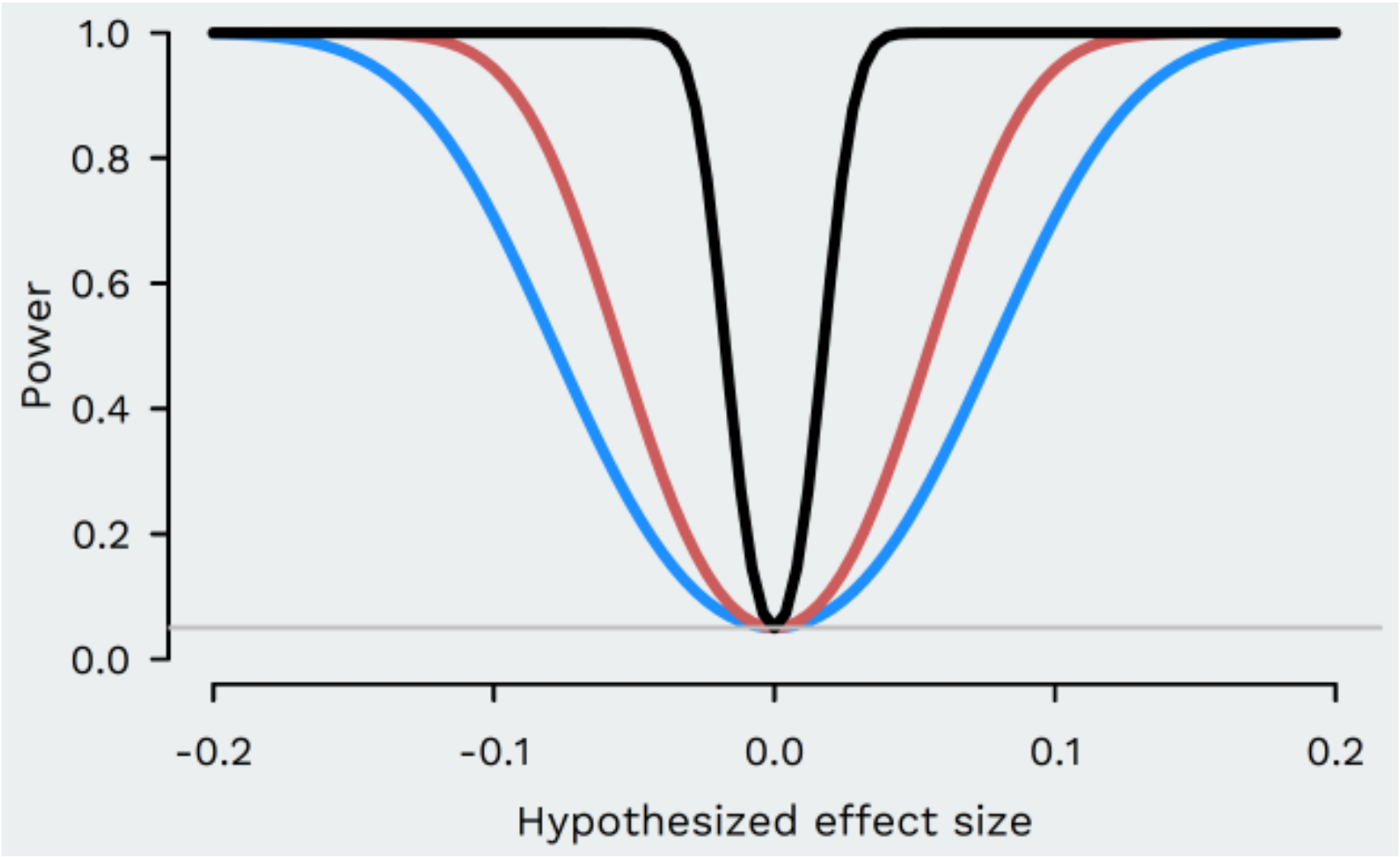

True Difference=0.05, Power=0.22
Reject
Retain
Reject
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
T

Wait! Does that mean I need to know the truth before I start to do power analysis?

Yeah. . . kind of. In practice, this is why we calculate under many possible configurations.

Power Curve

You can graph power for various possible effect sizes (here for 500, 1000, and 10000 samples).



Power calculations

- Power calculations depend on your assumed difference and the population variance, neither of which is typically known.
- Instead you may want to power depending on the minimal difference of interest with a conservative (high) estimate of population variance.
- In general power is favorable when:
  - large n
  - bigger difference (pushes the alternative distribution away)
  - smaller variance (it squeezes the distribution in)
- Power analysis is really important if you are planning experiments, but we will touch on it only cursorily in this class. The Gerber and Green Field Experiments book is an amazing resource for more on experiments in general.

1. Hypothesis Testing

2. Power

3. p-values and their Problems
   - Fun With Salmon
   - The Significance of Significance

4. Potential Outcomes

p-values

The appropriate level (α) for a hypothesis test depends on the relative costs of Type I and Type II errors.

What if there is disagreement about these costs?

We might like a quantity that summarizes the strength of evidence against the null hypothesis without making a yes or no decision.

Definition (*p*-value)

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

Under the null hypothesis, this corresponds to the probability of observing a test statistic as extreme or more extreme than the one in the observed data (where extreme is defined in terms of the alternative hypothesis).

Rejection Regions and p-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.
- How likely was it we would get a test statistics this extreme or more extreme under our null hypothesis?

