431 Class 21

Thomas E. Love

2017-11-09

Today's Agenda

- Answer Sketches for the Airline Etiquette Exercises
- Two "Exciting" New Results worth a closer look
  - Stents
  - Advil + Tylenol in the ED for acute pain
- A little more on the problems with p values
  - Statistical Significance doesn't have to be about p values
  - Confidence intervals as a partial solution journals like
  - Researcher Degrees of Freedom
  - The Garden of Forking Paths

Today's R Setup

```
library(Epi); library(magrittr)
library(forcats); library(tidyverse)

fly <- fivethirtyeight::flying %>%
  select(id = respondent_id, recline_frequency,
         recline_rude, unruly_child,
         have_kids = children_under_18) %>%
  mutate(have_kids = factor(have_kids)) %>%
  filter(complete.cases(.))

source("Love-boost.R")
```

Airplane Etiquette Survey: My Answers

Airplane Etiquette Survey

https://fivethirtyeight.com/features/airplane-etiquette-recline-seat/

```
summary(select(fly, unruly_child, have_kids,
               recline_rude, recline_frequency))
```

```
  unruly_child have_kids     recline_rude
No       :146   FALSE:657   No       :498
Somewhat:348   TRUE :188   Somewhat:279
Very    :351                Very    : 68


          recline_frequency
Never                 :166
Once in a while       :254
About half the time:116
Usually               :175
Always                :134
```

Exercise 1

1. Estimate a 90% confidence interval for the proportion of people answering either "Somewhat" or "Very" to the question of whether it is rude to knowingly bring an unruly child on a plane. What is the margin of error?

```
fly %$% table(unruly_child) %>% addmargins
```

```
unruly_child
      No Somewhat     Very      Sum
     146      348      351      845
```

Our sample probability of ("Somewhat" or "Very") is (348 + 351) / 845 = 699 / 845 = 0.827.

Exercise 1 (continued)

We could use `binom.test` to calculate the 90% CI.

```
prop.test(x = 699, n = 845, conf.level = 0.90)
```

```
    1-sample proportions test with continuity
    correction

data:  699 out of 845, null probability 0.5
X-squared = 360.6, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.8041724 0.8481192
sample estimates:
        p
0.8272189
```

Exercise 1 (continued)

In fact, we know of at least three reasonable approaches.

Approach	90% CI	half-width
`prop.test`	(0.804, 0.848)	0.022
`binom.test`	(0.804, 0.848)	0.022
`saifs.ci`	(0.805, 0.849)	0.022

In each case, the confidence interval's width is 0.044, and so the margin for error is approximately 0.022 (note that the confidence intervals we've fit aren't symmetric around the point estimate.)

Exercise 2

2. Does the proportion of people who feel it is "Somewhat" or "Very" rude to knowingly bring an unruly child on a plane show a significant association with whether or not they themselves have children under 18 years of age?

```
fly %$% table(have_kids, unruly_child) %>% addmargins
```

```
           unruly_child
have_kids   No Somewhat Very Sum
    FALSE   96      251  310 657
    TRUE    50       97   41 188
    Sum    146      348  351 845
```

We'd like to rearrange this by collapsing the "Somewhat" and "Very" categories and moving the result left, and it might be nice to move "TRUE" to the top row, so as to approximate standard epidemiological format.

Exercise 2 (data reshaping)

So, some data reshaping. . .

```
fly1 <- fly %>%
  mutate(kid_rude =
           fct_collapse(unruly_child,
                        yes = c("Somewhat", "Very"),
                        no = "No"),
         kid_rude = fct_relevel(kid_rude, "yes"),
         have_kids = fct_relevel(have_kids,
                                 "TRUE"))
```

Exercise 2 (revised table)

```
fly1 %$% table(have_kids, kid_rude) %>% addmargins
```

```
          kid_rude
have_kids yes  no Sum
    TRUE  138  50 188
    FALSE 561  96 657
    Sum   699 146 845
```

Now, we apply the twoby2 function from Epi...

```
twoby2(fly1 %$% table(have_kids, kid_rude))
```

Exercise 2 (`twoby2` results)

```
2 by 2 table analysis:
------------------------------------------------------
Outcome    : yes
Comparing : TRUE vs. FALSE

      yes no    P(yes) 95% conf. interval
TRUE  138 50    0.7340    0.6663   0.7923
FALSE 561 96    0.8539    0.8248   0.8789

                                   95% conf. interval
            Relative Risk:  0.8597    0.7844   0.9422
        Sample Odds Ratio:  0.4723    0.3200   0.6971
Conditional MLE Odds Ratio:  0.4728    0.3153   0.7139
    Probability difference: -0.1198   -0.1917  -0.0550

             Exact P-value: 3e-04
       Asymptotic P-value: 2e-04
------------------------------------------------------
```

Exercise 3

3. Given the actual data, what can you conclude about the true proportion of people who feel it is rude to recline your seat on a plane?

```
fly %>% count(recline_rude)
```

```
# A tibble: 3 x 2
  recline_rude     n
        <fctr> <int>
1           No   498
2     Somewhat   279
3         Very    68
```

It looks like 347 (279 + 68) respondents are in the "Somewhat" or "Very" category. That's 41.1% of the 845 respondents.

