431 Class 21

Thomas E. Love

2017-11-09

Today's Agenda

- Answer Sketches for the Airline Etiquette Exercises
- Reacting to Published Research: Design, Quality, Data, Analysis
    - Retrospective Power and Sample Size Calculations?
    - Type S and Type M Errors
    - Lots of material from Andrew Gelman and collaborators

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 Example

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)

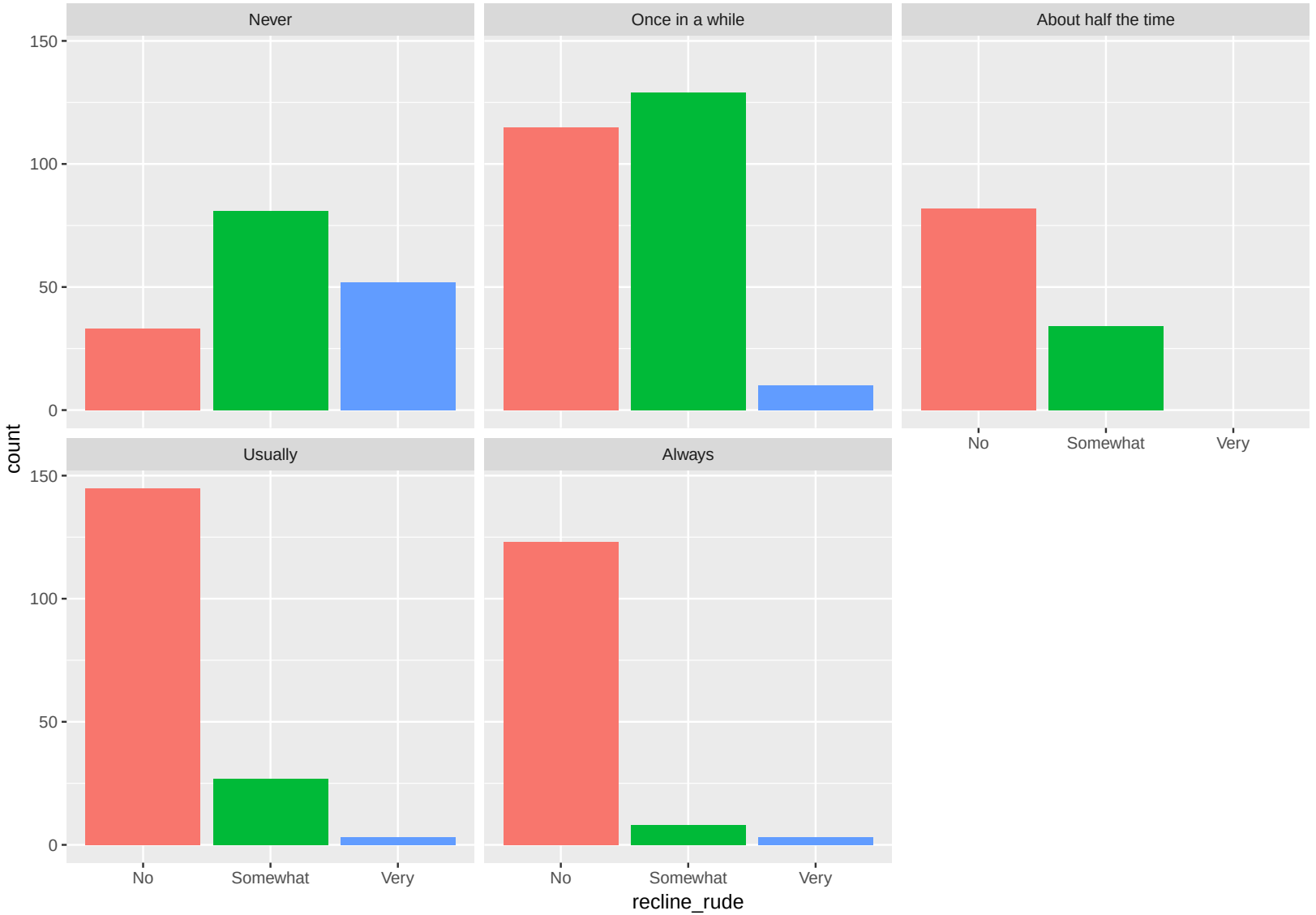
Exercise 4

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)

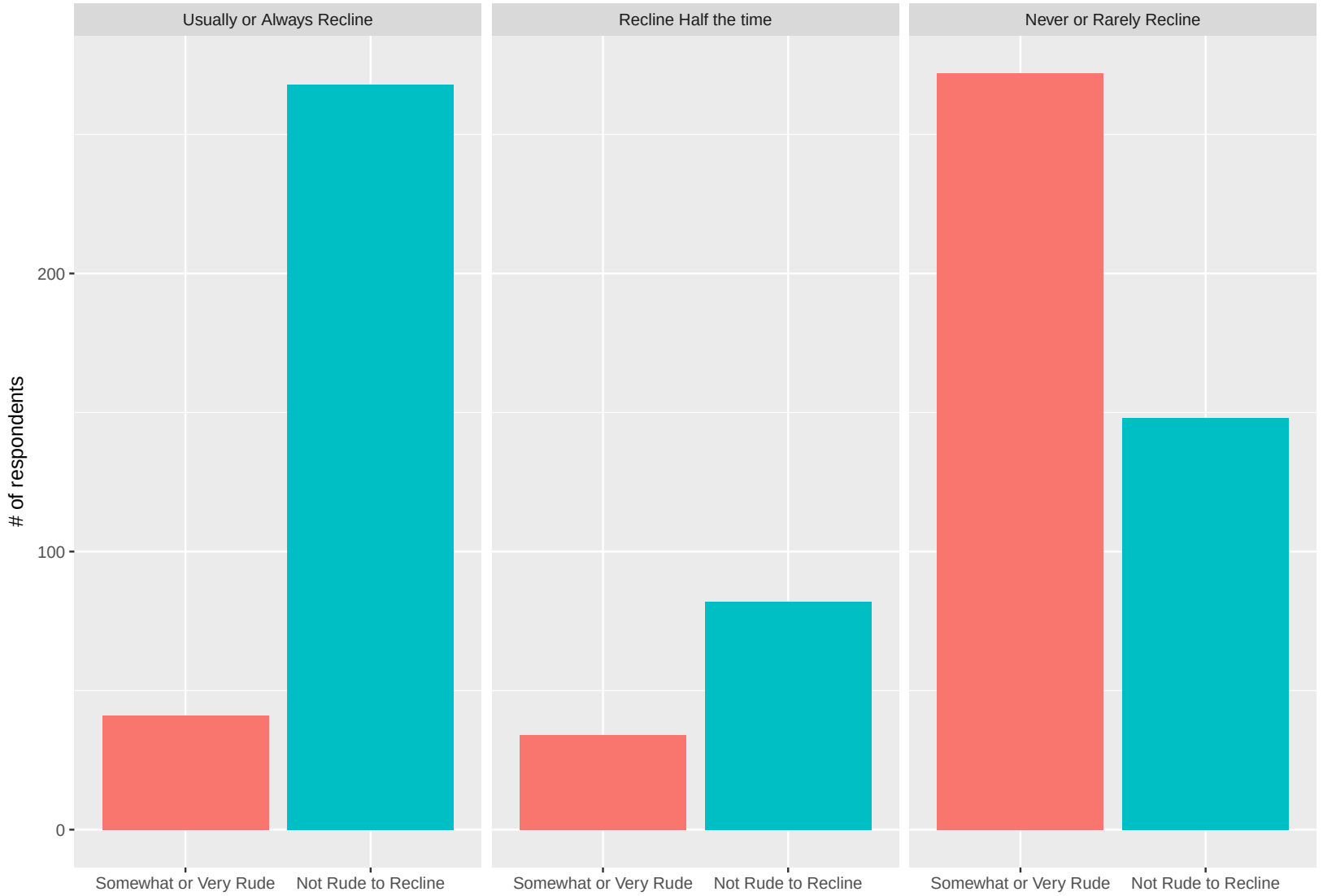
```
fly %$% table(recline_rude, recline_frequency) %>% chisq.test
```

```
    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
```

