p values, Reproducibility & Modern Science

github.com/THOMASELOVE/RCR2017

2017-11-08

Sorry about the title

I won't be doing any rigorous statistical analysis. Reproducibility is the actual theme.

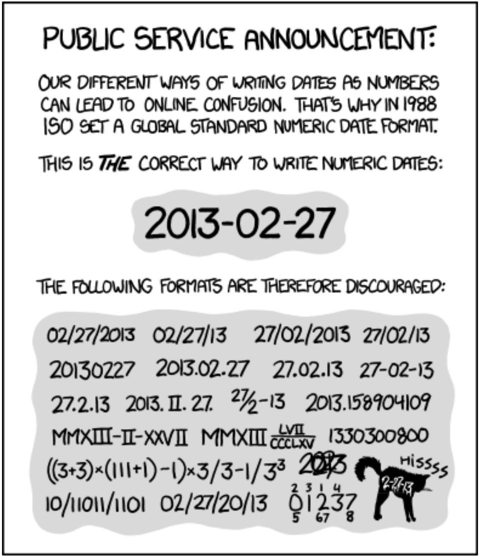
Today's slides are at https://github.com/THOMASELOVE/RCR2017.

Thomas E. Love, Ph.D.

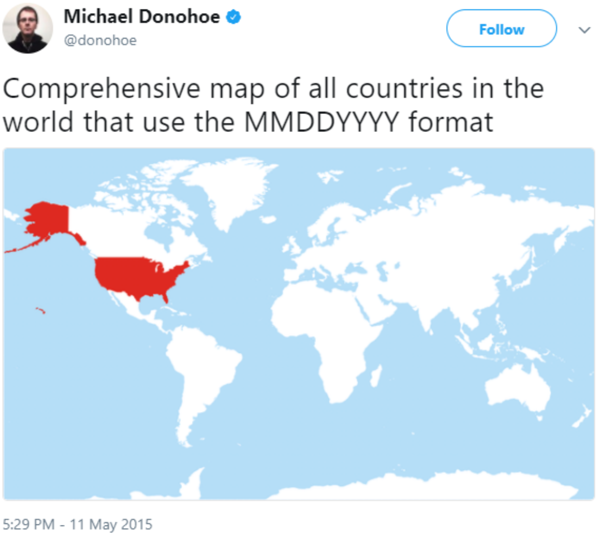
- Professor of Medicine, Population & Quantitative Health Sciences, CWRU School of Medicine
- Director of Biostatistics and Evaluation, Center for Health Care Research and Policy
- Chief Data Scientist and Past Data Director, Better Health Partnership
- Fellow, American Statistical Association

My email is Thomas dot Love at case dot edu.

Public Service Announcement from XKCD


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
HISSsss
2-27-13
10/11011/1101 02/27/20/13 2 3 1 4 01237 5 67 8

Dates



Michael Donohoe
@donohoe

Comprehensive map of all countries in the world that use the MMDDYYYY format

5:29 PM - 11 May 2015

https://twitter.com/donohoe/status/597876118688026624

What We'll Discuss Today

1. *p* values, p-hacking and the Reproducibility “Crisis” in Science
2. Doing Research More Effectively: The Modern Scientist's Toolbox
3. (maybe) Evaluating Research: A Formula for Decoding Health News

Let's talk about *p* values

fivethirtyeight.com video.

What I Taught for Many Years

On Reporting *p* Values

When reporting a *p* value and no rounding rules are in place from the lead author/journal/source for publication, follow these conventions. . .

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 $p = 1$.

Now what?

So sad. . .

