Banning the hypothesis test

Ben Klemens

12 March 2015

In case you missed it, a psychology journal, Basic and Applied Social Psychology, has banned the use of hypothesis tests[1].

Much has already been said about this, very little in support. The ASA[2] points out that this approach may itself have negative effects and a committee is already working on a study. Civil statistician[3], a person who is true to his nom de blog and never says anything uncivil about anything, is very annoyed. Nature[4] defended the party line.

This is an opportunity for statisticians to remind the world of what these p-values mean, exactly, and when they can or can not be trusted. A hypothesis test provides a sense of crisp, pass-or-fail clarity, but this can be a bad thing in situations where there is far too much complexity for crisp anything. How can we get readers to take p-values with the grain of salt that they must be taken with?

I agree with the dissenters, in the sense that if I were the editor of this journal, this is not something I would have done. If nothing else, smart people find a way to route around censorship. As noted by some of the commenters above, if you can only provide standard errors, I can do the mental arithmetic to double them to get the approximate 95% confidence intervals. Banning the reporting of the sample size and variance of the data would effectively keep me from solving for the confidence interval, but I doubt even these editors would contemplate such a ban.

The editors claim that p-values are 'invalid'. In the very narrow sense, this is kinda crazy. The values are based on theorems that work like any other mathematical theorem: given the assumptions and the laws of mathematics that we generally all accept, the conclusion holds with certainty. But once we look further at the context, things aren't so bright-lined:

- A p-value is a statement about the likelihood of a given claim about model parameters or a statistic of the data using the probability distribution defined by the model itself. We do not know the true probability distribution of the statistic, and so have to resort to testing the model using using the model we are testing (entry #034).

- A test is in the context of some data gathering or experimental design. Psychology is not Physics, and small details in the experimental design, such as the methods of

[1] http://www.tandfonline.com/toc/hbas20/current#.VQHlNWq1W01

[2] http://community.amstat.org/blogs/ronald-wasserstein/2015/02/26/asa-comment-on-a-journals-ban-on-null-hypothesis-statistical-testing

[3] http://civilstat.com/2015/02/journal-bans-null-hypothesis-testing-and-confidence-intervals/

[4] http://www.nature.com/news/psychology-journal-bans-p-values-1.17001?WT.mc_id=TWT_NatureNews

blinding and scoring, matter immensely and are not necessarily agreed upon. In my experience as a reader, I am far more likely to dismiss a paper because a point of design made the paper too situation-specific or too fallible than because they reject the null with only 89% confidence.

• We are Bayesian when reading papers, in the sense that we come in with prior beliefs about whether a given fact about the world is true, and use the paper to update our belief. At the extreme, a paper on ESP that proves its existence with 99.9% confidence might marginally sway me into thinking something might be there, but in my mind I'd be arguing with the methods and it'll take a hundred such papers before I take it seriously. A paper finding that colorblind people process colors differently from typically-sighted people would get a *well, yeah* from me even if the hypothesis test finds 85% confidence, and in my mind I'd think about how the experiment could have been improved to get better results next time.

A corollary to this last bullet point is that the editors are also Bayesian, and are inclined to believe some theories more than others. One editor may dislike the theory of rational addiction, for example, and then what keeps the editor from desk rejecting any papers that support the theory? Having only qualitative information means one less check on biases like these.

The full set of bullet points show how a crisp p-value can be misleading, in terms of the modeling, of the experiment as a whole, and the manner in which readers digest the information. Assuming Normally-distributed data, the p-value can be derived from first principles, but the statement that the reader should reject the null with $1 - p\%$ probability requires accepting that nothing went wrong in the context that led up to that number. (By the way, $1 - p$ is the q-value, and I sometimes wish that people reported it instead.)

Psychological problems Psychology is not physics, where context can be controlled much more easily. To talk meta-context, a psychology study has so many challenges and confounders that the typical p-value in a psychology journal is a qualitatively different thing from a p-value in a biology or physics journal. A p-value can be read as a claim about the odds of getting the same result if you repeat the experiment, but defining what goes into a correct replication is itself a challenge the psych literature is grappling with in the present day. [But yes, there are high-quality, simple, reproducible psych experiments and badly designed physics experiments.]

Second, your revolution in the understanding of *drosophila* is going to be upstaged in the papers by the most piddling result from a psychology lab, every time. There are press agents, journalists, pop psychologists, and people selling books who have very little interest in the complexities of a study in context, and every interest in finding a study that 'proves' a given agendum, and they have a large population of readers who are happy to eat it up.

Maybe you recall the study that obesity is contagious[5], perhaps because you read about it in Time magazine[6]. With lower likelihood, you saw the follow-up studies that questioned whether the contagion effect was real, or could be explained away by

[5] `http://ucsdnews.ucsd.edu/archive/newsrel/soc/07-07ObesityIK-.asp`

[6] `http://content.time.com/time/health/article/0,8599,1646997,00.html`

the simple fact that similar people like to hang out together (homophily). Much to their credit, Slate[7] did a write-up of some of the contrary papers. Or maybe you saw the later study that found that obesity contagion is reasonably robust[8] to homophily effects.

I'm not going to wade into whether obesity patterns show contagion effects beyond homophily here, but am going to acknowledge that finding the answer is an imperfect process that can't be summarized by any single statistic. Meanwhile, the journalists looking for the biggest story for Time magazine aren't going to wade into the question either, but will be comfortable stopping at the first step.

So I think it's an interesting counterfactual to ask what the journalists and other one-step authors would do if a psychology journal didn't provide a simple yes-or-no and had to acknowledge that any one study is only good for updating our beliefs by a step.

I commend the editors of the BASP, for being bold and running that experiment. It doesn't take much time in the psychology literature to learn that our brains are always eager to jump on cognitive shortcuts, yet it is the job of a researcher to pave the long road. No, if I were editor I would never ban p-values—I've pointed out a few arguments against doing so above, and the links at the head of this column provide many more valid reasons—but these editors have taken a big step in a discussion that has to happen about how we can report statistical results in the social sciences in a manner that accommodates all the uncertainty that comes before we get to the point where we can assume a t distribution.

[7] http://www.slate.com/articles/health_and_science/science/2011/07/disconnected.html

[8] http://m.smr.sagepub.com/content/40/2/240.abstract