Rethinking Statistical Significance

github.com/THOMASELOVE/rethink

2018-03-16

Public Service Announcement from XKCD

Dates


Michael Donohoe
@donohoe
Follow
Comprehensive map of all countries in the world that use the MMDDYYYY format
5:29 PM - 11 May 2015

Where To Get More Information

- Today's slides are at https://github.com/THOMASELOVE/rethink.
- You'll find all of the references there, as well, and some other sources.
- We'll also post the slides at chrp.org, too.

Thomas E. Love, Ph.D.

- Professor of Medicine, Population & Quantitative Health Sciences, CWRU School of Medicine
- Director of Biostatistics and Evaluation, Center for Health Care Research and Policy
- Chief Data Scientist, Better Health Partnership
- Fellow, American Statistical Association

My email is `Thomas dot Love at case dot edu`.

Today's Agenda

- What I Taught for Many Years
- Notions of Statistical Significance
  - *p* values
  - How the Journals have reacted
  - p-hacking, Researcher DF, Garden of Forking Paths
  - How much help are confidence intervals?
  - A couple of the noisier recent Suggestions
- Evaluating Research More Effectively?
  - Retrospective Design: Type S and Type M errors
  - Evaluating Health News: A primer
- What I Think I Think Now

"Not Even Scientists Can Easily Explain p Values" Video

What I Taught for Many Years

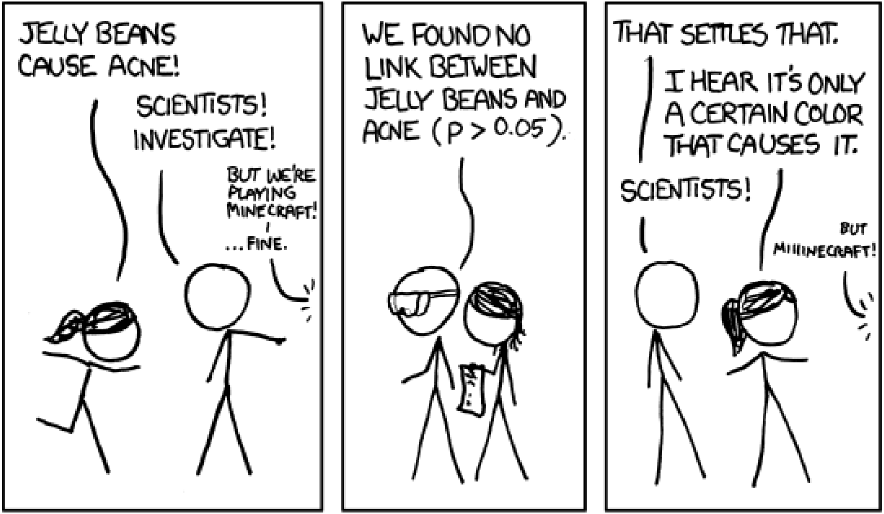
What I Taught for Many Years

- Null hypothesis significance testing is here to stay.
    - Learn how to present your p value so it looks like what everyone else does
    - Think about "statistically detectable" rather than "statistically significant"
    - Don't accept a null hypothesis, just retain it.
- Use point **and** interval estimates
    - Try to get your statements about confidence intervals right (right = just like I said it)
- Use Bayesian approaches/simulation/hierarchical models when they seem appropriate or for "non-standard" designs
    - But look elsewhere for people to teach/do that stuff
- Power is basically a hurdle to overcome in a grant application

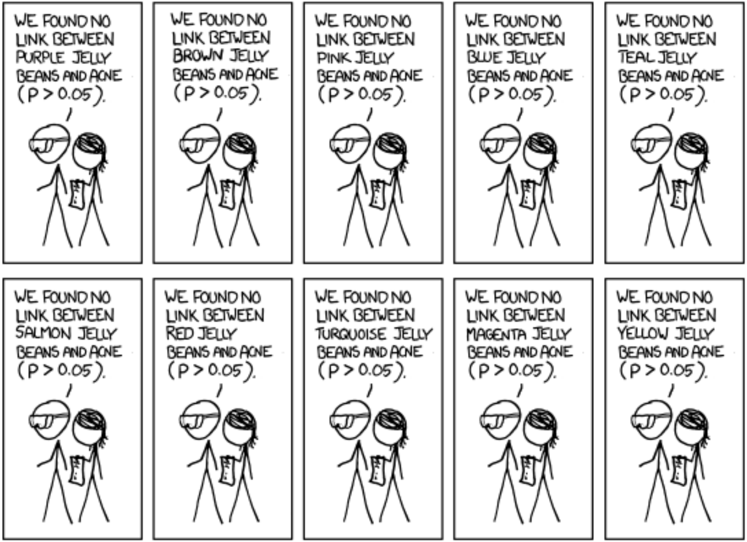
Conventions for Reporting *p* Values

1. Use an italicized, lower-case *p* to specify the *p* value. Don't use *p* for anything else.
2. For *p* values above 0.10, round to two decimal places, at most.
3. For *p* values near α, include only enough decimal places to clarify the reject/retain decision.
4. For very small *p* values, always report either $p < 0.0001$ or even just $p < 0.001$, rather than specifying the result in scientific notation, or, worse, as $p = 0$ which is glaringly inappropriate.
5. Report *p* values above 0.99 as $p > 0.99$, rather than $p = 1$.