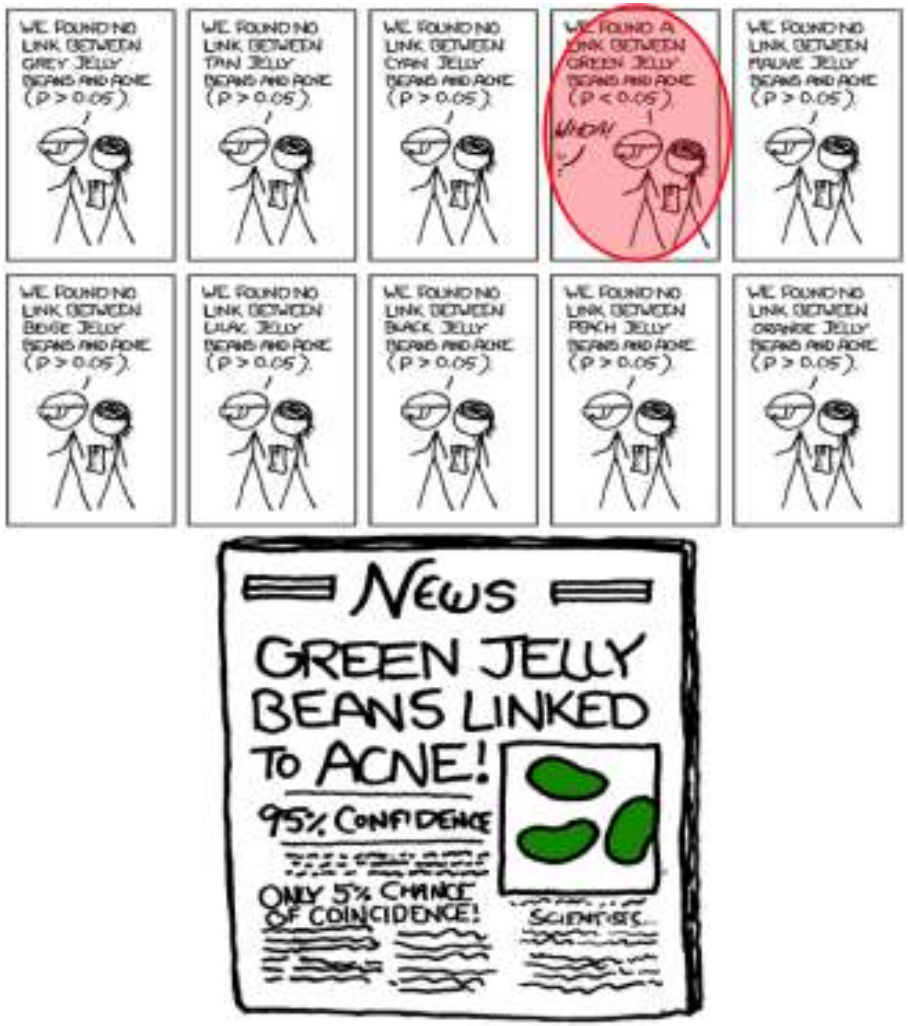
$$\begin{aligned} P_0(|T_n| > 18.5) &= P_0(T_n > 18.5) + P_0(T_n < -18.5) \\ &= 2P_0(T_n < -18.5) \end{aligned}$$

- We can get this in `R` with $2 * \texttt{pnorm}(-18.5)$
- That yields a p-value of 2.06×10^{-76}.
- By convention we would say it is statistically significant at level α for some α that the p-value is below.

All those guarantees on Type I error?

It only works for one test.

Star Chasing (aka there is an XKCD for everything)



Problem of Multiple Testing

- The multiple testing (or "multiple comparison") problem occurs when one considers a set of statistical tests simultaneously.
- Consider $k = 1, ..., m$ independent hypothesis tests (e.g. control versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast: $\alpha_{overall} = 1 - (1 - \alpha_k)^m$.
- That's right - it grows exponentially. E.g., given test 7 tests at $\alpha = .1$ level the overall type I error is .52.
- Even if all null hypotheses are true we will reject at least one of them with probability .52.
- Same for confidence intervals: probability that all 7 CI cover the true values simultaneously over repeated samples is $1 - .52$.
  So for each coefficient you have a .90 confidence interval, but overall a .52 percent confidence interval.

This all seems rather bad. Fixing this is an active research area for a lot of people.

A Sketch of Two Styles of Solutions:
(1) statistical
and
(2) procedural.

One Step Deeper: Statistical Paths Forward

- Control the Family-wise Error Rate:

$$P(\text{making at least one Type I error})$$

  - Bonferroni correction: reject the jth null hypothesis H_j if $p_j < \alpha/m$ where m is the total number of tests
    - ★ NB: VERY conservative

- Control the False Discovery Rate

$$E\Big[\frac{\#\text{ of false rejections}}{\max(\text{total }\#\text{ of rejections, } 1)}\Big]$$

  - Benjamini-Hochberg Procedure: Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
    - ★ Find the largest k such that $p_{(i)} \leq \frac{\alpha * k}{m}$ and call it k^*
    - ★ Reject all H_k for $k \leq k^*$

One Step Deeper: Benjamini-Hochberg Example

Pvalues: 0.011, 0.13, 0.06, 0.54, 0.008, 0.024, 0.001, 0.201, 0.78, 0.023

Step 1: Sort	Step 2: Find New Threshold	Step 3: Find k	Step 4: Reject
0.001	0.05*1/10 = 0.005	*	Reject
0.008	0.05*2/10 = 0.01	*	Reject
0.011	0.05*3/10 = 0.015	*	Reject
0.023	0.05*4/10 = 0.02		Reject
0.024	0.05*5/10 = 0.025	*	Reject
0.06	0.05*6/10 = 0.03		
0.13	0.05*7/10 = 0.035		
0.201	0.05*8/10 = 0.04		
0.54	0.05*9/10 = 0.045		
0.78	0.05*10/10 = 0.05		

One Step Deeper: Procedural Paths Forward

- Preregistration
  - in theory, forces people to pre-commit to analyses.
  - doesn't directly address multiple comparisons, but may credibly limit them.
  - doesn't work if pre-analysis plans aren't abided by or if people register many, many hypotheses.
- Sample Splits
  - Set-aside one sample for discovery where you can search over lots of different options.
  - A second sample can be used to test a small set of hypotheses.
  - Similar to preregistration in that it doesn't directly address multiple comparisons, but limits them.

