431 Class 22

Thomas E. Love

2017-11-14

Today's Agenda

- Quiz 2 Debrief
- *p* Hacking and other troubles with statistical significance
  - Researcher Degrees of Freedom
  - The Garden of Forking Paths
  - $p < 0.005$?
- Project Task C and The Course Survey
- A lot more on the problem of thinking about studies that are "finished"
  - Why *post hoc* power analysis doesn't really work
  - A Study of Facial Hair and Sexist Attitudes
  - Gelman and Carlin: Retrodesign
  - The Importance of Type S and Type M errors
  - The Beauty and Sex Ratios Study
  - The Ovulation and Voting Study

Today's R Setup

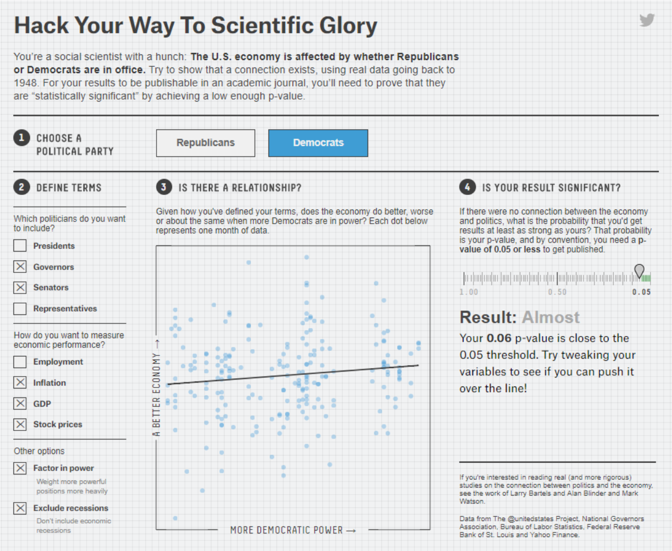
```
library(tidyverse)

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

Quiz 2

p Hacking and “Researcher Degrees of Freedom”

Hack Your Way To Scientific Glory

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



Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans [✓ Democrats]

2 DEFINE TERMS

Which politicians do you want to include?

□ Presidents

☒ Governors

☒ Senators

□ Representatives

How do you want to measure economic performance?

□ Employment

☒ Inflation

☒ GDP

☒ Stock prices

Other options

☒ Factor in power

Weight more powerful positions more heavily

☒ Exclude recessions

Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.

4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Almost

Your **0.06** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

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*.

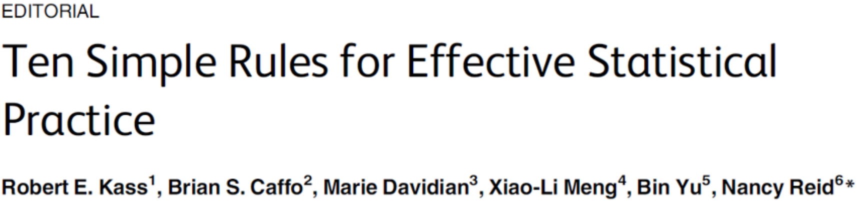
Lakens et al. Justify Your Alpha

"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*.

Being A More Effective / Transparent / Reproducible / Open Source Scientist

Ten Simple Rules for Effective Statistical Practice

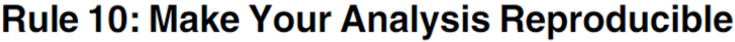
From *PLoS Comput Biol* link



EDITORIAL

Ten Simple Rules for Effective Statistical Practice

Robert E. Kass[1], Brian S. Caffo[2], Marie Davidian[3], Xiao-Li Meng[4], Bin Yu[5], Nancy Reid[6]*



Rule 10: Make Your Analysis Reproducible

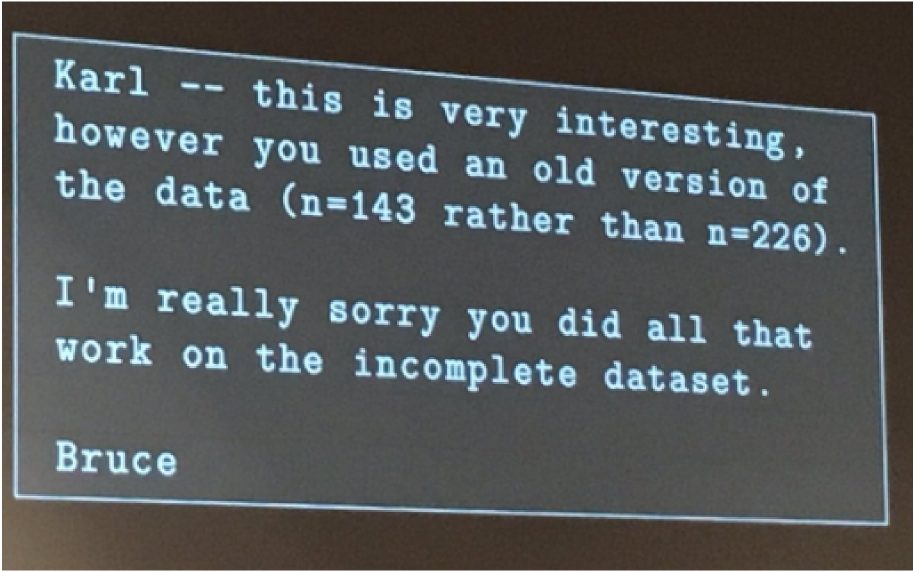
Goals of Reproducible Research