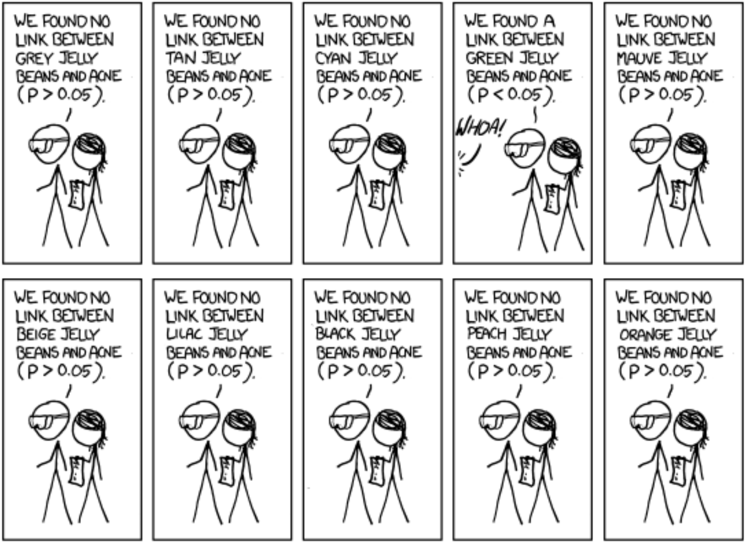
XKCD "Significance"



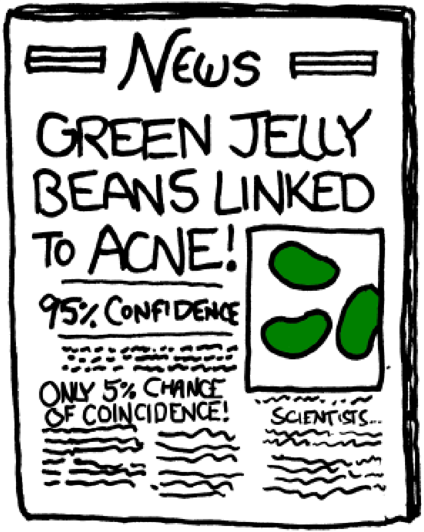
XKCD "Significance"



XKCD "Significance"



XKCD "Significance"



American Statistical Association to the rescue!?!

ASA Statement on *p* Values

ASA Statement: "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

fivethirtyeight.com "Not Even Scientists Can Easily Explain *p* Values"

... Try to distill the p-value down to an intuitive concept and it loses all its nuances and complexity, said science journalist Regina Nuzzo, a statistics professor at Gallaudet University. "Then people get it wrong, and this is why statisticians are upset and scientists are confused." **You can get it right, or you can make it intuitive, but it's all but impossible to do both.**

fivethirtyeight.com "Statisticians found one thing they can agree on"

A Few Comments on Significance

- **A significant effect is not necessarily the same thing as an interesting effect.** For example, results calculated from large samples are nearly always "significant" even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical interest. If not, why do the test?
- **A non-significant effect is not necessarily the same thing as no difference.** A large effect of real practical interest may still produce a non-significant result simply because the sample is too small.
- **There are assumptions behind all statistical inferences.** Checking assumptions is crucial to validating the inference made by any test or confidence interval.
- "**Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**"

ASA *statement* on p values

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$,

which morphed into a

- **rule** for journals: reject all articles that report p-values[1]

[1] http://www.nature.com/news/psychology-journal-bans-p-values-1.17001 describes the recent banning of null hypothesis significance testing by *Basic and Applied Psychology*.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$, which morphed into a
- **rule** for editors: reject the submitted article if $p > .05$, which morphed into a
- **rule** for journals: reject all articles that report p-values.

Bottom line: **Reject rules. Ideas matter.**

$p = 0.05$?

p = 0.05?

"For decades, the conventional p-value threshold has been 0.05," says Dr. Paul Wakim, chief of the biostatistics and clinical epidemiology service at the National Institutes of Health Clinical Center, "but it is extremely important to understand that this 0.05, there's nothing rigorous about it. It wasn't derived from statisticians who got together, calculated the best threshold, and then found that it is 0.05. No, it's Ronald Fisher, who basically said, 'Let's use 0.05,' and he admitted that it was arbitrary."

- NOVA "Rethinking Science's Magic Number" by Tiffany Dill 2018-02-28. See especially the video labeled "Science's most important (and controversial) number has its origins in a British experiment involving milk and tea."

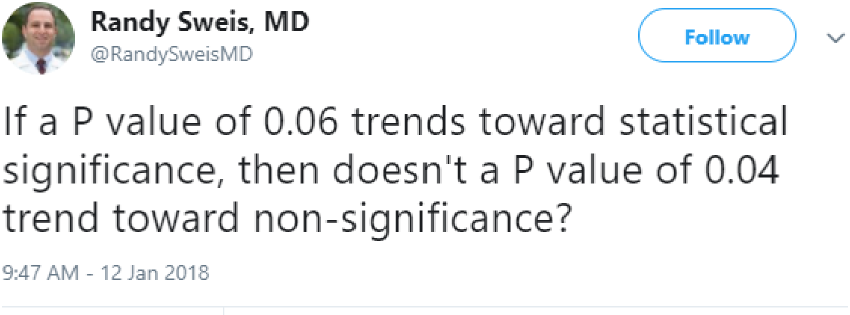
More from Dr. Wakim. . .

"People say, 'Ugh, it's above 0.05, I wasted my time.' No, you didn't waste your time." says Dr. Wakim. "If the research question is important, the result is important. Whatever it is."

- NOVA Season 45 Episode 6 Prediction by the Numbers 2018-02-28.