Exercise 3 (SAIFS and other confidence intervals)

```
saifs.ci(x = 347, n = 845)
```

```
Sample Proportion             0.025             0.975
            0.411             0.377             0.445
```

The 95% CI from the `prop.test` and `binom.test` (without Bayesian augmentation) are also (0.377, 0.445)

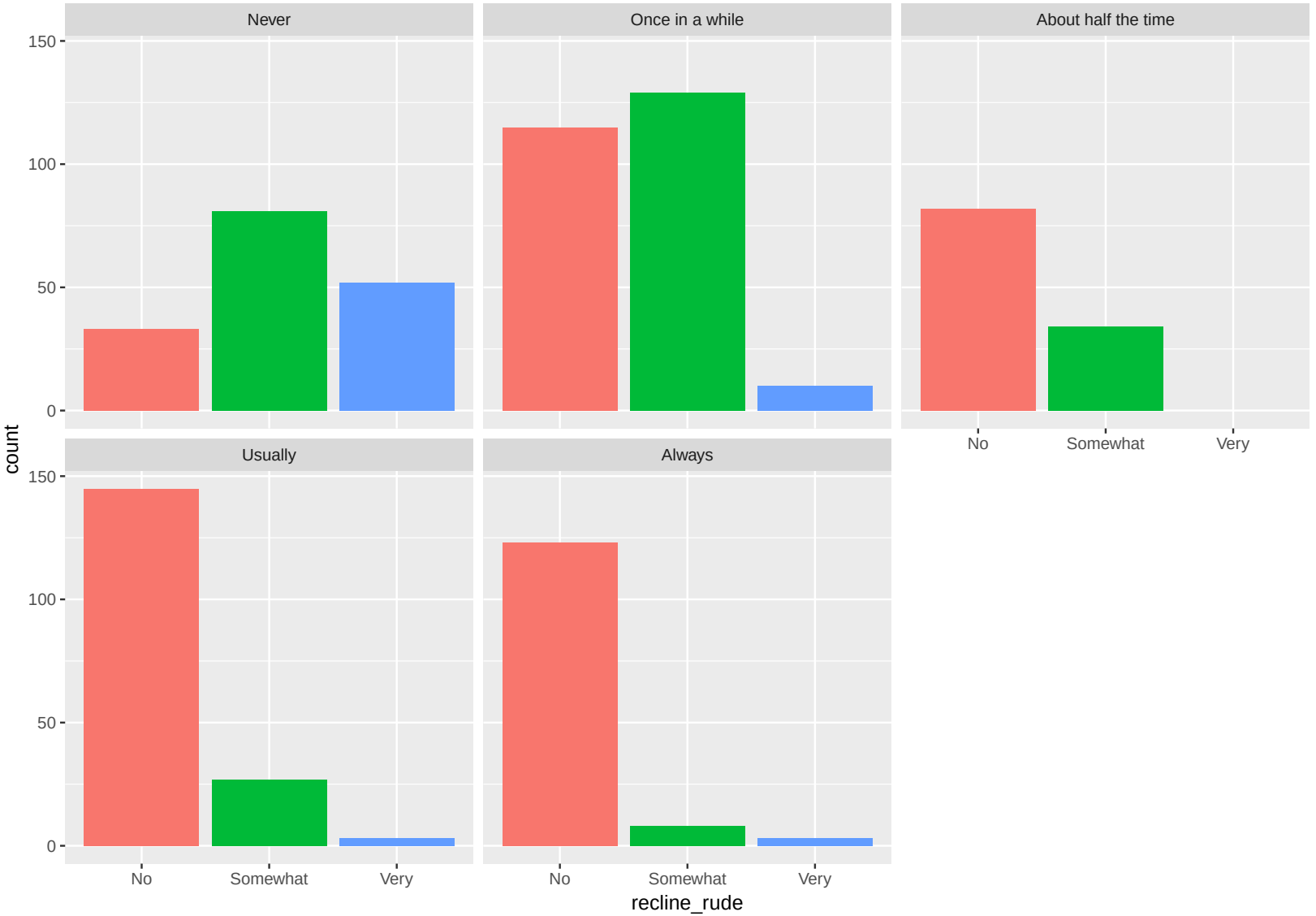
Exercise 4

4. Is there an association between how often you recline and your feelings about how rude it is?

```
fly %$% table(recline_rude, recline_frequency) %>% addmargins
```

```
             recline_frequency
recline_rude Never Once in a while About half the time
    No          33              115                  82
    Somewhat    81              129                  34
    Very        52               10                   0
    Sum        166              254                 116
             recline_frequency
recline_rude Usually Always Sum
    No           145    123 498
    Somewhat      27      8 279
    Very           3      3  68
    Sum          175    134 845
```

Exercise 4 (graph)



Exercise 4 (initial chi-square test)

```
fly %$% table(recline_rude, recline_frequency) %>% chisq.test
```