Fun With Salmon

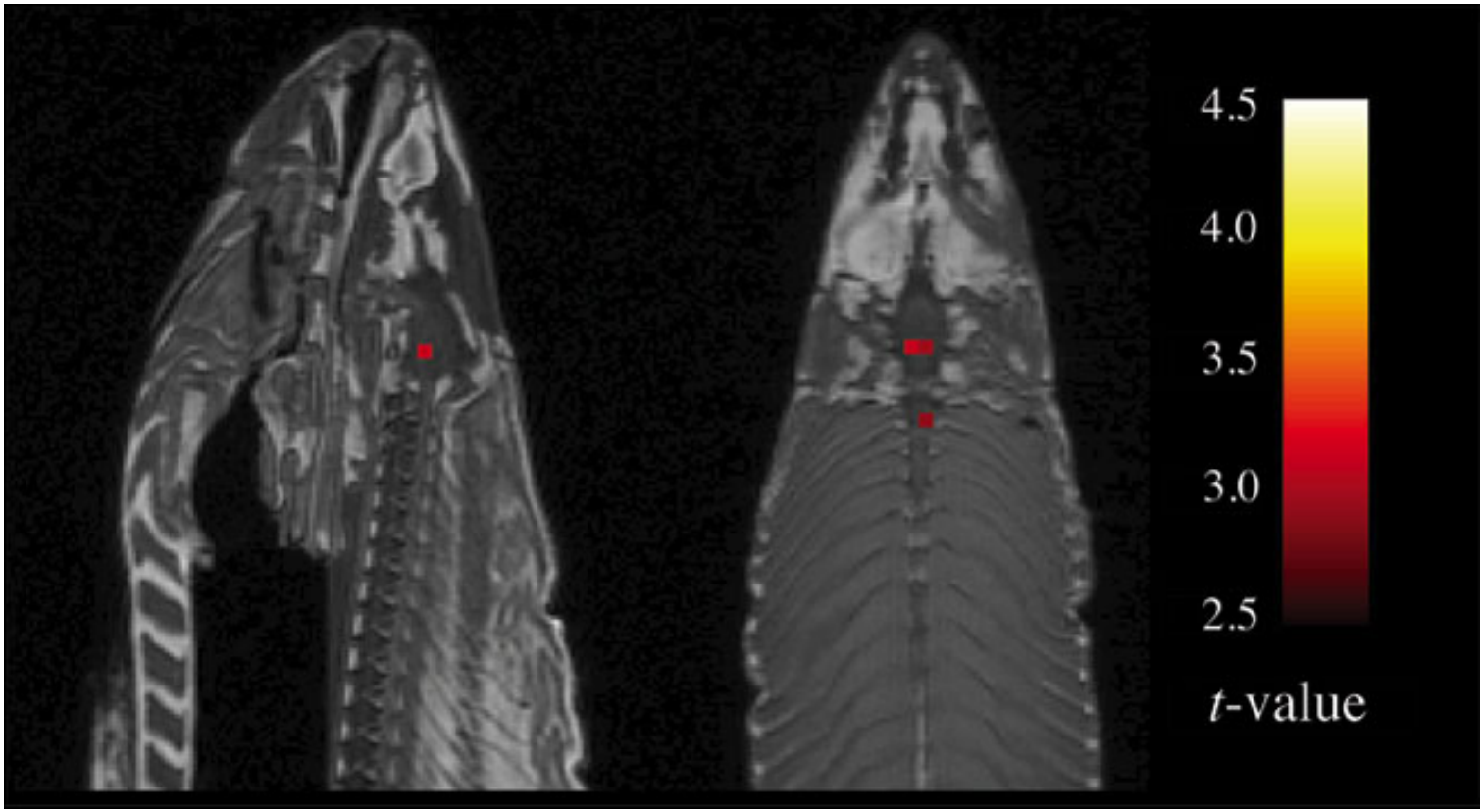
Bennett, Baird, Miller and Wolford. (2009). "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction."

Methods

(a.k.a. the greatest methods section of all time)

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."
- Design
  "Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest.A total of 15 photos were displayed. Total scan time was 5.5 minutes."

Results



"Several active voxels were discovered in a cluster located within the salmon's brain cavity. The size of this cluster was 81 mm^3 with a cluster-level significance of $p = .001$."

Okay, but what do they mean?

Practical versus Statistical Significance

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2/n}}$$

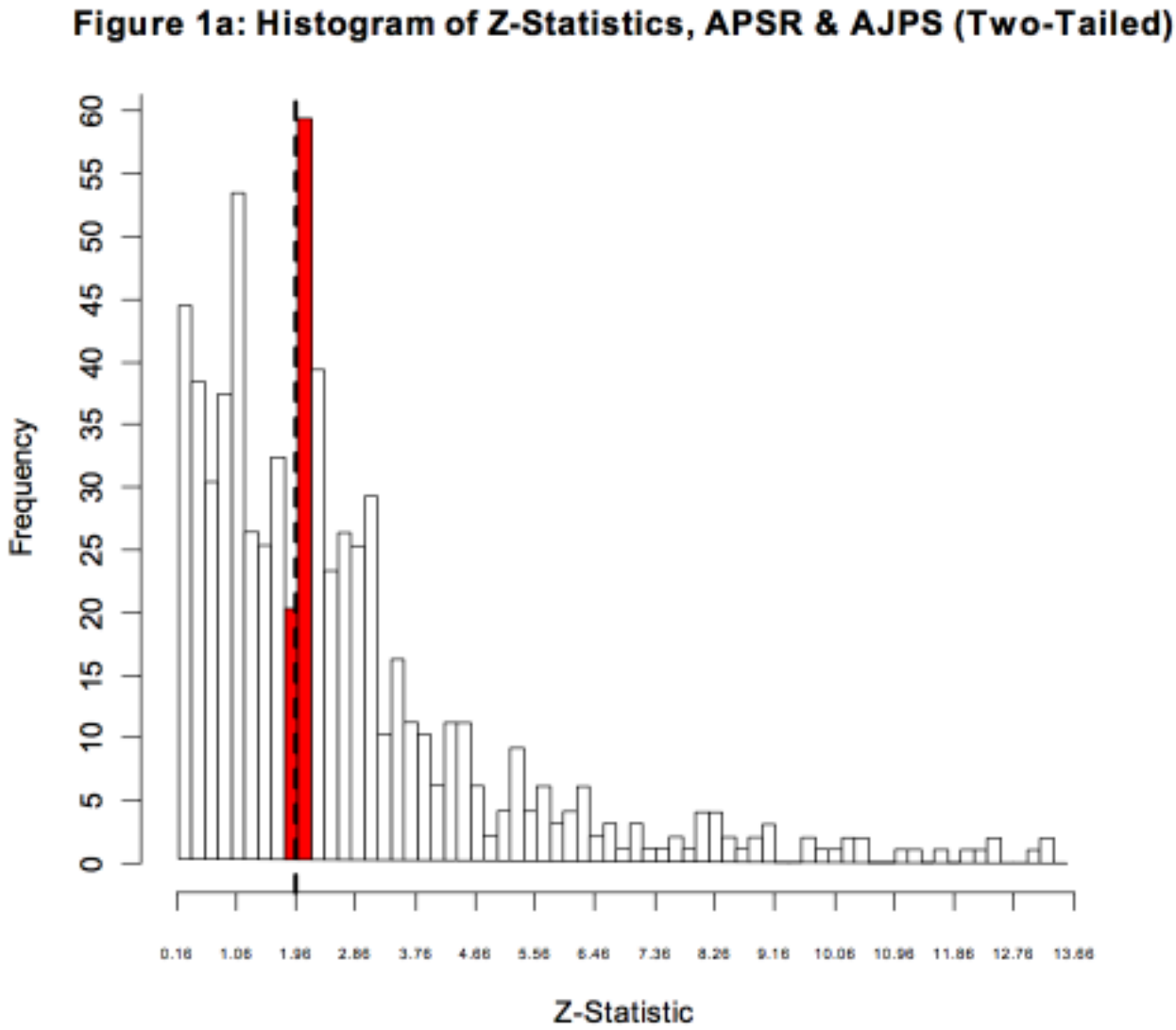
- What are the possible reasons for rejecting the null?
  1. $\overline{X} - \mu_0$ is large (big difference between sample mean and mean assumed by H_0)
  2. n is large (you have a lot of data so you have a lot of precision)
  3. $\hat{\sigma}^2$ is small (the outcome has low variability)
- We need to be careful to distinguish:
  - practical significance (e.g. a big effect)
  - statistical significance (i.e. we reject the null)
- In large samples even tiny effects will be significant, but the results may not be very important substantively. Always discuss both!