Now what?

p values don't trend...



George Cobb's Questions (with Answers)

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

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

A: Because that's **still** what the scientific community and journal editors use.

Q: Why do so many people still use $p = 0.05$?

A: Because that's what they were taught in college or grad school.

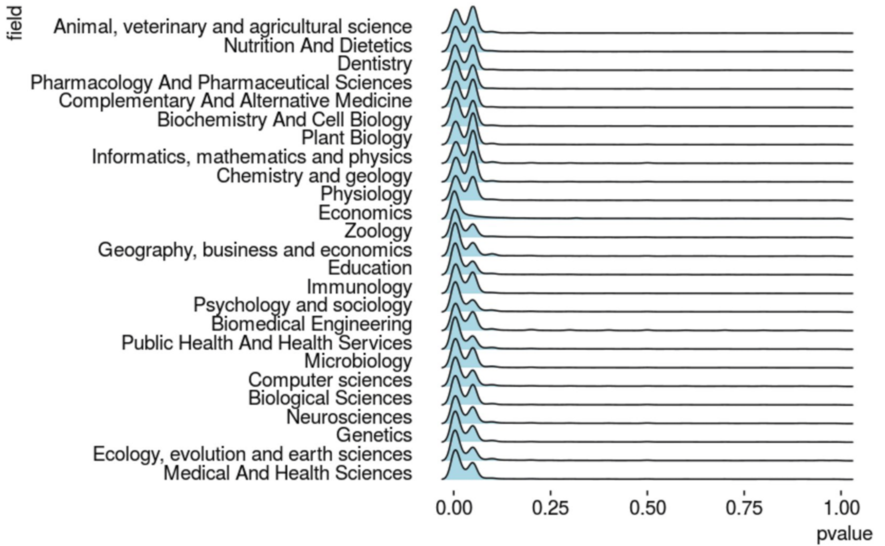
All the p values

The p-value is the most widely-known statistic. P-values are reported in a large majority of scientific publications that measure and report data. R.A. Fisher is widely credited with inventing the p-value. If he was cited every time a p-value was reported his paper would have, at the very least, 3 million citations - making it the most highly cited paper of all time.

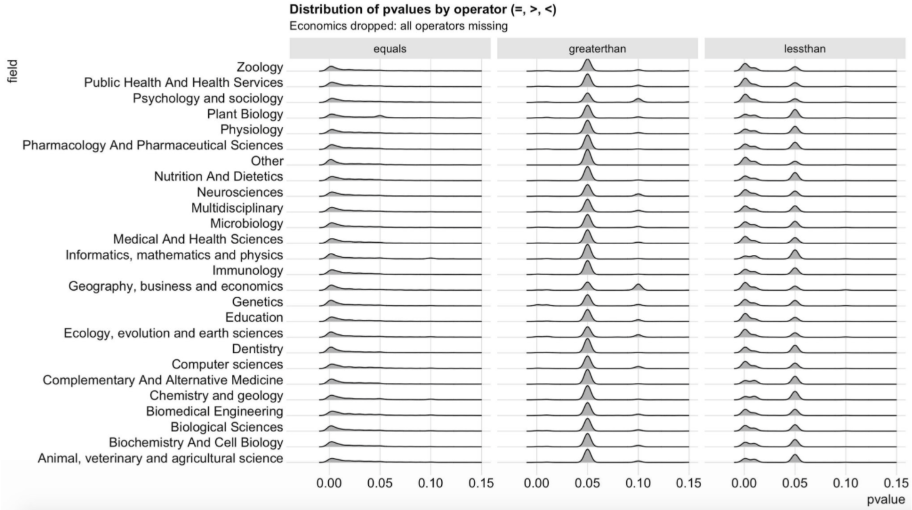
- Visit Jeff Leek's Github for tidypvals package
  - 2.5 million *p* values in 25 scientific fields

What do you suppose the distribution of those p values is going to look like?

2.5 million p values in 25 scientific fields: Jeff Leek



from Michael Lopez



The Journals React

JAMA Otolaryngology - Head and Neck Surgery

Mustafa Ascha Nov 22 (6 days ago)

to me

Hi Dr Love,

I wanted to share: JAMA-Otolaryngology just asked for a revision of a manuscript I'm writing. They specifically asked that I remove all p-values from the paper, and attached a statement from the journal editor that essentially repeated the ASA statement on p-values.

I'm pleasantly surprised by their request.

http://jamanetwork.com/journals/jamaotolaryngology/fullarticle/2546529

JAMA Otolaryngology - Head and Neck Surgery

Editorial

FREE

October 2016

Improving the Quality of the Reporting of Research Results

Jay F. Piccirillo, MD[1]

» Author Affiliations | Article Information

JAMA Otolaryngol Head Neck Surg. 2016;142(10):937-939. doi:10.1001/jamaoto.2016.2670

Problems with P Values

1. P values are inherently unstable
2. The p value, or statistical significance, does not measure the size of an effect or the importance of a result
3. Scientific conclusions should not be based only on whether a p value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. By itself, a p value does not provide a good measure of evidence regarding a model or hypothesis

http://jamanetwork.com/journals/jamaotolaryngology/fullarticle/2546529

Solutions to the *P* Value Problems

1. Estimation of the Size of the Effect
2. Precision of the Estimate (Confidence Intervals)
3. Inference About the Target Population
4. Determination of Whether the Results Are Compatible With a Clinically Meaningful Effect
5. Replication and Steady Accumulation of Knowledge

http://jamanetwork.com/journals/jamaotolaryngology/fullarticle/2546529

Importance of Meta-Analytic Thinking

In JAMA Otolaryngology: Head & Neck Surgery, we look to publish original investigations where the investigators planned the study with sufficient sample size to have adequate power to detect a clinically meaningful effect and report the results with effect sizes and CIs. Authors should interpret the effect sizes in relation to previous research and use CIs to help determine whether the results are compatible with clinically meaningful effects. And finally, we acknowledge that no single study can define truth and that the advancement of medical knowledge and patient care depends on the steady accumulation of reliable clinical information.

http://jamanetwork.com/journals/jamaotolaryngology/fullarticle/2546529

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.