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.**

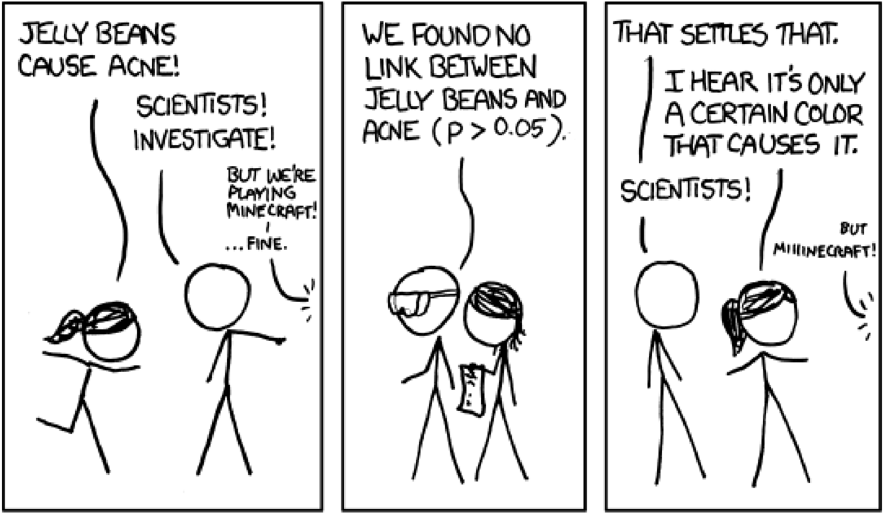
fivethirtyeight.com “Statisticians found one thing they can agree on”

A Few Comments on Significance

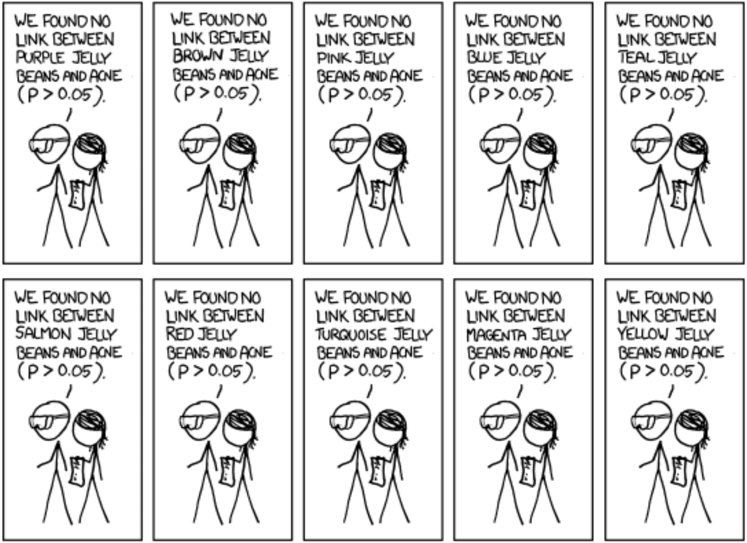
- **A significant effect is not necessarily the same thing as an interesting effect.** For example, results calculated from large samples are nearly always "significant" even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical 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

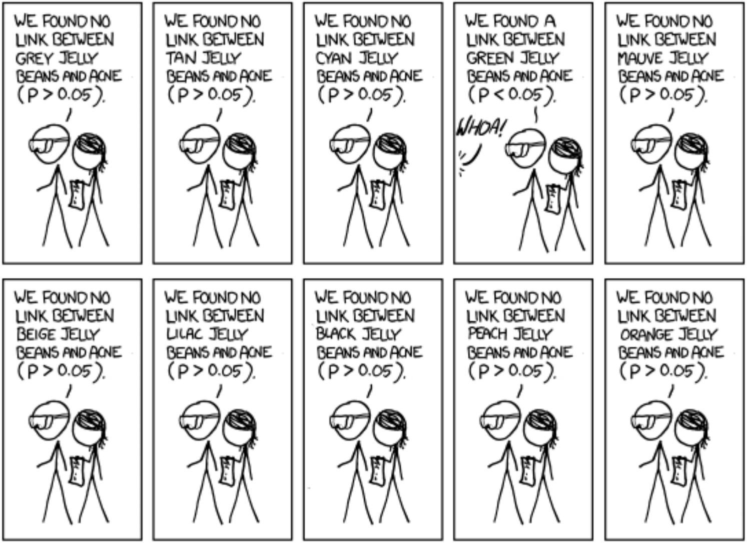
From XKCD (https://xkcd.com/882/)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!

