Statistical Modeling, Causal Inference, and Social Science

Response to some comments on “Abandon Statistical Significance”

Posted by Andrew on 2 October 2017, 1:00 pm

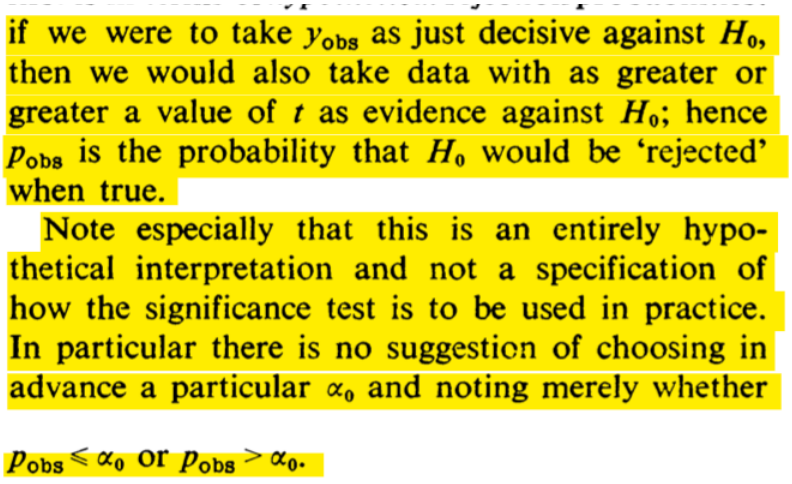
The other day, Blake McShane, David Gal, Christian Robert, Jennifer Tackett, and I wrote a paper, Abandon Statistical Significance, that began:

> In science publishing and many areas of research, the status quo is a lexicographic decision rule in which any result is first required to have a p-value that surpasses the 0.05 threshold and only then is consideration—often scant—given to such factors as prior and related 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. There have been recent proposals to change the p-value threshold, but instead we recommend abandoning the null hypothesis significance testing paradigm entirely, leaving p-values as just one of many pieces of information with no privileged role in scientific publication and decision making. We argue that this radical approach is both practical and sensible.

Since then we’ve received some feedback that we’d like to share and address.

1. Sander Greenland commented that maybe we shouldn’t label as “radical” our approach of removing statistical significance from its gatekeeper role, given that prominent statisticians and applied researchers have recommended this approach (abandoning statistical significance as a decision rule) for a long time.

Here are two quotes from David Cox et al. from a 1977 paper, “The role of significance tests”:



if we were to take y_{obs} as just decisive against H_0, then we would also take data with as greater or greater a value of t as evidence against H_0; hence p_{obs} is the probability that H_0 would be ‘rejected’ when true.

Note especially that this is an entirely hypothetical interpretation and not a specification of how the significance test is to be used in practice. In particular there is no suggestion of choosing in advance a particular α_0 and noting merely whether

$p_{obs} \leqslant \alpha_0$ or $p_{obs} > \alpha_0$.

separate. For then a separate measure of uncertainty need attaching to each conclusion, and consideration of so-called experiment-wise error rates is not required.

Here’s Cox from 1982 implicitly endorsing the idea of type S errors:

logical significance and practical significance. Statistical significance is concerned with whether, for instance, the direction of such and such an effect is reasonably firmly established by the data under analysis. Biological significance is concerned with

And here he is, explaining (a) the selection bias involved in any system in which statistical significance is a decision rule, and (b) the importance of measurement, a crucial issue in statistics that is obscured by statistical significance:

(vii) If a number of independent studies are done, possibly in different centres, and only those yielding an effect significant at say the 0.05 level are reported, it is clear that a distorted picture may emerge. The criterion for publication should be the achievement of

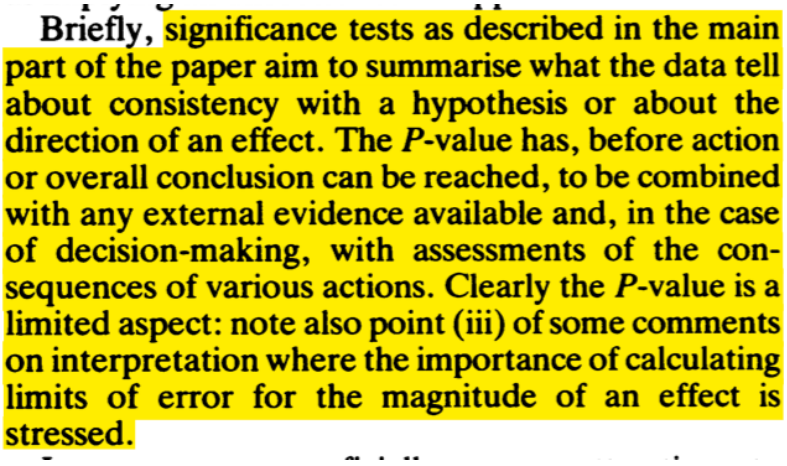
reasonable precision and not whether a significant effect has been found; it may be, however, that

Hey! He even pointed out that the difference between “significant” and “non-significant” is not itself statistically significant:

B. *Of two independent studies of the same topic, one gives no significant effect and the other gives significance at the 1% level. Thus the two studies are inconsistent.* Not necessarily so! A modest effect and

In this paper, Cox also brings up the crucial point that the “null hypothesis” is not just the assumption of zero effect (which is typically uninteresting) but also the assumption of zero systematic error (which is typically ridiculous).

And he says what we say, that the p-value tells us very little on its own:



Briefly, significance tests as described in the main part of the paper aim to summarise what the data tell about consistency with a hypothesis or about the direction of an effect. The *P*-value has, before action or overall conclusion can be reached, to be combined with any external evidence available and, in the case of decision-making, with assessments of the consequences of various actions. Clearly the *P*-value is a limited aspect: note also point (iii) of some comments on interpretation where the importance of calculating limits of error for the magnitude of an effect is stressed.