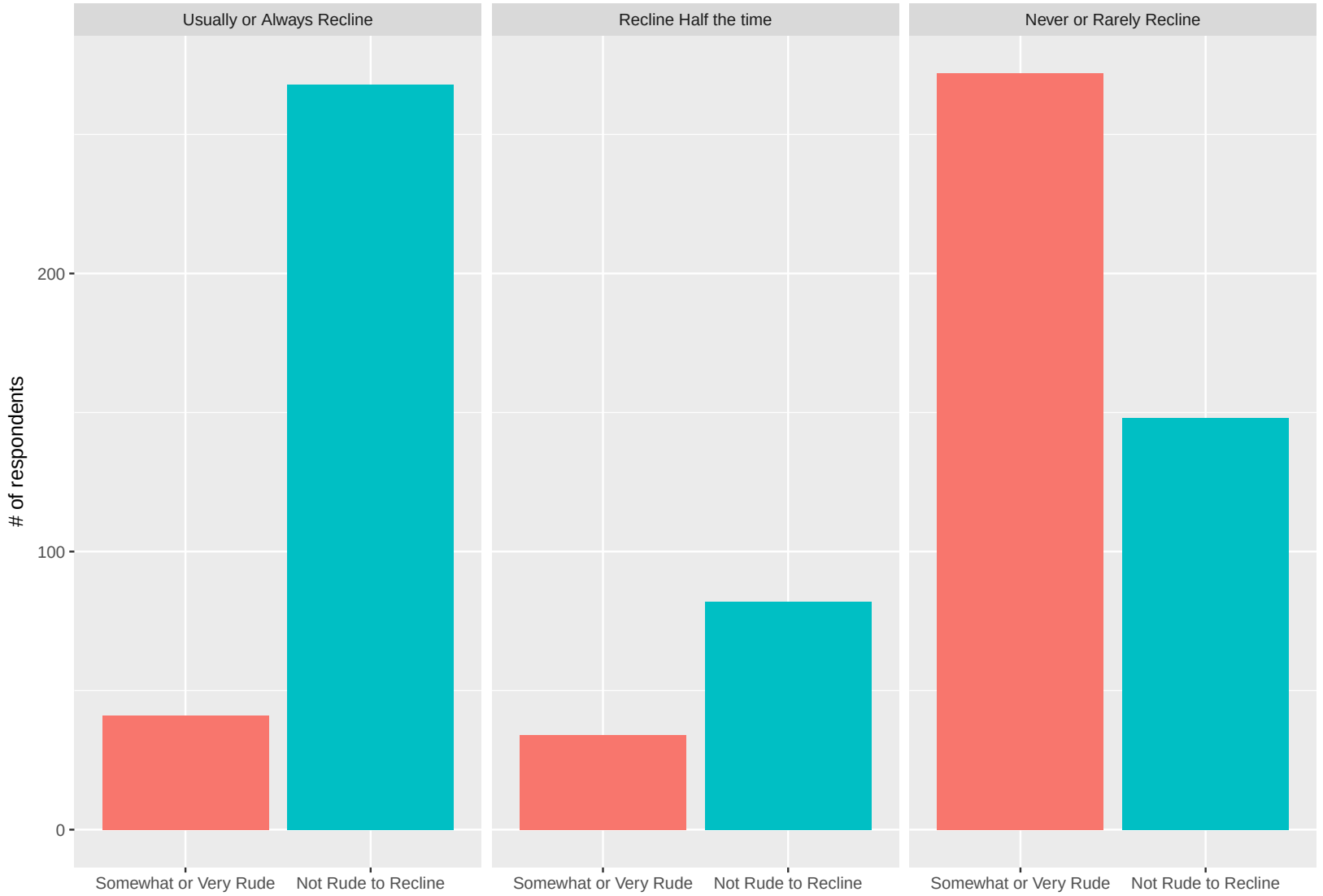
```
    Pearson's Chi-squared test

data:  .
X-squared = 319.42, df = 8, p-value < 2.2e-16
```

Exercise 4 (collapsing the table)

```
fly3 <- fly %>%
 mutate(rude =
  fct_collapse(recline_rude,
   "Somewhat or Very Rude" = c("Somewhat", "Very"),
   "Not Rude to Recline" = "No"),
  rude = fct_relevel(rude, "Somewhat or Very Rude"),
  behavior = fct_collapse(recline_frequency,
   "Usually or Always Recline" = c("Usually", "Always"),
   "Recline Half the time" = "About half the time",
   "Never or Rarely Recline" =
     c("Never", "Once in a while")),
  behavior = fct_relevel(behavior,
   "Usually or Always Recline",
   "Recline Half the time"))
```

Exercise 4 (graph, after collapsing)



Exercise 4 (table, after collapsing)

```
fly3 %$% table(behavior, rude) %>% addmargins
```

```
                                rude
behavior                         Somewhat or Very Rude
  Usually or Always Recline                         41
  Recline Half the time                             34
  Never or Rarely Recline                          272
  Sum                                              347
                                rude
behavior                         Not Rude to Recline Sum
  Usually or Always Recline                      268 309
  Recline Half the time                           82 116
  Never or Rarely Recline                        148 420
  Sum                                            498 845
```

OK - we're ready for a chi-square test.

Exercise 4 (chi-square test)

```
fly3 %$% table(behavior, rude) %>% chisq.test
```

```
    Pearson's Chi-squared test

data:  .
X-squared = 202.72, df = 2, p-value < 2.2e-16
```

Exercise 5

Suppose we wish to estimate the power a study will have to estimate the difference in proportion of people who feel that waking someone up to go for a walk is very or somewhat rude, comparing taller people to shorter people. Suppose we propose a new study, where we will collect data from 1200 tall and 1200 short people, and we look to declare as important any observed difference where one group is at 73% or more, while the other is at 70% or less.

5. Using a 10% significance level, what power will we have?

```
Two-sample comparison of proportions power calculation

              n = 1200
             p1 = 0.7
             p2 = 0.73
      sig.level = 0.1
          power = 0.4932237
```

Exercise 5 Result

```
power.prop.test(n = 1200, p1 = 0.70, p2 = 0.73,
                sig.level = 0.10)
```

```
     Two-sample comparison of proportions power calculation

              n = 1200
             p1 = 0.7
             p2 = 0.73
      sig.level = 0.1
          power = 0.4932237
    alternative = two.sided

NOTE: n is number in *each* group
```

