Confessions of a p-value lover

AUTHOR

Hieab H.H. Adams, MD PhD[1,2,3]

AFFILIATIONS

[1] Department of Epidemiology, Erasmus University Medical Center Rotterdam, the Netherlands

[2.] Department of Radiology and Nuclear Medicine, Erasmus University Medical Center Rotterdam, the Netherlands

MAIN TEXT

As an epidemiologist, I was trained to hate the p-value. During my scientific career, however, I have so far published over a trillion p-values. They have helped me interpret findings, determine which scientific leads to follow-up on, and which results are likely not worth the time and effort. Now, it is time for me to speak up for the statistical underdog that I have learned to love.

The hatred directed at p-values has more recently shifted to null hypothesis significance testing (NHST), the common practice of determining whether an association is statistically significant or not compared to the null hypothesis (typically, the absence of an effect). Many commentaries have been published bashing NHST, including the call of Amrhein et al. to ban it entirely. The logic behind it is as follows: the dichotomous nature of NHST has been misused to falsely claim presence or absence of effects, so therefore the whole concept of NHST should disappear from science. Specifically, they claim that 1) dichotomization is a cognitive disorder ("dichotomania"), 2) the null hypothesis is just an arbitrary choice, and 3) banning NHST is the solution to its misuse.

Dichotomania versus decisophobia

Dichotomization is natural and uncircumventable for decision making. Should a drug be taken further for development, is a putative risk factor worthy of follow-up studies, and does a certain

genetic variant warrant replication? This is not dichotomania. These are yes/no decisions for which there is no middle ground; the end result can be one of only two options. The authors call for embracing uncertainty, but fail to see that research is done to achieve exactly the opposite: we want to be as informed as possible when making yes/no scientific and policy decisions. Decisions can turn out to be wrong, because there will always remain uncertainty, but this should not paralyze research through decisophobia.

The null: more than just a number

Singling out the null makes perfect sense, since this is the default option for most studies. Most drugs in development won't have an effect on your particular outcome. Most putative risk factors will turn out not to be one. Most genetic variants do not influence your trait of interest. NHST is a crucial step in the process to arrive at certainty, as one of several criteria to decide whether a scientific lead is worth pursuing or not. While the scaremongering around NHST suggests so, in fact no healthcare policy has ever been based on a mere glance at whether $p<0.05$. Replication, cost-benefit analyses, and synthesis of information from various sources has been key and will remain so in the future. And NHST is a part of this.

Scapegoating of NHST

The authors argue that NHST should be banned to solve these problems, except for "specialized" situations – a caveat that will immediately make a careful reader question whether NHST is truly the cause of the problem. If some situations warrant NHST, then clearly NHST should not be blindly banned.

The authors mention many problems with NHST. First, there is not a big difference between drug A with $p=0.051$ and drug B with $p=0.049$, so why make the distinction? Obviously, comparing drug A and B head-to-head would not show a large difference, and any person could intuitively – and statistically – come to this conclusion. It is important to note that using NHST p-values, which test drugs against the null, to then compare drugs to one another is a misuse of NHST, since a separate statistical test would be required to test the *difference* between drugs A and B – and this would lead you to the correct conclusion of there being no difference between the two drugs. But putting this aside, are p-values falling just on opposite sides of the

significance threshold an argument for abandoning NHST altogether? Is a student who scored just below the passing grade (“p=0.051”) really much worse than another student scoring just above it (“p=0.049”)? Clearly not, that’s why grades are given (“p-values”), but this is not a valid argument for banning pass/fail exams (“NHST”). There needs to be a threshold for determining who needs to go back and study more, and who can continue.

A second problem the authors mention is conscious and subconscious P-hacking that researchers may embark upon in their journey to get results published. Although this is indeed a scientific problem, here too it is conflated with NHST: if students cheat on exams (“P-hacking”), would we then banish exams from schools, or try to enforce rules to prevent them from cheating (“pre-registration of studies”)? Clinical trials have definitively shown that the latter is the preferred solution.

The third problem the authors raised is the nonsensical proofs of the null, but does this not merely reflect misguided anger that should be directed at underpowered studies? Imagine a randomized controlled trial of a new drug that was performed in 1 million individuals, showing an odds ratio of 1.001 and p-value of 0.90. I will stick my neck out and say there is no treatment effect. It is possible to argue that I haven’t ruled out an effect of 1.00001, and you would be right, but this is nothing more than semantics since such a small effect is irrelevant. Now, when heavily underpowered studies publish null findings, this is an issue. But the solution is simply to perform well-powered studies, or to be aware that your study is underpowered and cannot prove the null; the solution is not to ban NHST. Here too, the analogy with exams fits perfectly. Imagine a one-question exam, upon which you base whether someone passes or not. Clearly, answering this question would hardly be sufficient to judge a student’s knowledge, but again, would this be an argument for banning exams? Or does this simply mean you should perform better powered exams with more questions?

P-values and NHST have gotten too much criticism for the faults committed by researchers who do not appropriately use them. Inadvertently, the authors have themselves stumbled upon yet another misuse of p-values and NHST: as a scapegoat for statistical malpractice.