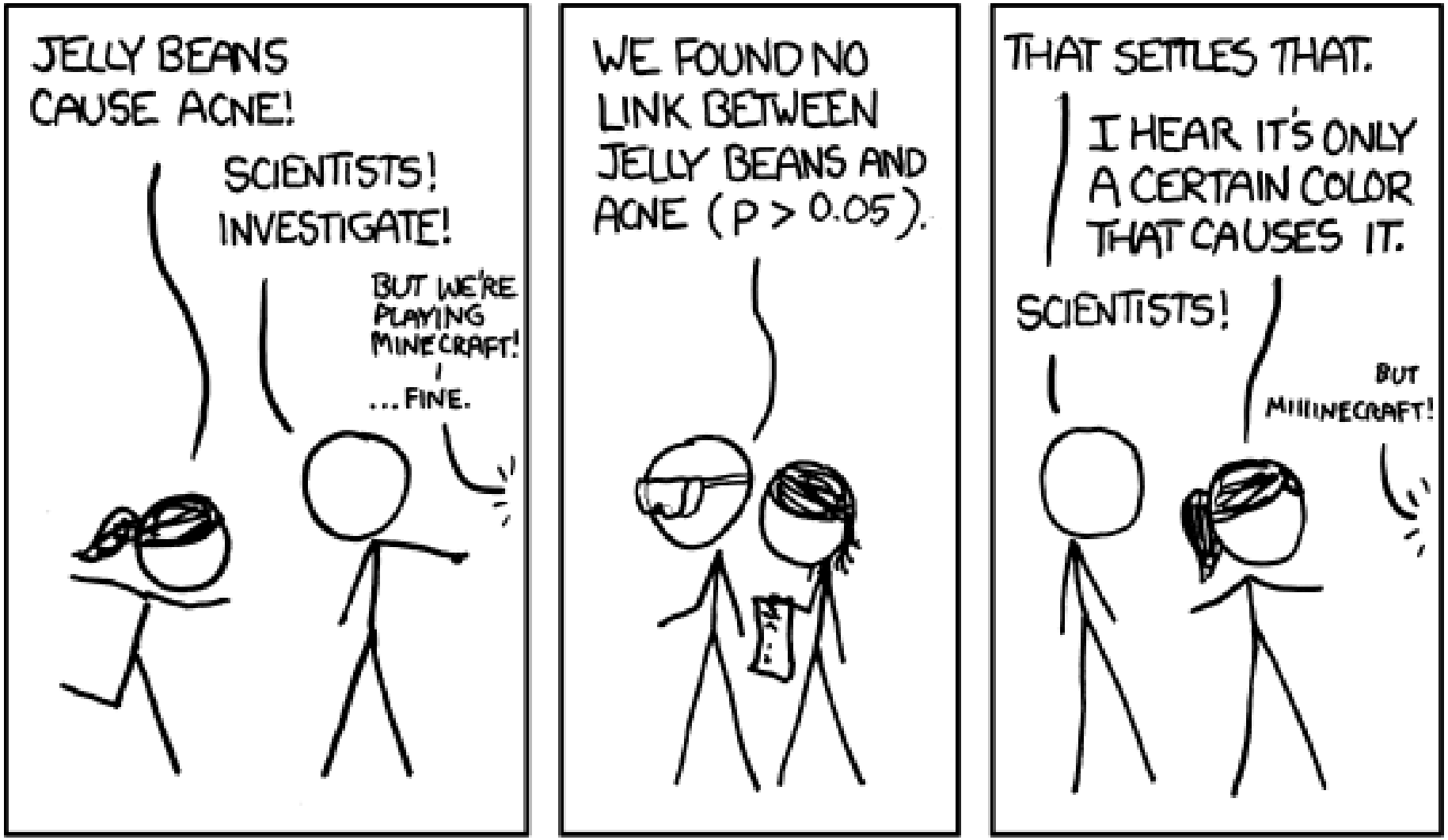
432 Class 22 Slides

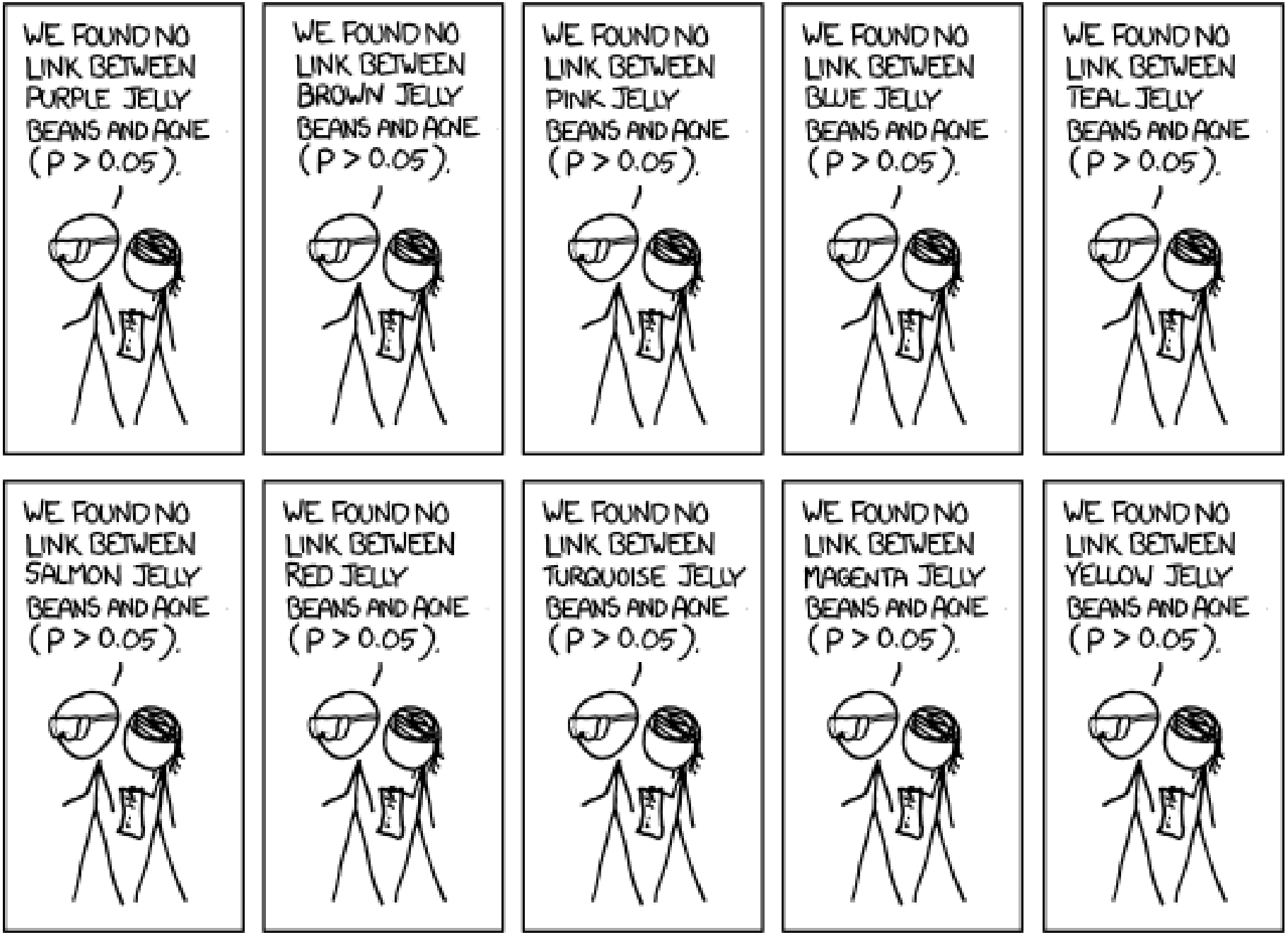
github.com/THOMASELOVE/2020-432

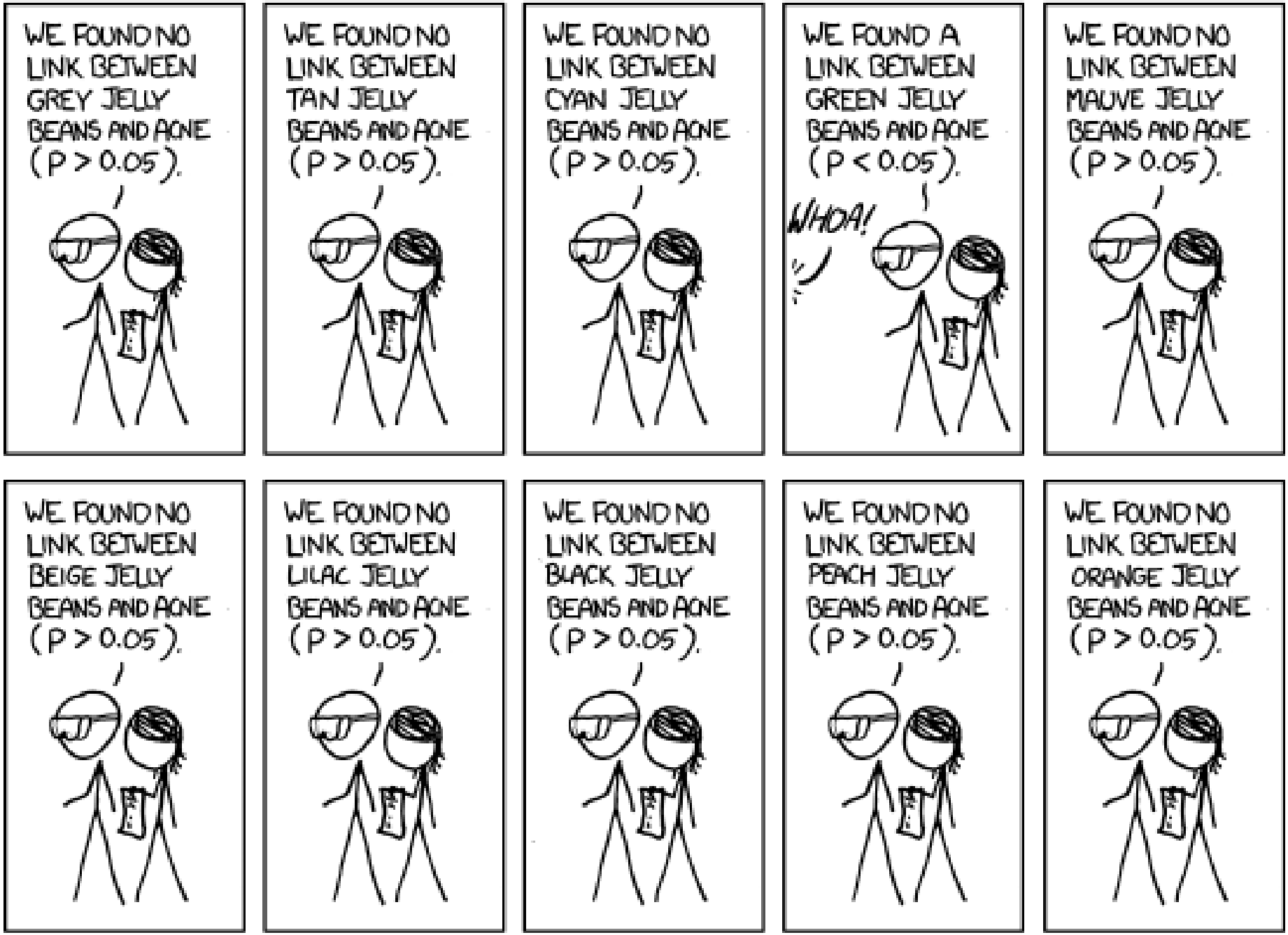
2020-04-14

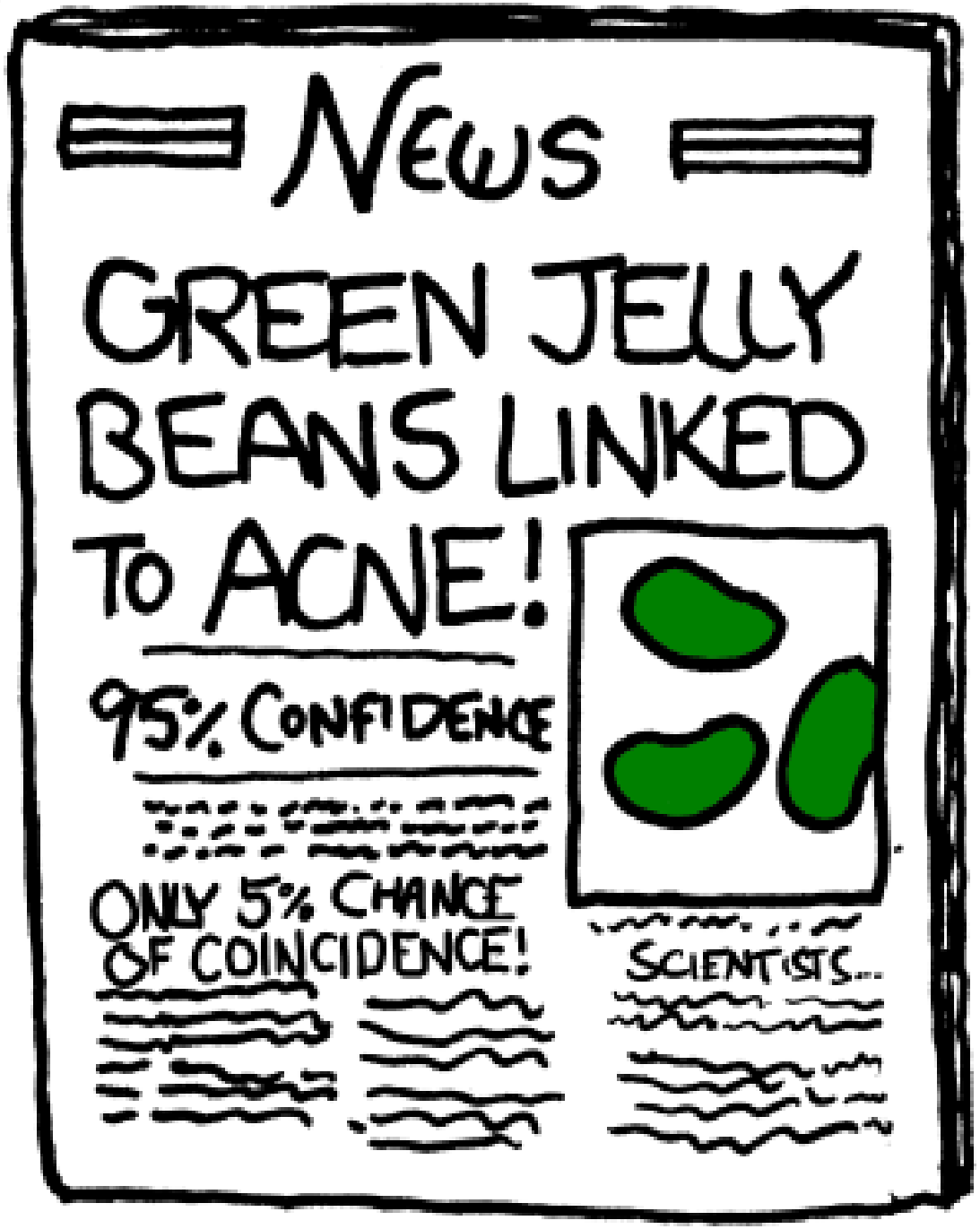
Replicable Research and the Crisis in Science

- ASA 2016 Statement on P values (Context, Process, Purpose)
- Is changing the p value cutoff the right strategy?
- Second-generation p values: A next step?
- ASA 2019 Statement on Statistical Inference in the 21st Century


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 MIIIINECRAFT!


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).


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).


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

Roger Peng's description of a successful data analysis

> *A data analysis is successful if the audience to which it is presented accepts the results.*

- "What is a Successful Data Analysis?" *simplystatistics.org* (2018-04-17).

So what makes a data analysis more believable / more acceptable?

The American Statistical Association

2016

- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

2019

- Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "p < 0.05", *The American Statistician*, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913.

Statistical Inference in the 21st Century

. . . a world learning to venture beyond "p < 0.05"

This is a world where researchers are free to treat "p = 0.051" and "p = 0.049" as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number.

In this world, where studies with "p < 0.05" and studies with "p > 0.05" are not automatically in conflict, researchers will see their results more easily replicated – and, even when not, they will better understand why.

The 2016 ASA Statement on P-Values and Statistical Significance started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times-an average of about 11 citations per week since its release. Now we must go further.

The American Statistical Association Statement on P values and Statistical Significance

The ASA Statement (2016) was mostly about what **not** to do.

The 2019 effort represents an attempt to explain what to do.

> *Some of you exploring this special issue of The American Statistician might be wondering if it's a scolding from pedantic statisticians lecturing you about what not to dowith p-values, without offering any real ideas of what to do about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.*

"Don't" is not enough.