Exercise 6

6. To obtain at least 80% power, how big a sample would we need?

```
power.prop.test(p1 = 0.70, p2 = 0.73,
                sig.level = 0.10, power = 0.80)
```

```
     Two-sample comparison of proportions power calculation

              n = 2798.621
             p1 = 0.7
             p2 = 0.73
      sig.level = 0.1
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

What's in the news?

Al-Lamee R et al. ORBITA: A Double-blind RCT of Cardiac Stents (*The Lancet* 2017-11-02)

SEARCH

The New York Times

Break 'h' to

Election Results Invigorate Medicaid Expansion Hopes

Are Mass Murderers Insane? Usually Not, Researchers Say

VOICES

A Gay Husband, a Dire Diagnosis and the Best-Laid Plans

HEALTH

'Unbelievable': Heart Stents Fail to Ease Chest Pain

Leer en español

By GINA KOLATA NOV. 2, 2017

Al-Lamee R et al *Lancet* 2017-11-02

Percutaneous coronary intervention in stable angina (ORBITA): a double-blind, randomised controlled trial

Rasha Al-Lamee, MRCP, David Thompson, MRCPI, Hakim-Moulay Dehbi, PhD, Sayan Sen, MRCP, Kare Tang, FRCP, John Davies, MRCP, Thomas Keeble, MRCP, Michael Mielewczik, PhD, Raffi Kaprielian, FRCP, Iqbal S Malik, FRCP, Sukhjinder S Nijjer, MRCP, Ricardo Petraco, MRCP, Christopher Cook, MRCP, Yousif Ahmad, MRCP, James Howard, MRCP, Christopher Baker, FRCP, Andrew Sharp, FRCP, Robert Gerber, FRCP, Suneel Talwar, MRCP, Ravi Assomull, MRCP, Prof Jamil Mayet, FRCP, Roland Wensel, MRCP, David Collier, PhD, Matthew Shun-Shin, MRCP, Prof Simon A Thom, FRCP, Dr Justin E Davies, MRCP, Prof Darrel P Francis, FRCP on behalf of the ORBITA investigators[†]

New York Times (2017-11-02) by Gina Kolata

For the study, Dr. Justin E. Davies, a cardiologist at Imperial College London, and his colleagues recruited 200 patients with a profoundly blocked coronary artery and chest pain severe enough to limit physical activity, common reasons for inserting a stent.

All were treated for six weeks with drugs to reduce the risk of a heart attack, like aspirin, a statin and a blood pressure drug, as well as medications that relieve chest pain by slowing the heart or opening blood vessels.

Then the subjects had a procedure: a real or fake insertion of a stent. This is one of the few studies in cardiology in which a sham procedure was given to controls who were then compared to patients receiving the actual treatment.

In both groups, doctors threaded a catheter through the groin or wrist of the patient and, with X-ray guidance, up to the blocked artery. Once the catheter reached the blockage, the doctor inserted a stent or, if the patient was getting the sham procedure, simply pulled the catheter out.

Al-Lamee R et al *Lancet* 2017-11-02

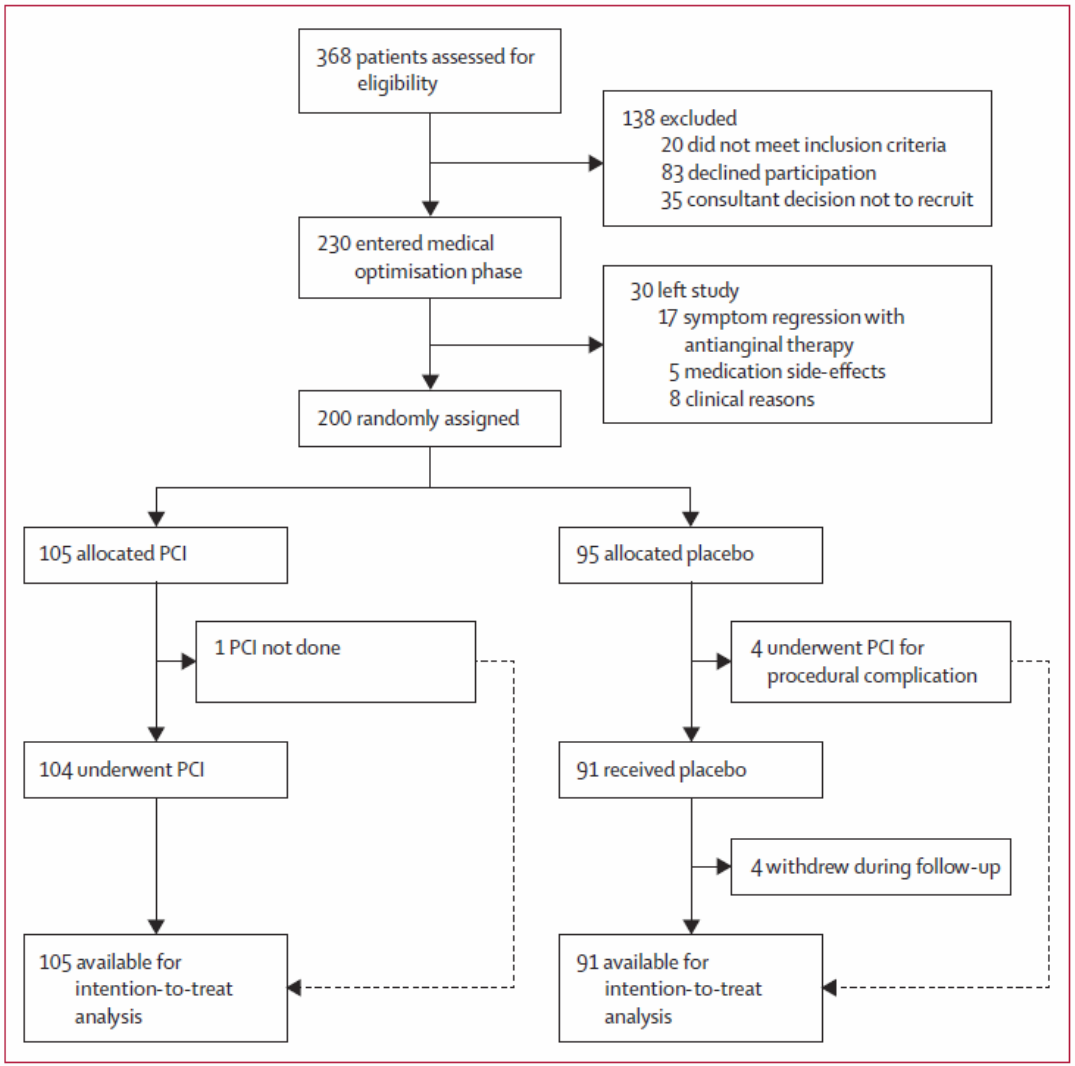


Figure 2: **Trial profile**
PCI=percutaneous coronary intervention.

Al-Lamee R et al *Lancet* 2017-11-02

Primary (pre-specified) endpoint: difference between PCI and placebo groups in the change in treadmill exercise time.

Design parameters: anticipated an effect size of 30 seconds (less than what you'd get from a single antianginal medication), and assumed a between-patient standard deviation of change in exercise time of 75 seconds.

What is the **power** of a sample of, say, 100 patients per group to detect such a difference between the PCI and placebo groups at the 5% significance level, with a two-sided test?

ORBITA trial power?