Andrew Gelman (and others) thinking hard about Study Design, Applied Statistics, and Evaluating Evidence

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

Editorial from JAMA Cardiology 2016-10-12

Editor's Note

Do Not Over (*P*) Value Your Research Article

Laine E. Thomas, PhD; Michael J. Pencina, PhD

***P* value** is by far the most prevalent statistic in the medical literature but also one attracting considerable controversy. Recently, the American Statistical Association[1] released a policy statement on *P* values, noting that misunderstanding and misuse of *P* values is an important contributing factor to the common problem of scientific conclusions that fail to be reproducible. Furthermore, reliance on *P* values may distract from the good scientific principles that are needed for high-quality research. Mark et al[2] delve deeper into the history and interpretation of the *P* value in this issue of *JAMA Cardiology*. Herein, we take the opportunity to state a few principles to help guide authors in the use and reporting of *P* values in the journal.

Related article

When the limitations surrounding *P* values are emphasized, a common question is, "What should we do instead?" Ron Wasserstein of the American Statistical Association explained: "In the post p<0.05 era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy.... Instead, journals [should evaluate] papers based on clear and detailed description of the study design, execution, and analysis, having conclusions that are based on valid statistical interpretations and scientific arguments, and reported transparently and thoroughly enough to be rigorously scrutinized by others."[3]

We suggest that researchers submitting manuscripts to *JAMA Cardiology* should also consider the following:

1. Data that are descriptive of the sample (ie, indicating imbalances between observed groups but not making inference to a population) should not be associated with *P* values. Appropriate language, in this case, would describe numerical differences and sample summary statistics and focus on differences of clinical importance.
2. In addition to summary statistics and confidence intervals, standardized differences (rather than *P* values) are a preferred way to exhibit imbalances between groups.
3. *P* values are most meaningful in the context of clear, a priori hypotheses that support the main conclusions of a manuscript.
4. Reporting stand-alone *P* values is discouraged, and preference should be given to presentation and interpretation of effect sizes and their uncertainty (confidence intervals) in the scientific context and in light of other evidence. Crossing a threshold (eg, $P < .05$) by itself constitutes only weak evidence.
5. Researchers should define and interpret effect measures that are clinically relevant. For example, clinical importance is often difficult to establish on the odds ratio scale but is clearer on the risk ratio or absolute risk difference scale.

In summary, following Mark et al,[2] we encourage researchers to focus on interpreting clinical research data in terms of treatment "effect" magnitude and precision, using *P* value only as one of many complementary tools in the statistical toolbox.

Mark, Lee, Harrell JAMA Cardiol 2016-10-12

Abstract

P values and hypothesis testing methods are frequently misused in clinical research. Much of this misuse appears to be owing to the widespread, mistaken belief that they provide simple, reliable, and objective triage tools for separating the true and important from the untrue or unimportant. The primary focus in interpreting therapeutic clinical research data should be on the treatment ("oomph") effect, a metaphorical force that moves patients given an effective treatment to a different clinical state relative to their control counterparts. This effect is assessed using 2 complementary types of statistical measures calculated from the data, namely, effect magnitude or size and precision of the effect size. In a randomized trial, effect size is often summarized using constructs, such as odds ratios, hazard ratios, relative risks, or adverse event rate differences. How large a treatment effect has to be to be consequential is a matter for clinical judgment. The precision of the effect size (conceptually related to the amount of spread in the data) is usually addressed with confidence intervals. *P* values (significance tests) were first proposed as an informal heuristic to help assess how "unexpected" the observed effect size was if the true state of nature was no effect or no difference. Hypothesis testing was a modification of the significance test approach that envisioned controlling the false-positive rate of study results over many (hypothetical) repetitions of the experiment of interest. Both can be helpful but, by themselves, provide only a tunnel vision perspective on study results that ignores the clinical effects the study was conducted to measure.

doi:10.1001/jamacardio.2016.3312

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

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).

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

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)

How To React to Published Research

Reacting to Published Research

My Sources include

Gelman and Carlin article at http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf

Gelman blogs for background and details:

- http://andrewgelman.com/2016/10/25/how-not-to-analyze-noisy-data-a-case-study/
- http://andrewgelman.com/2016/11/13/more-on-my-paper-with-john-carlin-on-type-m-and-type-s-errors/

The Impact of Study Design (AG)

Applied statistics is hard.

- Doing a statistical analysis is like playing basketball, or knitting a sweater. You can get better with practice.
- Incompetent statistics does not necessarily doom a research paper: some findings are solid enough that they show up even when there are mistakes in the data collection and data analyses. But we've also seen many examples where incompetent statistics led to conclusions that made no sense but still received publication and publicity.
- We should be thinking not just about data analysis, but also data quality.

What Kind of Errors? (from Gelman)

Consider: "The Association Between Men's Sexist Attitudes and Facial Hair" PubMed 26510427 (*Arch Sex Behavior* May 2016)

Headline Finding: A sample of ~500 men from America and India shows a significant relationship between sexist views and the presence of facial hair.

Excerpt 1:

> *Since a linear relationship has been found between facial hair thickness and perceived masculinity . . . we explored the relationship between facial hair thickness and sexism. . . . Pearson's correlation found no significant relationships between facial hair thickness and hostile or benevolent sexism, education, age, sexual orientation, or relationship status.*

Facial Hair and Sexist Attitudes (from Gelman)

Excerpt 2:

> *We conducted pairwise comparisons between clean-shaven men and each facial hair style on hostile and benevolent sexism scores. . . . For the purpose of further analyses, participants were classified as either clean-shaven or having facial hair based on their self- reported facial hair style . . . There was a significant Facial Hair Status by Sexism Type interaction . . .*

- So their headline finding appeared only because, after their first analysis failed, they shook and shook the data until they found something statistically significant.

Facial Hair and Sexist Attitudes (from Gelman)

Excerpt 2:

> *We conducted pairwise comparisons between clean-shaven men and each facial hair style on hostile and benevolent sexism scores. . . . For the purpose of further analyses, participants were classified as either clean-shaven or having facial hair based on their self- reported facial hair style . . . There was a significant Facial Hair Status by Sexism Type interaction . . .*

- So their headline finding appeared only because, after their first analysis failed, they shook and shook the data until they found something statistically significant.
- All credit to the researchers for admitting that they did this, but poor practice of them to present their result in the abstract to their paper without making this clear, and too bad that the journal got suckered into publishing this.

How do people specify effect sizes for power calculations?

1. **Empirical**: assuming an effect size equal to the estimate from a previous study or from the data at hand (if performed retrospectively).
    - generally based on small samples
    - when preliminary results look interesting, they are more likely biased towards unrealistically large effects
2. **On the basis of goals**: assuming an effect size deemed to be substantively important or more specifically the minimum effect that would be substantively important.
    - Can also lead to specifying effect sizes that are larger than what is likely to be the true effect.

- Both lead to performing studies that are too small or misinterpretation of findings after completion.

Gelman and Carlin

- The idea of a **design analysis** is to improve the design and evaluation of research, when you want to summarize your inference through concepts related to statistical significance.
- Type 1 and Type 2 errors are tricky concepts and aren't easy to describe before data are collected, and are very difficult to use well after data are collected.
- These problems are made worse when you have
  - Noisy studies, where the signal may be overwhelmed,
  - Small Sample Sizes
  - No pre-registered (prior to data gathering) specifications for analysis
- Top statisticians avoid "post hoc power analysis"...
  - Why? It's usually crummy.

Why not post hoc power analysis?