The goal of reproducible research is to tie specific instructions to data analysis so that scholarship can be recreated, better understood and verified. This is usually facilitated by literate programming: a document that combines content and data analytic code. Software? R and R Studio, mostly.

1. Be able to reproduce your own results and allow others to reproduce your results
2. Reproduce an entire report / manuscript / thesis / book / website with a single system command when changes occur (in operating system, statistical software, graphics engines, source data, derived variables, analysis, interpretation).
3. Save time.
4. Provide the ultimate documentation of work done.

Vanderbilt *Tutorial*

Why I Do This...



Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

Five Practical Tips

1. Document everything.
2. Everything is a (text) file.
3. All files should be human-readable.
4. Explicitly tie your files together.
5. Have a plan to organize, store and make your files available.

The papers/slideshows/abstracts are not the research. The research is the full software environment, code and data that produced the results. (Donoho, 2010). Separating research from its advertisement makes it hard for others to verify or reproduce our findings.

Your closest collaborator is you, six months ago, but you don't respond to emails. (Holder via Broman)

Karl Broman, Steps Towards Reproducible Research *link*

Build Tidy Data Sets

- Each variable you measure should be in one column.
- Each different observation of that variable should be in a different row.
- There should be one table for each "kind" of variable.
- If you have multiple tables, they should include a column in the table that allows them to be linked.
- Include a row at the top of each data table that contains real row names. `Age_at_Diagnosis` is a much much better name than `ADx`.
- Build useful codebooks.

Jeff Leek: "How to share data with a statistician" *link*

Wisdom from DL Donoho (2010) re: Open-Source

But other people will use my data and code to compete with me?

- True.

Wisdom from DL Donoho (2010) re: Open-Source

But other people will use my data and code to compete with me?

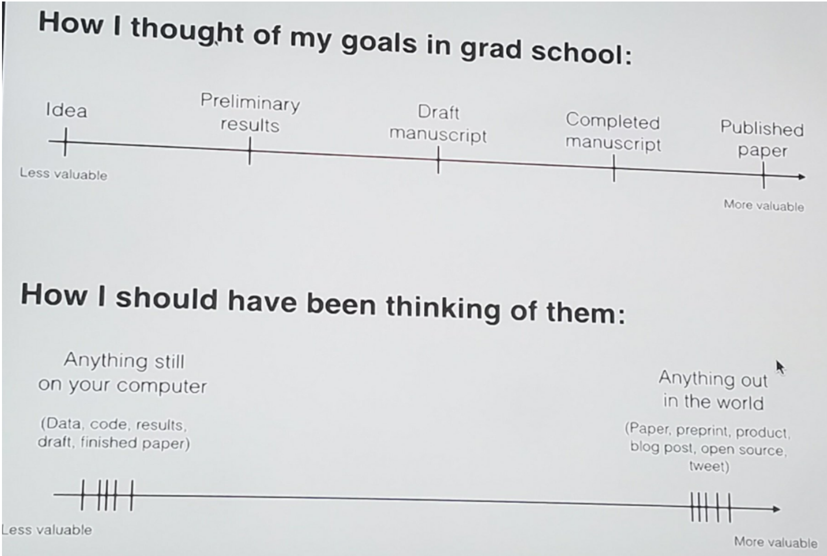
- True.
