Problem with p values: why p values do not tell you if your treatment is likely to work

Robert Price ,[1] Rob Bethune ,[2,3] Lisa Massey[2]

INTRODUCTION

Medicine has made remarkable progress within the lifetime of the oldest members of our society. Evidence from trials has come to replace expert opinion as the arbiter of treatment effectiveness. Following the work of Fisher (1925) and Neyman and Pearson (1933), null hypothesis significance testing (NHST) and the p value have become the cornerstones of clinical research. The attractions of a rule-based 'algorithm' approach are that it is easy to implement, permits binary decisions to be made and makes it simple for investigators, editors, readers and funding bodies to count or discount the work. But does that make it reliable? Even in the early days, this approach was controversial. Worse still, as these ideas became broadly adopted, fundamental misinterpretations were embedded in the literature and practice of biomedical research.[1]

Fisher originally proposed using the exact p value for a single trial as an indication of the credibility of the null hypothesis when considered together with all the available evidence.[2] It is worth noting that the null hypothesis does not necessarily mean no difference between groups, this is the nil null hypothesis.[3] Rather it is the hypothesis we aim to nullify by experiment, which can provide more powerful evidence if it includes a quantitative prediction of the expected difference. In the following decade, Neyman and Pearson developed the concepts of the alternative hypothesis, α, power, β, type I error and type II error, as a formal decision-making procedure to automate industrial production quality control. This method requires multiple samples and repeated analyses to control the long-term error rates, only using the p value as a binary decision-making threshold. Despite the inherent conflict in these two approaches, they have been fused into NHST for single trials, where a p value threshold is used to accept or reject the null hypothesis.[2]

The problem with p values and their misinterpretation

For several decades, there has been a failure by many authors to realise that all these probabilities (p, α, β, type I error, type II error) are conditional.[4] The order of terms matters in conditional probabilities, and they must be used according to a simple set of mathematical rules or their meanings are changed. However, the shorthand notation adopted in biomedical statistics assumes that everyone knows this. This makes it easy for authors to offer their own mistaken interpretations in the belief that they are providing clarity.

For example, the standard definition of the p value is 'the probability of having observed our data (or something more extreme) *given* the null hypothesis is true'.[5] In NHST, the p value is a conditional probability, conditioned on the null hypothesis always being true. Unfortunately, authors often replace *due to* with *given*, without realising that this makes it the probability of a completely different event. We calculate p values because it is easy, not because they are the probability we really want. We usually want the more intuitive probability; our hypothesis is true given the data. It is tempting to claim that this is the same as the p value and that if $p=0.05$ there is a 1 in 20 probability that the data arose by chance alone (the null hypothesis). This is incorrect, because the p value is conditioned on the null hypothesis being true, it therefore cannot be used to give a probability that the null hypothesis may be true or indeed false.[4 6]

This error is an example of the fallacy of the transposed conditional or the prosecutor's fallacy, so called because it may be used by the prosecution to exaggerate the weight of evidence against a defendant in a criminal trial. Similarly, this common misinterpretation of the p value exaggerates the weight of evidence against the null hypothesis. What we actually need is the false discovery rate (FDR), which is the proportion of reported discoveries that are false positives. Some authors prefer the term false positive risk to emphasise that this is the risk that, having observed a significant p value from a single experiment, it is a false positive.[7] The FDR can be calculated from the power, type I error rate and an estimate of the prevalence of real effects among the many ideas we may test.[8] It is crucial to note that the type I error rate is the long-term probability of a false positive for a single experiment repeated with exact replication. It is not the same as the FDR that applies to a single run of each experiment. Even when experiments are perfectly designed and executed without bias, missing data or multiple statistical testing, the FDR in NHST using a $p<0.05$ threshold has been estimated to be in the range 30%–60%, depending on the field of research and the power of the study.[9–12]

Measles and spots

Another example may help to clarify the problem of the transposed conditional. Consider the difference between the probability of having spots *given* you have measles, which is about 1, and the probability that the spots you have are *caused by* measles. In the latter, spots have become certain, the *given*, while measles has become an uncertain cause of the spots. So, by changing the word *given* to *caused by* we have inadvertently inverted the conditional probability and now have a probability which is much less than 1. There are many other causes of spots than measles, but if you have measles you are very likely to have spots. This example shows the danger of changing *given* to *caused by* which is what many researchers do.