So you collected data and analyzed the results. Now you want to do an after data gathering (post hoc) power analysis.

1. What will you use as your "true" effect size?
    - Often, point estimate from data - yuck - results very misleading - power is generally seriously overestimated when computed on the basis of statistically significant results.
    - Much better (but rarer) to identify plausible effect sizes based on external information rather than on your sparkling new result.
2. What are you trying to do? (too often)
    - get researcher off the hook (I didn't get $p < 0.05$ because I had low power - an alibi to explain away non-significant findings) or
    - encourage overconfidence in the finding.

Gelman and Carlin Broader Design Ideas

- A broader notion of design, though, can be useful before and after data are gathered.

Gelman and Carlin recommend design calculations to estimate

1. Type S (sign) error - the probability of an estimate being in the wrong direction, and
2. Type M (magnitude) error, or exaggeration ratio - the factor by which the magnitude of an effect might be overestimated.

- These can (and should) have value *both* before data collection/analysis and afterwards (especially when an apparently strong and significant effect is found.)
- The big challenge remains identifying plausible effect sizes based on external information. Crucial to base our design analysis with an external estimate.

The Building Blocks

You perform a study that yields estimate d with standard error s. Think of d as an estimated mean difference, for example.

- Looks significant if $|d/s| > 2$, which roughly corresponds to $p < 0.05$. Inconclusive otherwise.

The Building Blocks

You perform a study that yields estimate d with standard error s. Think of d as an estimated mean difference, for example.

- Looks significant if $|d/s| > 2$, which roughly corresponds to $p < 0.05$. Inconclusive otherwise.
- Now, consider a true effect size D (the value that d would take if you had an enormous sample)

The Building Blocks

You perform a study that yields estimate d with standard error s. Think of d as an estimated mean difference, for example.

- Looks significant if $|d/s| > 2$, which roughly corresponds to $p < 0.05$. Inconclusive otherwise.
- Now, consider a true effect size D (the value that d would take if you had an enormous sample)
- D is hypothesized based on *external* information (Other available data, Literature review, Modeling as appropriate, etc.)

The Building Blocks

You perform a study that yields estimate d with standard error s. Think of d as an estimated mean difference, for example.

- Looks significant if $|d/s| > 2$, which roughly corresponds to $p < 0.05$. Inconclusive otherwise.
- Now, consider a true effect size D (the value that d would take if you had an enormous sample)
- D is hypothesized based on *external* information (Other available data, Literature review, Modeling as appropriate, etc.)
- Define d^{rep} as the estimate that would be observed in a hypothetical replication study with a design identical to our original study.

Design Analysis (Gelman and Carlin)

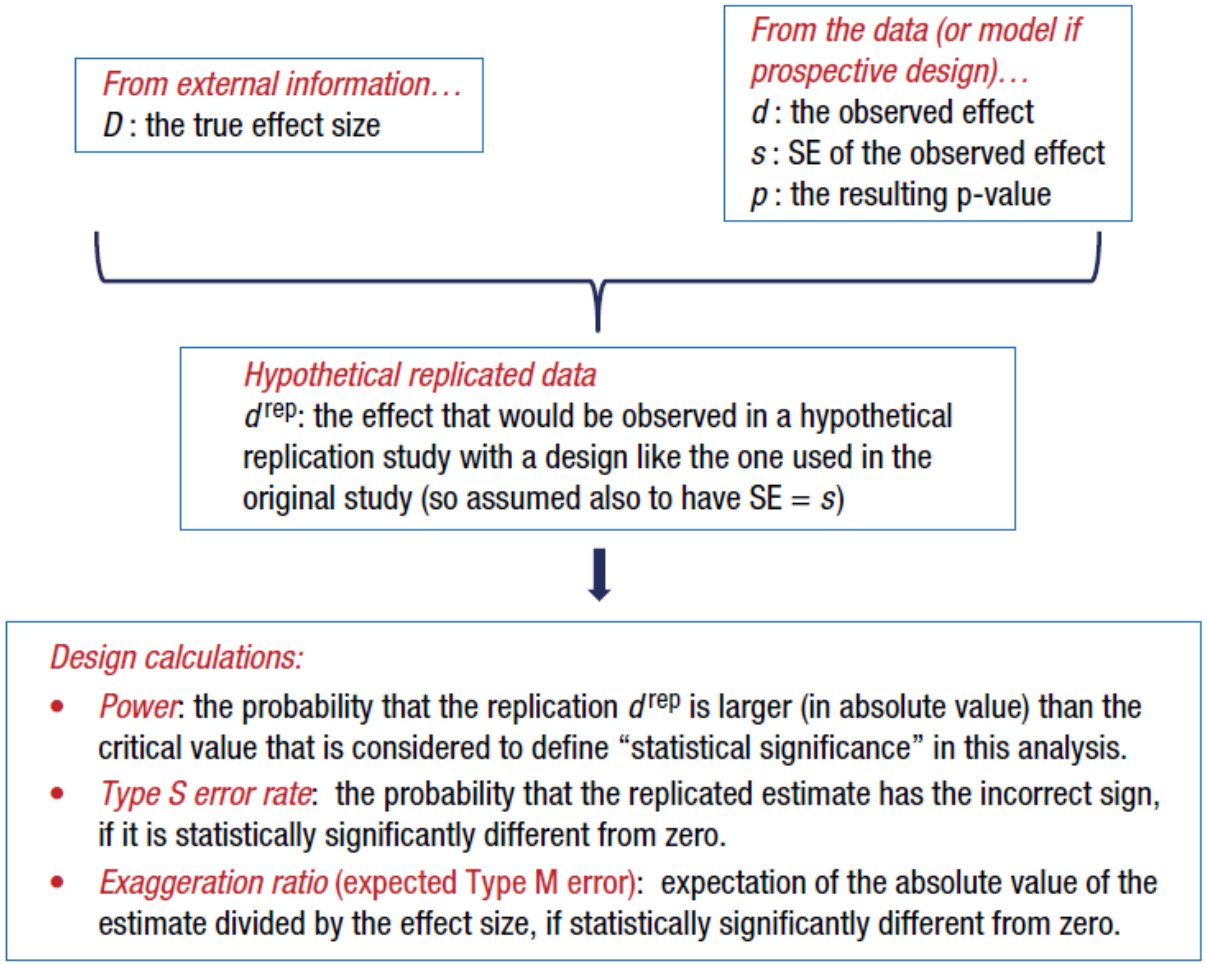


Figure 1. Diagram of our recommended approach to design analysis. It will typically make sense to consider different plausible values of D, the assumed true effect size.

The `retrodesign` function (shown on next slide)

Inputs to the function:

- D, the hypothesized true effect size (actually called A in the function)
- s, the standard error of the estimate
- alpha, the statistical significance threshold (default 0.05)
- df, the degrees of freedom (default assumption: infinite)

Output:

- the power
- the Type S error rate
- the exaggeration ratio

The retrodesign function (Gelman and Carlin)

```
retrodesign <- function(A, s, alpha=.05, df=Inf,
                        n.sims=10000){
    z <- qt(1-alpha/2, df)
    p.hi <- 1 - pt(z-A/s, df)
    p.lo <- pt(-z-A/s, df)
    power <- p.hi + p.lo
    typeS <- p.lo/power
    estimate <- A + s*rt(n.sims,df)
    significant <- abs(estimate) > s*z
    exaggeration <- mean(abs(estimate)[significant])/A
    return(list(power=power, typeS=typeS,
                exaggeration=exaggeration))
}
```

This is part of Love-boost.R

What if we have a beautiful, unbiased study?

Suppose we had a true effect that is 2.8 standard errors away from zero, in a study built to have 80% power to detect such an effect with 95% confidence.

```
retrodesign(A = 2.8, s = 1, alpha = .05)
```

```
$power
[1] 0.7995569

$typeS
[1] 1.210843e-06

$exaggeration
[1] 1.12382
```

- With the power this high (80%), we have a type S error rate of 1.2 x 10^{-6} and an expected exaggeration factor of 1.12.

