432 Class 23 Slides

github.com/THOMASELOVE/2019-432

2019-04-23

Today's Topics

- ATOM
- Replicable Research and the Crisis in Science
- Retrospective Power and why most smart folks avoid it
    - Type S and Type M error: Saying something more useful

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

- Statistical methods do not rid data of their uncertainty.

> Statistical methods do not rid data of their uncertainty. "Statistics," Gelman (2016) says, "is often sold as a sort of alchemy that transmutes randomness into certainty, an 'uncertainty laundering' that begins with data and concludes with success as measured by statistical significance." To accept uncertainty requires that we "treat statistical results as being much more incomplete and uncertain than is currently the norm" (Amrhein, Trafimow, and Greenland 2019). We must "countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error" (Calin-Jageman and Cumming 2019).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

> *We can make acceptance of uncertainty more natural to our thinking by accompanying every point estimate in our research with a measure of its uncertainty such as a standard error or interval estimate. Reporting and interpreting point and interval estimates should be routine.*
>
> *How will accepting uncertainty change anything? To begin, it will prompt us to seek better measures, more sensitive designs, and larger samples, all of which increase the rigor of research.*
>
> *It also helps us be modest ... [and] leads us to be thoughtful.*

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

3.2. Be Thoughtful

What do we mean by this exhortation to "be thoughtful"? Researchers already clearly put much thought into their work. We are not accusing anyone of laziness. Rather, we are envisioning a sort of "statistical thoughtfulness." In this perspective, statistically **thoughtful researchers** begin above all else with clearly expressed objectives. They recognize when they are doing exploratory studies and when they are doing more rigidly pre-planned studies. They invest in producing solid data. They consider not one but a multitude of data analysis techniques. And they think about so much more.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research looks ahead to prospective outcomes in the context of theory and previous research. Researchers would do well to ask, *What do we already know, and how certain are we in what we know?* And building on that and on the field's theory, *what magnitudes of differences, odds ratios, or other effect sizes are practically important?* These questions would naturally lead a researcher, for example, to use existing evidence from a literature review to identify specifically the findings that would be practically important for the key outcomes under study.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research includes careful consideration of the definition of a meaningful effect size. As a researcher you should communicate this up front, before data are collected and analyzed. Afterwards is just too late; it is dangerously easy to justify observed results after the fact and to overinterpret trivial effect sizes as being meaningful. Many authors in this special issue argue that consideration of the effect size and its "scientific meaningfulness" is essential for reliable inference (e.g., Blume et al. 2019; Betensky 2019). This concern is also addressed in the literature on equivalence testing (Wellek 2017).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Thoughtful research considers "related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain…without giving priority to *p*-values or other purely statistical measures" (McShane et al. 2019).

Thoughtful researchers "use a toolbox of statistical techniques, employ good judgment, and keep an eye on developments in statistical and data science," conclude Heck and Krueger (2019), who demonstrate how the *p*-value can be useful to researchers as a heuristic.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

In all instances, regardless of the value taken by p or any other statistic, consider what McShane et al. (2019) call the "currently subordinate factors"—the factors that should no longer be subordinate to "$p < 0.05$." These include relevant prior evidence, plausibility of mechanism, study design and data quality, and the real-world costs and benefits that determine what effects are scientifically important. The scientific context of your study matters, they say, and this should guide your interpretation.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

To **be open**, remember that one study is rarely enough. The words "a groundbreaking new study" might be loved by news writers but must be resisted by researchers. Breaking ground is only the first step in building a house. It will be suitable for habitation only after much more hard work.

Be open by providing sufficient information so that other researchers can execute meaningful alternative analyses. van Dongen et al. (2019) provide an illustrative example of such alternative analyses by different groups attacking the same problem.

Being open goes hand in hand with **being modest.**

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Being modest requires a reality check (Amrhein, Trafimow, and Greenland 2019). "A core problem," they observe, "is that both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct and of the reality being studied.

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Be modest in recognizing there is not a "true statistical model" underlying every problem, which is why it is wise to **thoughtfully** consider many possible models (Lavine 2019).

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

Be modest about the role of statistical inference in scientific inference. "Scientific inference is a far broader concept than statistical inference," says Hubbard, Haig, and Parsa (2019). "A major focus of scientific inference can be viewed as the pursuit of *significant sameness,* meaning replicable and empirically generalizable results among phenomena. Regrettably, the obsession with users of statistical inference to report *significant differences* in data sets actively thwarts cumulative knowledge development."