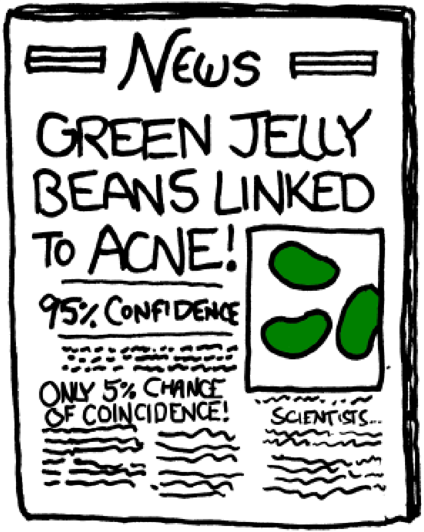
From XKCD (https://xkcd.com/882/)


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

From XKCD (https://xkcd.com/882/)


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE ($P < 0.05$).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE ($P > 0.05$).

From XKCD (https://xkcd.com/882/)


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

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.**

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.

Dividing Data Comparisons into Categories based on p values

Why Dividing Data Comparisons into Categories based on Significance Levels is Terrible.

> *The common practice of dividing data comparisons into categories based on significance levels is terrible, but it happens all the time.... so it's worth examining the prevalence of this error.*

Andrew Gelman's blog $2016-10-15$

Gelman on *p* values, 1

Let me first briefly explain why categorizing based on p-values is such a bad idea. Consider, for example, this division:

- "really significant" for $p < .01$,
- "significant" for $p < .05$,
- "marginally significant" for $p < .1$, and
- "not at all significant" otherwise.

Now consider some typical *p*-values in these ranges: say, $p = .005$, $p = .03$, $p = .08$, and $p = .2$.

Translate these two-sided *p*-values back into z-scores, which we can do in R via `qnorm(c(.005, .03, .08, .2)/2, lower.tail = FALSE)`

Gelman on *p* values, 2

Description	really sig.	sig.	marginally sig.	not at all sig.
p value	0.005	0.03	0.08	0.20
Z score	2.8	2.2	1.8	1.3

The seemingly yawning gap in p-values comparing the not at all significant *p*-value of .2 to the really significant *p*-value of .005, is only 1.5.

If you had two independent experiments with z-scores of 2.8 and 1.3 and with equal standard errors and you wanted to compare them, you'd get a difference of 1.5 with a standard error of 1.4, which is completely consistent with noise.

Gelman on *p* values, 3

From a **statistical** point of view, the trouble with using the p-value as a data summary is that the p-value is only interpretable in the context of the null hypothesis of zero effect, and (much of the time), nobody's interested in the null hypothesis.

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

From a **psychological** point of view, the trouble with using the p-value as a data summary is that this is a kind of deterministic thinking, an attempt to convert real uncertainty into firm statements that are just not possible (or, as we would say now, just not replicable).

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

Gelman on Statistical Significance

“. . . we use the term statistically significant in the conventional way, to mean that an estimate is **at least two standard errors away** from some”null hypothesis" or prespecified value that would indicate no effect present. An estimate is statistically insignificant if the observed value could reasonably be explained by simple chance variation, much in the way that a sequence of 20 coin tosses might happen to come up 8 heads and 12 tails; we would say that this result is not statistically significantly different from chance. More precisely, the observed proportion of heads is 40 percent but with a standard error of 11 percent - thus, the data are less than two standard errors away from the null hypothesis of 50 percent, and the outcome could clearly have occurred by chance. Standard error is a measure of the variation in an estimate and gets smaller as a sample size gets larger, converging on zero as the sample increases in size."

