432 Class 14

https://thomaselove.github.io/432-2025

2025-02-27

## Today’s Agenda

Evaluation of a Study through Retrospective Design

* Gelman and Carlin: Type S and Type M errors
* The retrodesign() function

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(tidyverse)  
theme\_set(theme\_bw())   
  
retrodesign <- function(D, s, alpha=.05, df=Inf, n.sims=10000)  
 {  
 z <- qt(1-alpha/2, df)  
 p.hi <- 1 - pt(z-D/s, df)  
 p.lo <- pt(-z-D/s, df)  
 power <- p.hi + p.lo  
 typeS <- p.lo/power  
 estimate <- D + s\*rt(n.sims,df)  
 significant <- abs(estimate) > s\*z  
 exaggeration <- mean(abs(estimate)[significant])/D  
 return(list(power=power, typeS=typeS,   
 exaggeration=exaggeration))  
}

## On “post hoc” power analysis

Suppose you read a study, and the result appears to not meet the standard of statistical significance that you wish it did. The idea is to show that a “non-significant” hypothesis test failed to achieve significance because it wasn’t powerful enough.

* Knowing what we have learned from this new study, what sort of power did the study have to detect the effect that we saw? (a *post hoc* power calculation)
* What sort of sample size might we use in a new study? (maybe useful?)
* Was this study dead in the water before we did it?

## To be clear, Post hoc power calculations are not useful

[Post Hoc Power Calculations Are Not Useful](https://library.virginia.edu/data/articles/post-hoc-power-calculations-are-not-useful) from University of Virginia’s Research Data Services cites these papers.

* [Althouse A (2021)](https://pubmed.ncbi.nlm.nih.gov/32814615/)
* [Goodman SN Berlin JA (1994)](https://pubmed.ncbi.nlm.nih.gov/8017747/)
* [Hoenig JM Heisey DM (2001)](https://www.tandfonline.com/doi/abs/10.1198/000313001300339897)

This last piece is from *The American Statistician* and is entitled “The abuse of power: The pervasive fallacy of power calculations for data analysis.” Much wisdom there.

## Andrew Gelman’s blog

Blog is [Statistical Modeling, Causal Inference, and Social Science](https://statmodeling.stat.columbia.edu/)

* [Statistics is like basketball, or knitting](https://statmodeling.stat.columbia.edu/2016/03/11/statistics-is-like-basketball-or-knitting/) from 2016-03-11
* The post discusses [this article](https://pubmed.ncbi.nlm.nih.gov/26510427/), shown on the next slide.

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



## Facial Hair and Sexist Attitudes

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 . . .

## Gelman, 2016-03-11 Blog

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

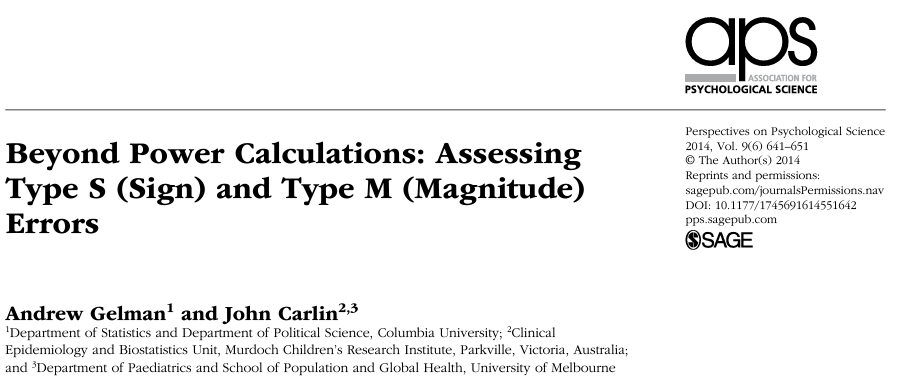
* All credit to the researchers for admitting that they did this, but poor practice of them to present their result in the abstract to their paper without making this clear, and too bad that the journal got suckered into publishing this.

## 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.

## Gelman and Carlin (2014) [PDF Link](http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf)



## Specifying effect sizes - how? (1/2)

When doing a power calculation, how do people specify an effect size of interest? Two main approaches…

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

## Specifying effect sizes - how? (2/2)

When doing a power calculation, how do people specify an effect size of interest? Two main approaches…

1. **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 approaches lead to performing studies that are too small or misinterpretation of findings after completion.

## What is a design analysis?

* 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.

## Why a design analysis?

* The previous slide’s 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?

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 - results very misleading - power is usually seriously overestimated when computed on the basis of “significant” results.
   * Much better (but rarer) to identify plausible effect sizes based on external information rather than on your sparkling new result.

## Why not post hoc power analysis?

1. 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.