ATOM: Accept uncertainty. Be Thoughtful, Open and Modest.

The nexus of openness and modesty is to report everything while at the same time not concluding anything from a single study with unwarranted certainty. Because of the strong desire to inform and be informed, there is a relentless demand to state results with certainty. Again, accept uncertainty and embrace variation in associations and effects, because they are always there, like it or not. Understand that expressions of uncertainty are themselves uncertain. Accept that one study is rarely definitive, so encourage, sponsor, conduct, and publish replication studies.

Be modest by encouraging others to reproduce your work. Of course, for it to be reproduced readily, you will necessarily have been thoughtful in conducting the research and open in presenting it.

Your Assignment for Today

I asked you to read one of these three pieces from the 2019 TAS supplement:

1. John P. A. Ioannidis What Have We (Not) Learnt from Millions of Scientific Papers with P Values?
2. Sherri Rose & Thomas G. McGuire (2019) Limitations of P-Values and R-squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment
3. Robert J. Calin-Jageman & Geoff Cumming (2019) The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known

You should have written:

1. In a sentence, restate the most important thing you learned from the article you read.
2. In a sentence, tell us how you could use this article to change your behavior, or the behavior of other people you are doing science with.

Evaluation through Retrospective Design

Gelman, 2016-03-11

Reviewing "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

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

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 should we react to this?

Gelman:

- 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.

Assessing Type S (Sign) and Type M (Magnitude) Errors

- Gelman and Carlin *Psychological Science* 2014 9(6): 641-651.

Thinking About Power

Specifying 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 on 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)

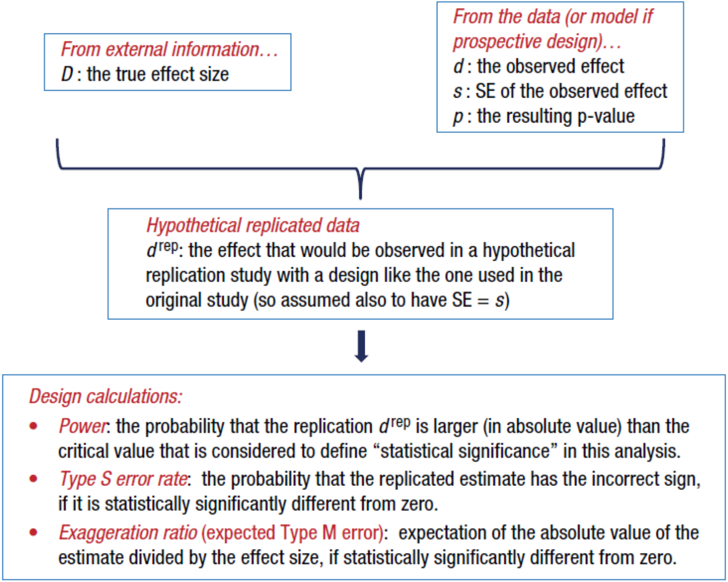


Figure 1. Diagram of our recommended approach to design analysis. It will typi-

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

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))
}
```

What if we have a beautiful, unbiased study?

Suppose the true effect that is 2.8 standard errors away from zero, in a study built to have 80% power for that effect with 95% confidence.

```
set.seed(201803161)
retrodesign(A = 28, s = 10, alpha = 0.05)
```

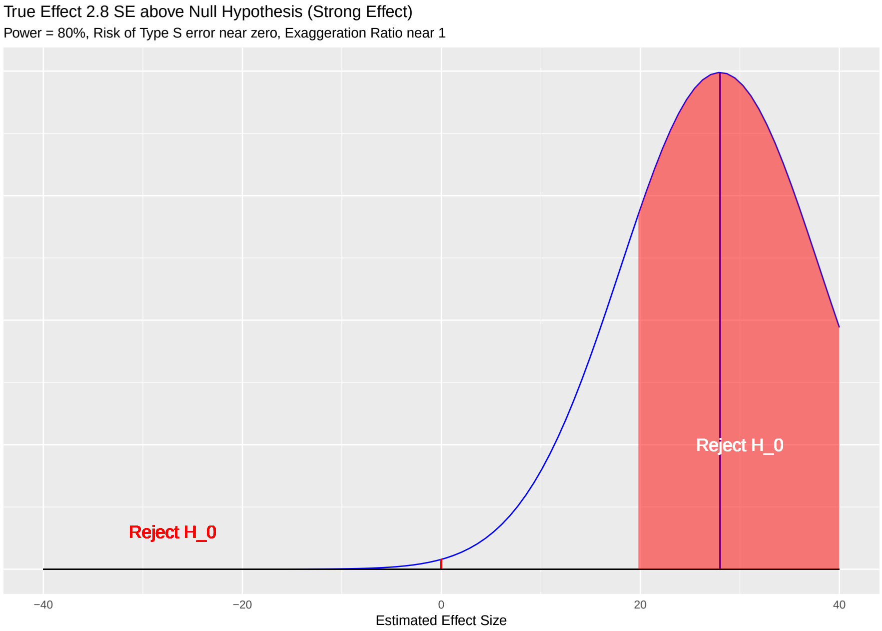
```
$power
[1] 0.7995569