Gelman $2017 - 10 - 28$

New (2014-2017) Proposals

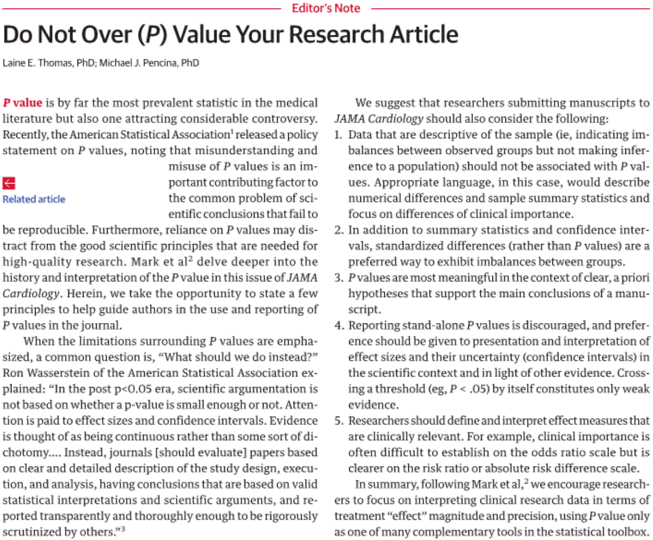
The Value of a *p*-Valueless Paper

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

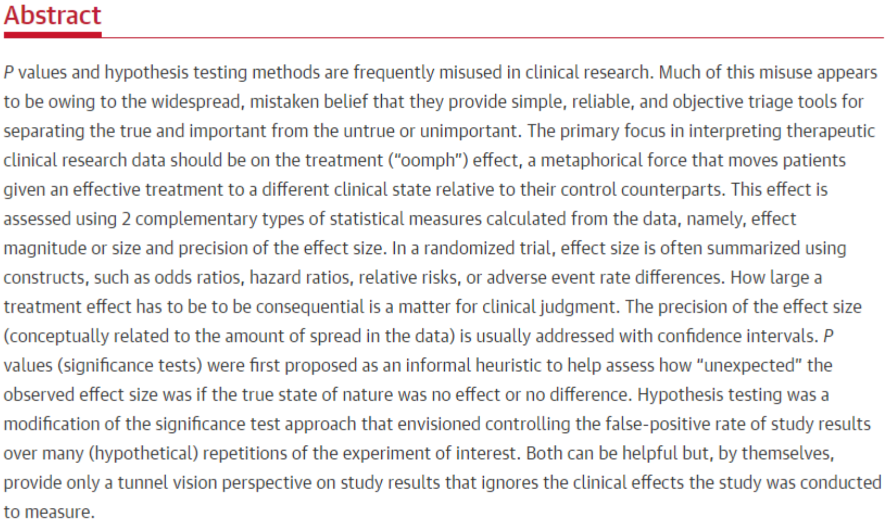
When the limitations surrounding *P* values are emphasized, a common question is, "What should we do instead?" Ron Wasserstein of the American Statistical Association explained: "In the post p<0.05 era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid 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

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"

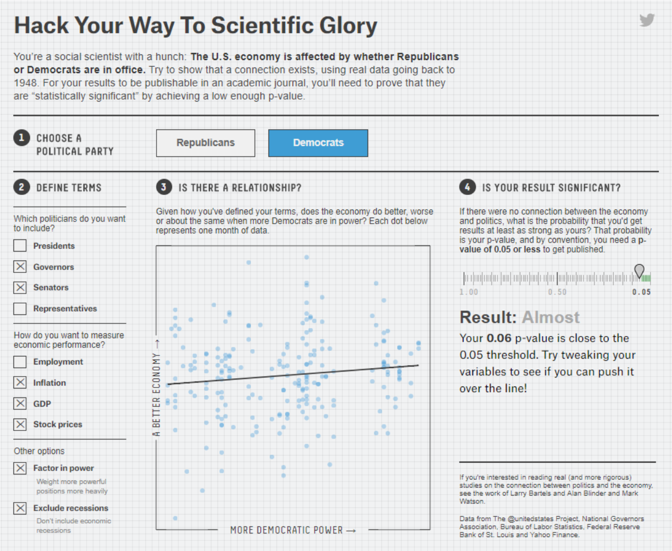
"Measurement error and variation are concerns even if your estimate is more than 2 standard errors from zero. Indeed, if variation or measurement error are high, then you learn almost nothing from an estimate even if it happens to be 'statistically significant.' "

Read the whole paper *here*

p-Hacking

Hack Your Way To Scientific Glory

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



Hack Your Way To Scientific Glory

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

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- □ Presidents
- ☒ Governors
- ☒ Senators
- □ Representatives

How do you want to measure economic performance?

