■ EDITORIAL

Statistical significance and p-values

GUIDELINES FOR USE AND REPORTING

N. Parsons,
R. Carey-Smith,
M. Dritsaki,
X. Griffin,
D. Metcalfe,
D. Perry,
D. Stengel,
M. Costa

From Warwick Medical School, University of Warwick, Coventry, United Kingdom

Correspondence should be sent to N. Parsons; email: nick.parsons@warwick.ac.uk

doi:10.1302/0301-620X.101B10.BJJ-2019-0890 $2.00

Bone Joint J
2019;101-B:1179–1183.

In recent years, there has been much debate on the merits and pitfalls of reporting p-values, and more generally on the hypothesis-testing framework that supports published research.[1-10] The unrest reflects longstanding issues at the root of statistics and a response to widespread concerns about the lack of reproducibility of many scientific findings.[11-14] How should those of us actively engaged in clinical research react to all this? A recent editorial in *Nature* suggests that we should go as far as to retire statistical significance altogether.[15] The Research Methods Group and the Editorial Board of *The Bone & Joint Journal* here discuss the issues and the expectations of authors submitting to our Journal.

The issues are deceptively simple. The methodology that we now refer to as 'null hypothesis significance testing' originated in the early 20th century with the ideas of Fisher and Neyman and Pearson.[16,17] It has become a rather inconsistent mixture of ideas from both Fisherian and frequentist (Neyman–Pearson) approaches, which were much disputed between the parties at the time and for many subsequent years, on philosophical grounds.[18,19]

The cornerstone of null hypothesis significance testing is the p-value, which is the probability, under the null hypothesis (i.e. no effect or no difference), of obtaining a result as extreme or more extreme than that observed. It could be described as an objective measure of the discrepancy between the observed data and the null hypothesis. Consider a study that aims to compare two groups. The null hypothesis, which we state at the start of the study, might be that there is no difference in the means between the groups, measured as the outcome of the study (test statistic). Figure 1 shows the probability density curve of every possible outcome of the study, under the null hypothesis, H_0 (i.e. assuming the null hypothesis is true).

The area under the curve in Figure 1 is such that it adds up to one; therefore, every possible outcome must lie somewhere in the distribution. The larger the value of the probability density, the more likely the observation. Under the null hypothesis, the most likely outcome is zero (no difference between groups), and the least likely outcomes are those at the extreme ends of the curve.

A type I or false positive error rate (α), which is set at the start of the study, is the probability at which a result is declared to be significant, the shaded area of the curve to the right of the critical value (Z_α). This is often called the rejection region (the null hypothesis is rejected if the test statistic falls in this region) and is conventionally set at 0.05 (5% level). The p-value is the probability of getting the observed test statistic (Y), or more extreme, from the study data if the null hypothesis were true (i.e. the dark shaded tail error of the curve). If the p-value is less than α, the result is declared significant and the null hypothesis is rejected at level α.

In response to ongoing debate between and among scientists and statisticians, the Board of Directors of the American Statistical Association (ASA) published the following six principles, in 2016, accompanied by statements from leading statisticians, to help nonstatisticians understand statistical significance and p-values:[20,21] 1) "p-values can indicate how incompatible the data are with a specified statistical model"; 2) "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"; 3) "scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold"; 4) "proper inference requires full reporting and transparency"; 5) "a p-value, or statistical significance, does not measure the size of an effect or the importance of a result"; and 6) "by itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis".[20]

Some of these statements follow immediately from the definition of the p-value, while others are part of the bedrock of statistical teaching. It is worth emphasizing the importance of point 3 – "scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold"[20] – and recalling that the p-value was by derivation and definition proposed as an informal index to quantify the discrepancy between observed data and the null hypothesis. It was not intended to be used in place of sound scientific reasoning. The ASA statement concludes with an elegant sentence summarizing this, which we would all do well to remember: "no single index should substitute for scientific reasoning".[20] However, despite highlighting numerous issues with misinterpretation and misuse of