Link

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

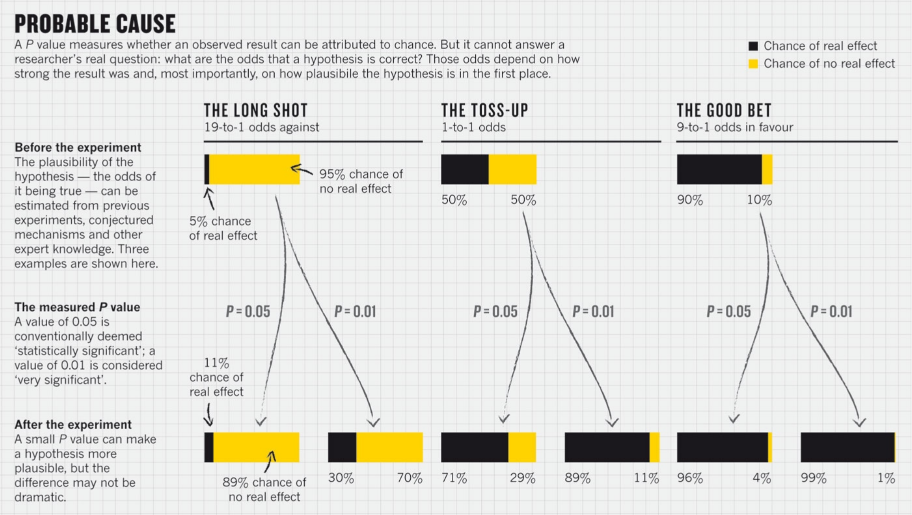
Abstract

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

Dividing Data Comparisons into Categories based on p values

Regina Nuzzo in Nature on Statistical Errors



Gelman on p values, 1

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.

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. . .

Gelman 2016-10-15

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 a z score of 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.

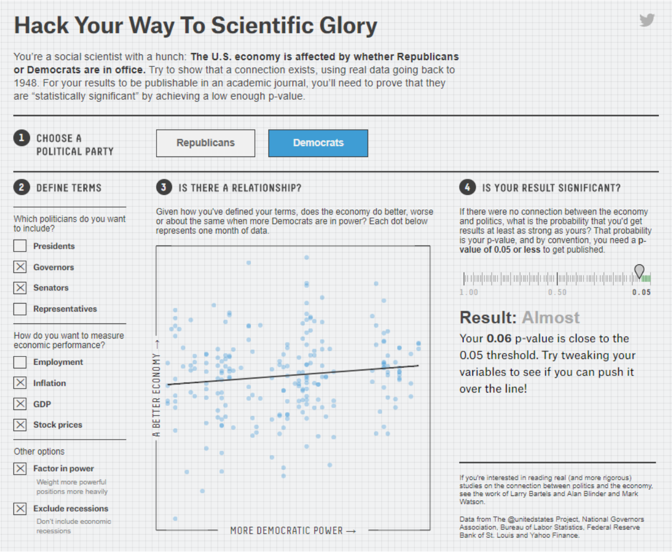
Indeed, once you see comparisons between large, marginal, and small effects, the null hypothesis is irrelevant, as you want to be comparing effect sizes.

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.

p-Hacking

Hack Your Way To Scientific Glory (fivethirtyeight)



"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

Do Confidence Intervals get us out of this mess?

Confidence Intervals - do they solve our problem?

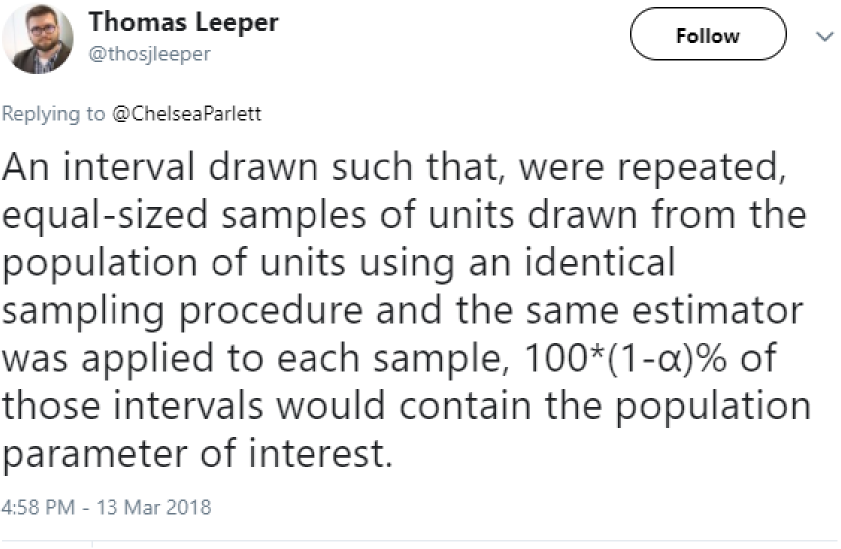
Chelsea Parlett Pelleriti
@ChelseaParlett

Follow

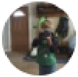
Hey Stats folk, what's your 280 character definition of a confidence interval? 🤔

4:30 PM - 13 Mar 2018

Confidence Intervals - do they solve our problem?