$typeS
[1] 1.210843e-06

$exaggeration
[1] 1.12875
```

What if we have a beautiful, unbiased study?

power	typeS	exaggeration
0.79956	1.2×10^{-6}	1.13

- With the power this high (80%), we have a type S error rate of 1.2×10^{-6} and an expected exaggeration factor of 1.13.
- Nothing to worry about with either direction of a statistically significant estimate and the overestimation of the magnitude of the effect will be small.
- What does this look like?

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



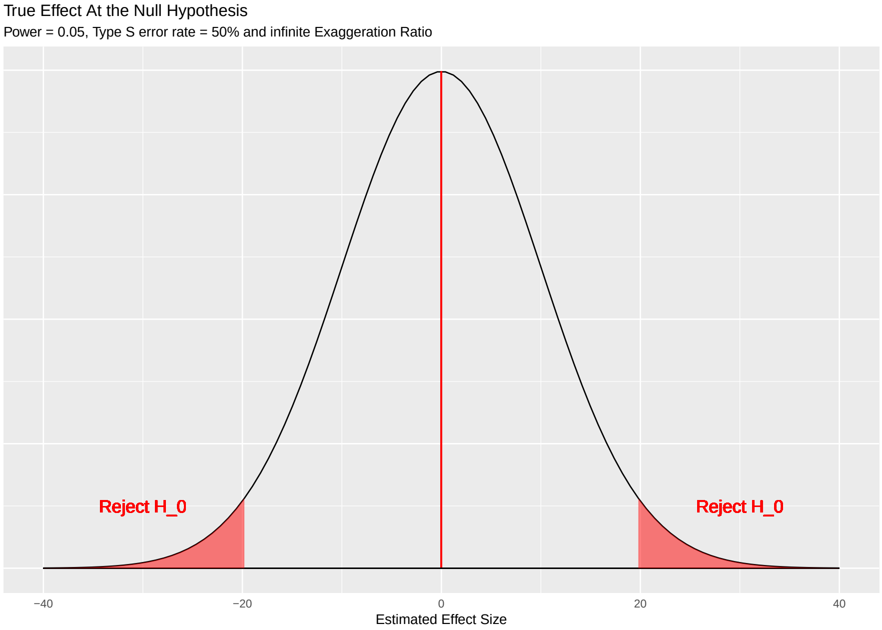
`retrodesign` for Zero Effect