- □ Employment
- ☒ Inflation
- ☒ GDP
- ☒ Stock prices

Other options

- ☒ Factor in power
  Weight more powerful positions more heavily
- ☒ Exclude recessions
  Don't include economic recessions

3 IS THERE A RELATIONSHIP?

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

4 IS YOUR RESULT SIGNIFICANT?

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

Result: Almost

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

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

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

“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

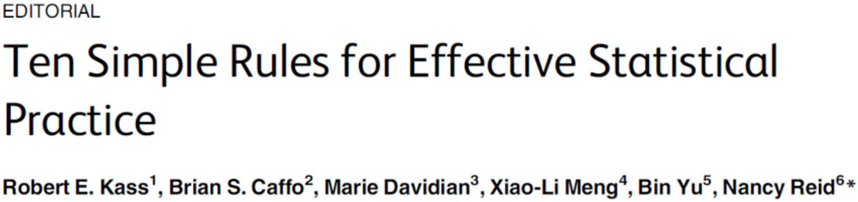
> *Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the 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

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

Ten Simple Rules for Effective Statistical Practice

From *PLoS Comput Biol* link



EDITORIAL

Ten Simple Rules for Effective Statistical Practice

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

Rule 10: Make Your Analysis Reproducible

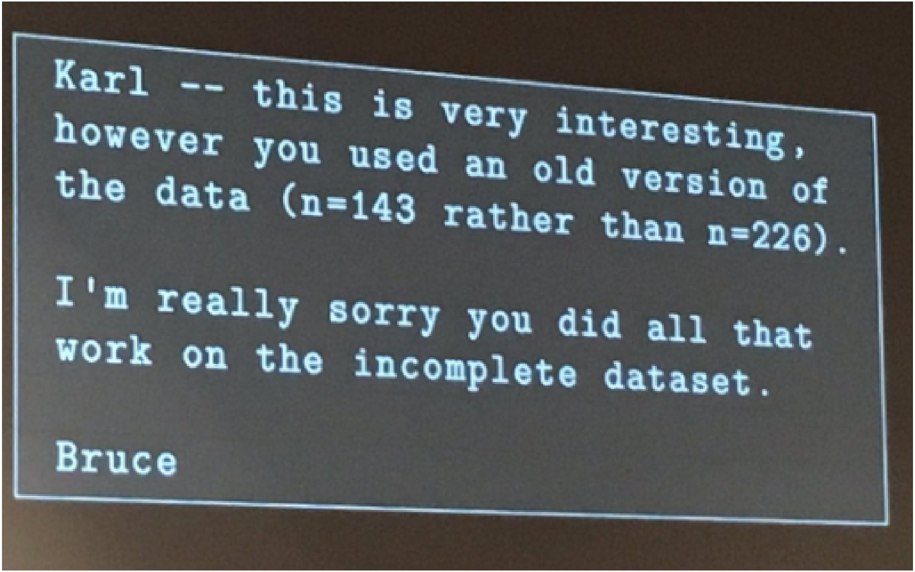
Goals of Reproducible Research

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

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

Vanderbilt *Tutorial*

Why I Do This...



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

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

Bruce

Five Practical Tips

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

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

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

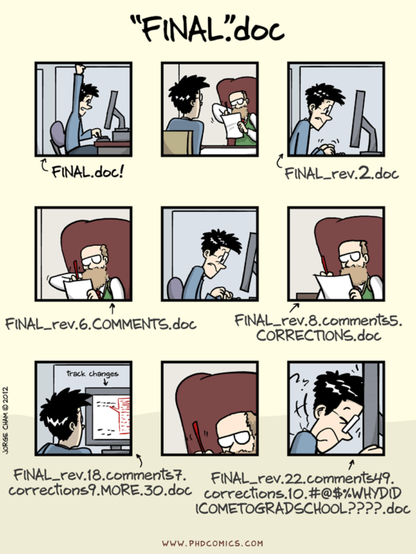
Karl Broman, Steps Towards Reproducible Research *link*

Build Tidy Data Sets

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

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

Choose good names for things.