Confidence Intervals - do they solve our problem?



Joran Elias
@joranelias

Follow

A confidence interval is a meaure of uncertainty such that all definitions of it elicit corrections from Bayesians.

(Didn't need all 280.)

Confidence Intervals - do they solve our problem?

Jenny Bryan
@JennyBryan

Following

Pedantry about the definition of a confidence interval ... why is this the hill statisticians choose to die on? Every time you feel the urge, go convert a table to a figure. It is likely to do more good.

Confidence Intervals - do they solve our problem?

Frank Harrell @f2harrell · 28 Dec 2017

Tables and figures are important but so is this. We need to get this right. Too many faulty conclusions being drawn with frequentist statistical analysis. If one is going to be a frequentist one should make exactly correct interpretations.

2 10

Jenny Bryan
@JennyBryan

Following

Replying to @f2harrell

I just feel like the people we're often trying to reach aren't making informed comparisons of frequentist vs Bayesian methods, they're still struggling with decision making under uncertainty

1:02 PM - 28 Dec 2017

Using Bayesian Ideas: Confidence Intervals

> *My current favorite (hypothetical) example is an epidemiology study of some small effect where the point estimate of the odds ratio is 3.0 with a 95% conf interval of [1.1, 8.2].*
>
> *As a 95% conf interval, this is fine (assuming the underlying assumptions regarding sampling, causal identification, etc. are valid).*

(but on some level you need to deal with the fact that. . .)

> *. . . real-world odds ratios are much more likely to be near 1.1 than to be near 8.2.*

See Gelman 2014-12-11.

Uncertainty intervals?

I've (Gelman) become increasingly uncomfortable with the term "confidence interval" for several reasons:

- The well-known difficulties in interpretation (officially the confidence statement can be interpreted only on average, but people typically implicitly give the Bayesian interpretation to each case.)
- The ambiguity between confidence intervals and predictive intervals.
- The awkwardness of explaining that confidence intervals are big in noisy situations where you have less confidence, and confidence intervals are small when you have more confidence.

> *So here's my proposal. Let's use the term "uncertainty interval" instead. The uncertainty interval tells you how much uncertainty you have.*

See Gelman 2010-12-21.

Some Noisy Recent Suggestions

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.

Motivations:

- links to Bayes Factor intepretation
- 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*.

Abandon Statistical Significance

Gelman blog 2017 – 09 – 26 on "Abandon Statistical Significance"

"Measurement error and variation are concerns even if your estimate is more than 2 standard errors from zero. Indeed, if variation or measurement error are high, then you learn almost nothing from an estimate even if it happens to be 'statistically significant.' "

Read the whole paper *here*

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.

Some Ways of 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)

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.

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?

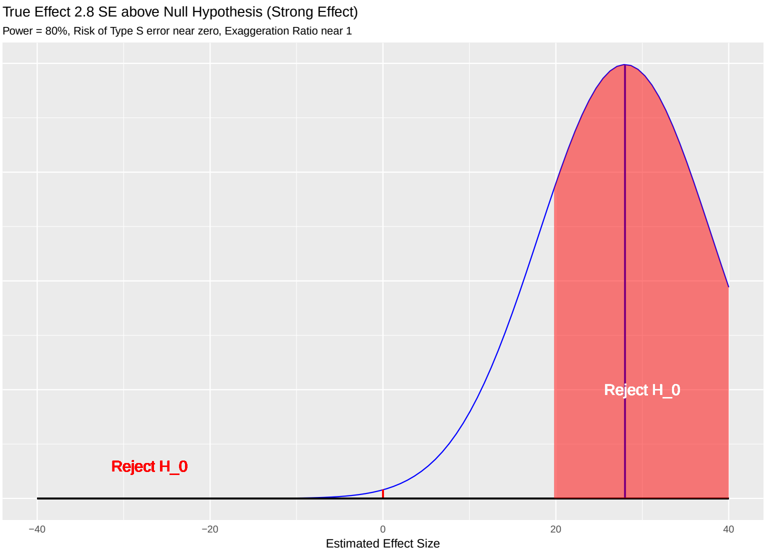
Suppose the true effect that is 2.8 standard errors away from zero, in a study built to have 80% power for that effect with 95% confidence.

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

```
$power
[1] 0.7995569

$typeS
[1] 1.210843e-06

$exaggeration
[1] 1.12875
```

- 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 1.13.