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

## Broader Design Ideas

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.

## The Value of Type S and Type M error

These ideas 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.

## Building Blocks (1/2)

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

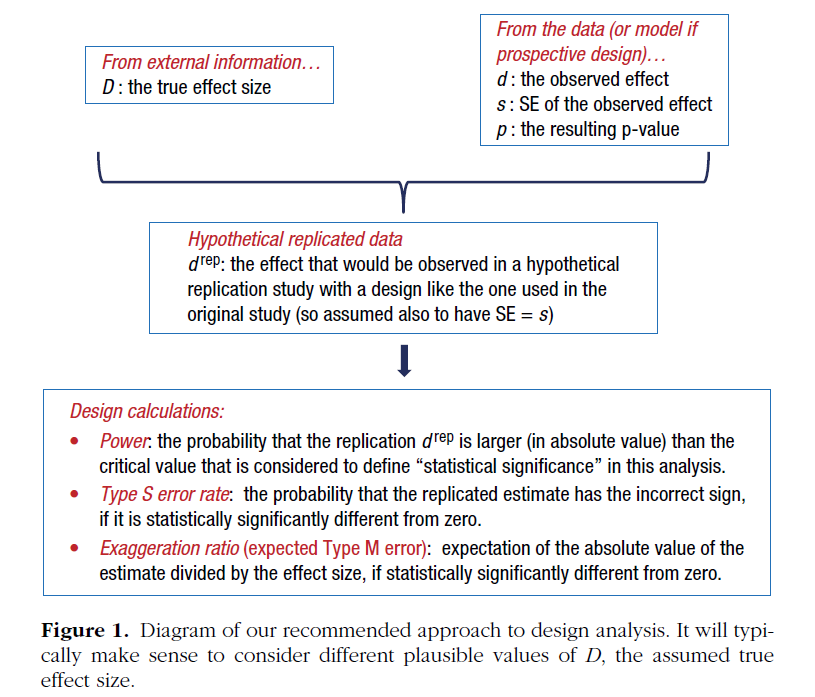
* Looks significant if , which roughly corresponds to *p* < 0.05. Inconclusive otherwise.
* Now, consider a true effect size *D* (the value that *d* would take if you had an enormous sample)
* *D* is hypothesized based on *external* information (Other available data, Literature review, sometimes Modeling,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

## Building Blocks (2/2)

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

* Define 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)



## Retrodesign function (R code coming)

Inputs to the function:

* D, the hypothesized true effect size
* s, the standard error of the estimate
* alpha, the statistical significance threshold (default 0.05)
* df, the degrees of freedom (default assumption: infinite)

## Retrodesign function (R code coming)

Output:

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

## Retrodesign function

* built by Gelman and Carlin

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

## A beautiful, unbiased study?

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

set.seed(4320227)  
retrodesign(D = 28, s = 10, alpha = 0.05)

$power  
[1] 0.7995569  
  
$typeS  
[1] 1.210843e-06  
  
$exaggeration  
[1] 1.126261

## A beautiful, unbiased study?

| power | typeS | exaggeration |
| --- | --- | --- |
| 0.79956 |  | 1.13 |

* With the power this high (80%), we have a type S error rate of 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 Standard Errors above the value…

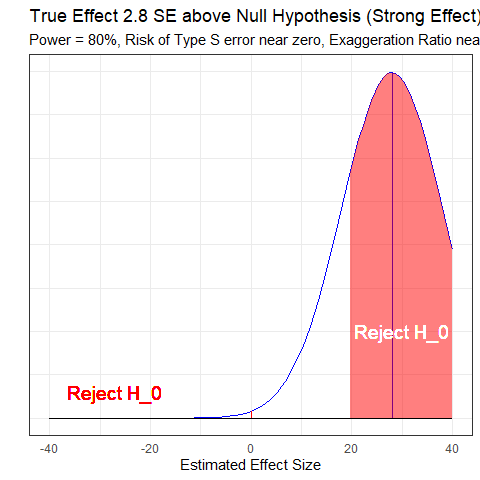
x <- seq(-40, 40, length = 100)  
hx0 <- dnorm(x, mean = 0, sd = 10)  
hx3 <- dnorm(x, mean = 3, sd = 10)  
hx12 <- dnorm(x, mean = 12, sd = 10)  
hx28 <- dnorm(x, mean = 28, sd = 10)  
hx2215 <- dnorm(x, mean = 22.15, sd = 10)  
dat <- data.frame(x, hx0, hx3, hx12, hx28, hx2215)

ggplot(dat, aes(x, hx28)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 28, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 28, xend = 28, y = 0, yend = dnorm(28, mean = 28, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx28), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx28), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.01, label = "Reject H\_0", col = "white", size = 5) +  
 geom\_text(x = -27, y = 0.003, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 2.8 SE above Null Hypothesis (Strong Effect)", subtitle = "Power = 80%, Risk of Type S error near zero, Exaggeration Ratio near 1")

Warning in geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 28, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = 28, xend = 28, y = 0, yend = dnorm(28, mean = 28, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)): All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.