Source

Great advice on choosing good filenames

https://speakerdeck.com/jennybc/how-to-name-files from **Jenny Bryan**

Good names are:

- machine-readable (easy to search, easy to extract info),
- human readable (name contains info on content, delimited so eyes don't bleed)
- and play well with default ordering (something numeric first, left pad other numbers with zeros).

```
01-marshal_data.md
02-pre-dea-filtering.md
2013-04-26_BRAFASSAY_Plasmid-Cell-100-1MutantFraction_H01.csv
2013-04-26_BRAFASSAY_Plasmid-Cell-100-1MutantFraction_H02.csv
```

File organization and naming are powerful weapons against chaos. Go forward and use awesome filenames!

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

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

- True.

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

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

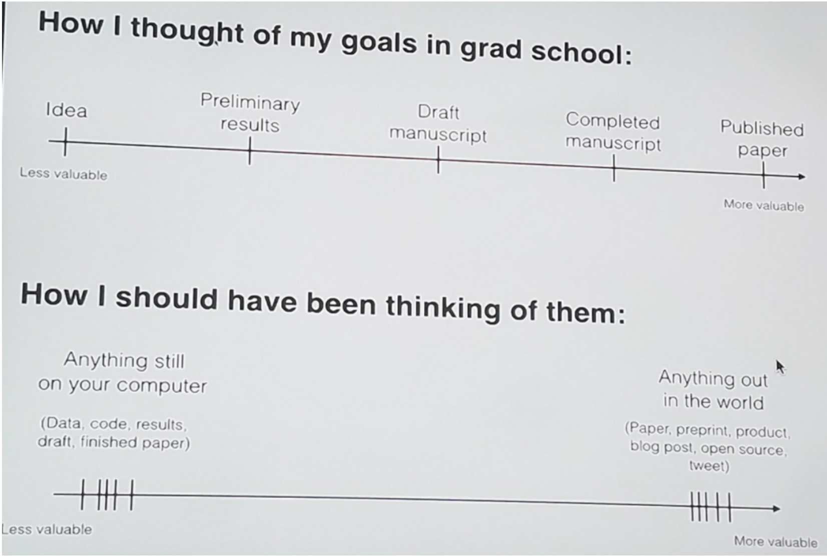
- True.
- But competition means that strangers will read your work, try to learn from you, cite you, and try to do things even better.

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

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

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

A Most Important Tip (@drob)



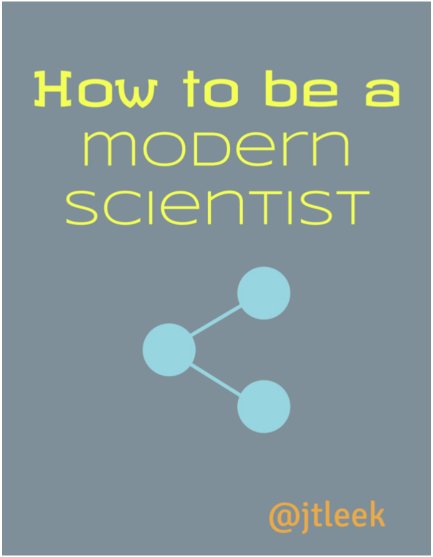
How I thought of my goals in grad school:

Idea
Preliminary results
Draft manuscript
Completed manuscript
Published paper
Less valuable
More valuable

How I should have been thinking of them:

Anything still on your computer
(Data, code, results, draft, finished paper)
Anything out in the world
(Paper, preprint, product, blog post, open source, tweet)
Less valuable
More valuable

Read This Book



https://leanpub.com/modernscientist

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)
Source: 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 — 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*