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



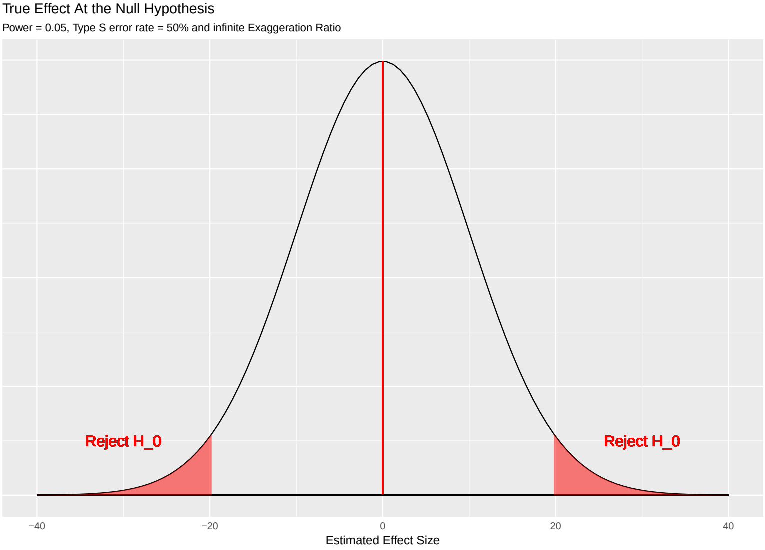
`retrodesign` for Zero Effect

```
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

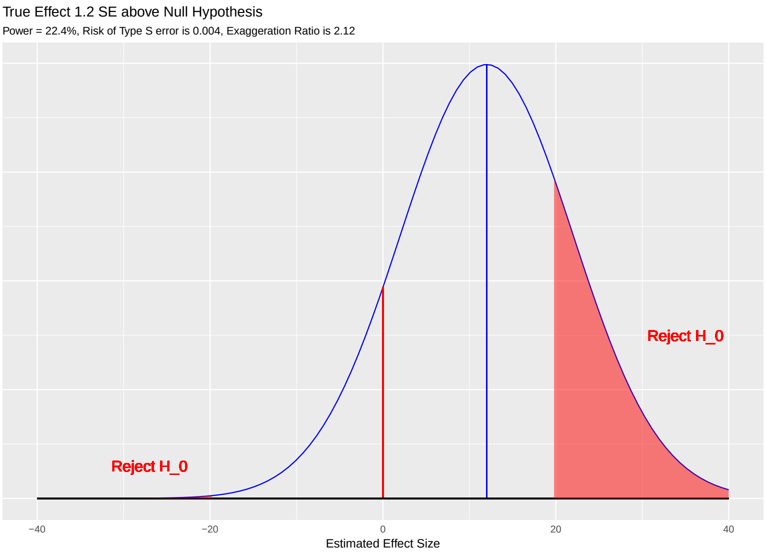


Retrodesign for a true effect 1.2 SE above H_0

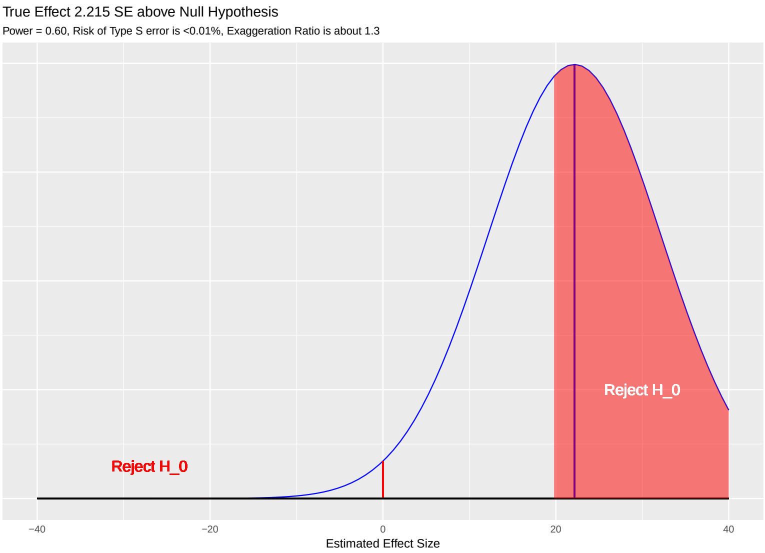
```r
set.seed(201803163)
retrodesign(A = 12, s = 10)
```

```
$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

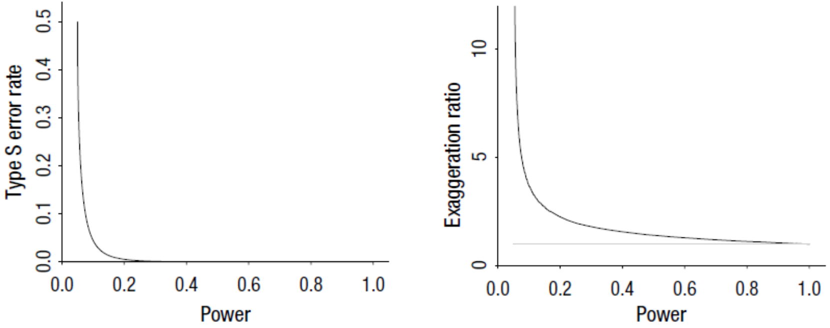


Figure 2. Type S error rate and exaggeration ratio as a function of statistical power for unbiased estimates that are normally distributed. If the estimate is unbiased, the power must be between 0.05 and 1.0, the Type S error rate must be less than 0.5, and the exaggeration ratio must be greater than 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)` $= 2.432$
  - $Z = \text{estimate}/\text{SE}$, and if estimate $= 8$ and $Z = 2.432$, then $\text{SE} = 8/2.432 = 3.29$