Example: Beauty and Sex Ratios

Kanazawa study of 2972 respondents from the National Longitudinal Study of Adolescent Health

- Each subject was assigned an attractiveness rating on a 1-5 scale and then, years later, had at least one child.
- Of the first-born children with parents in the most attractive category, 56% were girls, compared with 48% girls in the other groups.
- So the estimated difference was 8 percentage points with a reported $p = 0.015$
- Kanazawa stopped there, but Gelman and Carlin don't.

Beauty and Sex Ratios

We need to postulate an effect size, which will not be 8 percentage points. Instead, Gelman and colleagues hypothesized a range of true effect sizes using the scientific literature.

> *There is a large literature on variation in the sex ratio of human births, and the effects that have been found have been on the order of 1 percentage point (for example, the probability of a girl birth shifting from 48.5 percent to 49.5 percent). Variation attributable to factors such as race, parental age, birth order, maternal weight, partnership status and season of birth is estimated at from less than 0.3 percentage points to about 2 percentage points, with larger changes (as high as 3 percentage points) arising under economic conditions of poverty and famine. (There are) reliable findings that male fetuses (and also male babies and adults) are more likely than females to die under adverse conditions.*

So, what is a reasonable effect size?

- Small observed differences in sex ratios in a multitude of studies of other issues (much more like 1 percentage point, tops)
- Noisiness of the subjective attractiveness rating (1-5) used in this particular study

So, Gelman and colleagues hypothesized three potential effect sizes (0.1, 0.3 and 1.0 percentage points) and under each effect size, considered what might happen in a study with sample size equal to Kanazawa's study.

How big is the standard error?

- From the reported estimate of 8 percentage points and p value of 0.015, the standard error of the difference is 3.29 percentage points.
  - If p value = 0.015 (two-sided), then Z score = `qnorm(p = 0.015/2, lower.tail=FALSE)` = 2.432
  - Z = estimate/SE, and if estimate = 8 and Z = 2.432, then SE = 8/2.432 = 3.29

Retrodesign Results: Option 1

- Assume true difference D = 0.1 percentage point (probability of girl births differing by 0.1 percentage points, comparing attractive with unattractive parents).
- Standard error assumed to be 3.29, and $\alpha = 0.05$

```
retrodesign(.1, 3.29)
```

```
$power
[1] 0.05010584

$typeS
[1] 0.4645306

$exaggeration
[1] 76.70491
```

Option 1 Conclusions

Assuming the true difference is 0.1 means that probability of girl births differs by 0.1 percentage points, comparing attractive with unattractive parents.

If the estimate is statistically significant, then:

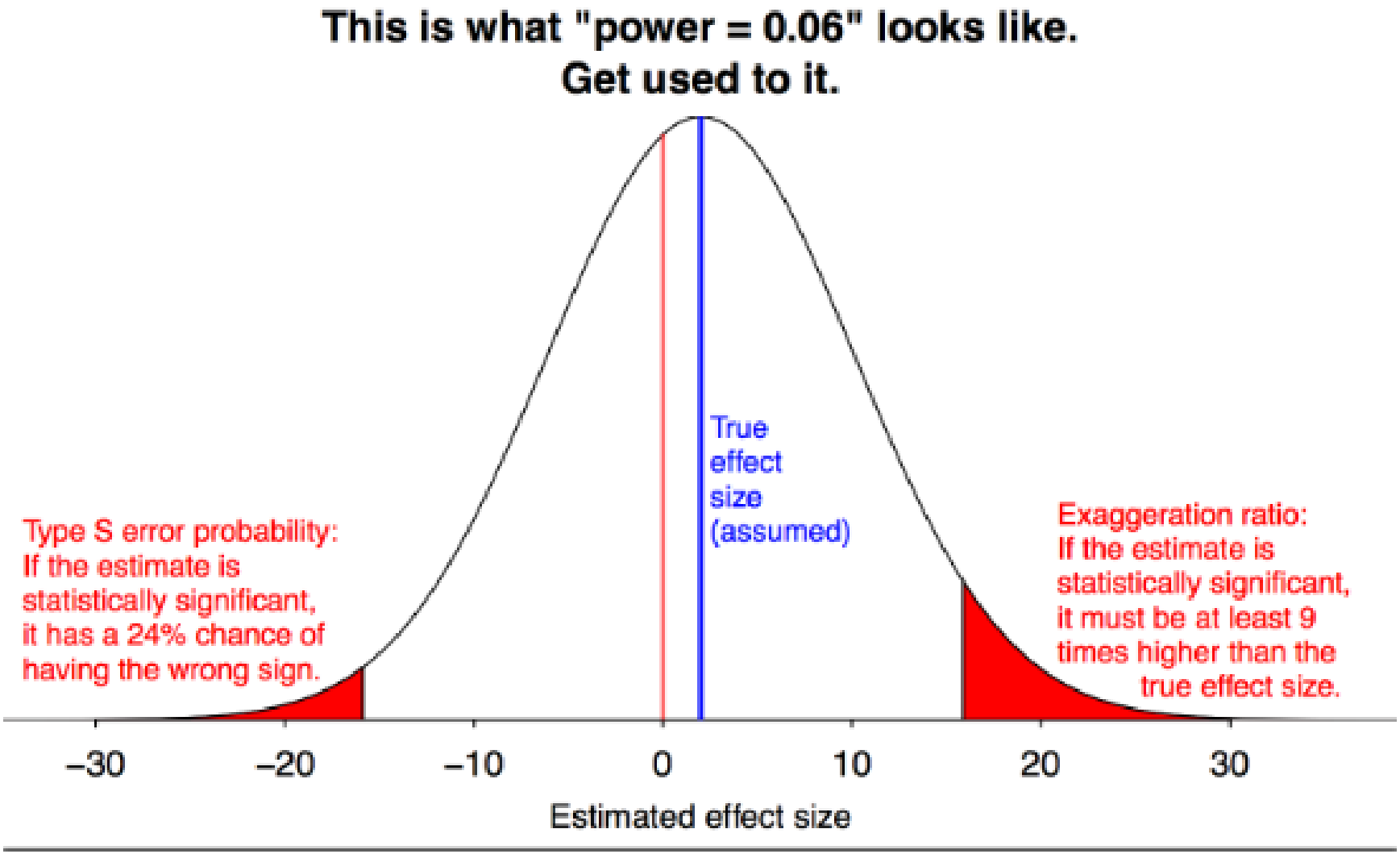
1. There is a 46% chance it will have the wrong sign (from the Type S error rate).
2. The power is 5% and the Type S error rate of 46%. Multiplying those gives a 2.3% probability that we will find a statistically significant result in the wrong direction.
3. We thus have a power - 2.3% = 2.7% probability of showing statistical significance in the correct direction.
4. In expectation, a statistically significant result will be 78 times too high (the exaggeration ratio).

Retrodesign Results: Options 2 and 3

Assumption	Power	Type S	Exaggeration Ratio
D = 0.1	0.05	0.46	78
D = 0.3	0.05	0.39	25
D = 1.0	0.06	0.19	7.8

- Under a true difference of 1.0 percentage point, there would be
  - a 4.9% chance of the result being statistically significantly positive and a 1.1% chance of a statistically significantly negative result.
  - A statistically significant finding in this case has a 19% chance of appearing with the wrong sign, and
  - the magnitude of the true effect would be overestimated by an expected factor of 8.