In the last decade, attempts have been made in social science, psychology, medicine and pharmacology to replicate important experiments with very disappointing results. This has led to increasing concern about the validity of much scientific work. Attention has inevitably turned to the statistical techniques and our interpretation of them.[2 11]

In biomedical research, we often do not know the effect size as it is frequently small, sampling is difficult and variance is often large and poorly known. Crucially, we only do the experiment once and have only a one-point estimate of the p value. Additionally, our theories are only weakly predictive and do not generate precise numerical quantities that can be checked in quantitative experiments as is possible

[1] Anaesthetics, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK
[2] Surgery, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK
[3] South West Academic Health Science Network, Exeter, UK

Correspondence to Dr Robert Price, Anaesthetics, Exeter, Devon, UK; robert.price1@nhs.net

Table 1 Fraction of correct definitions of the p value found in medical textbooks by library subject classification

Subject area	Fraction of correct and unambiguous definitions for the p value
Evidence-based medicine	3/15
Exam revision	3/17
Research	8/13
Statistics	4/4
Total overall subjects	18/49

in the physical sciences. NHST does not perform well under these circumstances. Added to this are the frequent misunderstandings about the meaning of the p value. To reiterate, the p value is strictly the probability of obtaining data as extreme or more so, *given the null hypothesis is true*.

To investigate how common this misunderstanding is in medical textbooks, we audited the definition of the term p value in books held in the medical library of the University of Exeter Medical School at the Royal Devon and Exeter Foundation Trust (table 1). We included texts under the subject groups; statistics, research, evidence-based medicine and examination revision books (see online supplementary data file 1).

The most common error was to claim that the p value is the probability that the data was generated by chance alone. This definition has been frequently and vigorously refuted in the statistics literature but is very persistent in medical textbooks and journals.[4–6]

Beyond misinterpretations of p values, there are also widespread problems with multiple testing, sometimes inadvertant, which grossly inflates the proportion of false positive results. This is known as 'p-hacking' or 'data dredging' and allows researchers to selectively report spurious results as significant.[13]

DISCUSSION

Given the scale of this problem, what should be done? There are two main areas to address. First of all, we need to teach the correct statistical interpretation of NHST because of the huge volume of trials already published. This has already been attempted without success for at least the last 40 years. Second, we need to move to statistical models that are better suited to current research problems and address some of the shortcomings of NHST. Both of these issues will need the entire research community to change. Research funding agencies, universities and journals must recognise that they have played a key role in promoting a culture where the p value has had primacy over reason. Researchers must resist redefining statistical quantities to suit their own arguments, because it is mathematically wrong to do so. Possible statistical approaches include the use of effect size estimation accompanied by 95% CIs, a reduction in the p value threshold for significance,[14] rigorous trial pre-registration and the mandatory publication of negative studies. However, CIs are also misinterpreted and often used to produce identical results to a $p<0.05$ significance test. Reducing the p value threshold will reduce the false positive rate at the cost of an increase in the false negative rate, particularly in under-powered studies. A more intuitive methodology to use is Bayesian statistics; this calculates the probability that a hypothesis is true given the data; this is mostly what researches and readers actually assume the p value to be. The mathematical and logical basis of Bayes theory has been established by Cox,[15] Jaynes[16] and others.[17] However, it is not without its problems and if misapplied can be just as misleading as NHST. It is possible to directly compare the probabilities of different hypotheses being true and updating knowledge as new data arrives. Because Bayesians use probability distributions, they can easily calculate a sensible measure of uncertainty on a parameter, the Bayesian credible interval. This is much more meaningful than the frequentist CI, which is again based on performance over many repetitions but is measured only once. In the past, the two major objections to Bayesian methods have been the difficulty of calculating intractable integrals and the use of prior probabilities. The first is a practical point, while the second is philosophical. On the first point computing power, Markov Chain Monte Carlo algorithms, Gibbs samplers and open-source Bayesian statistics software have made numerical solutions to the integrals required for Bayesian data analysis relatively easy to do, which was not the case in the past. On the second point, we already know that prior probabilities influence NHST FDRs, even though this is ignored in the data analysis.[8 9 12] It would be much better to state the priors explicitly and test the analysis for sensitivity to different assumptions, as can be done in Bayesian data analysis.