There are also more recent papers that say what McShane et al. and I say; for example, Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth wrote:

> The widespread use of ‘statistical significance’ as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process. We review why degrading p-values into ‘significant’ and ‘nonsignificant’ contributes to making studies irreproducible, or to making them seem irreproducible. A major problem is that we tend to take small p-values at face value, but mistrust results with larger p-values. In either case, p-values can tell little about reliability of research, because they are hardly replicable even if an alternative hypothesis is true. . . . Data dredging, p-hacking and publication bias should be addressed by removing fixed significance thresholds. Consistent with the recommendations of the late Ronald Fisher, p-values should be interpreted as graded measures of the strength of evidence against the null hypothesis. Also larger p-values offer some evidence against the null hypothesis, and they cannot be interpreted as supporting the null hypothesis, falsely concluding that ‘there is no effect’. . . . We further discuss potential arguments against removing significance thresholds, such as ‘we need more stringent decision rules’, ‘sample sizes will decrease’ or ‘we need to get rid of p-values’. We conclude that, whatever method of statistical inference we use, dichotomous threshold thinking must give way to non-automated informed judgment.

Damn! I liked that paper when it came out, but now that I see it again, I realize how similar our points are to theirs.

Also this recent letter by Valentin Amrhein and Sander Greenland, “Remove, rather than redefine, statistical significance” which, again, has a very similar perspective to ours.

2. In the park today I ran into a friend who said that he’d read our recent article. He expressed the opinion that our plan might be good in some ideal sense but it can’t work in the real world because it requires more time-consuming and complex analyses than researchers are willing or able to do. If we get rid of p-values, what would we replace them with?

I replied: No, our plan is eminently realistic! First off, we don’t recommend getting rid of p-values; we recommend treating them as one piece of evidence. Yes, it can be useful to see that a given data pattern could or not plausibly have arisen purely by chance. But, no, we don’t think that publication of a result, or further research in an area, should require a low p-value. Depending on the context, it can be

completely reasonable to report and follow up on a result that is interesting and important, even if the data are weak enough that the pattern could’ve been obtained by chance: that just tells us we need better data. Report the p-value and the confidence interval and other summaries; don’t use them to decide what to report. And definitely don’t use them to partition results into “significant” and “non-significant” groups.

I also remarked that it’s not like the current system is so automatic. Statistically significance, in most cases, a requirement for publication, but journals still have to decide what to do with the zillions of “p less than 0.05” papers that get sent to them every month. So we’re just saying that, at a start, that journals can use whatever rules they’re currently using to decide which of these papers to publish.

Then I launched into another argument . . . but at this point my friend gave me a funny look and started to back away. I think he’d just mentioned my article and his reaction as a way to say hi, and he wasn’t really asking for a harangue in the middle of the park on a nice day.

But I’m pretty sure that most of you reading this blog are sitting in your parent’s basement eating Cheetos, with one finger on the TV remote and the other on the Twitter “like” button. So I can feel free to rant away.

3. There’s a paper, “Redefine statistical significance,” by Daniel Benjamin et al., who recognize that the p=0.05 threshold has lots of problems (I don’t think they mention air rage, himmicanes, ages ending in 9, fat arms and political attitudes, ovulation and clothing, ovulation and voting, power pose, embodied cognition, and the collected works of Satoshi Kanazawa and Brian Wansink, but they could have) and promote a revised p-value threshold of 0.005. As we wrote in our article (which was in part a response to Benjamin et al.):

> We believe this proposal is insufficient to overcome current difficulties with replication . . . In the short term, a more stringent threshold could reduce the flow of low quality work that is currently polluting even top journals. In the medium term, it could motivate researchers to perform higher-quality work that is more likely to crack the 0.005 barrier. On the other hand, a steeper cutoff could lead to even more overconfidence in results that do get published as well as greater exaggeration of the effect sizes associated with such results. It could also lead to the discounting of important findings that happen not to reach it. In sum, we have no idea whether implementation of the proposed 0.005 threshold would improve or degrade the state of science as we can envision both positive and negative outcomes resulting from it. Ultimately, while this question may be interesting if difficult to answer, we view it as outside our purview because we believe that p-value thresholds (as well as those based on other statistical measures) are a bad idea in general.

4. And then yet another article, this one by Lakens et al., “Justify your alpha.” Their view is closer to ours in that they do not want to use any fixed p-value threshold, but they still seem to recommend that statistical significance be used for decision rules: “researchers justify their choice for an alpha level before collecting the data, instead 2of adopting a new uniform standard.” We agree with most of what Lakens et al. write, especially things like, “Single studies, regardless of their p-value, are never enough to conclude that there is strong evidence for a theory” and their call

to researchers to provide “justifications of key choices in research design and statistical practice.”

We just don’t see any good reason to make design, analysis, publication, and decision choices based on “alpha” or significance levels. As we write:

> Various features of contemporary biomedical and social sciences—small and variable effects, noisy measurements, a publication process that screens for statistical significance, and research practices—make null hypothesis significance testing and in particular the sharp point null hypothesis of zero effect and zero systematic error particularly poorly suited for these domains. . . .
>
> Proposals such as changing the default p-value threshold for statistical significance, employing confidence intervals with a focus on whether or not they contain zero, or employing Bayes factors along with conventional classifications for evaluating the strength of evidence suffer from the same or similar issues as the current use of p-values with the 0.05 threshold. In particular, each implicitly or explicitly categorizes evidence based on thresholds relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error.

5. E. J. Wagenmakers, one of the authors of the Benjamin et al. paper that motivated a lot of this recent discussion, wrote a post on his new blog (E. J. has a blog now! Cool. Will he start posting on chess?), along with Quentin Gronau, responding to our recent article.

E. J. and Quentin begin their post with five places where they agree with us. Then, in true blog fashion, they spends most of the post elaborating on three places where they disagree with us. Fair enough.

I’ll go through them one at a time:

E. J. and Quentin’s disagreement 1. E. J. says that our general advice (studying and reporting the totality of their data and relevant results) is eminently sensible, but it is not sufficiently explicit to replace anything. Rightly or wrongly, the p-value offers a concrete and unambiguous guideline for making key claims; the Abandoners [that’s us!] wish to replace it with something that can be summarized as ‘transparency and common sense.’”