This is what Power = 0.06 looks like (Gelman)



Design Analysis (Gelman and Carlin, Figure 1)

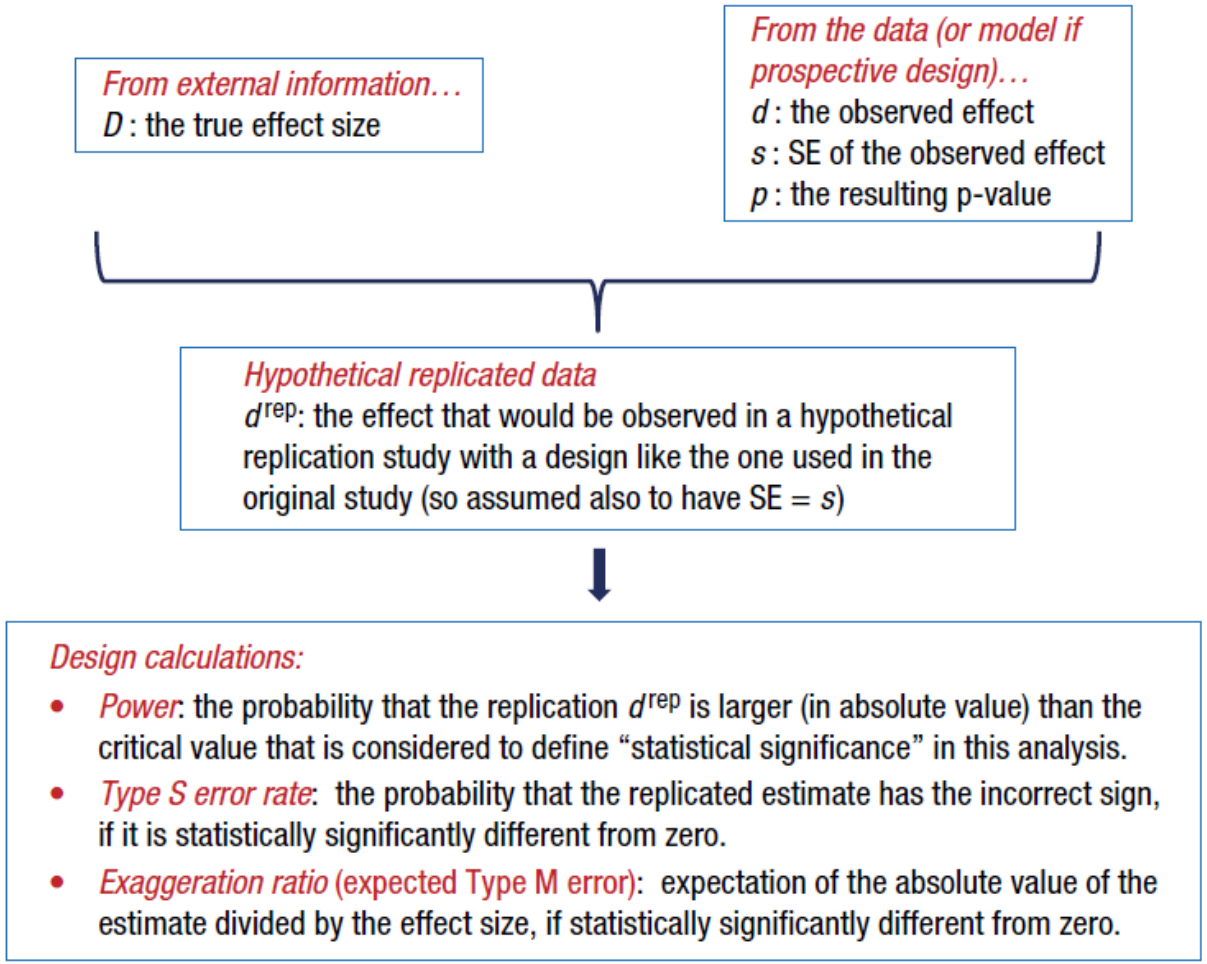


Figure 1. Diagram of our recommended approach to design analysis. It will typically make sense to consider different plausible values of D, the assumed true effect size.

The Ovulation and Voting study (Gelman)

Durante K et al. "The Fluctuating Female Vote: Politics, Religion and the Ovulatory Cycle" *Psychological Science* (reported then retracted from CNN under the title "Study looks at voting and hormones: Hormones may influence female voting choices.")

Abstract on next slide

Abstract for Ovulation and Voting Study

Each month many women experience an ovulatory cycle that regulates fertility. Whereas research finds that this cycle influences women's mating preferences, we propose that it might also change women's political and religious views. Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single versus married women. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious, and more likely to vote for Mitt Romney. In addition, ovulatory-induced changes in political orientation mediated women's voting behavior. Overall, the ovulatory cycle not only influences women's politics, but appears to do so differently for single versus married women.

What Do They Report? (see Gelman)

A bunch of comparisons and p values, some of which were statistically significant, and then lots of stories.

The problem is that there are so many things that could be compared, and all we see is some subset of the comparisons. And some of the effects are much too large to be plausible.

- For example, among women in relationships, 40% in the ovulation period supported Romney, compared to 23% in the non-fertile part of their cycle.
- Given that surveys find very few people switching their vote preferences during the campaign for any reason, I just don't buy it.
- The authors might respond that they don't care about the magnitude of the difference, just the sign, but (a) with a magnitude of this size, we're talking noise noise noise, and (b) one could just as easily explain this as a differential nonresponse - easy enough to come up with a story about that!

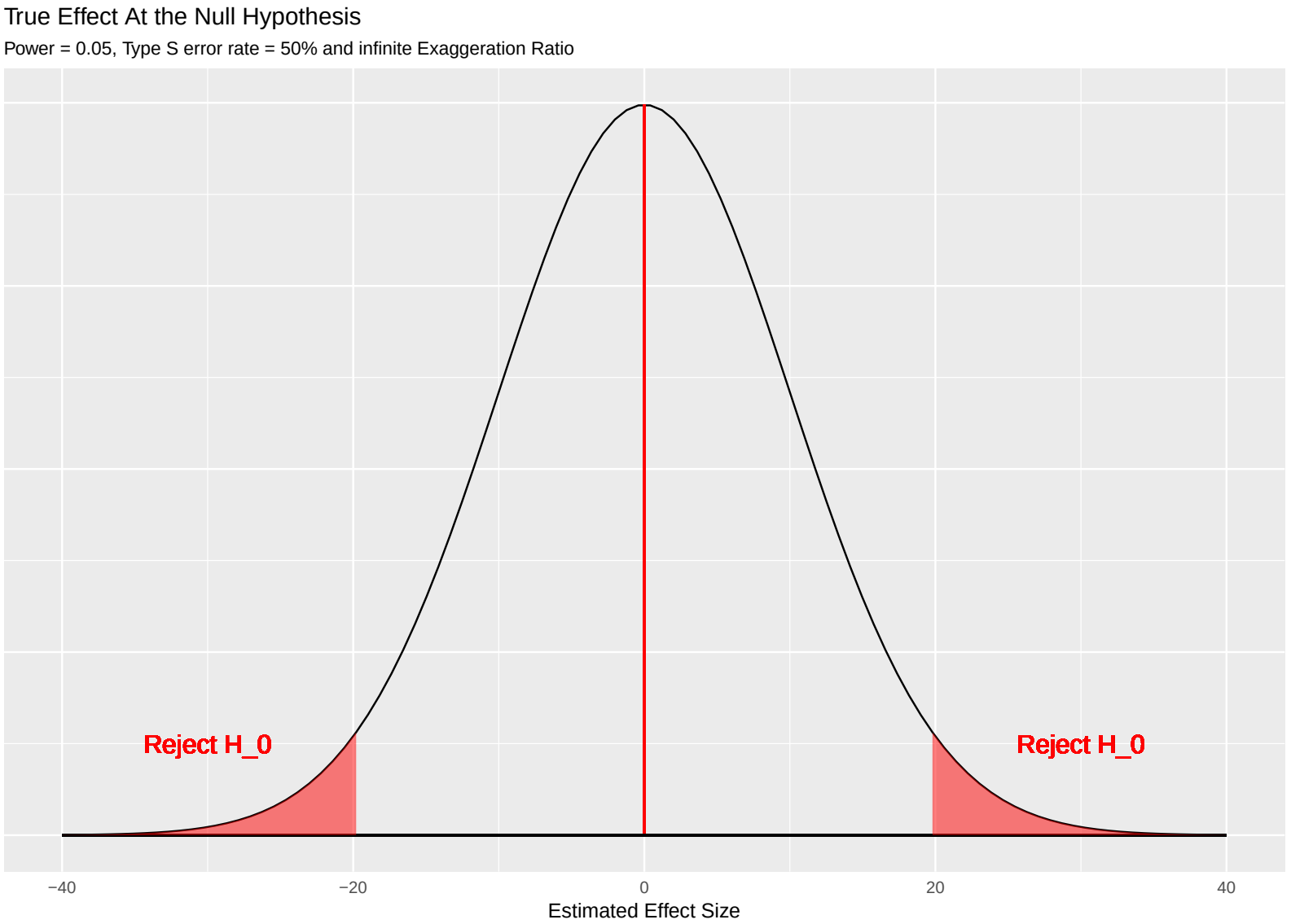
What to do?