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 $\alpha = 0.05$

```
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:

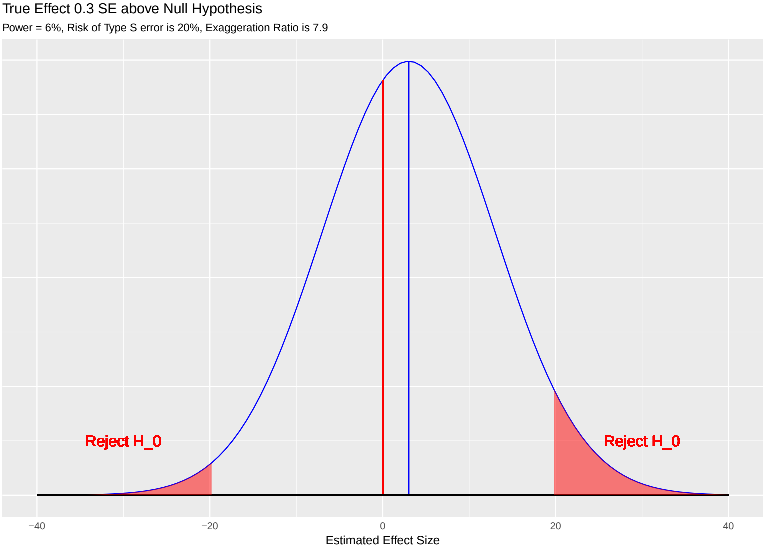
1. There is a 46% chance it will have the wrong sign (from the Type S error rate).
2. The power is 5% and the Type S error rate of 46%. Multiplying those gives a 2.3% probability that we will find a statistically significant result in the wrong direction.
3. We thus have a power - 2.3% = 2.7% probability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

Evaluating Research More Effectively

A Formula for Decoding Health News

Health Headlines are Advertising

Think about this headline: "Hospital checklist cut infections, saved lives."

- Suppose you are a little surprised that a checklist could really save lives. If you think say the odds of this being true are 1 in 4, you would set your initial gut feeling to 1/4. Because this number is less than one, it means initially you're less likely to believe the study.

Bayes' Rule

Final opinion = (initial gut feeling) * (study support for headline)
Jeff Leek, *fivethirtyeight.com*

Assessing Study Support for a Headline

1. Was the study a clinical study in humans?
2. Was the outcome of the study something directly related to human health like longer life or less disease? Was the outcome something you care about, such as living longer or feeling better?
3. Was the study a randomized, controlled trial (RCT)?
4. Was it a large study - at least hundreds of patients?
5. Did the treatment have a major impact on the outcome?
6. Did predictions hold up in at least two separate groups of people?

Assessing Study Support

Support for Headline: Multiply by 2 for every yes, and 1/2 for every no.

Evaluating A Research Article

Intensive care units (ICUs) at Michigan hospitals implemented a new strategy for reducing infections through training, a daily goals sheet and a program to improve the culture of safety in the ICUs. The doctors measured the rate of infection before and after implementing this safety program.

1. Was the study a clinical study in humans?
   - The study was done in humans in ICUs (+)
2. Was the outcome of the study something directly related to human health like longer life or less disease? Was the outcome something you care about, such as living longer or feeling better?
   - The outcome was the rate of infections after surgery. According to the article, these infections cost U.S. hospitals up to $2.3 billion annually. (+)

Evaluating a Research Article

3. Was the study a randomized, controlled trial (RCT)?
   - The study compared the same hospitals before and after a change in ICU policy. This is an example of a crossover study, which is not as strong as a randomized trial but does account for some of the differences among hospitals because the same ICUs were measured before and after using the checklist. (-)
4. Was it a large study - at least hundreds of patients?
   - The study looked at more than 100 ICUs over 1,981 months. In total, it followed patients for 375,757 catheter-days. (A catheter-day means watching one patient for one day while she is on a catheter.) This is a huge number of days to monitor patients for potential infections. (+)
5. Did the treatment have a major impact on the outcome?
   - The study showed a sustained drop of up to 66 percent in infections. (+)
6. Did predictions hold up in at least two separate groups of people?
   - The study looked at 103 hospitals in Michigan. (+)

So we have 5 + and 1 - in our evaluation of this study.

Final Opinion?