I disagree!

First, the p-value does *not* offer “a concrete and unambiguous guideline for making key claims.” Thousands of experiments are performed every month (maybe every day!) with “p less than 0.05” results, but only a very small fraction of these make their way into JPSP, Psych Science, PPNAS, etc. P-value thresholds supply an illusion of rigor, and maybe in some settings that’s a good idea, by analogy to “the consent of the governed” in politics, but there’s nothing concrete or unambiguous about their use.

Second, yes I too support “transparency and common sense,” but that’s *not* all we’re recommending. Not at all! Recall my recent paper, Transparency and honesty are not enough. All the transparency and common sense in the world—even with

preregistered replication—won't get you very far in the absence of accurate and relevant measurement. Hence the last paragraph of this post.

E. J. and Quentin's disagreement 2. I'll let my coauthor Christian Robert respond to this one. And he did!

E. J. and Quentin's disagreement 3. They write, "One of the Abandoners' favorite arguments is that the point-null hypothesis is usually neither true nor interesting. So why test it? This echoes the opinion of researchers like Meehl and Cohen. We believe, however, that Meehl and Cohen were overstating their case."

E. J. and Quentin begin with an example of a hypothetical researcher comparing the efficacies of unblended or blended whisky as a treatment of snake bites. I agree that in this case the point null hypothesis is worth studying. This sort of example has come up in some recent comment threads so I'll repeat what I said there:

> I don't think that point hypotheses are *never* true; I just don't find them interesting or appropriate in the problems in social and environmental science that I work on and which we spend a lot of time discussing on this blog.
>
> There are some problems where discrete models make sense. On commenter gave the example of a physical law; other examples are spell checking (where, at least most of the time, a person was intending to write some particular word) and genetics (to some reasonable approximation). In such problems I recommend fitting a Bayesian model for the different possibilities. I still don't recommend hypothesis testing as a decision rule, in part because in the examples I've seen, the null hypothesis also bundles in a bunch of other assumptions about measurement error etc. which are not so sharply defined.

I'm happy to (roughly) discretely divide the world into discrete and continuous problems, and to use discrete methods when studying the effects of snakebites, and ESP, and spell checking, and certain problems in genetics, and various other problems of this sort; and to use continuous methods when studying the effects of educational interventions, and patterns of voting and opinion, and the effects of air pollution on health, and sex ratios and hurricanes and behavior on airplanes and posture and differences between gay and straight people and all sorts of other topics that come up all the time. And I'm also happy to use mixture models with some discrete components; for example, in some settings in drug development I expect it makes sense to allow for the possibility that a particular compound has approximately no effect (I've heard this line of research is popular at UC Irvine right now). I don't want to take a hard line, nothing-is-ever-approximately-zero position. But I do think that comparisons to a null model of absolutely zero effect and zero systematic error are rarely relevant.

E. J. and Quentin also point out that if an effect is very small compared to measurement/estimation error, then it doesn't matter, from the standpoint of null hypothesis significance testing, whether the effect is exactly zero. True. But we don't particularly care about null hypothesis significance testing! For example, consider "embodied cognition." Embodied cognition is a joke, and it's been featured in lots of junk science, but I don't think that masked messages have zero or even

necessarily tiny effects. I think that any effects will vary a lot by person and by context. And, more to the point, if someone wants to do research in this topic, I don’t think that a null hypothesis significance test should be a screener for what results are considered worth looking at, and I think that it’s a mistake to use a noisy data summary to selecting a limited subset of results to report.

Summary

We’re in agreement with just about all the people in this discussion on the following key point: We’re unhappy with the current in which “p less than 0.05” is used as the first step in a lexicographic decision rule in deciding which results in a study should be presented, which studies should be published, and which lines of research should be pursued.

Beyond this, here are the different takes:

Benjamin et al. recommend replacing 0.05 by 0.005, not because they think a significance-testing-based lexicographic decision rule is a good idea, but, as I understand them, because they think that 0.005 is a stringent enough cutoff that it will essentially break the current system. Assuming there is a move to reduce uncorrected researcher degrees of freedom and forking paths, it will become very difficult for researchers to reach the 0.005 threshold with noisy, useless studies. Thus, the new threshold, if applied well, will suddenly cause the stream of easy papers to dry up. Bad news for Ted, NPR, and Susan Fiske, but good news for science, as lots of journals will either have to get a lot thinner or will need to find some interesting papers outside the usual patterns. In the longer term, the stringent threshold (if tied to control of forking paths) could motivate researchers to do higher-quality studies with more serious measurement tied more carefully to theory.

Lakens et al. recommend using p-value thresholds but with different thresholds for different problems. This has the plus of moving away from automatic rules but has the minus of asking people to “justify their alpha.” I’d rather have scientists justifying their substantive conditions by delineating reasonable ranges of effect sizes (see, for example, section 2.1 of this paper) rather than having them justify a scientifically meaningless threshold, and I’d prefer that statisticians and methodologists evaluate frequency properties of type M and type S errors rather than p-values. But, again, we agree with Lakens et al., and with Benjamin et al., on the key point that what we need is better measurement and better science.

Finally, our perspective, shared with Amrhein, Korner-Nievergelt, and Roth, as well as Amrhein and Greenland, is that it’s better to just remove null hypothesis significance testing from its gatekeeper role. That is, instead of trying to tinker with the current system (Lakens et al.) or to change the threshold so much that the system will break (Benjamin et al.), let’s just discretize less and display more.

We have some disagreements regarding the relevance of significance tests and null hypotheses but we’re all roughly on the same page as Cox, Meehl, and other predecessors.

