Five tips for avoiding P-value potholes

April 26 2016, by Hilda Bastian

The hunt for p-values less than 0.05 has left many of science's roadways riddled with potholes.

More than 50% of them in the biomedical literature are wrong, according to one reckoning, maybe 30% or more wherever it's used, according to another. It's high, but we can't know for sure how high, according to a third. It's definitely part of the reason more than half of studies in psychology can't be replicated. And the lowest estimate of wrong-ness I've seen is 14% in biomedical research – which is still pretty high.

The American Statistical Association (ASA) is trying to stem the tide of misuse and misinterpretation. They issued a statement on p-values this year. It's the first time they ever took a position on a statistical practice. They did so because, they said, it's an important cause of science's reproducibility crisis.

Perhaps the ASA's intervention will help stop the p-value's seemingly unstoppable advance in the sciences. In psychology, according to a study by Hubbard and Ryan in 2000, statistical hypothesis testing – the test that calculates statistical significance – was being used by around 17% of studies in major journals in the 1920s. It pretty much exploded in the 1940s and 1950s, spreading to 85% of studies by 1960, and passing 90% in the 1970s.

How did this get so out of hand? Hubbard and Ryan argue it's because was simple and appealing for researchers, and there was "widespread unawareness" of the test's limitations. There is no simple alternative to replace it with, and some argued for it fiercely. So it was easier to let it take over than fight it. Hubbard and Ryan call out "the failure of professional statisticians to effectively assist in debunking the appeal of these tests".

The ASA takes a similar line: the "bright line" of p-values at 0.05 is taught because the scientific community and journals use it so much. And they use it so much because that's what they're taught.

The result is a bumpy ride in the literature. Here's my choice of the top 5 things to keep in mind to avoid p-value potholes.

1. "Significant" in "statistically significant" doesn't mean "important."

You can have a statistically significant p-value of an utterly trivial difference – say, getting better from a week-long cold 10 minutes faster. You could call that "a statistically significant difference", but it's no reason to be impressed.

Back in Shakespeare's day, significance still didn't have the connotation of importance. "Signify" only referred to meaning something. And as Regina Nuzzo explains, that's all the developers of these tests meant in the 1920s, too: a p-value less than 0.05 just signified a result worth studying further.

2. A p-value is only a piece of a puzzle: it cannot prove whether a hypothesis is true or not.

This gets to the heart of misuse of p-values. The ASA statement could not be clearer on this:

P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

A statistical hypothesis test is measuring an actual result against a theoretical set of data. It doesn't know if the hypothesis it is "testing" is true or not: it just assumes that it is not true (the null hypothesis). And then it measures whether or not the result is far enough away from this theoretical null, to be worth more attention.

Steve Goodman says the most common misconception about a p-value of 0.05 is that "the null hypothesis has only a 5% chance of being true" [PDF]. You can't rest a case on it alone, that's for sure. It's hard to even nail exactly what the p-value's implications actually are – which doesn't mean that it it's always useless [PDF]. So where does that leave us?

There isn't a simple answer. The best ways to analyze data are specific to the situation. But in general, you need to be looking for several things:

Methodological quality of the research: No amount of statistics in the world can make up for a study that is the wrong design for the question you care about. What or who is included can skew the value of the results too, if you want to apply the knowledge to another situation.

Effect size: You need to understand exactly what is being measured and how big the apparent effect is to be able to get a result into perspective.

Understanding the uncertainty: If there's a p-value, you need to know exactly what it was, not only that it is under 0.05 – is it just under, or are there more zeros? (Or how much is over 0.05.) But even that's not enough. In fact, you don't really need the p-value. You need better ways to understand the uncertainty of the estimate: and that means standard deviations, margin of error, or confidence/credible intervals.

Certainty doesn't come from a one-off – and especially not from a surprising one. This is why we need systematic reviews and meta-analysis.

Some argue, on the other hand, that "the answer" is simply to have a more stringent level of statistical significance than 0.05 (which is the 95% mark, or 1 in 20). Particle physicists have taken this the furthest, expecting to get a 5 sigma result (and at least twice) before being sure. In p-value terms, that would be is 0.0000003, or 1 in 3.5 million.

Very high levels are going to be unachievable for many kinds of research. The limits to the feasible number you can have in a study for something as complicated as the effects of a drug on human beings wouldn't come close to enough certainty anyway.

That's not the only option, though. Bayesian statistics offer more options, with the ability to incorporate what's known about the probability of a hypothesis being true into analysis.

3. More is not necessarily better: more questions or bigger datasets increase the chances of p-value potholes.

The more common an event is, the more likely it is to reach p.

The more tests are run on a data set, the higher the risk of p-value false alarms gets. There are tests to try to account for this.

An alternative here is for researchers to use the false discovery rate (FDR), which is one way of trying to achieve what people think the test for statistical significance does. That said, Andrew Gelman described the FDR as just "trying to make the Bayesian omelette without breaking the eggs" [PDF].

As if this whole road isn't already hard enough to negotiate, an awful lot of researchers are putting a lot of effort into digging potholes for the rest of us. It's so common, that many don't even realize what they are doing is wrong.

It's called p-hacking or data-dredging: hunting for p-values here. Christie Aschwanden and FiveThirtyEight have provided a great interactive tool for you to see how you can p-hack your glory, too.

4. A p-value higher than 0.05 could be an absence of evidence – not evidence of absence.

This one is tricky terrain, too. People often choose the statistical hypothesis test as their main analysis – but then want to have their cake and eat it too if the result isn't statistically significant. Matthew Hankins nails this practice of trying to disguise a non-statistically different result "as something more interesting" here.

On the other hand, if it's important or possible that something is making a difference, you need something stronger than non-significance in a single study too (especially if it's a small one).

5. Some potholes are deliberately hidden: shining the light only on p's less than 0.05.

This is a form of bias called selective outcome reporting. You can see it often in abstracts, where p-values

In the biomedical literature, the number of studies reporting p-values in the abstract alone rose from 7% in 1990 to over 15% in 2014, almost always claiming at least one p-value below 0.05 – and that's not a good sign.

This is not always easy to spot, as researchers sometimes go to considerable lengths to hide it. Clinical trial hypotheses and planned outcome assessment are meant to be published before the trial is done and analyzed to prevent this – as well as make it obvious when a trial's results are not published at all. (More on this at All Trials.) But even that isn't enough to end the practice of biased selective reporting.

Ben Goldacre and colleagues from Oxford's Centre for Evidence-Based

Medicine are systematically studying and calling out outcome-switching in clinical trials. You can read more about this at COMPARE.

Pre-registering research is spreading from trials into other fields as well: read more about the campaign for registered reports. You can see the impact of unpublished negative results in psychology in an example in Neuroskeptic's post this week.

In many ways, once you get the hang of it, the types of potholes that are out in plain sight are easier to handle – just like potholes in real life.

But especially because so many are hidden, it's better to always go slowly and look for more solid roadway.

Wherever there are p-values, there can always be potholes.

More information: Ronald L. Wasserstein et al. The ASA's statement on p-values: context, process, and purpose, *The American Statistician* (2016). DOI: 10.1080/00031305.2016.1154108

This story is republished courtesy of PLOS Blogs: blogs.plos.org.

Provided by Public Library of Science

Citation: Five tips for avoiding P-value potholes (2016, April 26) retrieved 2 May 2024 from https://phys.org/news/2016-04-p-value-potholes.html