If you're just arriving to the debate, here's a sampling of what not to do.

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p* value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

One More Don't...

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless. Made

A label of statistical significance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.

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

Link

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

Link

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.

Link

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 re-

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

Mark, Lee, Harrell JAMA Cardiol 2016-10-12

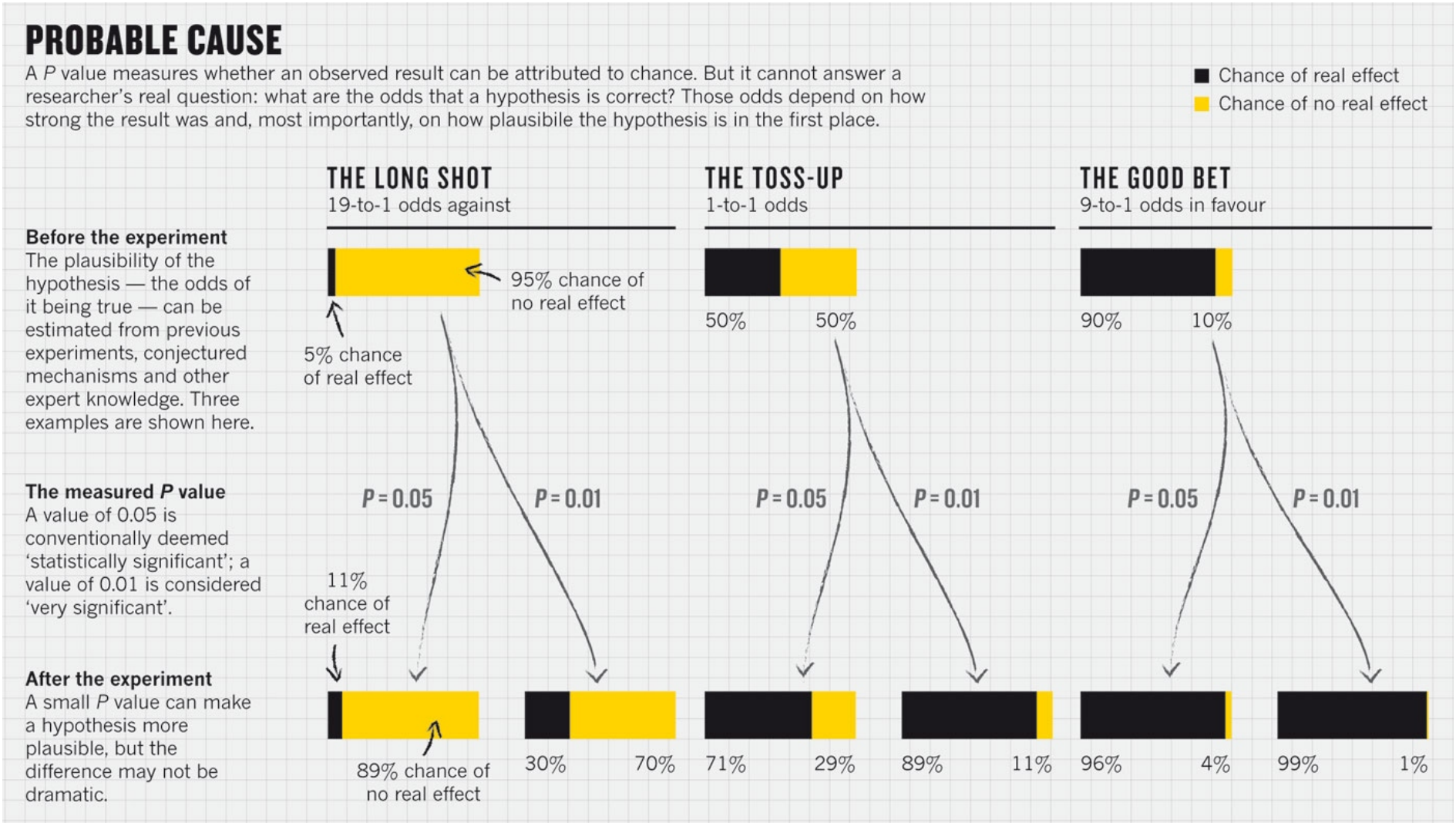
Abstract

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.