Filed under Decision Theory, Miscellaneous Statistics, Multilevel Modeling | Permalink

47 Comments

1. *Stephen Martin* says:
October 2, 2017 at 2:01 pm

I am a coauthor on the Lakens et al. paper, and I agree with everything you say here. There was some heterogeneity in opinion on that commentary. From my perspective, I wish to get away from discretizing evidence and hypothesis testing in most cases (Some thought droppings here: http://srmart.in/thought-droppings-substantive-statistical-hypotheses/). There’s rarely a case where substantive hypotheses map very well onto statistical hypotheses, and because of that, I’d rather not use statistical hypothesis testing to make decisions about substantive hypotheses. Instead, we should build strong models, and make inferences about substantive claims in a more continuous manner, free of thresholds.

I’ve been railing against the Benjamin proposal /primarily/ because I pretty well hate thresholds. So long as thresholds exist, people will try their damndest to dive past it. So long as we have this “past the threshold, evidence; not past the threshold, no evidence” mentality, the career-incentives + publication practices essentially mandate that people will threshold-dive and mischaracterize evidence (i.e., p-hacking). And any inferential quantity can be ‘hacked’; I’ve shown that it can be done, of course, with p-values, CIs, credible intervals, BFs, whatever else. It’s not particularly hard, regardless of the threshold. I even wrote a script that will find subsets where some threshold is met, just to demonstrate. And of course, it’s arbitrary; why one line can delineate “true” from “not true”, or “publication-worthy” from “not publication worthy” is silly.

Finally, my argument about the Benjamin proposal was that it’s unrealistically optimistic: “In an imperfect world where p-values are used, use .005” but then they say “.005 should not be used as a publication threshold”; this is a sticking point for me as an ECR. In the same imperfect world where people misinterpret p-values and dichotomize evidence, they will use .005 as a pub threshold instead of .05 — Because it’s an imperfect world. Saying one should use .005 < p < .05 as 'suggestive' and < .005 as 'significant' but "don't use p-value as pub-worthiness" is too "idealistic" for me; of COURSE people will just now trichotomize evidence, and those people are now being told "only .005 signifies evidence", therefore they will just s/.05/.005 in their pub-worthiness evaluation. TDLR; why would people who judge pub-worthiness based on whether p is less than some evidentiary threshold now stop doing so with .005 if .005 is the new evidentiary threshold?

Even if I am on the 'justify your alpha' paper, I will say that I just don't like thresholds, period. But for me, IF we are going to use thresholds, they should be justified, and hence my contribution to the paper.

Aside from the 'where should the threshold be' question, which I personally think is moot… as I said on twitter:
Things that caused psych problems: HARKing, QRPs, threshold diving, publication bias, poor stats understandings, bad incentive structure. Incentive structure requiring novel, sexy findings, 10 papers a year, only publishable if beyond threshold. No replications permitted. Things that don't need fixing: An arbitrary threshold for rejecting a hypothesis noone believes based on a fictitious universe of events.

2. *Phil* says:
October 2, 2017 at 2:04 pm

> First off, we don’t recommend getting rid of p-values; we recommend treating them as one piece of evidence.

Perhaps “abandon” is the wrong word to use then.

- *Stephen Martin* says:
October 2, 2017 at 2:07 pm

Abandon statistical significance != abandon p-values.

  - *Peter Erwin* says:
October 5, 2017 at 9:34 am

“Abandon statistical significance != abandon p-values.”

The problem is that people rather easily jump to that conclusion. For example, when the journal *Basic and Applied Social Psychology* made their 2015 shift to “no more significance testing”, the editorial specifically said (in answer to the question “Will manuscripts with p-values be desk-rejected automatically?”), “No. … But prior to publication, the authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about ‘significant’ differences or lack thereof, and so on).”

3. *Jonathan (another one)* says:
October 2, 2017 at 2:19 pm

The “late Ronald Fisher?” He’s been dead for 65 years… That’s about 5 standard deviations past the consensus of how long one can use this locution… (P<<.01)

- *Jonathan (another one)* says:
October 2, 2017 at 2:20 pm

Sorry… 55 years. I’ll give you a standard deviation back.

4. *Bob* says:
October 2, 2017 at 2:47 pm

I think there are political-economic obstacles to the implementation of this

idea. Speaking from a social science perspective, what you’re arguing for is a kind of idealism, or change that derives from shifting people’s perspective of the world. But p-values are deeply embedded in people’s material interests in which they have been claiming for decades that their research is correct because p<0.05. Of course it's still possible to change this, but there has to be a coalition built that is capable of creating substantial disincentives to the current system, such as by considering research that uses discretized p-values "second-class".

In sum, perhaps a new journal that adopts these standards may be the best way to get pragmatic change by offering researchers a chance to showcase their more enlightened analyses.

- *Stephen Martin* says:
  October 2, 2017 at 3:03 pm

  Yup; I think the problem is multi-rooted.

  Bad statistical training invites lazy assessment of ‘evidence’.
  Journals get tons of submissions, and they need some easy manner of deciding on pub-worthiness (not that evidence should even factor into that equation).
  The entire notion of “hypothesis testing” is so strongly embedded into psychology, it’s going to be hard to uproot. Hypothesis testing sounds more “scientific” than “making the best inference from the data we have” or “trying to recover the DGP with out-of-sample prediction as a goal”, even if the latter two are really much harder [and more useful, more informative, and arguably more scientific].

  I fear the next move will be:
  1) P-values? Boo hiss. Let’s use Bayes factors.
  2) Oh, Bayes factors don’t condition on the data either? The priors are actually prior predictive hypotheses? That’s not what I want.
  3) Bayesian posteriors! Huzzah! If the credible interval excludes zero, then H1 is supported! Oh, that has many of the same problems as p-values?
  4) Prediction! Let’s use predictive utility as a goal, and based on that construct better predictive hypothesis tests. Oh, that requires more work, and about 10 people on earth understand how to do that….
  5) Let’s have everyone report everything. p-value, BF, informed BF, predictive utility metrics, posterior credible intervals. Ah, but that’s confusing and hard to interpret.

  The statisticians: Exactly. Data are messy. Stop thresholding. Lots of metrics needed for decisions and inferences.