## retrodesign for Zero Effect

set.seed(43202272)  
retrodesign(D = 0, s = 10)

$power  
[1] 0.05  
  
$typeS  
[1] 0.5  
  
$exaggeration  
[1] Inf

* Power = 0.05,
* Pr(Type S error) = 0.50,
* Exaggeration Ratio is infinite.

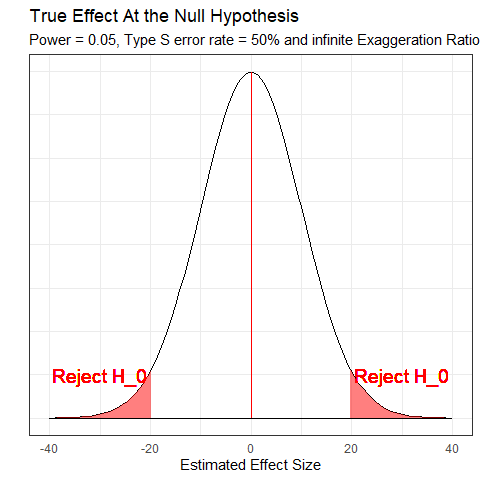
## Power, Type S and Type M Errors

* Zero Effect

ggplot(dat, aes(x, hx0)) +  
 geom\_line() +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 0, sd = 10)), col = "red") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx0), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx0), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 geom\_text(x = -30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect At the Null Hypothesis", subtitle = "Power = 0.05, Type S error rate = 50% and infinite Exaggeration Ratio")

Warning in geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 0, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)): All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.



## Retrodesign (moderate effect)

* Suppose the true effect was 1.2 Standard Errors above

set.seed(43202273)  
retrodesign(D = 12, s = 10)

$power  
[1] 0.224427  
  
$typeS  
[1] 0.003515367  
  
$exaggeration  
[1] 2.120127

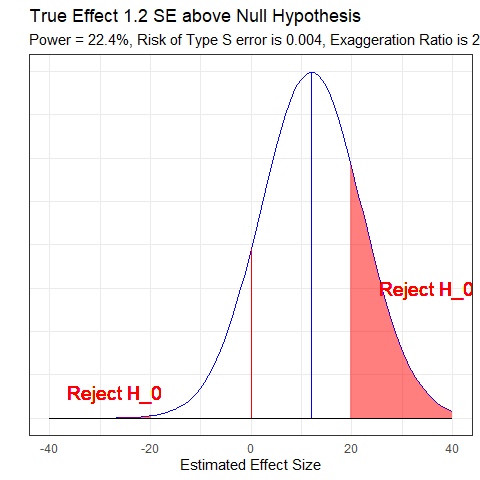
## What 22.4% power looks like…

ggplot(dat, aes(x, hx12)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 12, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 12, xend = 12, y = 0, yend = dnorm(12, mean = 12, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx12), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx12), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 35, y = 0.015, label = "Reject H\_0", col = "red", size = 5) +  
 geom\_text(x = -27, y = 0.003, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 1.2 SE above Null Hypothesis", subtitle = "Power = 22.4%, Risk of Type S error is 0.004, Exaggeration Ratio is 2.12")

Warning in geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 12, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = 12, xend = 12, y = 0, yend = dnorm(12, mean = 12, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)): All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.



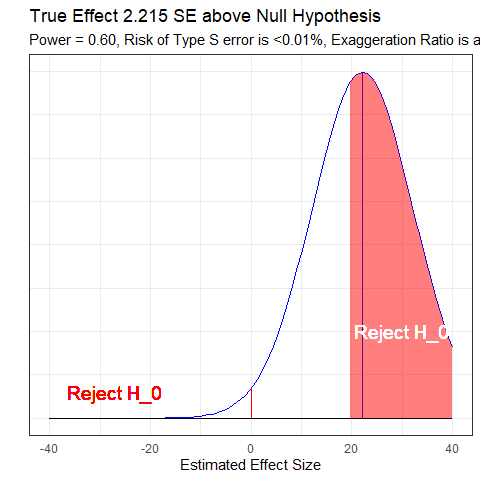
## What 60% Power Looks Like

ggplot(dat, aes(x, hx2215)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 22.15, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 22.15, xend = 22.15, y = 0, yend = dnorm(22.15, mean = 22.15, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx2215), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx2215), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.01, label = "Reject H\_0", col = "white", size = 5) +  
 geom\_text(x = -27, y = 0.003, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 2.215 SE above Null Hypothesis", subtitle = "Power = 0.60, Risk of Type S error is <0.01%, Exaggeration Ratio is about 1.3")