Problems with *p*-values

- *p*-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values do not:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.
  - measure the probability that the null hypothesis is true.
  - indicate how predictive a variable is of another.
  - provide a good indication of whether a paper should be published.
- a large *p*-value could mean either that we are in the null world OR that we had insufficient power.

See also: The ASA's (American Statistical Association) Statement on *p*-Values: Context, Process, and Purpose (`http://dx.doi.org/10.1080/00031305.2016.1154108`)

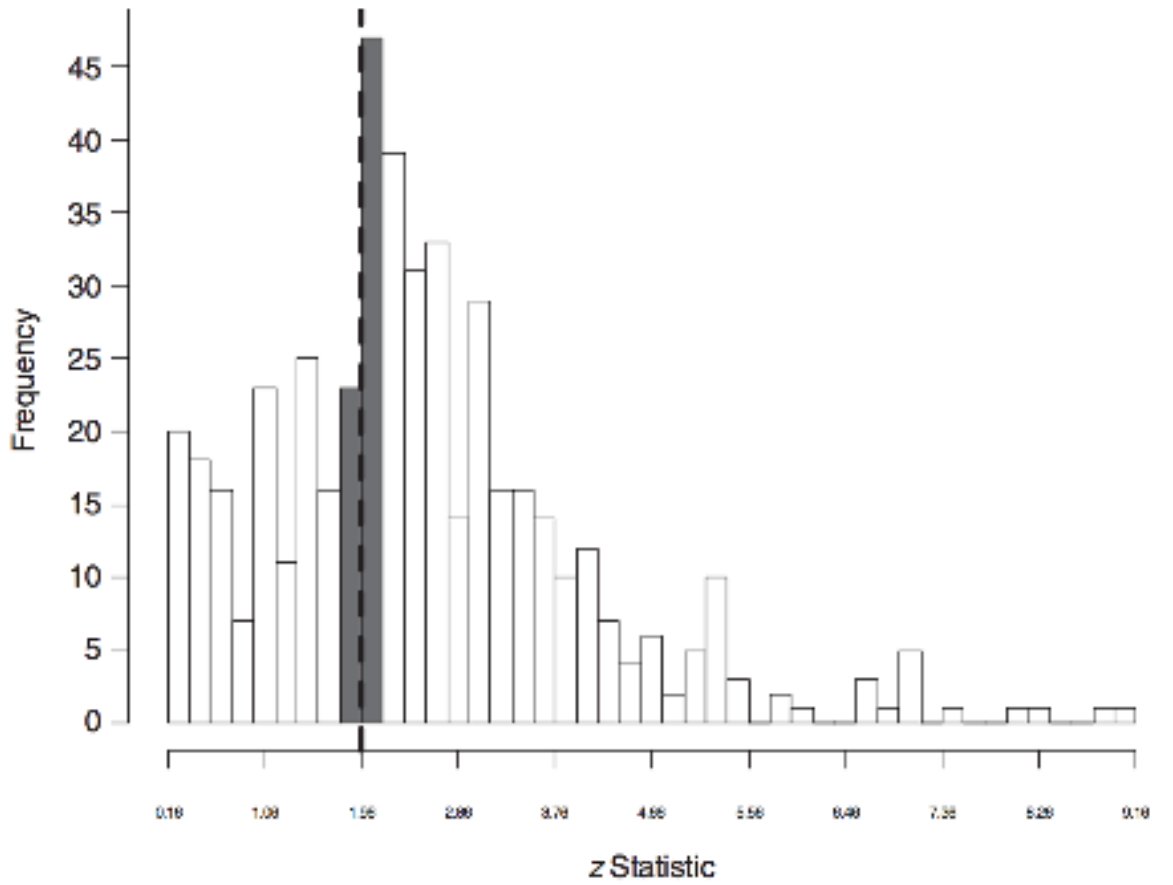
Arbitrary Publication Cutoffs



Gerber and Malhotra (2006) Top Political Science Journals

Arbitrary Publication Cutoffs

Figure 1
Histogram of z Statistics From the *American Sociological Review*, the *American Journal of Sociology*, and *The Sociological Quarterly* (Two-Tailed)