- But competition means that strangers will read your work, try to learn from you, cite you, and try to do things even better.

Wisdom from DL Donoho (2010) re: Open-Source

But other people will use my data and code to compete with me?

- True.
- But competition means that strangers will read your work, try to learn from you, cite you, and try to do things even better.
- If you prefer obscurity, why are you publishing?

A Most Important Tip (@drob)



How I thought of my goals in grad school:

Idea

Preliminary results

Draft manuscript

Completed manuscript

Published paper

Less valuable

More valuable

How I should have been thinking of them:

Anything still on your computer

(Data, code, results, draft, finished paper)

Anything out in the world

(Paper, preprint, product, blog post, open source, tweet)

Less valuable

More valuable

Project Task C discussion

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.

Design Analysis instead of Post Hoc Power

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.

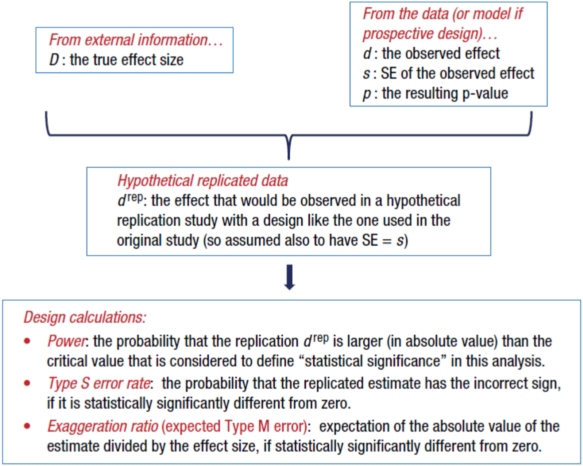
- Looks significant if $|d/s| > 2$, which roughly corresponds to $p < 0.05$. Inconclusive otherwise.
- Now, consider a true effect size D (the value that d would take if you had an enormous sample)
- D is hypothesized based on *external* information (Other available data, Literature review, Modeling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



From external information…
D : the true effect size

From the data (or model if prospective design)…
d : the observed effect
s : SE of the observed effect
p : the resulting p-value

Hypothetical replicated data
d^{rep}: the effect that would be observed in a hypothetical replication study with a design like the one used in the original study (so assumed also to have SE = s)

Design calculations:

- *Power*: the probability that the replication d^{rep} is larger (in absolute value) than the critical value that is considered to define "statistical significance" in this analysis.
- *Type S error rate*: the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero.
- *Exaggeration ratio* (expected Type M error): expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero.

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)