  - *Solomon Kurz* says:
    October 2, 2017 at 7:42 pm

    That’s one way to approach the problem. And yet, as they say,

science progresses one funeral at a time. Which implies it’s the youngsters—like me—who should be the easiest to target. And you all have targeted me well. Here was one of the most effective ways I was targeted: high quality intro stats books focusing on all the cool things you can do outside of the NHST paradigm. My prime example is McElreath’s “Statistical Rethinking.” Sure, some of his code is a little spaghetti-ish, but now we have the tidyverse and brms to help with future attempts. Perhaps the 2nd edition of Gelman and Hill will qualify, too.

Also, YouTube tutorials have been tremendous in my stats education. Again, McElreath’s lecture series is a great example. The internet needs more. But they don’t need to be that polished. There are tons of low-production value screenshot only YouTube videos on classical statistics in SPSS and so forth. The grad student audience is hungry for the Bayesian analogues featuring Stan, rstanarm, brms, and so forth.

And of course, blogs. But it appears Andrew and many of the rest of y’all have that one covered. [I love this community.]

So, yes, we can attempt top-down approached. And we should. But the current youngsters are primed and ready. Reach us with more of these.

- *Ben Prytherch* says:
  October 2, 2017 at 8:18 pm

  +1 Solomon Kurz

- *Ben Goodrich* says:
  October 3, 2017 at 12:25 am

  My lectures are up at
  https://www.youtube.com/channel
  /UCBiO111B17hlhtRY5Cg3V_Q

5. *Sameera Daniels* says:
   October 2, 2017 at 3:59 pm

   Well NHST should never have expanded its criteria to nearly all disciplines, like economics, medicine, & public health. And it’s not hard to see that it substitutes as a marketing tool & thinking more generally.

6. *Sameera Daniels* says:
   October 2, 2017 at 4:02 pm

   As a non-statistician, however, I haven’t yet found a cogent explanation for why p-values should be regarded as continuous measure either. More broadly, the quality of insights is the real problem.

7. *Sameera Daniels* says:
   October 2, 2017 at 4:05 pm

   Sander Greenland is absolutely right in positing that non-standard & standard definitions of p-values as well as other statistics terms have lent their abuses & misuses in practice & of epistemics.

8. *Keith O'Rourke* says:
   October 2, 2017 at 4:14 pm

   > but they {Lakens et al] still seem to recommend that statistical significance be used for decision rules:

   Maybe not – “Our recommendation is similarly twofold. First, when describing results, we recommend that the label ‘statistically significant’ simply no longer be used.”

   Then later “Second, when designing studies, we propose that authors transparently specify their design choices. These include (where applicable) the alpha level,”

   Now for “when studying the effects of snakebites, and ESP, and spell checking,” the alpha level should be justified and hence varied by application but what purpose does it serve other than to ignore their first recommendation?

   AG> “But I do think that comparisons to a null model of absolutely zero effect and zero systematic error are rarely relevant.”
   With regard to the zero systematic error rarely being relevant – I raised that in a comment perhaps too late in the editing phase “there does not seem to be anything on systematic error or confounding”

   In general, I think there are number of authors that are agreeing more than is suggested by their making distinctions out of differences they discern in each others papers.

   Perhaps the various papers should be partially pooled towards Valentin Amrhein and Sander Greenland very concise letter – to better pin point what the important distinctions really are?

9. *Daniel Lakeland* says:
   October 2, 2017 at 4:19 pm

   I think there’s not enough emphasis placed on the critical idea in this blog post, which is:

   ‘But I’m pretty sure that most of you reading this blog are sitting in your parent’s basement eating Cheetos, with one finger on the TV remote and the other on the Twitter “like” button. So I can feel free to rant away.’

   ;-)

- *Joachim* says:
  October 2, 2017 at 4:29 pm

  It does make one wonder how Andrew thinks we read his papers.

  - *Keith O'Rourke* says:
    October 2, 2017 at 5:43 pm

    As I understand it many readers never read any comments (nor make any).

    So my rough estimate is that most readers are just looking for distractions that don’t require much commitment.

- *Sameera Daniels* says:
  October 2, 2017 at 5:00 pm

  Actually Daniel I’m eating chips & salsa. And I’d rather post here than watching numbing tv.

- *Corey* says:
  October 2, 2017 at 7:25 pm

  For the record, I only sit in my parent’s basement on Sundays.

  - *Martha (Smith)* says:
    October 3, 2017 at 1:03 am

    Gee, my parents never had a TV in their basement, let alone one with a remote. (Come to think of it, my house has neither a basement nor a TV.)

10. *Jacob* says:
    October 2, 2017 at 9:25 pm

    You raise an interesting point with regard to the fact that sometimes we might be justified to interpret p > .05 results as interesting and meriting further research (which is as far as most social scientists are willing to go in scientific outlets in describing their findings). Some of the reformism has often felt like it originated in the idea that p = roughly .05 results just aren’t compelling, but we’d prefer not to have to come up with a new, lower threshold. That’s why I wasn’t surprised by the new suggested .005 threshold.

    Of course, there are many reasons to dislike the p < .05 threshold that go beyond "it's not sufficiently conservative." But what I have never seen made clear is what we should do with higher p values. In this new world that doesn't believe in thresholds, is there value in p = .10? It's often said that in the garden of forking paths, there's always a scientific explanation for your results. Do we trust scientific reasoning about theory enough to accept results that are statistically weaker as part of this movement that is about science improvement?

- *Andrew* says:
  October 2, 2017 at 9:44 pm

  Jacob:

  The quick answer is, yes, lots of little data can become big data, and I think it’s fine for people to publish and analyze what data they have, and others can build on this. It can also be valuable to publish and recognize uncertainty, to say: Here’s what seemed like a promising line of research, but the data are too noisy to learn much of anything useful.

  Finally, sometimes we have real decisions to make and we can’t wait until there is any sort of near-certainty. Again, better to give the information that is available, in all its ambiguity. In a field such as empirical macroeconomics, the questions are too important and the data too sparse for us to wait for statistical significance in any form.