Gerber and Malhotra (2008) Top Sociology Journals

Arbitrary Publication Cutoffs

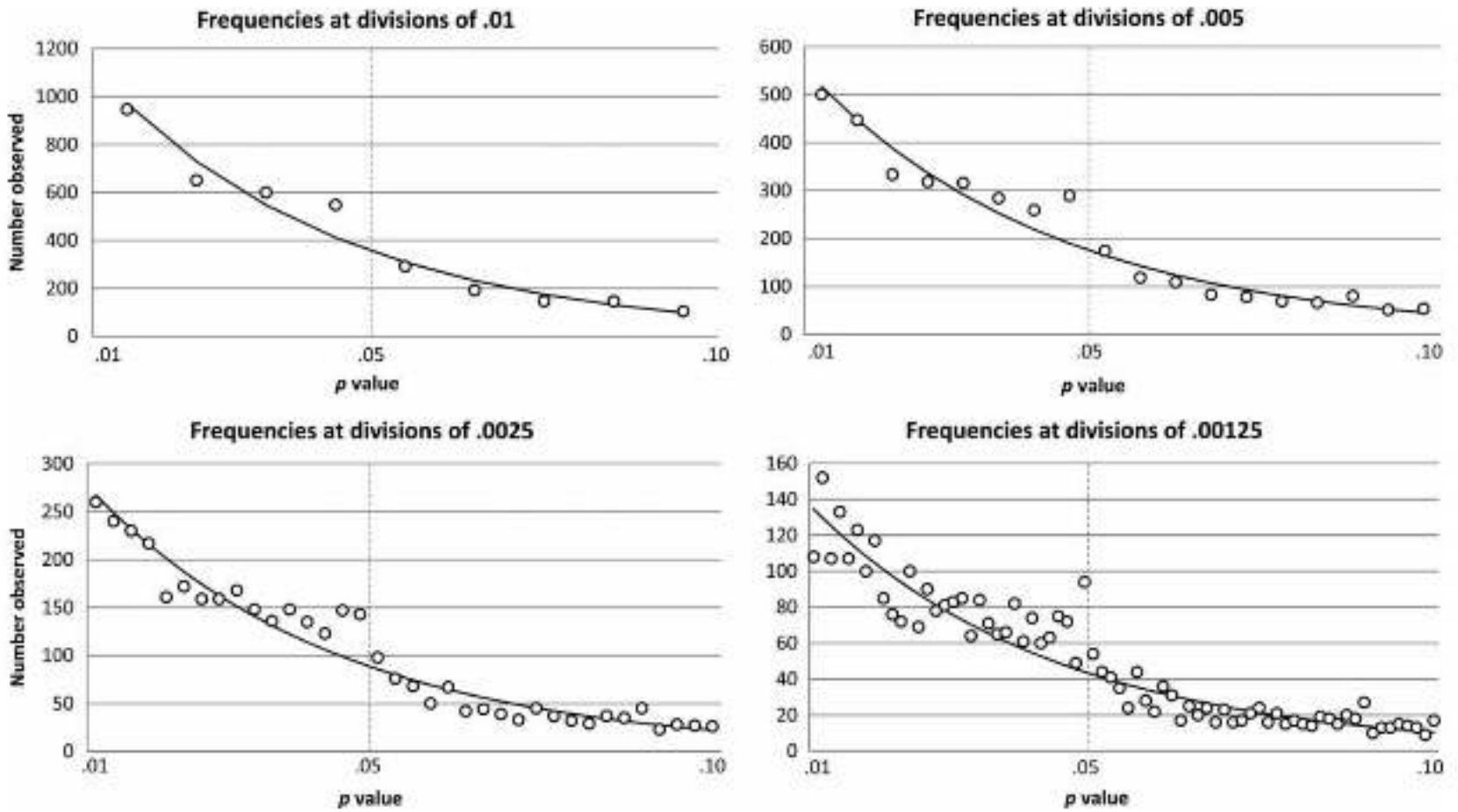


Figure 1. *The graphs show the distribution of 3,627* p *values from three major psychology journals.*

Masicampo and Lalande (2012) Top Psychology Journals

Still Not Convinced?
The Real Harm of Misinterpreted *p*-values

ELSEVIER

Accident Analysis and Prevention 36 (2004) 495–500

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Viewpoint

The harm done by tests of significance

Ezra Hauer*

35 Merton Street, Apt. 1706, Toronto, Ont., Canada M4S 3G4

Abstract

Three historical episodes in which the application of null hypothesis significance testing (NHST) led to the mis-interpretation of data are described. It is argued that the pervasive use of this statistical ritual impedes the accumulation of knowledge and is unfit for use.

Keywords: Significance; Statistical hypothesis; Scientific method

Example from Hauer: Right-Turn-On-Red

Table 1
The Virginia RTOR study

	Before RTOR signing	After RTOR signing
Fatal crashes	0	0
Personal injury crashes	43	60
Persons injured	69	72
Property damage crashes	265	277
Property damage (US$)	161243	170807
Total crashes	308	337

The Point in Hauer

- Two other interesting examples in Hauer (2004)
- Core issue is that lack of significance is not an indication of a zero effect, it could also be a lack of power (i.e. a small sample size relative to the difficulty of detecting the effect)
- On the opposite end, large tech companies rarely use significance testing because they have huge samples which essentially always find some non-zero effect. But that doesn't make the finding significant in a colloquial sense of important.

What if I need to show evidence of a zero effect?

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

An Equivalence Approach to Balance and Placebo Tests

Erin Hartman University of California Los Angeles
F. Daniel Hidalgo Massachusetts Institute of Technology

Abstract: *Recent emphasis on credible causal designs has led to the expectation that scholars justify their research designs by testing the plausibility of their causal identification assumptions, often through balance and placebo tests. Yet current practice is to use statistical tests with an inappropriate null hypothesis of no difference, which can result in equating nonsignificant differences with significant homogeneity. Instead, we argue that researchers should begin with the initial hypothesis that the data are inconsistent with a valid research design, and provide sufficient statistical evidence in favor of a valid design. When tests are correctly specified so that difference is the null and equivalence is the alternative, the problems afflicting traditional tests are alleviated. We argue that equivalence tests are better able to incorporate substantive considerations about what constitutes good balance on covariates and placebo outcomes than traditional tests. We demonstrate these advantages with applications to natural experiments.*

Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/RYNSDG.

One Step Deeper: Equivalence Tests

Solution: Flip the hypotheses:

$$H_0 : \theta_T - \theta_C \leq \epsilon_L \text{ or } \theta_T - \theta_C \geq \epsilon_U$$

versus

$$H_A : \epsilon_L < \theta_T - \theta_C < \epsilon_U$$

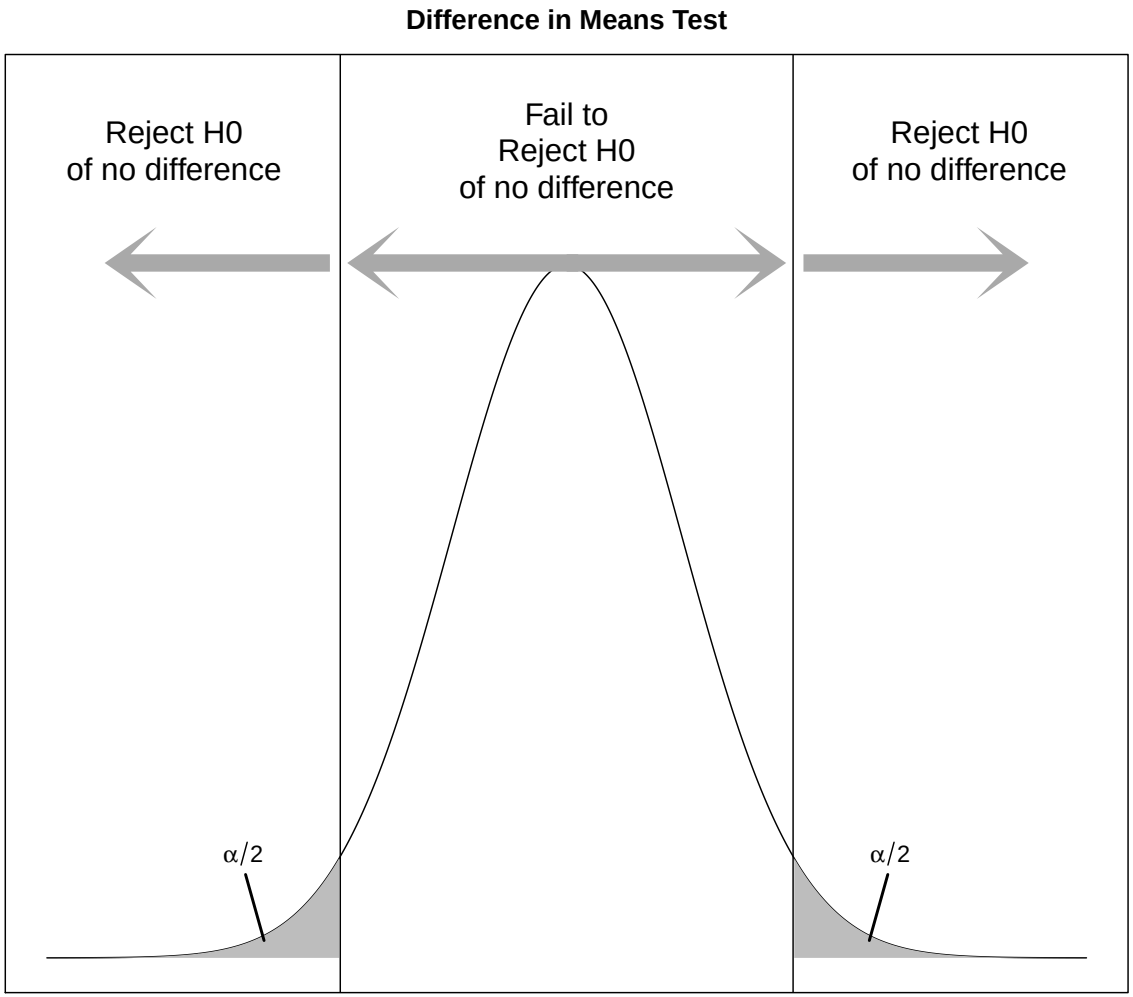
Tests of Difference vs. Equivalence Tests