```r
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.

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

```
$power
[1] 0.7995569

$typeS
[1] 1.210843e-06

$exaggeration
[1] 1.129545
```

- 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.

An example: Beauty and Sex Ratios

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] 76.50475
```

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:

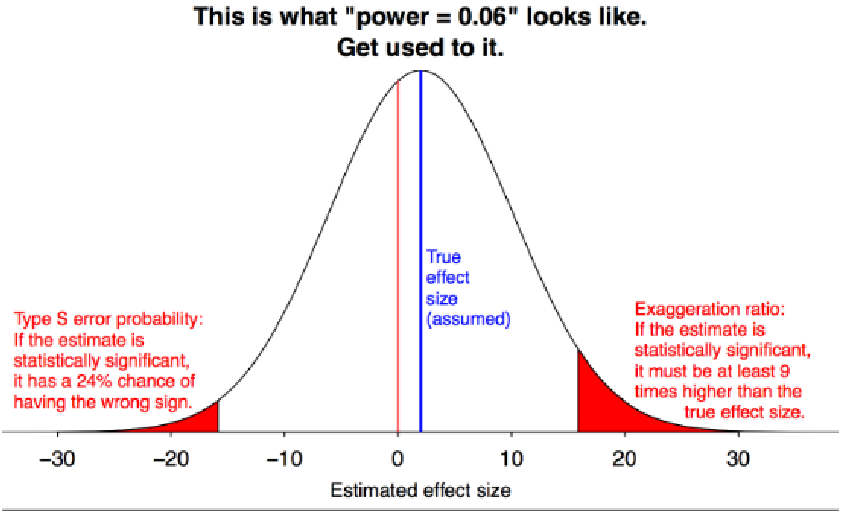
1. There is a 46% chance it will have the wrong sign (from the Type S error rate).
2. The power is 5% and the Type S error rate of 46%. Multiplying those gives a 2.3% probability that we will find a statistically significant result in the wrong direction.
3. We thus have a power - 2.3% = 2.7% probability 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

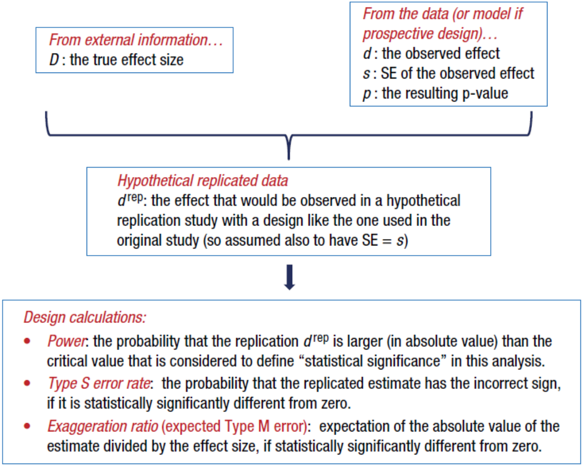
Assumption	Power	Type S	Exaggeration Ratio
D = 0.1	0.05	0.46	78
D = 0.3	0.05	0.39	25
D = 1.0	0.06	0.19	7.8

- 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)



From external information…
D : the true effect size

From the data (or model if prospective design)…
d : the observed effect
s : SE of the observed effect
p : the resulting p-value

Hypothetical replicated data
d^{rep}: the effect that would be observed in a hypothetical replication study with a design like the one used in the original study (so assumed also to have SE = s)

Design calculations:

- *Power*: the probability that the replication d^{rep} is larger (in absolute value) than the critical value that is considered to define “statistical significance” in this analysis.
- *Type S error rate*: the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero.
- *Exaggeration ratio* (expected Type M error): expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero.

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.

Another example: Ovulation and Voting

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!

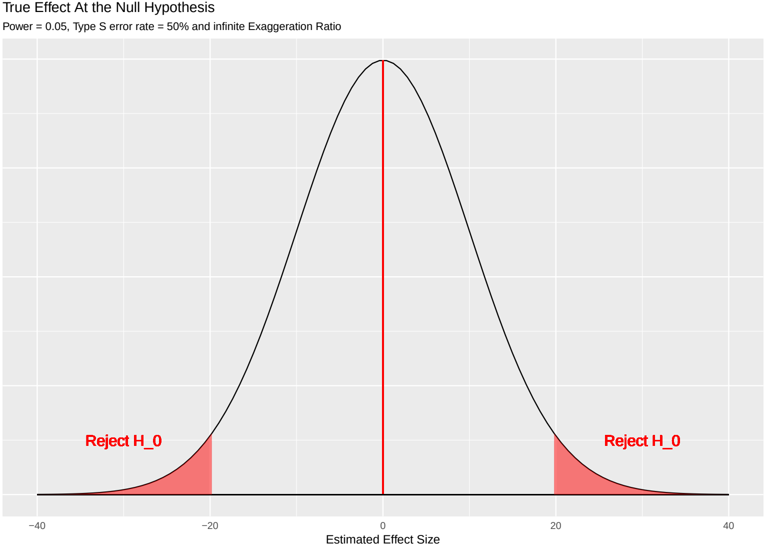
What to do?

- 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


True Effect At the Null Hypothesis

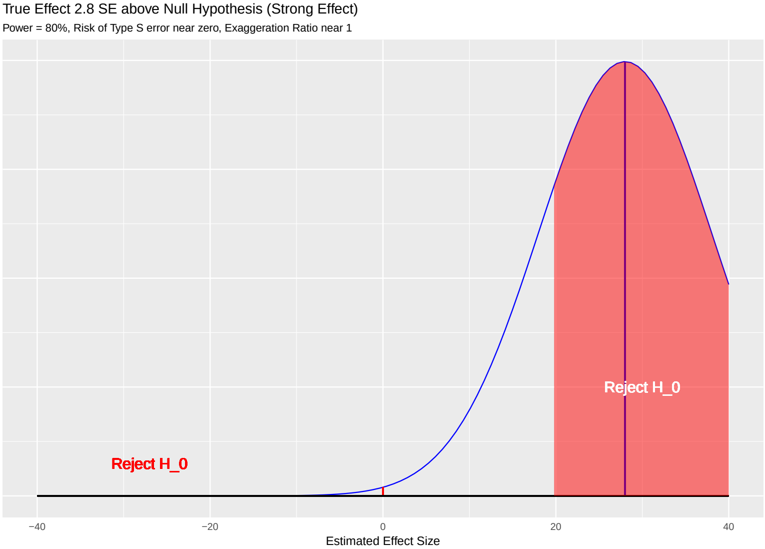
Power = 0.05, Type S error rate = 50% and infinite Exaggeration Ratio

retrodesign for Zero Effect