  In addition to all of the above, there’s also the question of incentives. So much bad work and data misrepresentation is done because of the implicit requirement that claims be presented as near-certain. I’d like to remove that burden, so the Susan T. Fiskes of the world can publish speculative work without the pressure to misrepresent the results as conclusive.

- *Ben Prytherch* says:
  October 5, 2017 at 8:56 pm

  Jacob, regarding this question:

  “But what I have never seen made clear is what we should do with higher p values. In this new world that doesn’t believe in thresholds, is there value in p = .10?”

  This is one reason I see merit in encouraging people to switch out their p-values for confidence intervals when possible. I know this gets pooh-poohed on the grounds that a 95% CI is just an inverted null hypothesis test using p<0.05, but one huge advantage to the interval is that its width tells you something. p=0.10 and a wide CI means that your results are consistent with there being "zero effect" or "a large effect", so this isn't very informative. p=0.10 and a narrow CI means that your results are consistent with there being "zero effect" or a "small effect", but a large effect is pretty much out (at least contingent on all the other assumptions that went into the analysis).

  I think this is a useful distinction, but with just a p-value you can't make it. And of course it also gives perspective to those p = 0.02 results where zero is just barely outside some really wide CI.

  More broadly, I think "statistically weak" results are still useful; if the study was worth doing then the results are worth reporting, even if the

conclusion is that we can't conclude anything. "Turns out there's more noise here than we thought" is a result worth sharing, not least because it can point to where improvements need to be made in measurement and design and modeling.

I don't know how many journal editors would buy into this, but maybe our new world that doesn't believe in thresholds will also be more open to the value of "unpublished" research.

- *Martha (Smith)* says:
  October 5, 2017 at 10:24 pm

  ” “Turns out there’s more noise here than we thought” is a result worth sharing, not least because it can point to where improvements need to be made in measurement and design and modeling.”

  +1

- *Valentin Amrhein* says:
  October 6, 2017 at 3:36 am

  We made a very similar point in our review on significance thresholds (see, e.g., fig. 1 for a comparison of p-values with CI, and how CI may be meaningfully interpreted even though p-values are relatively large).
  https://peerj.com/articles/3544/

  - *Ben Prytherch* says:
    October 7, 2017 at 5:02 pm

    Thanks Valentin, I like the homage to Cohen in your title, and the plot of difference CIs along with their p-values really drives the point home regarding how much more informative a CI is relative to a p-value. Not very many people who use p-values know what they are. And maybe not many people who use CIs really know what they are either, but I don’t think it matters as much with CIs. The picture of the interval basically tells the story, while the p-value remains mysterious.

- *Huw Llewelyn* says:
  October 6, 2017 at 6:03 pm

  Ben. I like your suggestion of using confidence intervals. The British Medical Journal has also been campaigning for their use as you suggest for years and look more favourably on papers that include them. The ’pooh-poohing’ suggests that critics of CIs think that the inversion is invalid. I have a special interest in this because I have shown that random sampling is an interesting special case where the prior probabilities of all its possible specified outcomes are equally probable (i.e. they are of necessity uniform). I explain this in

my Oxford University Press blog: https://blog.oup.com/2017/06/suspected-fake-results-in-science/ . I argue that any non-uniform Bayesian prior distribution is actually a posterior distribution formed by ‘normalising’ a likelihood distribution (based on real or pseudo-data) by assuming a uniform ‘base-rate’ distribution. This posterior probability distribution then becomes a Bayesian prior to be used with new data.

If a symmetrical distribution (e.g. Student’s ‘t’ or Gaussian) is used to model data, the 95% confidence intervals become 95% credibility intervals and that the null hypothesis corresponds to one of the confidence limits. However, if the CI applies to proportions, then the likelihood distribution will often be asymmetrical so that the simple relationship between confidence and credibility intervals does not hold. In this situation, the likelihood distributions and credibility intervals have to be calculated ‘exactly’ using hyper-geometric distributions using binomial coefficients (by analogy with ‘Fisher’s exact tests).

By applying the ‘principle of uniform distributions’, it is possible to provide a proper posterior likelihood distribution that allows the scientist to specify any credibility interval of his or her choosing and then for the statistician to estimate the probability of a mean or proportion falling inside that range after making an infinite number of observations. Prior data distributions can also be incorporated as a form of meta-analysis. This would be the probability of (long term) replication; I prefer to use ‘chosen replication range’ rather than CI. This would be much more intuitively appealing to scientists (and doctors whom I teach). The lines of ‘statistical significance’ (or not) could be marked on the distribution and treated with a pinch of salt. I also discuss some of the issues related to this approach elsewhere on Andrew’s blog: http://statmodeling.stat.columbia.edu/2017/10/04/worry-rigged-priors/#comment-578758).

- *Huw Llewelyn* says:
  October 6, 2017 at 7:59 pm

  Erratum: In the first sentence of the 3rd paragraph the phrase “proper posterior likelihood distribution” should have been “proper posterior PROBABILITY distribution”.

- *Simon Gates* says:
  October 7, 2017 at 12:07 pm

  Most medical journals require reporting of confidence intervals now, so people do generally report them (though not all the time). But when it comes to interpreting the results, it’s a different story – the usual thing is dichotomisation into there’s an effect/no effect based on a significance test.

Some examples on Frank Harrell’s blog
http://www.fharrell.com/2017/04/statistical-errors-in-medical-literature.html

I have loads more if you want.

- *Huw Llewelyn* says:
  October 7, 2017 at 8:10 pm

  Thank you. You, Frank Harrell and others are preaching to the converted! However, a ‘null hypothesis’ of zero difference between treatment and placebo may be of little interest to a doctor who may wish to know the estimated probability of at least a 25% cure rate (for example) before starting a new treatment policy. He would therefore specify a ‘replication range’ of >25% (the interval not having an upper bound) and wish to know the probability of achieving this conditional on the existing data alone. This would require a Bayesian estimate with uniform priors. In order to plan further studies it might help to use the existing study result to calculate the power required to achieve a specified higher probability of such a ‘ >25% replication range’ based on the distribution suggested by the existing data and also perhaps together with a sensitivity analysis using various subjectively estimated likelihood distributions that are different to the one provided by an existing study. I think that this would be easier for doctors to understand than P values and even fixed (e.g. 95%) confidence intervals. I completely agree with Ben and everyone that P-values are inadequate.