```
power.t.test(n = 100, delta = 30, sd = 75, sig.level = 0.05)
```

```
     Two-sample t test power calculation

              n = 100
          delta = 30
             sd = 75
      sig.level = 0.05
          power = 0.8036466
    alternative = two.sided

NOTE: n is number in *each* group
```

Al-Lamee R et al *Lancet* 2017-11-02

Statistical analysis

The primary endpoint of ORBITA was the difference between PCI and placebo groups in the change in treadmill exercise time. Single antianginal agents have been found to increase treadmill exercise time by 48–55 s more than placebo.[18,19] We designed ORBITA conservatively, to detect an effect size from invasive PCI of 30 s, smaller than that of a single antianginal agent. We calculated that, from the point of randomisation, a sample size of 100 patients per group had more than 80% power to detect a between-group difference in the increment of exercise duration of 30 seconds, at the 5% significance level, using the two-sample t test of the difference between groups. This calculation assumed a between-patient standard deviation of change in exercise time of 75 s. There have been no previous placebo-controlled trials of PCI. We therefore initially allowed for a one-third dropout rate in the 6-week period of medical optimisation between enrolment and randomisation and therefore planned to enrol 300 patients. In fact, the dropout rate was much lower so only 230 patients had to be enrolled before 200 participants had been randomised.

ORBITA primary result?

- n = 105 randomized to PCI and 95 randomized to placebo (ITT).
- Observed difference between the groups was 16.6 (95% CI 8.9 - 42.0)

	PCI	Placebo
Exercise time (s)		
Patients assessed	104	90
Pre-randomisation	528·0 (178·7)	490·0 (195·0)
Follow-up	556·3 (178·7)	501·8 (190·9)
Increment (pre-randomisation to follow-up)	28·4 (95% CI 11·6 to 45·1)	11·8 (95% CI –7·8 to 31·3)
Difference in increment between groups	16·6 (95% CI –8·9 to 42·0)	..
p value	0·200	..

A Secondary Outcome (CCS angina grade)

	PCI	Placebo	p value
Enrolment to pre-randomisation	..	..	0·916
Patients assessed	105	95	..
No change or deterioration	63 (60%)	59 (62%)	..
1 class improvement	27 (26%)	22 (23%)	..
≥2 class improvement	15 (14%)	14 (15%)	..
Pre-randomisation to follow-up	..	..	0·633
Patients assessed	105	91	..
No change or deterioration	51 (49%)	50 (55%)	..
1 class improvement	27 (26%)	22 (24%)	..
≥2 class improvement	27 (26%)	19 (21%)	..

Data are n (%) unless otherwise specified. PCI=percutaneous coronary intervention. CCS=Canadian Cardiovascular Society.

χ^2 for Pre-randomization vs. follow-up

```
ccs <- matrix(c(51, 27, 27, 50, 22, 19),
              byrow=TRUE, nrow = 2, ncol = 3)
rownames(ccs) <- c("PCI", "Placebo")
colnames(ccs) <- c("No change", "Improved 1 class",
                   "Improved 2+")
```

```
> ccs

        No change Improved 1 class Improved 2+
PCI            51                27          27
Placebo        50                22          19

> chisq.test(ccs)