```
set.seed(201803162)
retrodesign(A = 0, s = 10)
```

```
$power
[1] 0.05

$typeS
[1] 0.5

$exaggeration
[1] Inf
```

- Power = 0.5, Pr(Type S error) = 0.5, Exaggeration Ratio is infinite.

Power, Type S and Type M Errors: Zero Effect

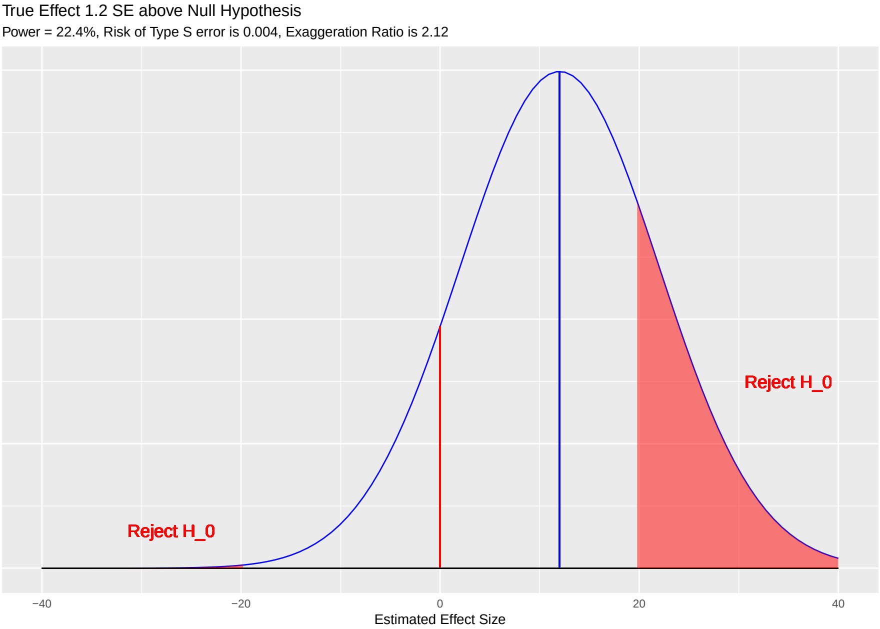


Retrodesign for a true effect 1.2 SE above H_0

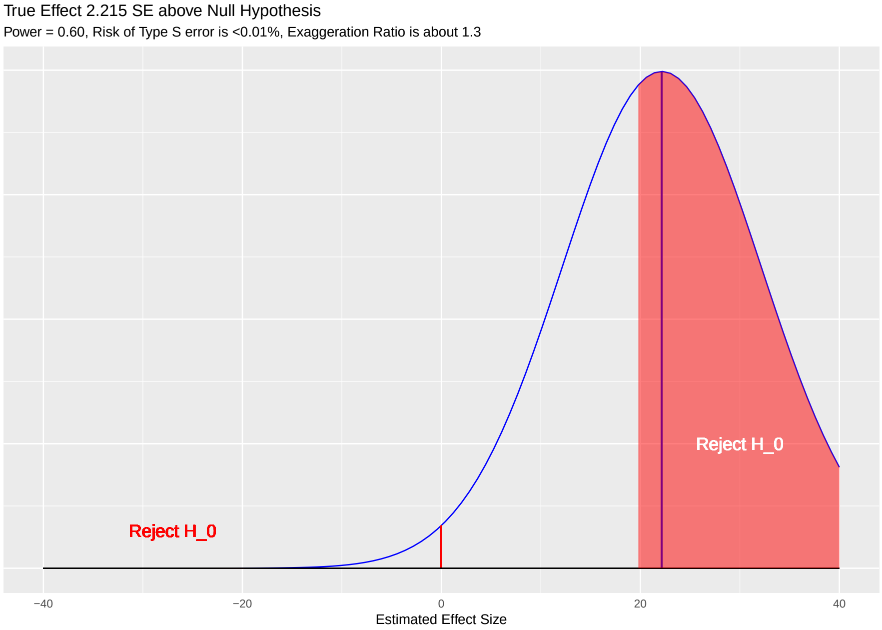
```
set.seed(201803163)
retrodesign(A = 12, s = 10)
```

```
$power
[1] 0.224427

$typeS
[1] 0.003515367

$exaggeration
[1] 2.117846
```

What 22.4% power looks like...



What 60% Power Looks Like



Gelman & Carlin, Figure 2

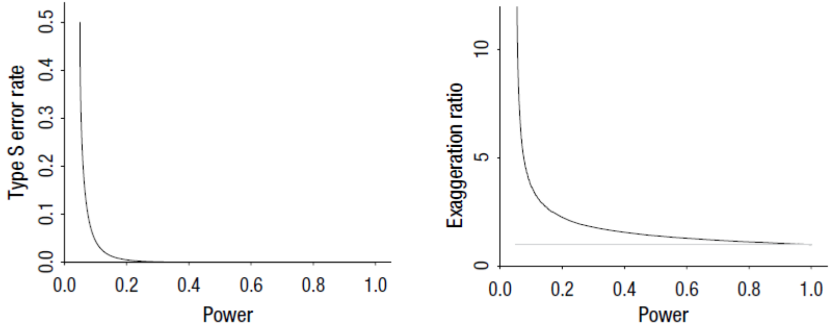


Figure 2. Type S error rate and exaggeration ratio as a function of statistical power for unbiased estimates that are normally distributed. If the estimate is unbiased, the power must be between 0.05 and 1.0, the Type S error rate must be less than 0.5, and the exaggeration ratio must be greater than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

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 $\alpha = 0.05$