- Analyze *all* your data.
  - For most of their analyses, the authors threw out all the data from participants who were PMS-ing or having their period. (We also did not include women at the beginning of the ovulatory cycle (cycle days 1-6) or at the very end of the ovulatory cycle (cycle days 26-28) to avoid potential confounds due to premenstrual or menstrual symptoms.) That's a mistake. Instead of throwing out one-third of their data, they should've just included that other category in their analysis.
- Present *all* your comparisons, not just a select few.
  - A big table, or even a graph, is what you want.
- Make your data public.
  - If the topic is worth studying, you should want others to be able to make rapid progress.

So what is a plausible size for the effect under study?

Maybe it's 30% of a standard error, tops. What does that mean, exactly?

Understanding Power, Type S and Type M Errors. Zero Effect

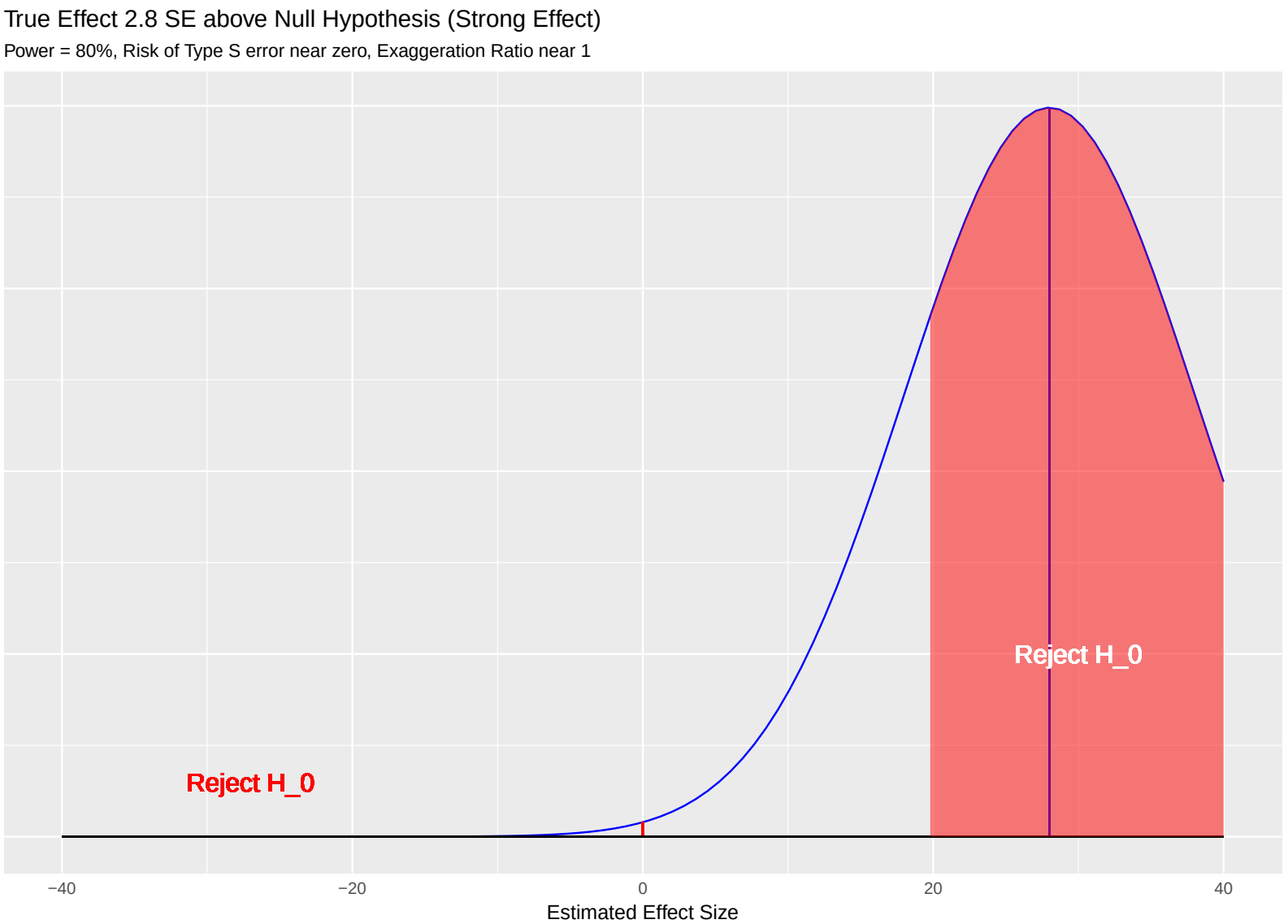


retrodesign for Zero Effect

```
retrodesign(A = 0, s = 10)
```

```
$power
[1] 0.05

$typeS
[1] 0.5

$exaggeration
[1] Inf
```

80% power; large effect (2.8 SE above H_0)

