Editorial

Are the fallacies of the P value finally ended?

Luca Bertolaccini, Andrea Viti, Alberto Terzi

Thoracic Surgery Unit, Sacro Cuore Don Calabria Research Hospital—Cancer Care Center, Negrar Verona, Italy

Correspondence to: Luca Bertolaccini, MD, PhD, FCCP. Thoracic Surgery Unit, Sacro Cuore Don Calabria Research Hospital—Cancer Care Center, Via Don Angelo Sempreboni 5, 37024 Negrar Verona, Italy. Email: luca.bertolaccini@gmail.com.

Submitted Mar 15, 2016. Accepted for publication Mar 31, 2016.
doi: 10.21037/jtd.2016.04.48
View this article at: http://dx.doi.org/10.21037/jtd.2016.04.48

"*Fallax vulpes illis mendacibus verbis corvum cum facilitate de eius pulchritudine fallit…* "

As well acknowledged in this fable of Aesop (shooting then from Phaedrus, and later by Jean de la Fontaine), the fallacy of the fox is like the fallacy of the P value, the ubiquitous, misunderstood, misinterpreted, and miscalculated index of research. Ronald Aylmer Fisher, who introduced the P value as a formal research tool, could not explain exactly its inferential meaning. The classical definition of the P value is the probability of the observed result, plus more extreme results if the null hypothesis (H_0) were true, or in a more formal notation:

$$Prob(X \geq x \mid H_0)$$

where X is a random variable corresponding to some way of summarizing data (such as a mean or proportion), and x is the observed value of that summary in the current data. Because the P value is not part of any formal calculus of inference, its meaning is elusive and difficult to interpret correctly (1). When a researcher performs a hypothesis test, a P value could help to determine the significance of results. Hypothesis tests are used to test the validity of a claim made about a population. This assertion is called the H_0. The alternative hypothesis is the one you would believe if the H_0 were concluded to be untrue. All hypotheses use a P value to weigh the strength of the evidence (what the data are telling you about the population). Nevertheless, in the Fisher's paper, the P value was to be used as an algebraic guide to the strength of the evidence against the H_0, but there was no mention of error rates or the rejection of the hypothesis (2).

We well know that the P value ranges between zero and one and is widely interpreted in three-way: a P value ≤0.05 indicates strong evidence against the H_0 (the H_0 could be rejected); a P value >0.05 indicates weak as evidence against the H_0 (the H_0 could be failed to reject); P values close to the cut-off are marginal.

In an example modified from Nuzzo (3), suppose a city courier claims their delivery times are 1 hour or less on average but you think it is more than that. You conduct a hypothesis test because you believe the H_0 that the mean delivery time is 1 hour max is incorrect. Your alternative hypothesis (H_a) is that the mean time is greater than 1 hour. You randomly sample some delivery times and run the data through the hypothesis test, and your P value turns out to be 0.001, which is much less than 0.05. In real terms, there is a probability of 0.001 that you will mistakenly reject the city courier's claim that their delivery time is less than or equal to 1 hour. Since typically we are willing to reject the H_0 when this probability is less than 0.05, you conclude that the courier is wrong; their delivery times are in fact more than 1 hour on average (3).

Nevertheless, P values have always had critics. When Fisher introduced the P value, he did not mean it to be a definitive test, but he intended it simply as an informal way to judge whether the evidence was significant in the old-fashioned sense (3).

In the biostatistics or in the medical statistics fields, there are three common mistakes in the interpretation of P value: potentially medically important differences observed in small studies with P value <0.05 are denoted as not significant and ignored; all P value <0.05 findings assumed to results from real treat effects, and all P value <0.05 findings considered to be of medical importance (4).

Therefore, although a significant result in a large study may sometimes not be clinically important, a far greater problem arises from misinterpretation of non-significant findings. Randomized controlled clinical trials that do not show a significant difference between the treatments being

compared is often called negative. This term wrongly implies that the study has shown that there is no difference, whereas usually all that has been demonstrated is an absence of evidence of a difference (5).

Recently, in a survey of academic scientist about the interpretations of P values, many researchers do not know how to interpret it correctly, indicating that scientists are not immune to erroneous interpretations. Problems in understanding the P value influence the conclusions that professionals draw from their data and jeopardize the quality of the results of psychological research. The value of the evidence depends on the quality of the statistical analyzes and their interpretation (6).

There is a long troublesome circularity in the use and misuse of P value and it is a process that feeds on itself: we teach it because it is what we do; we do it because it is what we teach. Since the P value is commonly misused and misinterpreted, this has led to some scientific journals discouraging the use of P values, and some scientists and statisticians recommending their abandonment. Fortunately, on March 2016, the American Statistical Association (ASA) published the statement on statistical significance and P values. This report does not seek to resolve all the issues relating to statistical practice, but articulates in non-technical terms few selected principles. A brief, albeit not exhaustive, synthesis of the principles could be as follow. P values can indicate how the data are incompatible with a specified statistical model, and this incompatibility can be interpreted providing evidence against the H_0. P values do not measure the probability that the studied hypothesis is correct, or the likelihood that the data were produced by random chance alone. Scientific conclusions should not be based only on whether a P value passes a specific threshold. Proper inference requires full reporting and transparency. A P value, or statistical significance, does not measure the size of an effect or the importance of a result (7).

In conclusion, the concept of a P value is not simple, and any statement associated with it must be considered cautiously. It is also important to reemphasize that, if the result of a hypothesis test is that difference was not statistically significant, it does not mean that there is no difference between the treatment groups in the target population. Moreover, we hope that the ASA statement may offer a basis to improve the use of statistical inferences in biostatistics.

Cite this article as: Bertolaccini L, Viti A, Terzi A. Are the fallacies of the P value finally ended? J Thorac Dis 2016;8(6):1067-1068. doi: 10.21037/jtd.2016.04.48

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008;45:135-40.
2. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993;137:485-96; discussion 497-501.
3. Nuzzo R. Scientific method: statistical errors. Nature 2014;506:150-2.
4. Kirkwood B, Sterne J. Essential Medical Statistics. 2nd Edition. Wiley-Blackwell, 2003.
5. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
6. Badenes-Ribera L, Frías-Navarro D, Monterde-i-Bort H, et al. Interpretation of the p value: A national survey study in academic psychologists from Spain. Psicothema 2015;27:290-5.
7. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016. Available online: http://dx.doi.org/10.1080/00031305.2016.1154108