  - *Andrew* says:
    October 7, 2017 at 8:19 pm

    Huw:

    A Bayesian inference with uniform priors is what it is, and it can be useful as a data summary. But I generally wouldn’t want to use this posterior distribution to make decisions as it can be wildly optimistic.

    - *Huw Llewelyn* says:
      October 8, 2017 at 6:26 am

      Andrew

      I agree that other information needs to be taken into account to avoid over-optimism. I suggest regarding a Bayesian posterior probability based on the study data alone and a uniform prior as an

upper bound (e.g. see 2nd paragraph of https://blog.oup.com/2017/06/suspected-fake-results-in-science/).

One approach is to explore the effect of various pessimistic subjective priors (or ‘posterior likelihood’ distributions) as part of a sensitivity analysis to model the possible posterior probabilities of replication when a repeat study is conducted in other centres. Is your view on the over-optimistic nature of a posterior distribution based on a uniform prior down to Nosek et al’s study on ‘Estimating the reproducibility of psychological science’ or some other rationale?

- *Mark Schaffer* says:
  October 7, 2017 at 10:14 am

  +1 from me too. And maybe worth rewording one of Ben’s examples to cover another case: “p=0.10 and a wide CI that doesn’t include zero means that your results are consistent with there being “small effect” or “a large effect”, so this isn’t very informative either”.

  This is the trap that so many fall into. H0:b=0 and p=0.01 but the CI is huge … woo-hoo. So you think maybe it isn’t zero … so what? (Yes, I know, interpreting realized CIs is tricky, but still this is an improvement over mindless p-value citing.)

  But I agree with Ben and Martha that the answer to “So what?” can be “There’s more noise here than we thought.” Indeed may well be a result worth sharing. A CI can convey this.

  - *Carlos Ungil* says:
    October 7, 2017 at 10:39 am

    > p=0.10 and a wide CI that doesn’t include zero means that …

    …this is not a 95% confidence interval or the null hypothesis used to calculate the p-value is not H0:b=0.

    > This is the trap that so many fall into. H0:b=0 and p=0.01 but the CI is huge … woo-hoo. So you think maybe it isn’t zero … so what?

    The case when the CI is wide is may be more interesting (for the same p-value). If you have H0:b=0 and p=0.01 and the CI is narrow then maybe b isn’t zero but surely it’s small… so what?

    - *Mark Schaffer* says:
      October 7, 2017 at 11:44 am

“> p=0.10 and a wide CI that doesn’t include zero means that …

…this is not a 95% confidence interval or the null hypothesis used to calculate the p-value is not H0:b=0.”

This is nitpicky or I’m missing something very obvious so apologies (fighting a cold so I will get my excuses in early). Ben mentions 95% CI, then switches to p=0.10 for his example … then I used p=0.01 … does it matter? I think (hope?) the point is clear. If you are going to fixate on a particular cutoff for “significance”, you can tell us a lot more with the CI than with the (usually almost useless) fact of whether or not zero is inside it or not.

- *Carlos Ungil* says:
  October 7, 2017 at 2:37 pm

  > p=0.10 and a wide CI means that your results are consistent with there being “zero effect” or “a large effect”, so this isn’t very informative.
  > p=0.10 and a narrow CI means that your results are consistent with there being “zero effect” or a “small effect”, but a large effect is pretty much out
  p=0.10 means that the interval does cover zero. In his words, the results are consistent with there bing zero effect. There is no “another case”.

  You are right that a CI tells us more than a p-value (a p-value is one single piece on information, the CI contains two pieces of information).
  But consider these two possible 95% confidence intervals for H0:b=0 with (two-sided) p=0.01:

  Wide CI: 6 < b < 45
  Narrow CI: 0.006 < b < 0.045

  I am not sure the first one warrants a "so what?" reaction more than the second one.

- *Ben Prytherch* says:
  October 7, 2017 at 4:58 pm

  Thanks Carlos, that’s a good point and I agree. p=0.01 can be consistent with a precisely estimated small effect, or an imprecisely estimated large effect… but precisely estimated large effects produce p-values of 0.00000….whatever.

- *Mark Schaffer* says:
  October 7, 2017 at 5:35 pm

Ah … my fault, I wasn’t clear (I blame that cold). The “so what?” was in reference to the rejection of a point null of b=0, which we all agree is on its own usually pretty pointless and uninformative. CIs are far more informative, as the comments by Carlos and Ben (and me) illustrate.

- *Carlos Ungil* says:
  October 7, 2017 at 6:46 pm

  Ben, Mark, if we all agree this is no fun! ;-)

  - *Ben Prytherch* says:
    October 7, 2017 at 8:00 pm

    Carlos, now that I read back over my previous posts, I forgot to add that I think 95% CIs are not only more informative than p-values, but that they should also be treated as the ultimate arbiter of Truth. Scientific papers should be accepted for publication if and only if a 95% CI for something or another is reported to exclude zero. All hail the 95% CI!

11. *Sameera Daniels* says:
    October 3, 2017 at 7:32 am

    I would think that the preprint & pregisteration platforms should give us better insight into the reasoning processes of researchers. Simply carefully rereading of journal articles have yielded reconsideration of findings as well.
    As an aside I think that some researchers should reread Dr. Ioannidis work more carefully. It is paraphrased incorrectly & somewhat correctly. But not interpreted precisely & correctly. Either the applications are over-generalized or referred to apply too narrowlly.

12. *Harlan* says:
    October 6, 2017 at 12:18 am

    I’m putting out a different idea, going in the opposite direction– let’s rely even more on significance and further the dichotomization of evidence.
    https://arxiv.org/abs/1710.01771

    - *Anoneuoid* says:
      October 6, 2017 at 8:16 am

      Step 1- Calculate a (1 – alpha_1)% Confidence Interval for theta.