Pearson's Chi-squared test data:  ccs
X-squared = 0.91608, df = 2, p-value = 0.6325
```

New York Times (2017-11-02) by Gina Kolata

Neither the patients nor the researchers assessing them afterward knew who had received a stent. Following the procedure, both groups of patients took powerful drugs to prevent blood clots.

The stents did what they were supposed to do in patients who received them. Blood flow through the previously blocked artery was greatly improved.

When the researchers tested the patients six weeks later, both groups said they had less chest pain, and they did better than before on treadmill tests.

But there was no real difference between the patients, the researchers found. Those who got the sham procedure did just as well as those who got stents.

Chang AK et al. Opioids vs. Non-Opioids for Acute Extremity Pain in the ED (JAMA 2017-11-07)

E SEARCH

The New York Times

lternatives to Opioids for
ain Relief

THE SWEET SPOT
An Addict Brother's Death; a Sister's Guilt-Ridden Grief

MODERN LOVE
On the Path to Empathy, Some Forks in the Road

Women More L
Men to Die in F
After Heart Atta

WELL | LIVE

Alternatives to Opioids for Pain Relief

By NICHOLAS BAKALAR NOV. 8, 2017

A combination of Tylenol and Advil worked just as well as opioids for relief of pain in the emergency room, a randomized trial has found.

Advil vs. Opioids for acute pain in ED (JAMA 2017-11-07)

November 7, 2017

Effect of a Single Dose of Oral Opioid and Nonopioid Analgesics on Acute Extremity Pain in the Emergency Department

A Randomized Clinical Trial

Andrew K. Chang, MD, MS[1]; Polly E. Bijur, PhD[2]; David Esses, MD[2]; et al

Chang et al. Abstract

DESIGN, SETTINGS, AND PARTICIPANTS Randomized clinical trial conducted at 2 urban EDs in the Bronx, New York, that included 416 patients aged 21 to 64 years with moderate to severe acute extremity pain enrolled from July 2015 to August 2016.

INTERVENTIONS Participants (104 per each combination analgesic group) received 400 mg of ibuprofen and 1000 mg of acetaminophen; 5 mg of oxycodone and 325 mg of acetaminophen; 5 mg of hydrocodone and 300 mg of acetaminophen; or 30 mg of codeine and 300 mg of acetaminophen.

MAIN OUTCOMES AND MEASURES The primary outcome was the between-group difference in decline in pain 2 hours after ingestion. Pain intensity was assessed using an 11-point numerical rating scale (NRS), in which 0 indicates no pain and 10 indicates the worst possible pain. The predefined minimum clinically important difference was 1.3 on the NRS. Analysis of variance was used to test the overall between-group difference at $P = .05$ and 99.2% CIs adjusted for multiple pairwise comparisons.

Chang et al. Sample Size

Sample Size Calculation

The following parameters were used to calculate the sample size: an overall 2-sided significance level of .05 (.008 for all pairwise comparisons using the Bonferroni correction),[18] 80% power, between-group difference for change in mean NRS pain score of 1.3, and a within-group SD of 2.6 based on estimates of variability from our prior work.[3-5] Using these parameters, we estimated that 100 patients would be needed per group for a total of 400 patients.

Chang et al. Results

Table 2. Numerical Rating Scale (NRS) Pain Scores and Decline in Pain Scores by Treatment Group

	NRS Pain Score, Mean (95% CI)[a]				
	Ibuprofen and Acetaminophen[b]	Oxycodone and Acetaminophen[c]	Hydrocodone and Acetaminophen[d]	Codeine and Acetaminophen[e]	*P* Value[f]
No. of patients[g]	101	104	103	103	
Primary end point: decline in score to 2 h	4.3 (3.6 to 4.9)	4.4 (3.7 to 5.0)	3.5 (2.9 to 4.2)	3.9 (3.2 to 4.5)	.053
Baseline score	8.9 (8.5 to 9.2)	8.7 (8.3 to 9.0)	8.6 (8.3 to 9.0)	8.6 (8.2 to 8.9)	.47
Score at 1 h	5.9 (5.3 to 6.6)	5.5 (4.9 to 6.2)	6.2 (5.6 to 6.9)	5.9 (5.2 to 6.5)	.25
Score at 2 h	4.6 (3.9 to 5.3)	4.3 (3.6 to 5.0)	5.1 (4.5 to 5.8)	4.7 (4.0 to 5.4)	.13

Table 3. Between-Group Difference in Mean Change in Numerical Rating Scale (NRS) Pain Scores

	Between-Group Difference in Mean Change in NRS Pain Score (99.2% CI)[a]	
Comparison	From Baseline to 1 h	From Baseline to 2 h
Ibuprofen and acetaminophen vs oxycodone and acetaminophen	−0.2 (−1.0 to 0.6)	−0.1 (−1.0 to 0.8)
Ibuprofen and acetaminophen vs hydrocodone and acetaminophen	0.5 (−0.3 to 1.3)	0.8 (−0.2 to 1.7)
Ibuprofen and acetaminophen vs codeine and acetaminophen	0.2 (−0.6 to 1.0)	0.4 (−0.6 to 1.3)
Oxycodone and acetaminophen vs hydrocodone and acetaminophen	0.7 (−0.1 to 1.5)	0.9 (−0.1 to 1.8)
Oxycodone and acetaminophen vs codeine and acetaminophen	0.4 (−0.4 to 1.2)	0.5 (−0.4 to 1.4)
Hydrocodone and acetaminophen vs codeine and acetaminophen	−0.3 (−1.1 to 0.5)	−0.4 (−1.3 to 0.6)