```r
set.seed(201803164)
retrodesign(A = 0.1, s = 3.29, alpha = 0.05)
```

```
$power
[1] 0.05010584

$typeS
[1] 0.4645306

$exaggeration
[1] 76.93614
```

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:

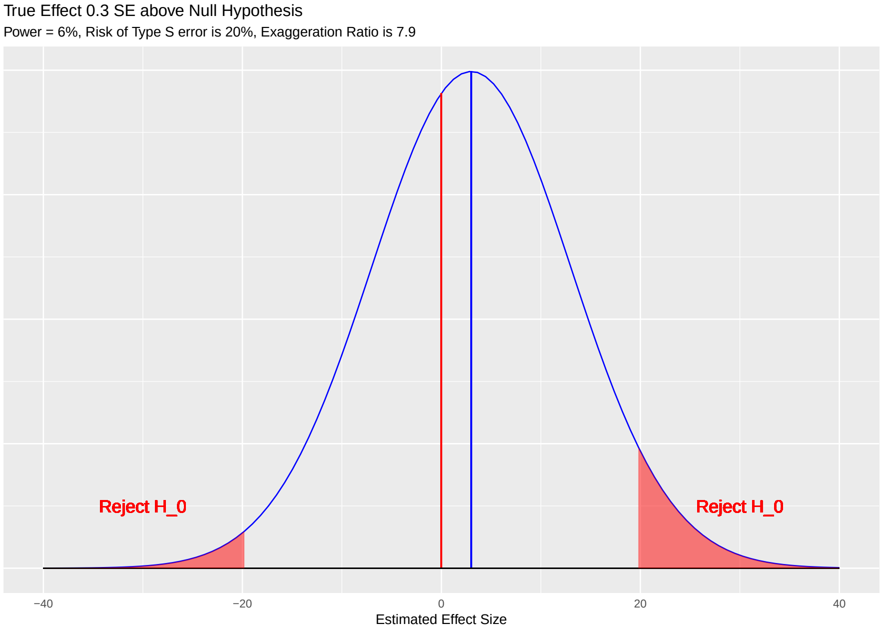
1. There is a 46% chance it will have the wrong sign (from the Type S error rate).
2. The power is 5% and the Type S error rate of 46%. Multiplying those gives a 2.3% probability that we will find a statistically significant result in the wrong direction.
3. We thus have a power - 2.3% = 2.7% probability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

What 6% power looks like. . .



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

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.

What to do?

In advance, **and** after the fact, think hard about what a plausible effect size might be.

Then...

- Analyze *all* your data.
- 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.

But I do studies with 80% power?

Based on some reasonable assumptions regarding main effects and interactions (specifically that the interactions are half the size of the main effects), you need **16 times** the sample size to estimate an interaction that you need to estimate a main effect.

> *And this implies a major, major problem with the usual plan of designing a study with a focus on the main effect, maybe even preregistering, and then looking to see what shows up in the interactions.*
>
> *Or, even worse, designing a study, not finding the anticipated main effect, and then using the interactions to bail you out. The problem is not just that this sort of analysis is "exploratory"; it's that these data are a lot noisier than you realize, so what you think of as interesting exploratory findings could be just a bunch of noise.*

- Gelman 2018-03-15

What I Think I Think Now

- Null hypothesis significance testing is much harder than I thought.
    - The null hypothesis is almost never a real thing.
    - Rather than rejiggering the cutoff, I would mostly abandon the p value as a summary
    - Replication is far more useful than I thought it was.
- Some hills aren't worth dying on.
    - Think about uncertainty intervals more than confidence or credible intervals
    - Retrospective calculations about Type S (sign) and Type M (magnitude) errors can help me illustrate ideas.
- Which method to use is far less important than finding better data
    - The biggest mistake I make regularly is throwing away useful data
    - I'm not the only one with this problem.
- The best thing I do most days is communicate more clearly.
    - When stuck in a design, I think about how to get better data.
    - When stuck in an analysis, I try to turn a table into a graph.
- I have A LOT to learn.

Next Time. . .

- Guest Speaker Nik Krieger in the early part of class
- Some Thoughts on Matching in Observational Studies to estimate Causal Effects - why you might want to take the PQHS 500 / CRSP 500 course
- Quiz 2