      Step 2- If this C.I. excludes theta, then declare a positive result. Otherwise, if theta is within the C.I., proceed to Step 3.

      Step 3- Calculate a (1 – 2*alpha_2)% Confidence Interval for

theta.

Step 4- If this C.I. is entirely within delta, declare a negative result. Otherwise, proceed to Step 5.

Step 5- Declare an inconclusive result. There is insufficient evidence to support any conclusion.

What you seem to be saying is a result can be insignificant either because the estimated difference from zero is very small or the uncertainty is very large (for arbitrary, yet exact, definitions of small and large). In the former case you want to call the results “negative”, in the latter you want to call them “inconclusive”.

If this new terminology is adopted you expect people will at least publish the subset of insignificant results that are “negative”, but still leave out those that are “inconclusive”. Currently both are sitting in file drawers, so you consider this an improvement.

Is that right? If so, I don’t think this proposal actually solves the real issues with NHST. Can you give a “real life” example of how this would be used and interpreted? For example:

“BioLab X is testing a new amyloid-beta clearing drug in mice to see if it may cure Alzheimer’s. To determine memory deficit, they measure how long it takes treated and control groups to gather all the food from a familiar maze. To determine amyloid-beta levels, they split these mice into high/low categories based on their olfactory habituation test performance (deficits in learning smells have been previously linked to Alzheimer’s disease).”

Include whatever sample sizes, effect sizes, as desired.

- *Harlan* says:
  October 8, 2017 at 2:17 pm

  Thanks for the feedback.
  “Currently both are sitting in file drawers, so you consider this an improvement. Is that right?”– Indeed that is right! Consider the proportion of all published studies that are actually true (Ioannidis, 2005). Including more “negative results” will up this proportion. If this policy provides extra incentive for researchers to use larger sample sizes, this proportion is increased further. While the improvement may still be somewhat small, these are achievable, impactful gains. Expecting a journal to publish all null results is a lot to ask- I think those file drawers are pretty damn full! So why not start will the “high-quality null results”.

  Example:
  “BioLab X is testing a new drug in mice to see if it may cure

Alzheimer’s. To determine memory deficit, they measure how long it takes treated and control groups to gather all the food from a familiar maze. Due to operational/ethical/budgetary reasons (let’s be honest, these are expensive mice!), the sample size is restricted to 32 per arm. This sample size provides ~88% power to detect a large effect (d=0.8) and ~50% power to detect a medium effect (d=0.5).

Additional clinical research would only be justified if the true effect were substantial (i.e. d>0.5). As such, determining with certainty that d is less than 0.5, would be an important finding. The equivalence margin is therefore = [-0.5, 0.5]. Under the assumption that the true effect size is 0, this sample size provides ~52% chance of reaching a “negative result”, ~43% chance of reaching an “inconclusive result”, and ~5% chance of a (false)-“positive result”, (alpha_1=0.05, alpha_2=0.10).

I agree that it’s not a perfect solution to all issues with NHST. However, I think publication policies like this one can and should be some part of the solution.

- *Anoneuoid* says:
  October 8, 2017 at 2:34 pm

  > memory deficit, they measure how long it takes treated and control groups to gather all the food from a familiar maze.

  The “effect” may be due to the treatment making the mice more/less hungry, faster/slower, stressed out by being put in the maze, aggressive (thus influencing how much food they get, or how the handler deals with them), their sense of smell/sight /hearing, etc. It may have nothing to do with memory, or memory may be only part of the story.

  To deal with this you need to eventually come up with a prediction of what the results should look like if one explanation is correct vs the others. Just knowing there is “an effect” is not helpful. I literally do not care if you detect an effect or not (I always assume there is one), this type of data/analysis gives me nothing to work with.

- *Keith O'Rourke* says:
  October 6, 2017 at 10:22 am

  So with what Anoneuoid wrote – all studies get statistically labeled as better, not too different or too uncertain to discern whether better, worse or not too different.

  I believe it is really hard to anticipate how journals, authors and review committees will react and evolve under such as system.

Additionally, you are disregarding some shared insights (revised proposed list below) especially item 4 as your step 5 arguable applies to any single isolated study analysis “There is insufficient evidence to support any conclusion”

1. A p_value is just one view of what to assess about an experiment/study – that being how consistent is the data with a specific bundle of assumptions (which includes the null hypothesis). Furthermore, what to make of such an assessment as being rare, is seldom obvious or clearly spelled out. For instance, if it is from the first study – this may suggest further studies are likely not wasteful. Whereas, if they can be usually brought about in repeated studies – this may support the effect being real (replicable). On the other hand, it might be more prudent to simply take it to suggest that estimation based on the bundle of assumptions (which now also includes the alternative hypothesis) may be completely misleading (i.e. the assumed model is just too wrong). At least, that is, if estimation is considered as an essential step in answering the real scientific question.

2. Consider p_values as continuous assessments and be wary of any thresholds it may or may not be under (or targeted alpha error levels).

3. Keep in mind that p_value assessments are based on the possibly questionable assumption of zero effect and zero systematic error as well as additional ancillary assumptions.

4. Realize that the real or penultimate inference considers the ensemble of studies (completed, ongoing and future), individual studies are just pieces in that, which only jointly allows the assessment of real uncertainty.

5. Be aware that informative prior (beyond the ensemble of studies) information, even if informally brought in as categorical qualifications (e.g. in large well done RCTs with large effects the assumption of zero systematic error is not problematic) maybe unavoidable – learning how to peer review priors so that they are not just seen personal opinion may also be unavoidable.

6. The above considerations must be highly motivated towards discerning what experiments suggest/support as well as quantifying the uncertainties in that, as all of them can be gamed for publication and career advantage. It seems the importance of this cannot be over-estimated nor the need to repeatedly mention it in teaching and writing about statistics.

7. All of this simply cannot be entrusted to single individuals or groups no matter how well meaning they attempt to be – bias and error are unavoidable and random audits may be the only way to overcome these.

8. ???

9. ???