We urge authors and editors to demote the prominence of p values in journal articles, have the actual null hypothesis formally stated at the beginning of the article and promote the use of the more intuitive (but harder to calculate) Bayesian statistics.

To the point

- The p value in null hypothesis significance testing is conditioned on the null hypothesis being true.
- This means that a p value of 0.05 *does not mean* that the probability our data arose by chance alone is 1 in 20.
- In fact, the chance of us mistakenly rejecting the null hypothesis and concluding we have a successful treatment is more in the region of 30%–60%.
- Scientific journals and textbooks need to be explicit on how p values are used and defined.
- Use of the more intuitive Bayesian statistics should become more widespread.

Correction notice This article has been corrected since it was published Online First. This paper is now Open Access.

Contributors RP and RB conceived the idea of writing the article. RP was the primary author and conducted the audit; RB and LM contributed to drafting and revising the article, in particular helping to clarify the text and make it suitable for a wider medical audience.

The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/postgradmedj-2019-137079).

To cite Price R, Bethune R, Massey L. *Postgrad Med J* 2020;**96**:1–3.

Received 28 August 2019
Revised 14 October 2019
Accepted 17 October 2019
Published Online First 29 October 2019

Postgrad Med J 2020;**96**:1–3.
doi:10.1136/postgradmedj-2019-137079

ORCID iDs

Robert Price http://orcid.org/0000-0002-1933-8728
Rob Bethune http://orcid.org/0000-0002-1855-0639

REFERENCES

1 Westover MB, Westover KD, Bianchi MT. Significance testing as perverse probabilistic Reasoning. *BMC Med* 2011;9:20.
2 Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. *Front Hum Neurosci* 2017;11:390.
3 Cohen J. The earth is round (P<0.05). *American Psychologist* 1994;49:997–1003.
4 Goodman S. A dirty dozen: twelve P-Value misconceptions. *Semin Hematol* 2008;45:135–40.
5 Altman DG. *Practical statistics for medical research 1st edition*. London: Chapman and Hall, 1991: 167–71.
6 Wasserstein RL, Lazar NA. The ASA Statement on *p* -Values: Context, Process, and Purpose. *Am Stat* 2016;70:129–33.
7 Colquhoun D. The reproducibility of research and the misinterpretation of *p* -values. *R. Soc. open sci.* 2017;4.
8 Vidgen B, Yasseri T. P-Values: misunderstood and misused. *Frontiers in Physics* 2016;4:1–5.
9 Kirkwood BR, Sterne JAC. *Essential medical statistics*. 2nd edition. Massachusetts: Blackwell Science Ltd, 2003: 426–8.
10 Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005;2:e124–701.
11 Begley CG, Ioannidis JPA. Reproducibility in science. improving the standard for basic and preclinical research. *Circulation Research* 2015;116:116–26.
12 Nuzzo R. Statistical errors. *Nature* 2014;506:150–2.
13 Davey Smith G, Ebrahim S. Data dredging, bias, or confounding. they can all get you into the BMJ and the Friday papers. *BMJ* 2002;325:1437–8.
14 Benjamin DJ, Berger JO, Johannesson M, *et al*. Redefine statistical significance. *Nat Hum Behav* 2018;2:6–10.
15 Cox RT. Probability, frequency and reasonable expectation. *Am J Phys* 1946;14:1–13.
16 Jaynes ET, Theory P. *Probability theory. The logic of science*. Cambridge University Press, 2013.
17 Terinin A, Draper D. Cox's Theorem and Jaynesian Interpretation of Probability 2017. arXiv:1507.06597v2 [math.ST].