- So, a large study showed a major drop in infections that is directly medically important. For the sake of the exercise, let's multiply by two every time we see a *yes* answer and by 1/2 every time we see a *no* answer. I would say this study's result is about 16 times more likely (five out of six *yes* answers so 2 x 2 x 2 x 2 x 2 x (1/2) = 16) if checklists really do reduce infections than if they don't. I set study support for headline = 16.
- Multiply to get final opinion on headline = 1/4*16 = 4, also expressed as 4/1. I would say that my updated odds are 4 to 1 that the headline is true. The strength of the study won over my initially skeptical gut feeling.

Bayes' Rule

Final opinion = (initial gut feeling) * (study support for headline)
Source: Jeff Leek, *fivethirtyeight.com*

Evaluating Health News: For Consumers

1. Watch out for single source stories. They're usually based on a press release, which will have a hidden agenda.
2. Beware of stories that don't mention cost. It's crucial information. (If the cost of the great, new treatment is out of reach – it's not that great, is it?)
3. Headline percentages are misleading. If something "reduces your risk of X by 50%" chances are that number doesn't mean what you think it means.
4. If it sounds too good to be true, it probably is. If a report presents only or primarily the benefits of a new treatment, it's a bad report. ALL healthcare interventions have trade-offs.
5. Patient anecdotes are not data. Beware of stories that rely on them. Anecdotes are used to compensate for data that are unavailable or flawed.

Source: *NPR*

Evaluating Health News: For Consumers

6. A "simple screening test" is never simple. The decision to take one is one of the most complex and difficult decisions a health consumer can make.
7. Watch out for hyperbolic language. "Breakthrough", "first-of-its-kind", and "game-changer" are red flags. When you read "it may become…" substitute "it may not become…"
8. Newer isn't always better. Often the latest test, treatment or procedure is no better than what already exists, just pricier.
9. Beware of disease-mongering. Risk factors, symptoms for diseases, or data can be exaggerated in a way that causes needless worry, and expense.
10. The latest treatment may not exist yet, or ever. "Awaiting FDA approval" or "in pre-clinical trial phase" means it's still a pipe dream.
11. There is a leap from mice to men. Getting from rodent trials to human use is a very, very long road, that may in fact lead nowhere.

Source: *NPR*

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

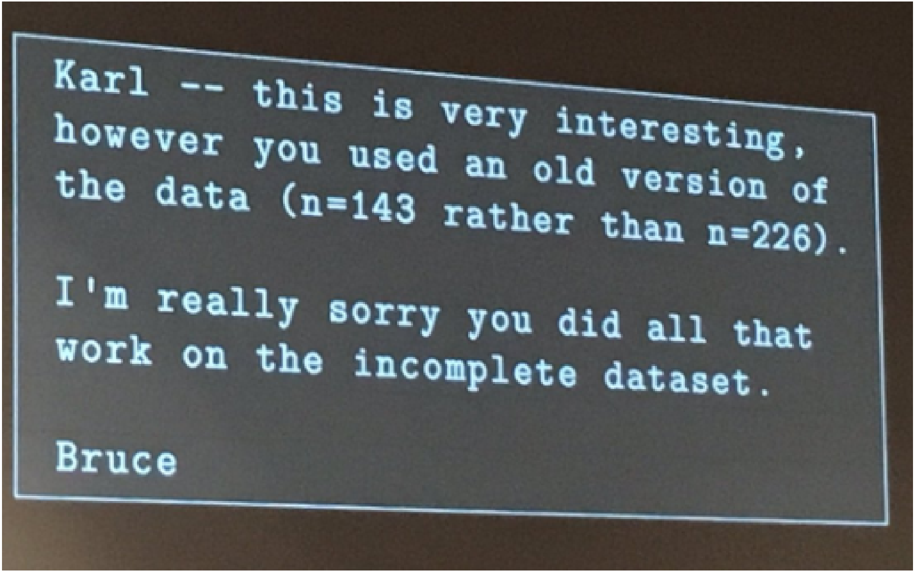
Goals of Reproducible Research

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

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?

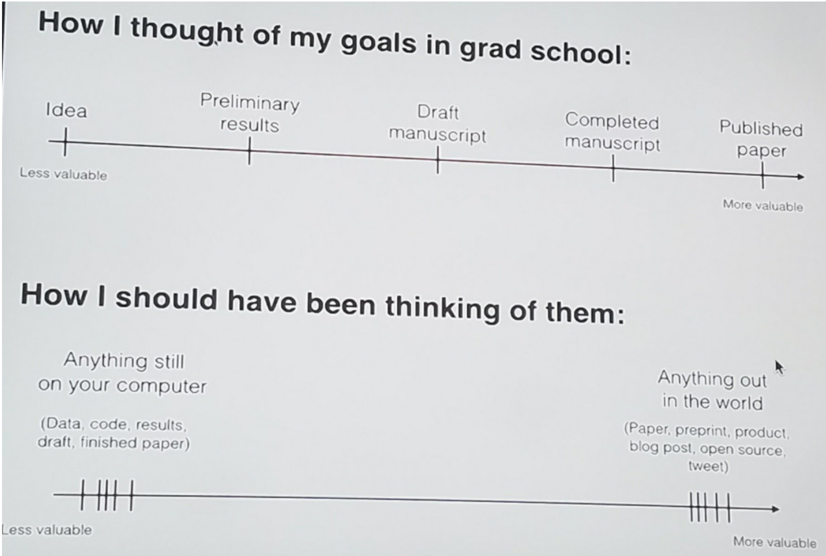
- True.
- But competition means that strangers will read your work, try to learn from you, cite you, and try to do things even better.

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

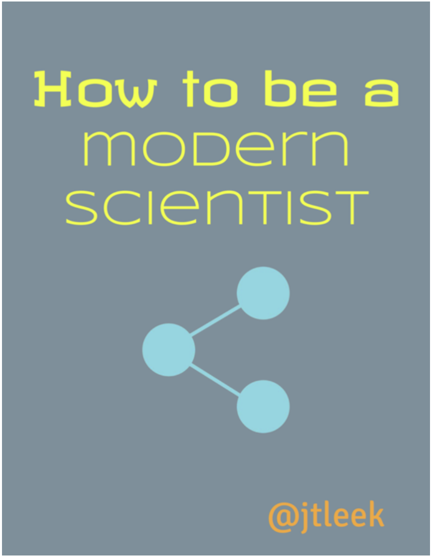
But other people will use my data and code to compete with me?

- 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 (from David Robinson @drob)



Read This Book (Jeff Leek)



https://leanpub.com/modernscientist

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.

Where To Get More Information

- Today's slides are at https://github.com/THOMASELOVE/rethink.
- You'll find all of the references there, as well.
- We'll also post the slides on the CHRP.org site.

Thomas E. Love, Ph.D.

- Professor of Medicine, Population & Quantitative Health Sciences, CWRU School of Medicine
- Director of Biostatistics and Evaluation, Center for Health Care Research and Policy
- Chief Data Scientist, Better Health Partnership
- Fellow, American Statistical Association

My email is `Thomas dot Love at case dot edu`.