$$H_0 : \frac{\mu_T - \mu_C}{\sigma} = 0 \qquad \text{versus} \qquad H_A : \frac{\mu_T - \mu_C}{\sigma} \neq 0$$

Type I Error

Test has α probability of declaring the two means different when they are, in fact, the same.

Problem: Controlling for the incorrect type of error if we're trying to provide evidence in favor of equivalence.



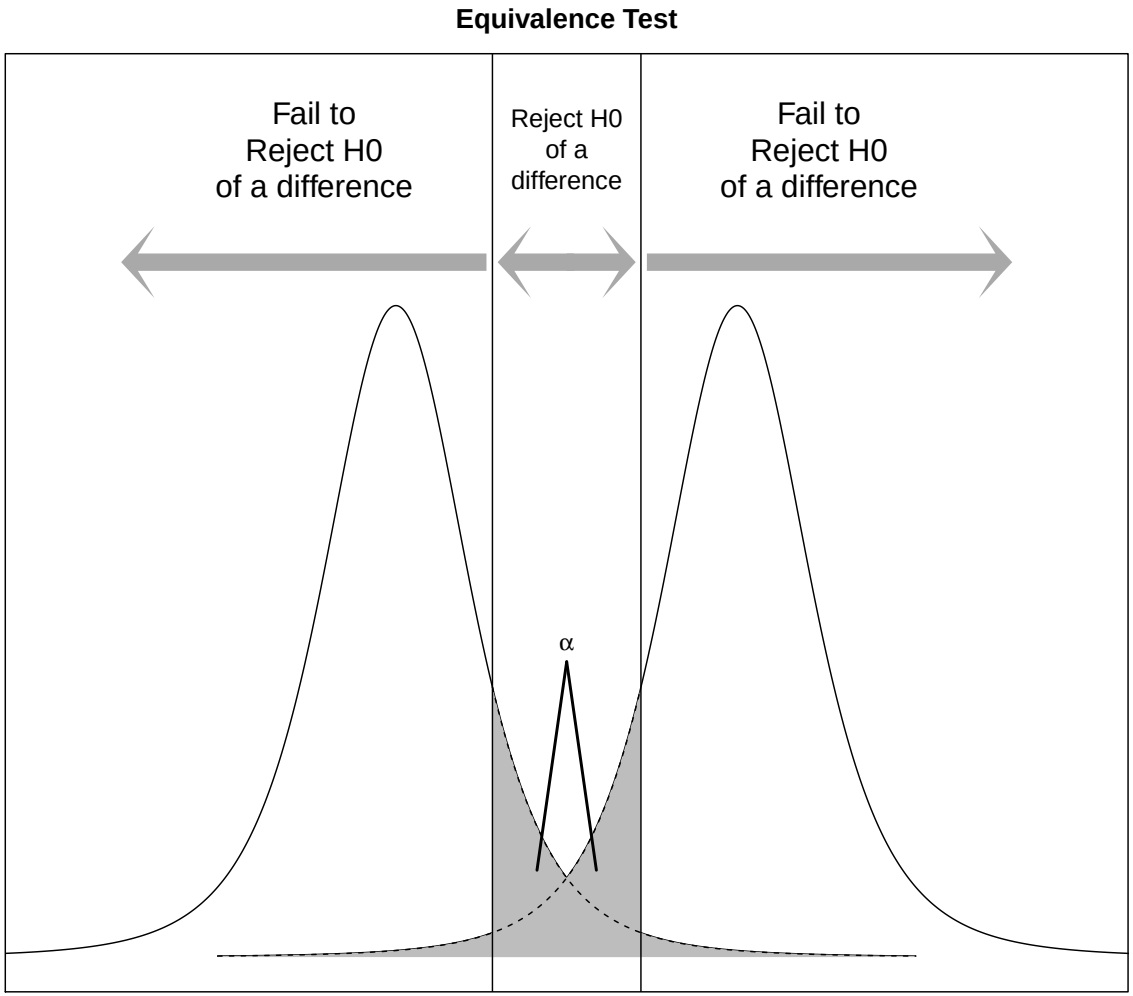
Tests of Difference vs. **Equivalence Tests**

$H_0 : \frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U$ or $\frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L$ versus $H_A : \epsilon_L < \frac{\mu_T - \mu_C}{\sigma} < \epsilon_U$

Type I Error

Test has α probability of declaring the two means equivalent when they are, in fact, the different.

Solution: Now control for the correct type of false positive.



p-values

General Message:

Don't misinterpret, or rely too heavily, on your p-values. They are evidence against your null, not evidence in favor of your alternative.

We Covered

- hypothesis testing
- power
- p-values
- multiple testing
- the problems with p-values

Next Time: Causal Inference

Bonus reading for those interested to learn more:

- Hauer. 2004 "The harm done by tests of significance." *Accident Analysis & Prevention*.
- Gigerenzer. 2004. "Mindless statistics." *Journal of Socio-Economics*.
- Nuzzo. 2014. "Statistical Errors." *Nature*
- Ward et al. 2010. "The perils of policy by p-value: Predicting civil conflicts." *Journal of Peace Research*
- Cohen. 1994. "The Earth is Round ($p < 0.05$)." *American Psychologist*
- Schwab. 2011. "Researchers should make thoughtful assessments instead of null-hypothesis significance tests." *Organizational Science*.