```r
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

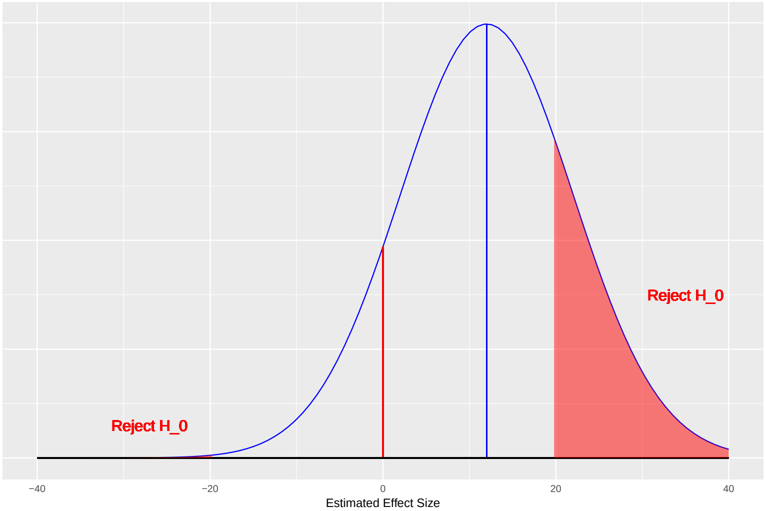
```r
retrodesign(A = 28, s = 10)
```

```
$power
[1] 0.7995569

$typeS
[1] 1.210843e-06

$exaggeration
[1] 1.123588
```

What 23% power looks like...