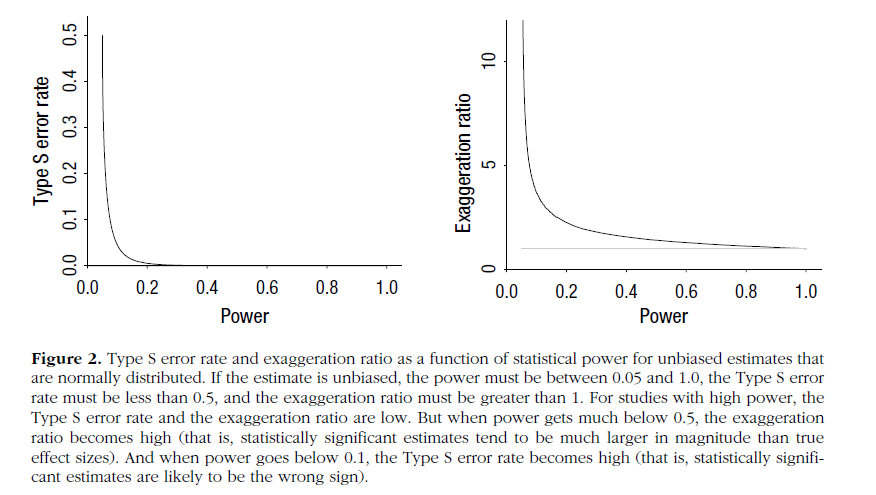
Warning in geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 22.15, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = 22.15, xend = 22.15, y = 0, yend = dnorm(22.15, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)): All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.



## Gelman & Carlin, Figure 2



## 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).

## More from Gelman et al.

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 = 0.05

set.seed(201803164)  
retrodesign(D = 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).

## 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. 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.

## 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. We thus have a power - 2.3% = 2.7% probability of showing statistical significance in the correct direction.

## 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. In expectation, a statistically significant result will be 77 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 | 77 |
| D = 0.3 | 0.05 | 0.39 | 25 |
| D = 1.0 | 0.06 | 0.19 | 7.8 |

## What if true D = 1.0 point?

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.

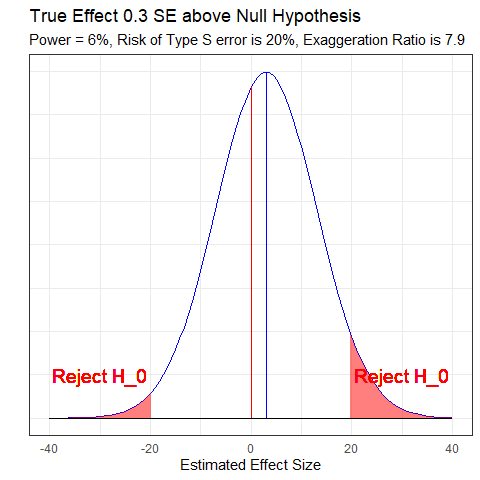
## What 6% power looks like…

ggplot(dat, aes(x, hx3)) +  
 geom\_line(col = "blue") +  
 geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 3, sd = 10)), col = "red") +  
 geom\_segment(aes(x = 3, xend = 3, y = 0, yend = dnorm(3, mean = 3, sd = 10)), col = "blue") +  
 geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)) +  
 geom\_ribbon(data = subset(dat, x > 19.6), aes(ymax = hx3), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_ribbon(data = subset(dat, x < -19.6), aes(ymax = hx3), ymin = 0, fill = "red", col = NA, alpha = 0.5) +  
 geom\_text(x = 30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 geom\_text(x = -30, y = 0.005, label = "Reject H\_0", col = "red", size = 5) +  
 theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) +  
 labs(x = "Estimated Effect Size", y = "", title = "True Effect 0.3 SE above Null Hypothesis", subtitle = "Power = 6%, Risk of Type S error is 20%, Exaggeration Ratio is 7.9")

Warning in geom\_segment(aes(x = 0, xend = 0, y = 0, yend = dnorm(0, mean = 3, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = 3, xend = 3, y = 0, yend = dnorm(3, mean = 3, : All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.

Warning in geom\_segment(aes(x = -40, xend = 40, y = 0, yend = 0)): All aesthetics have length 1, but the data has 100 rows.  
ℹ Please consider using `annotate()` or provide this layer with data containing  
 a single row.



## 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.

## But I do studies with 80% power?

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](http://andrewgelman.com/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/)