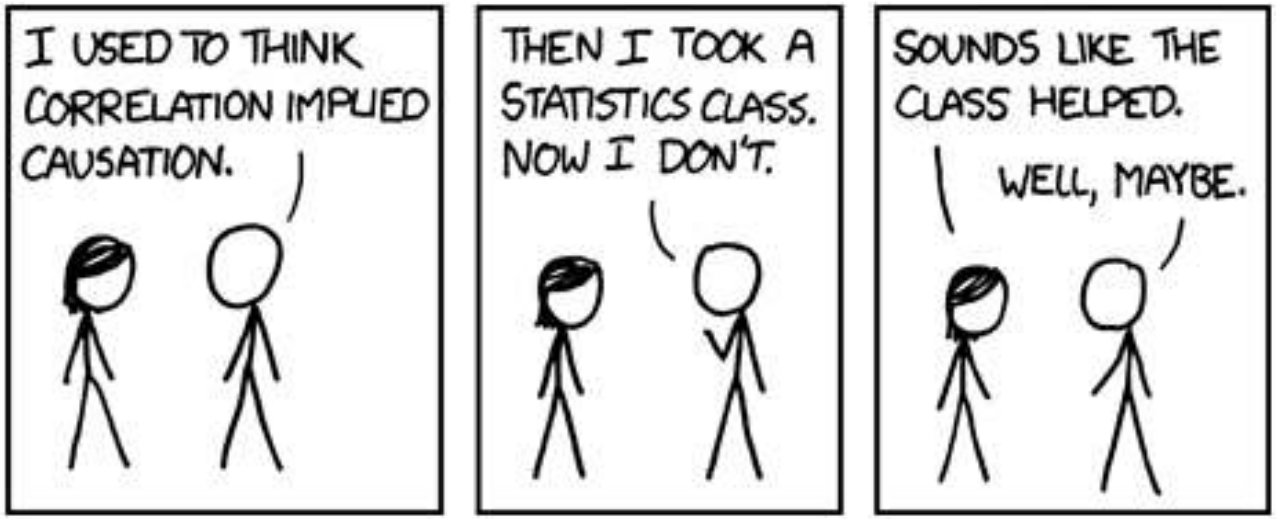
1 Hypothesis Testing

2 Power

3 p-values and their Problems
- Fun With Salmon
- The Significance of Significance

4 Potential Outcomes

Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Causal Questions

- Will advertising this product to this customer make a sale that otherwise wouldn't have happened?
- Will implementing this policy help people?
- Would sending police to this location lower crime?
- Does getting an A in this class help me get a job?

What is Causal Inference?

- Loosely, causal inference is a statement about counterfactuals—it is a statement about the difference between what would happen with an intervention and what would happen without the intervention.
- We can predict what didn't (or wouldn't) happen but, by definition, this comes from a different distribution what did (or would).
- The core puzzle of causal inference is how you get the information to predict outcomes under both states of the intervention.

Why Can't I Just Use Standard Inference?

- The distribution of what we want to know (the counterfactual) might be different from we do know (the observed).
- Consider the question of whether attending college increases lifetime earnings. We want to know:

$$\text{Average}\,(\text{Earnings if College} - \text{Earnings if not College})$$

- We could use prediction, but what are we using to do the prediction? Take person A who went to college.
  - We have: Earnings if College
  - We need: Earnings if not College
  - To predict Earnings if not College, we only have people who actually didn't go to college
- When is this okay?
- The key trick of causal inference: finding sets of data where distribution of observed is equal to distribution of counterfactual.

Fundamental Problem of Causal Inference

- The Fundamental Problem of Causal Inference Holland (1986) is that we never get to see all the values of interest because some counterfactuals are always unobserved $\rightsquigarrow$ assumptions and careful design are the only way out of this problem.
- By convention we often call the counterfactual levels we care about treated and control and we start with binary treatment variables (continuous things get much more complicated!).
- "Potential Outcomes" (more formally the Neyman-Rubin Causal Model) is one of two major frameworks that we will consider for doing causal inference.
- It is a notational system for counterfactuals and the assumptions required to make statements about them.

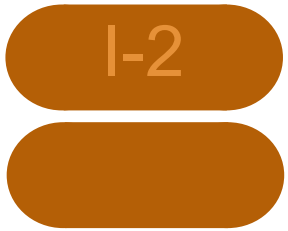
Potential Outcomes—Aspirin Example

Definitions:

T_i: Unit assigned to:

$$T_i = \begin{cases} 1 & \text{Receive Aspirin} \\ 0 & \text{Receive Placebo} \end{cases}$$

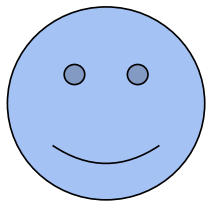
$(T_i = 1)$



$(T_i = 0)$

Y_i: Outcome for unit i – Patient has headache, or not

$(Y_i = 1)$



$(Y_i = 0)$

Potential outcomes for unit i:

$$Y_i(T_i) = \begin{cases} Y_i(1) & \text{Headache (or not) for unit } i \text{ with Aspirin} \\ Y_i(0) & \text{Headache (or not) for unit } i \text{ with placebo} \end{cases}$$

τ_i: The treatment effect

$$\tau_i = Y_i(1) - Y_i(0)$$

Illustrated potential outcomes here and later courtesy of Erin Hartman

Causal Inference is a Missing Data Problem

Example: Asprin's Impact on Headaches

Patient i		Pill T_i	Headache Status $Y_i(0)$	$Y_i(1)$	Y_i	Age X_{1i}	Academic X_{2i}
1		1	0	0	0	25	Y
2			0	1		55	N
3			1	1		62	Y
4		0	1	1	1	80	N
5		1	0	1	1	32	Y
6			1	0		45	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n		0	0	0	0	71	N

Built-in Assumptions

The notation builds in three related assumptions:

- No simultaneity
- No interference
  - We are implicitly stating that the potential outcomes for that unit are unaffected by the treatment status of other units
  - If this is not true, the number of potential outcomes for unit i grows
  - Ex: in an experiment with 3 units, if the potential outcomes for unit i depend on the treatment assignment of units j and k, the potential outcomes for unit i are defined by $Y(i, j, k)$:

$$\begin{array}{ll} Y(1,0,0) & Y(0,0,0) \\ Y(1,1,0) & Y(0,1,0) \\ Y(1,0,1) & Y(0,0,1) \\ Y(1,1,1) & Y(0,1,1) \end{array}$$

- Same version of the treatment

How Do We Proceed?