Link

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 can only be interpreted 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)

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
- [x] Governors
- [x] Senators
- [ ] Representatives

How do you want to measure economic performance?

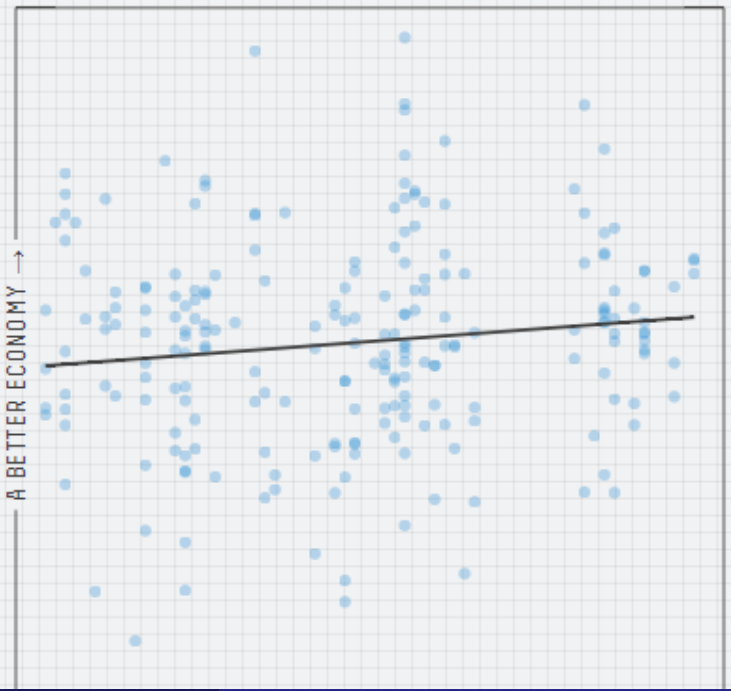
- [ ] Employment
- [x] Inflation
- [x] GDP
- [x] Stock prices

Other options

- [x] Factor in power
  Weight more powerful positions more heavily
- [x] Exclude 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.

"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

Are P values all that bad?

Grumpy Old Health Stats Dude
@healthstatsdude

Following

"If you never use another p-value, you will have improved medicine."

-me, to clinicians

#statstwitter #medtwitter #epitwitter

12:36 AM - 4 Mar 2019

Grumpy Old Health Stats Dude
@healthstatsdude

Following

Replying to @healthstatsdude @EugeneDayDSc and 2 others

My main reason for being overtly/in public anti p-values is this:

P values
f overall
analyses
partly
atistical
even if
group,
wer, due
l distri-
fference
specific
ll death

mate of a 5% decrease in 10-year survival with watchful waiting, 750 men might have died prematurely as a result.

A mistake in the operating room can threaten the life of one patient; a mistake in statistical analysis or interpretation can lead to hundreds of early deaths. So it is perhaps odd that, while we allow a doctor to conduct surgery only after years of training, we give SPSS® (SPSS, Chicago, IL) to almost anyone. Moreover, whilst only a surgeon would comment on surgical technique, it seems that anybody, regardless of statistical training,

day); a
that on
risk of t
in many
that the
event is

Comp
The aut
no com

7:59 PM - 19 Apr 2019

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?

Thomas Leeper
@thosjleeper

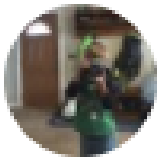
Follow

Replying to @ChelseaParlett

An interval drawn such that, were repeated, equal-sized samples of units drawn from the population of units using an identical sampling procedure and the same estimator was applied to each sample, 100*(1-α)% of those intervals would contain the population parameter of interest.

4:58 PM - 13 Mar 2018

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

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 interpretation
- 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*

JAMA 2018-04-10

VIEWPOINT

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Meta-Research Innovation Center at Stanford, Departments of Medicine, Health Research and Policy, Biomedical Data Science, and Statistics, Stanford University, Stanford, California.

The Proposal to Lower *P* Value Thresholds to .005

***P* values and accompanying methods** of statistical significance testing are creating challenges in biomedical science and other disciplines. The vast majority (96%) of articles that report *P* values in the abstract, full text, or both include some values of .05 or less.[1] However, many of the claims that these reports highlight are likely false.[2] Recognizing the major importance of the statistical significance conundrum, the American Statistical Association (ASA) published[3] a statement on *P* values in 2016. The status quo is widely believed to be problematic, but how exactly to fix the problem is far more contentious. The contributors to the ASA statement also wrote 20 independent, accompanying commentaries focusing on different aspects and prioritizing different solutions. Another large coalition of 72 methodologists recently proposed[4] a specific, simple move: lowering the routine *P* value threshold for claiming statistical significance from .05 to .005 for new discoveries. The proposal met with strong endorsement in some circles and concerns in others.