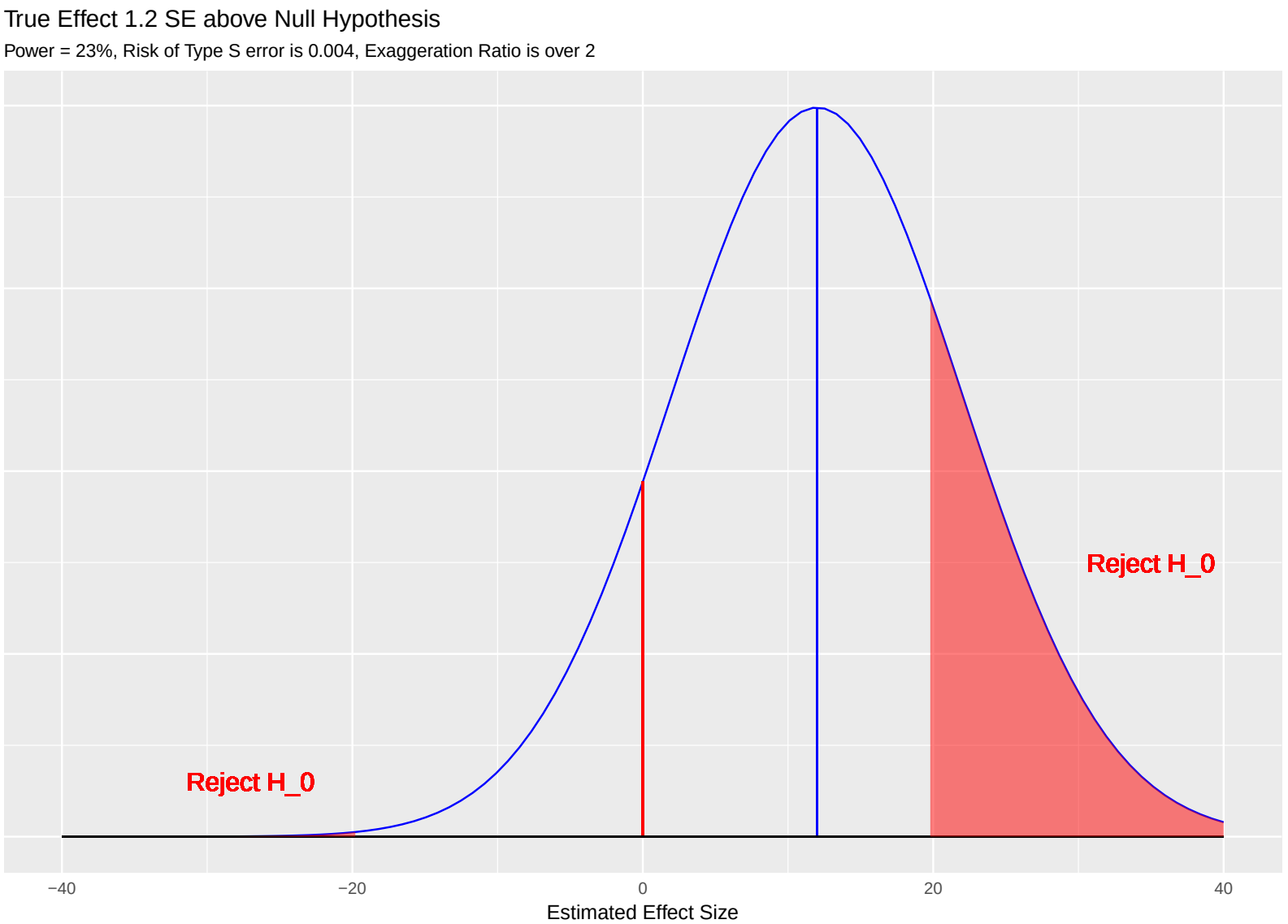


retrodesign for 2.8 SE effect size

```
retrodesign(A = 28, s = 10)
```

```
$power
[1] 0.7995569

$typeS
[1] 1.210843e-06

$exaggeration
[1] 1.122182
```

What 23% power looks like. . .



retrodesign for a true effect 1.2 SE above H_0

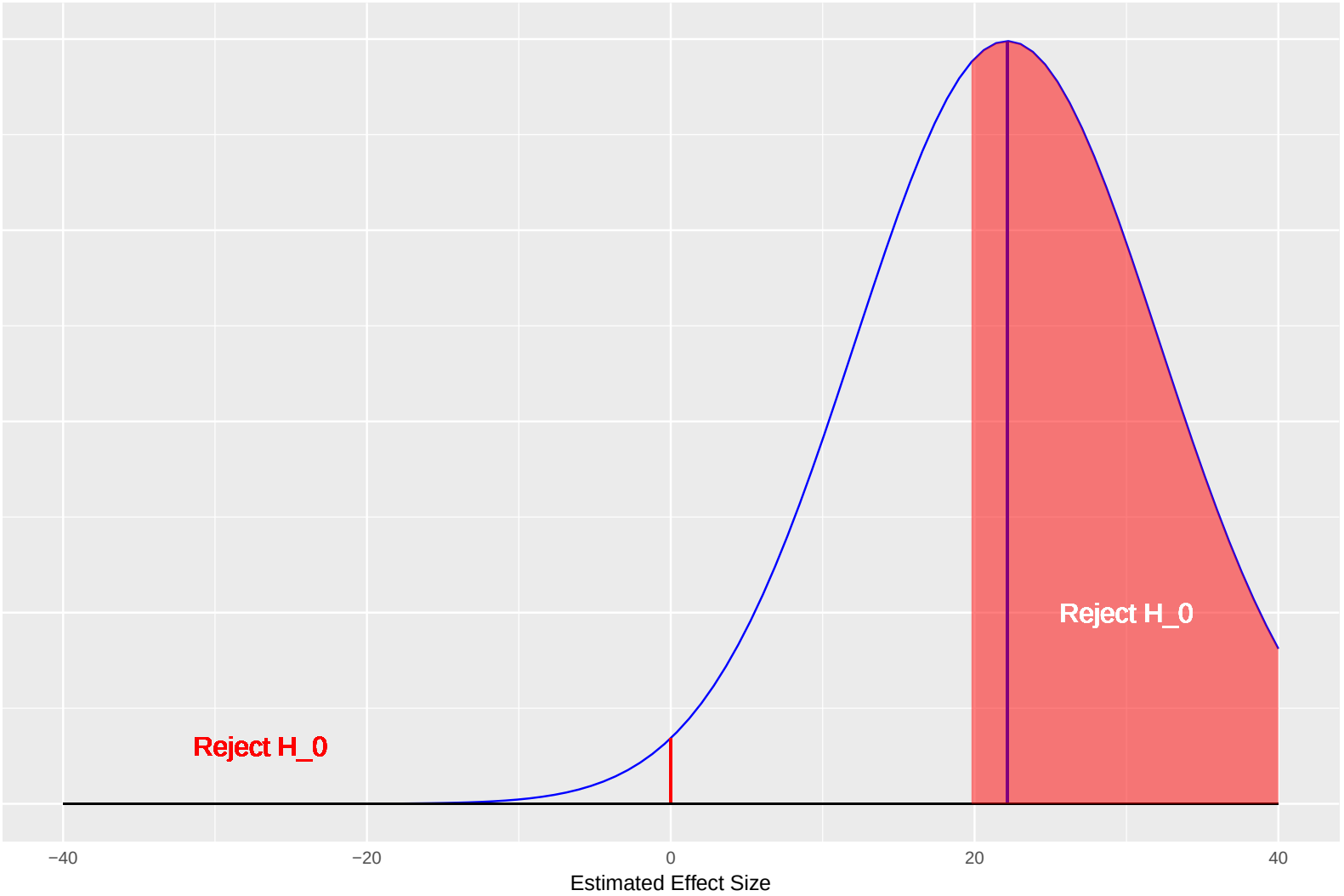
```
retrodesign(A = 12, s = 10)
```

```
$power
[1] 0.224427

$typeS
[1] 0.003515367

$exaggeration
[1] 2.107904
```