Combined, the previous assumptions give us

- **S**table **U**nit **T**reatment **V**alue **A**ssumption (SUTVA)
- Potential violations:
  - feedback effects
  - spill-over effects, carry-over effects
  - different treatment administration

Some Estimands of Interest

- **Sample average treatment effect (SATE)**
  $\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$
- **Population average treatment effect (PATE)**
  $\frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$
- **Population average treatment effect for the treated (PATT)**
  $\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$
- **Population conditional average treatment effect (CATE)**
  $\mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)$
- Treatment effect heterogeneity: Zero ATE doesn't mean zero effect for everyone

Identification vs. Estimation

Identification: How much can you learn about the estimand if you have an infinite amount of data?

> Suppose there is a counterfactual distribution $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), T_i, X_i\}$ and observational distribution $\mathbb{P}$ of $\{Y_i, R_i, X_i\}$, then a causal query $\Psi(\mathbb{P}^*)$ is identified when we can write quantity as a function of $\mathbb{P}$

Estimation: How much can you learn about the estimand from a finite sample?

> Estimation is about our ability to estimate the necessary function of $\mathbb{P}$ from finite data.

Identification precedes estimation and is based on (hopefully minimal) assumptions.

Ignorability

Identification by randomization:

- If treatment is randomized, then treatment is unrelated to any and all underlying characteristics, observed and unobserved (and even unknown)
- Randomization therefore means treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, i.e.

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$$

- This is sometimes called unconfoundedness, ignorability, or exchangeability.
- Another way of thinking of it: The distributions of the potential outcomes $(Y_i(1),\ Y_i(0))$ are the same for the treatment and control group.

How Do We Proceed?

Identification by randomization:

- Assume a completely randomized trial (fixed number of units assigned to treatment)
- Assumption: for all $i = 1, \ldots, n, \sum_{i=1}^{n} T_i = n_t$ and

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i, \qquad Pr(T_i = 1) = \frac{n_t}{n}, \qquad n = n_t + n_c$$

- Estimand of interest: SATE
- Estimator: Difference-in-means:

$$\begin{aligned}\hat{\tau} &= \frac{1}{n_t}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_c}\sum_{i=1}^{n}(1 - T_i) Y_i \\ &= \frac{1}{n_t}\sum_{i=1}^{n} T_i Y_i(1) - \frac{1}{n_c}\sum_{i=1}^{n}(1 - T_i) Y_i(0)\end{aligned}$$

Unbiasedness holds in sample just by randomization and SUTVA!

Design-based Inference for SATE

Variance of $\hat{\tau}$:

$$V[\hat{\tau}] = \frac{1}{n}\left(\frac{n_c}{n_t}S_1^2 + \frac{n_t}{n_c}S_0^2 + 2S_{0,1}\right)$$

where for $t = 0, 1$,

$$S_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(t) - \overline{Y(t)})^2 \qquad \text{sample variance of } Y_i(t)$$

$$S_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(Y_i(1) - \overline{Y(1)})(Y_i(0) - \overline{Y(0)})\Big)$$

Is the variance of our estimator is identifiable?

Conservative estimator:

$$\hat{V}[\hat{\tau}] \leq \frac{\hat{S}_1^2}{n_t} + \frac{\hat{S}_0^2}{n_c}$$

How Do We Proceed Without Randomization?

Identification by conditional independence:

- If treatment is not randomized, then treatment may be related underlying characteristics, observed and unobserved, which are related to the potential outcomes
- Therefore, we need to assume that treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, conditional on some pre-treatment characteristics X, i.e.

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$$

- Conditioning set should yield $Y_i(0)$, $Y_i(1)$ and T_i conditionally independent. (This is next week's topic).
- This is conditional ignorability.

The Selection Problem

- Why is this difficult? selection bias
- The core idea is that the people who get treatment might look different from those who get control and thus they are not good counterfactuals for each other.
- Let's look at what we get from a naive difference in means with a binary treatment:

$$
\begin{aligned}
&E[Y_i|T_i=1] - E[Y_i|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=1] + E[Y_i(0)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = \underbrace{E[Y_i(1) - Y_i(0)|T_i=1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|T_i=1] - E[Y_i(0)|T_i=0]}_{\text{selection bias}}
\end{aligned}
$$

- Naive estimator = Average Treatment Effect on Treated + Selection Bias
- Selection bias: how different the treated and control groups are in terms of their potential outcome under control.

Selection Makes Us Care About Assignment Mechanisms

Assignment Mechanism

"The process that determines which units receive which treatments, hence which potential outcomes are realized and thus can be observed, and, conversely, which potential outcomes are missing."
(Imbens and Rubin, 2015, p. 31)

Types of assignment mechanisms:

- random assignment
- selection on observables
- selection on unobservables

What Gets to Be a Cause?

We can imagine a world where individual i is assigned to treatment and control conditions

What is the Hypothetical Experiment?

Problem: Immutable (or difficult to change) characteristics

- Effect of gender on promotion
- Effect of race on traffic stops

Consider causal effect of race on traffic stops:

- Do we mean effect of officer perceiving a certain race?
- Do we mean randomly assigning race at birth?
- manipulating perceptions is a lot different from manipulating the characteristic

No Causation Without Manipulation

Caveats and Implications

- Does not dismiss claims of discrimination on immutable characteristics as illegitimate
  - Pervasive effects of racism/sexism in society
  - Suggests: we need a different empirical strategy to evaluate claims
  - What facet of institutionalized racism (or its consequences) causes racial disparities?
- Correlation problem :
  - Regression models can estimate coefficients for immutable characteristics
  - But are necessarily imprecise: what do scholars have in mind in models?
- Design Principle:
  - Pretend you could design any experiment
  - If that experiment does not exist, be concerned about interpretation

Summing Up: Neyman-Rubin causal model

- Useful for studying the "effects of causes", less so for the "causes of effects".
- No assumption of homogeneity, allows for causal effects to vary unit by unit
  - No single "causal effect", thus the need to be precise about the target estimand. (This is true even for perfect experiments.)
- Distinguishes between observed outcomes and potential outcomes.
- Causal inference is a missing data problem: we typically make assumptions about the assignment mechanism to go from descriptive inference to causal inference.

This Week in Review

- Hypothesis Testing!
- P-Values!
- Potential Outcomes!

Going Deeper:

> Aronow and Miller (2019) *Foundations of Agnostic Statistics*. Cambridge University Press. Chapter 4.

Next week: More frameworks for causal inference.