True Effect 1.2 SE above Null Hypothesis

Power = 23%, Risk of Type S error is 0.004, Exaggeration Ratio is over 2

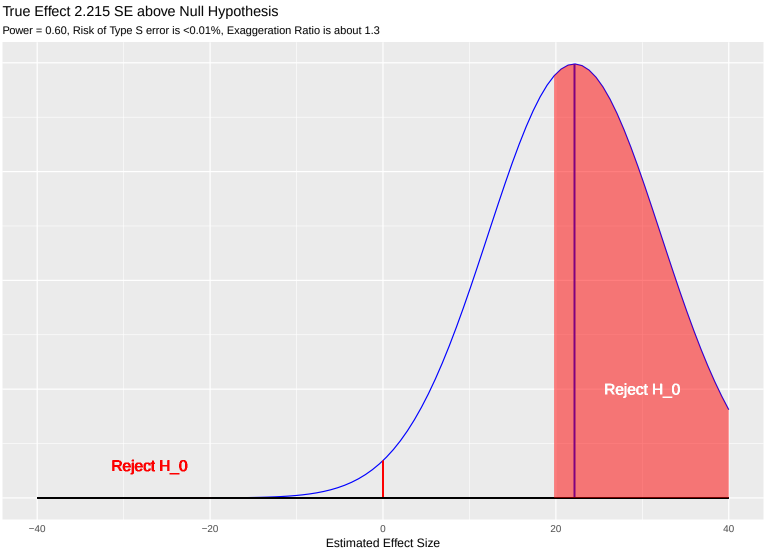


retrodesign for a true effect 1.2 SE above H_0

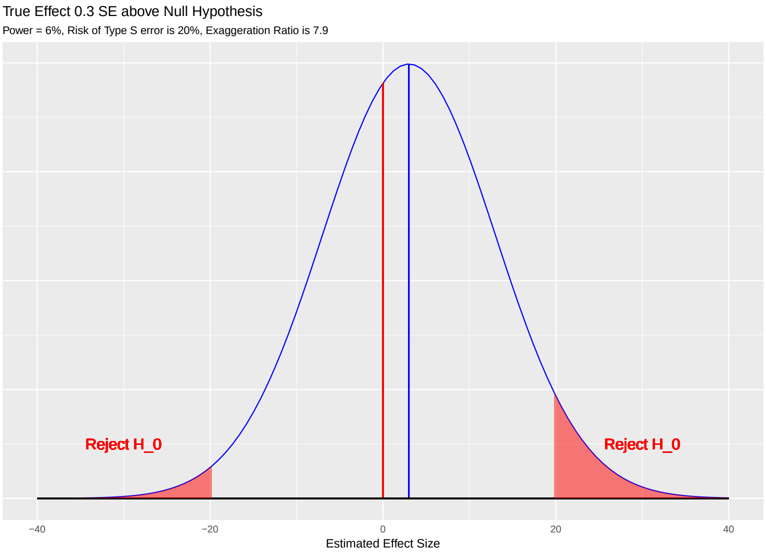
```r
retrodesign(A = 12, s = 10)
```

```
$power
[1] 0.224427

$typeS
[1] 0.003515367

$exaggeration
[1] 2.115428
```

What 60% Power Looks Like



What 6% power looks like. . .



Gelman & Carlin, Figure 2 (again)

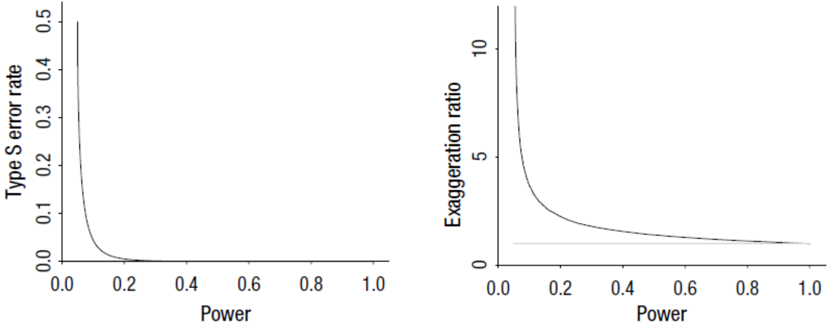


Figure 2. Type S error rate and exaggeration ratio as a function of statistical power for unbiased estimates that are normally distributed. If the estimate is unbiased, the power must be between 0.05 and 1.0, the Type S error rate must be less than 0.5, and the exaggeration ratio must be greater than 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