P values are misinterpreted, overtrusted, and misused. The language of the ASA statement enables the dis-

fully considered how low a *P* value should be for a research finding to have a sufficiently high chance of being true. For example, adoption of genome-wide significance thresholds ($P < 5 \times 10^{-8}$) in population genomics has made discovered associations highly replicable and these associations also appear consistently when tested in new populations. The human genome is very complex, but the extent of multiplicity of significance testing involved is known, the analyses are systematic and transparent, and a requirement for $P < 5 \times 10^{-8}$ can be cogently arrived at.

However, for most other types of biomedical research, the multiplicity involved is unclear and the analyses are nonsystematic and nontransparent. For most observational exploratory research that lacks preregistered protocols and analysis plans, it is unclear how many analyses were performed and what various analytic paths were explored. Hidden multiplicity, nonsystematic exploration, and selective reporting may affect even experimental research and randomized trials. Even though it is now more common to have a preexisting protocol and statistical analysis plan and preregistration of

Blume et al. PLoS ONE (2018) 13(3): e0188299

RESEARCH ARTICLE

Second-generation *p*-values: Improved rigor, reproducibility, & transparency in statistical analyses

Jeffrey D. Blume[1]*, Lucy D'Agostino McGowan[2], William D. Dupont[3], Robert A. Greevy, Jr.[1]

Second-generation p values

Verifying that a statistically significant result is scientifically meaningful is not only good scientific practice, it is a natural way to control the Type I error rate. Here we introduce a novel extension of the p-value—a second-generation p-value (p_δ)–that formally accounts for scientific relevance and leverages this natural Type I Error control. The approach relies on a pre-specified interval null hypothesis that represents the collection of effect sizes that are scientifically uninteresting or are practically null. The second-generation p-value is the proportion of data-supported hypotheses that are also null hypotheses. As such, second-generation p-values indicate when the data are compatible with null hypotheses ($p_\delta = 1$), or with alternative hypotheses ($p_\delta = 0$), or when the data are inconclusive ($0 < p_\delta < 1$). Moreover, second-generation p-values provide a proper scientific adjustment for multiple comparisons and reduce false discovery rates. This is an advance for environments rich in data, where traditional p-value adjustments are needlessly punitive. Second-generation p-values promote transparency, rigor and reproducibility of scientific results by *a priori* specifying which candidate hypotheses are practically meaningful and by providing a more reliable statistical summary of when the data are compatible with alternative or null hypotheses.

Nature P values are just the tip of the iceberg!

COMMENT

P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

OK, so what SHOULD we do?

The American Statistician Volume 73, 2019, Supplement 1

Articles on:

1. Getting to a Post "$p < 0.05$" Era
2. Interpreting and Using p
3. Supplementing or Replacing p
4. Adopting more holistic approaches
5. Reforming Institutions: Changing Publication Policies and Statistical Education

- Note that there is an enormous list of "things to do" in Section 7 of the main editorial, too.

Statistical Inference in the 21st Century

The American Statistician

Taylor & Francis
Taylor & Francis Group

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p<0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

To link to this article: https://doi.org/10.1080/00031305.2019.1583913

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

- Statistical methods do not rid data of their uncertainty.

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

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

We can make acceptance of uncertainty more natural to our thinking by accompanying every point estimate in our research with a measure of its uncertainty such as a standard error or interval estimate. Reporting and interpreting point and interval estimates should be routine.

How will accepting uncertainty change anything? To begin, it will prompt us to seek better measures, more sensitive designs, and larger samples, all of which increase the rigor of research.

It also helps us be modest . . . [and] leads us to be thoughtful.

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

3.2. Be Thoughtful

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Grim Reality

- Editorial, Educational and Other Institutional Practices Will Have to Change
- It Is Going to Take Work, and It Is Going to Take Time
- Why Will Change Finally Happen Now?

Next Time

Gelman and Carlin (2014) Power and Retrospective Design