[a] Indicates mean change in pain of first analgesic minus mean change in pain from second analgesic. Pain intensity was assessed using an 11-point NRS in which a score of 0 indicates no pain and a score of 10 indicates the worst possible pain.

Chang et al. Abstract

RESULTS Of 416 patients randomized, 411 were analyzed (mean [SD] age, 37 [12] years; 199 [48%] women; 247 [60%] Latino). The baseline mean NRS pain score was 8.7 (SD, 1.3). At 2 hours, the mean NRS pain score decreased by 4.3 (95% CI, 3.6 to 4.9) in the ibuprofen and acetaminophen group; by 4.4 (95% CI, 3.7 to 5.0) in the oxycodone and acetaminophen group; by 3.5 (95% CI, 2.9 to 4.2) in the hydrocodone and acetaminophen group; and by 3.9 (95% CI, 3.2 to 4.5) in the codeine and acetaminophen group (P = .053). The largest difference in decline in the NRS pain score from baseline to 2 hours was between the oxycodone and acetaminophen group and the hydrocodone and acetaminophen group (0.9; 99.2% CI, −0.1 to 1.8), which was less than the minimum clinically important difference in NRS pain score of 1.3. Adverse events were not assessed.

CONCLUSIONS AND RELEVANCE For patients presenting to the ED with acute extremity pain, there were no statistically significant or clinically important differences in pain reduction at 2 hours among single-dose treatment with ibuprofen and acetaminophen or with 3 different opioid and acetaminophen combination analgesics. Further research to assess adverse events and other dosing may be warranted.

TRIAL REGISTRATION clinicaltrials.gov Identifier: NCT02455518

New York Times (2017-11-08) by Nicholas Bakalar

Researchers studied 416 men and women who arrived in the E.R. with moderate to severe pain in their arms or legs from sprains, strains, fractures or other injuries. They randomly assigned them to an oral dose of acetaminophen (Tylenol) with either ibuprofen (Advil) or the opioids oxycodone, hydrocodone or codeine.

Two hours later, they questioned them using an 11-point pain scale (higher scores = more pain).

- The average score was 8.7 before taking medicine.
- That score decreased by:
  - 4.3 points with ibuprofen and Tylenol,
  - 4.4 with oxycodone and Tylenol,
  - 3.5 with hydrocodone and Tylenol, and
  - 3.9 with codeine and Tylenol.

In other words, there was no significant difference, either statistically or clinically, among any of the four regimens.

On p values

George Cobb's Questions (with Answers)

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

- Q: Why do so many colleges and grad schools teach $p = 0.05$?

George Cobb's Questions (with Answers)

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

- Q: Why do so many colleges and grad schools teach $p = 0.05$?
- A: Because that's still what the scientific community and journal editors use.

George Cobb's Questions (with Answers)

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

- Q: Why do so many colleges and grad schools teach $p = 0.05$?
- A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use $p = 0.05$?

George Cobb's Questions (with Answers)

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

- Q: Why do so many colleges and grad schools teach $p = 0.05$?
- A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use $p = 0.05$?
- A: Because that's what they were taught in college or grad school.

Now what?

So sad. . .

Gelman on Statistical Significance

"... we use the term statistically significant in the conventional way, to mean that an estimate is **at least two standard errors away** from some"null hypothesis" or prespecified value that would indicate no effect present. An estimate is statistically insignificant if the observed value could reasonably be explained by simple chance variation, much in the way that a sequence of 20 coin tosses might happen to come up 8 heads and 12 tails; we would say that this result is not statistically significantly different from chance. More precisely, the observed proportion of heads is 40 percent but with a standard error of 11 percent - thus, the data are less than two standard errors away from the null hypothesis of 50 percent, and the outcome could clearly have occurred by chance. Standard error is a measure of the variation in an estimate and gets smaller as a sample size gets larger, converging on zero as the sample increases in size."

Gelman's blog (2017-10-28)

The Value of a *p*-Valueless Paper

Jason T. Connor (2004) *American J of Gastroenterology* 99(9): 1638-40.

Abstract: As is common in current bio-medical research, about 85% of original contributions in *The American Journal of Gastroenterology* in 2004 have reported *p*-values. However, none are reported in this issue's article by Abraham et al. who, instead, rely exclusively on effect size estimates and associated confidence intervals to summarize their findings. **Authors using confidence intervals communicate much more information in a clear and efficient manner than those using *p*-values. This strategy also prevents readers from drawing erroneous conclusions caused by common misunderstandings about *p*-values**. I outline how standard, two-sided confidence intervals can be used to measure whether two treatments differ or test whether they are clinically equivalent.