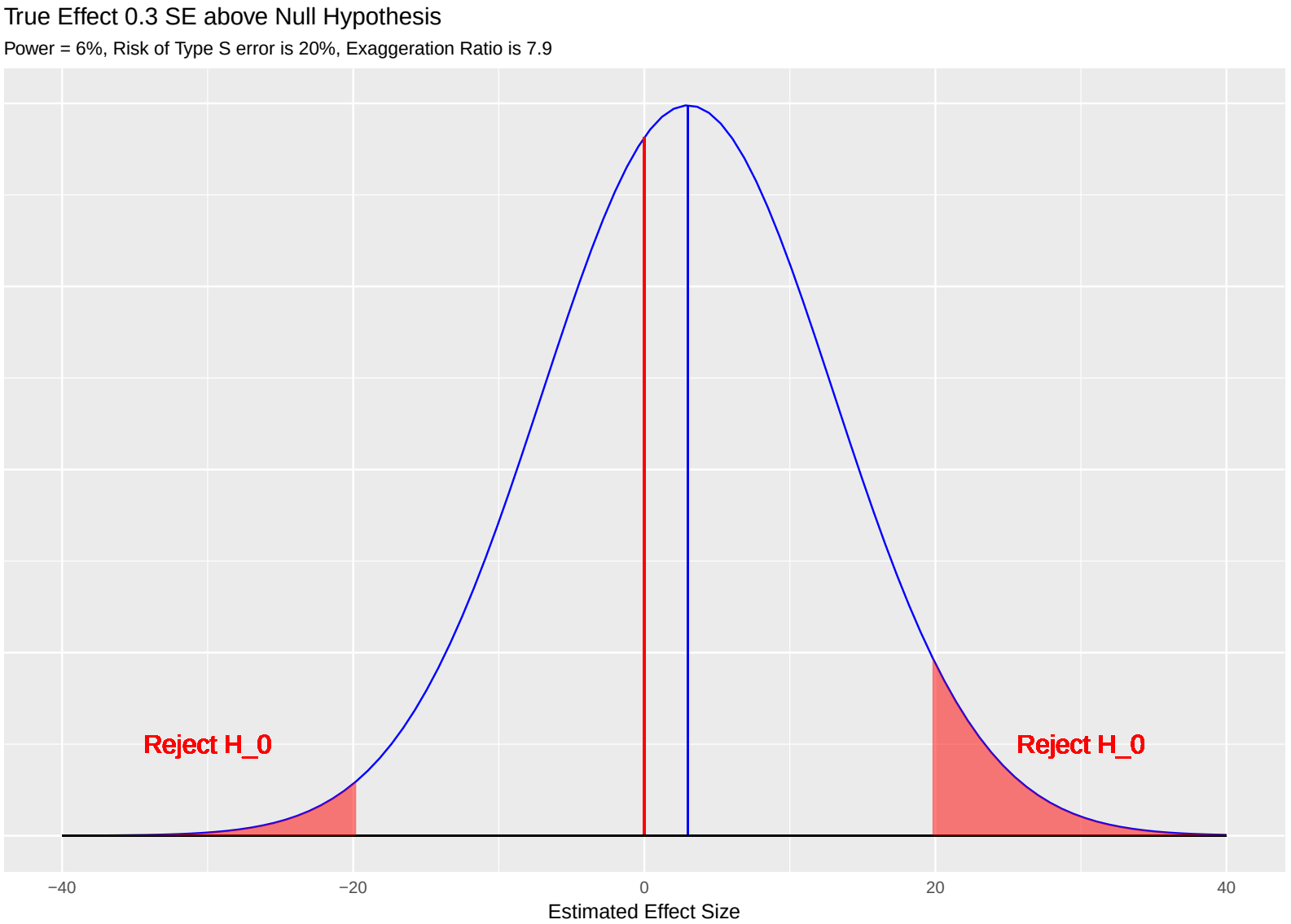
What 60% Power Looks Like

True Effect 2.215 SE above Null Hypothesis

Power = 0.60, Risk of Type S error is <0.01%, Exaggeration Ratio is about 1.3



What 6% power looks like. . .



Gelman & Carlin, Figure 2 (again)

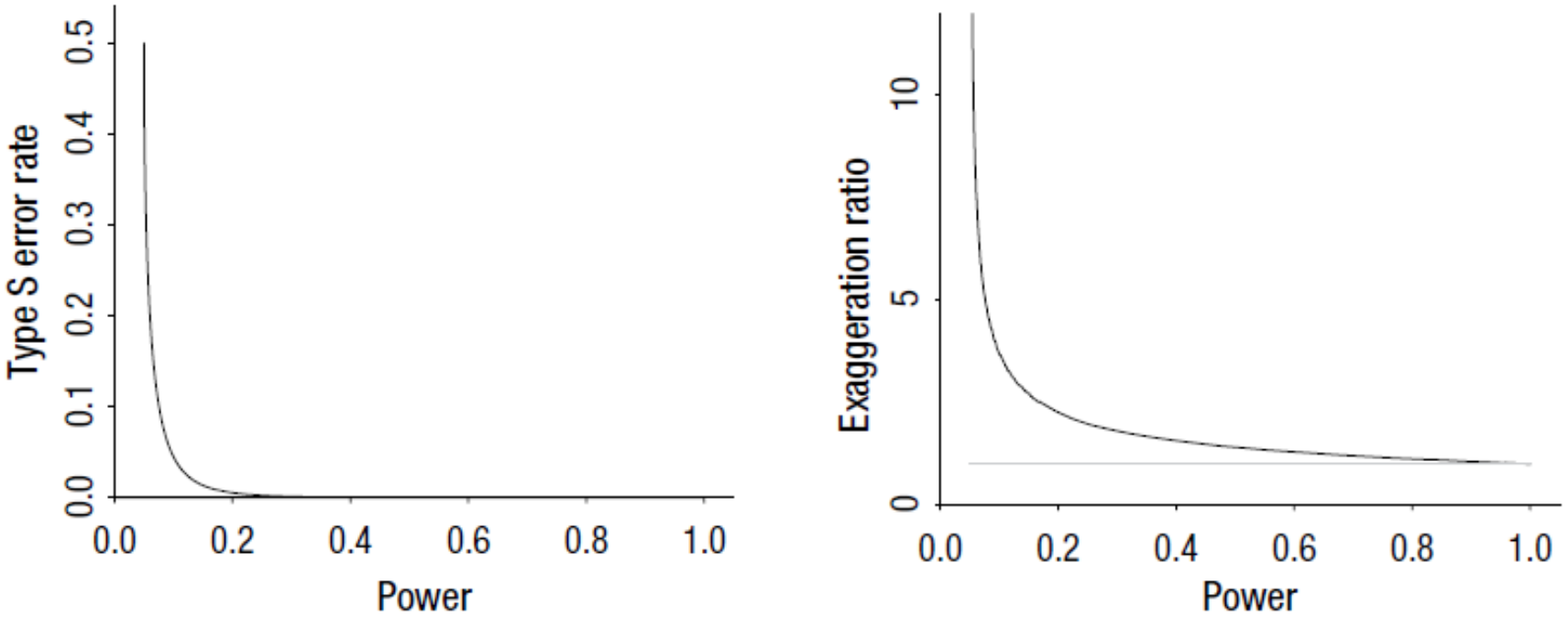


Figure 2. Type S error rate and exaggeration ratio as a function of statistical power for unbiased estimates that are normally distributed. If the estimate is unbiased, the power must be between 0.05 and 1.0, the Type S error rate must be less than 0.5, and the exaggeration ratio must be greater than 1. For studies with high power, the Type S error rate and the exaggeration ratio are low. But when power gets much below 0.5, the exaggeration ratio becomes high (that is, statistically significant estimates tend to be much larger in magnitude than true effect sizes). And when power goes below 0.1, the Type S error rate becomes high (that is, statistically significant estimates are likely to be the wrong sign).

Gelman's Chief Criticism: 6% Power = D.O.A.

My criticism of the ovulation-and-voting study is ultimately quantitative. Their effect size is tiny and their measurement error is huge. My best analogy is that they are trying to use a bathroom scale to weigh a feather . . . and the feather is resting loosely in the pouch of a kangaroo that is vigorously jumping up and down.

More from Gelman

How should we react to this?

- Statisticians such as myself should recognize that the point of criticizing a study is, in general, to shed light on statistical errors, maybe with the hope of reforming future statistical education.
- Researchers and policymakers should not just trust what they read in published journals.

http://andrewgelman.com/2016/03/11/statistics-is-like-basketball-or-knitting/

What I Think of as a Fun Read

How To Lie To Yourself and Others with Statistics

Eric Ravenscraft 2016-10-25 at Lifehacker

- Choose the Analysis That Supports Your Ideas
- Make Charts That Only Emphasize Your Pre-Conceived Conclusion
- Obscure Your Sources at All Costs
- Gather Sample Data that Adds Bias to Your Findings
    - Self-Selection Bias, Convenience Sampling, Non-Response Bias, Open-Access Polls

http://lifehacker.com/
how-to-lie-to-yourself-and-others-with-statistics-1788184031

For more on these ideas. . .

http://andrewgelman.com/2015/04/21/feather-bathroom-scale-kangaroo/

http://andrewgelman.com/2014/11/17/power-06-looks-like-get-used/

http://andrewgelman.com/2013/05/17/how-can-statisticians-help-psychologists-do-their-research-better/

http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Quiz 2 Setup

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