DOI: 10.1111/j.1572-0241.2004.40592.x

Why Dividing Data Comparisons into Categories based on Significance Levels is Terrible.

The common practice of dividing data comparisons into categories based on significance levels is terrible, but it happens all the time. . . . so it's worth examining the prevalence of this error.

Link to Andrew Gelman's blog, 2016-10-15

Gelman on p values, 1

Let me first briefly explain why categorizing based on p-values is such a bad idea. Consider, for example, this division:

- "really significant" for $p < .01$,
- "significant" for $p < .05$,
- "marginally significant" for $p < .1$, and
- "not at all significant" otherwise.

Now consider some typical p-values in these ranges: say, $p = .005$, $p = .03$, $p = .08$, and $p = .2$.

Translate these two-sided p-values back into z-scores, which we can do in R via `qnorm(c(.005, .03, .08, .2)/2, lower.tail = FALSE)`

Gelman on *p* values, 2

Description	really sig.	sig.	marginally sig.	not at all sig.
p value	0.005	0.03	0.08	0.20
Z score	2.8	2.2	1.8	1.3

The seemingly yawning gap in p-values comparing the not at all significant *p*-value of .2 to the really significant *p*-value of .005, is only 1.5.

If you had two independent experiments with z-scores of 2.8 and 1.3 and with equal standard errors and you wanted to compare them, you'd get a difference of 1.5 with a standard error of 1.4, which is completely consistent with noise.

Gelman on *p* values, 3

The key point: The difference between statistically significant and NOT statistically significant is not, generally, statistically significant.

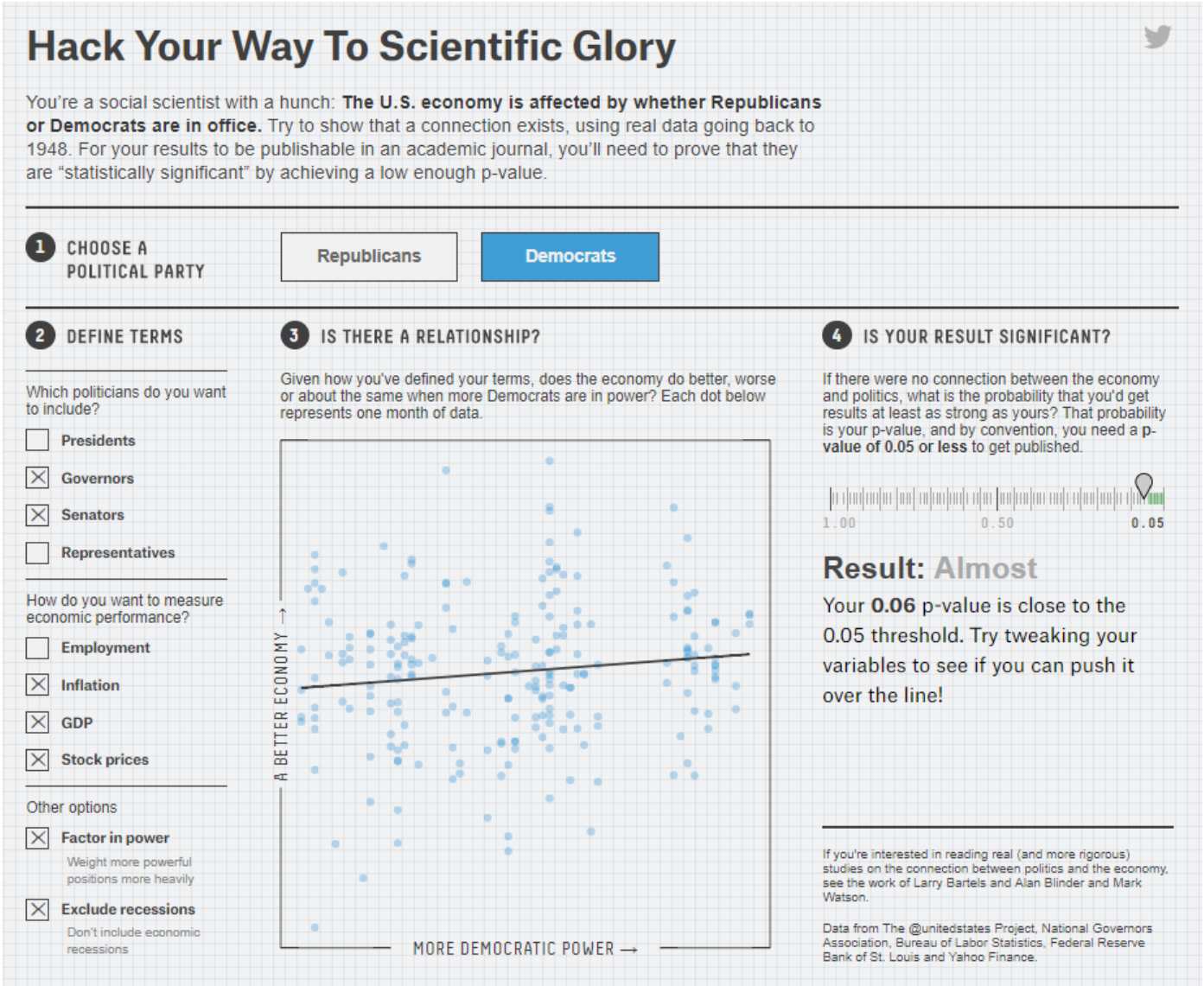
From a **statistical** point of view, the trouble with using the p-value as a data summary is that the p-value is only interpretable in the context of the null hypothesis of zero effect, and (much of the time), nobody's interested in the null hypothesis. Indeed, once you see comparisons between large, marginal, and small effects, the null hypothesis is irrelevant, as you want to be comparing effect sizes.

From a **psychological** point of view, the trouble with using the p-value as a data summary is that this is a kind of deterministic thinking, an attempt to convert real uncertainty into firm statements that are just not possible (or, as we would say now, just not replicable).

p Hacking and "Researcher Degrees of Freedom"

Hack Your Way To Scientific Glory

https://fivethirtyeight.com/features/science-isnt-broken



What can you get?

I was able to get

- $p < 0.01$ (positive effect of Democrats on economy)
- $p = 0.01$ (negative effect of Democrats)
- $p = 0.03$ (negative effect of Democrats)
- $p = 0.03$ (positive effect of Democrats)

but also . . .

- $p = 0.05, 0.06, 0.07, 0.09, 0.17, 0.19, 0.20, 0.22, 0.23, 0.47, 0.51$

without even switching parties, exclusively by changing my definitions of terms (section 2 of the graphic.)

"Researcher Degrees of Freedom", 1

[I]t is unacceptably easy to publish "statistically significant" evidence consistent with any hypothesis.

The culprit is a construct we refer to as ***researcher degrees of freedom****. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?*

Simmons et al. *link*

"Researcher Degrees of Freedom", 2

> *. . . It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields statistical significance, and to then report only what worked. The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.*

For more, see

- Gelman's blog $2012 - 11 - 01$ "Researcher Degrees of Freedom",
- Paper by *Simmons* and others, defining the term.

And this is really hard to deal with. . .

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time

> *Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.*

- *Link* to the paper from Gelman and Loken

Benjamin et al 2017 Redefine Statistical Significance

We propose to change the default P-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

- 0.005 is stringent enough to "break" the current system - makes it very difficult for researchers to reach threshold with noisy, useless studies.

Visit the main *article*. Visit an explanatory piece in *Science*.

Lakens et al. Justify Your Alpha

"In response to recommendations to redefine statistical significance to $p \leq .005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level." Visit *link*.

Quiz 2 Setup

- Quiz 2 will be yours by 5 PM today.
  - It's now due Tuesday Nov 14 at **8 AM**.