Three inferential questions, two types of P-value

Michael J. Lew

February 29, 2016

Many users of P-values don't understand why their standard practices are derided as 'P-value hacking' and when and why they should abstain from 'cherry picking'. Such confusion is predictable as statistical methods and behaviours appropriate in some circumstances are inappropriate, dangerous, and verging on dishonest in others. The ASA statement doesn't describe such circumstances and so I will introduce some of them here. This commentary should be read more as an extension of the statement rather than a commentary on its contents, with which I agree with few reservations.

Richard Royall noted that there are three types of inferential question that can be answered with the help of statistical methods (Royall, 1997).

1. What do these data say?
2. What should I believe now that I have these data?
3. What should I do or decide now that I have these data?

Those questions might seem so obvious that it is superfluous to mention them. However they are so rarely mentioned as to remain novel, and they provide a good scaffold for understanding the roles of P-values in scientific inference. Misuse of P-values often involves an implicit assumption that they provide answers to all three questions, but P-values cannot, by themselves, tell an investigator what to believe or what to decide.

P-values answer the first question by being an index to the evidential meaning of the data within a statistical model. As noted in the ASA statement, P-values are anchored to a single hypothetical value of the parameter of interest, the 'null hypothesis', within a particular statistical model, so they are not always the *best* way to answer the first question. A likelihood

function gives a richer depiction of the evidence in the data about parameter values than does a P-value from the same statistical model, as the likelihood function allows comparison of the evidential support for all values of the parameter of interest. Nevertheless, P-values are a useable and defensible answer to the question of what the data say—at least when they are accompanied by adequate demonstration of the observed effect size and relevant experimental and analytical details.

An answer to "what should I believe now that I have these data?" should meld what the data say with what was known or believed beforehand. Bayesian methods formally answer that question with a prior probability distribution to represent the pre-data information or belief. The question "what should I do or decide now that I have these data?" requires consideration of what the data say in conjunction with the benefits and costs of correct and incorrect decisions or actions. In other words, a decision process requires a loss function in addition to the data. The classical Neyman & Pearson hypothesis test is probably the most widely used decision theory approach, and its loss function is built into the designed balance between false positive and false negative error rates, α and β. For example, if α is set to a smaller value than β in the pre-data study design then the loss function reflects a greater cost of false positive than false negative errors.

Researchers should be aware of the distinction between the questions answered by the exact P-value and the conventionally dichotomized hypothesis test result. To reflect what the data say, P-values have to be treated in a non-dichotomous manner, as the evidence is not simply present or absent, but is graded. Converting P-values into 'significant' and 'not significant' can be appropriate when answering the third question with a Neyman & Pearson hypothesis test procedure, as long as a pre-study power analysis for sample size determination has been done. Unfortunately, such a power analysis is rarely performed or reported in publications of basic biomedical science (Strasak et al., 2007), and if you dichotomise a P-value by taking P $\leqslant 0.05$ as 'significant' without having designed the loss function, then you are using the mechanical "bright line" rule deprecated in the ASA statement. The absence of a loss function does not preclude exact P-values from serving as an answer to the first question, and a dichotomizing hypothesis test is not the only basis for a scientific conclusion (Lew, 2012).

The choice of analytical procedures should be informed by the nature of the study because if you restrict your attention to answering the first question you can identify the areas where cherries are most numerous and ripe without picking them. Data from preliminary or exploratory studies intended to

determine fruitful directions of enquiry can be interrogated repeatedly and intensively and results can sensibly be assessed and communicated on the basis of observed P-values, even if the study involves many comparisons, even if the comparisons are unplanned, and even if the sampling rules were ill-defined or flexible. No 'correction' of those P-values for multiplicity of comparisons is necessary—or desirable—because what the data say about one hypothetical effect is not influenced by whether the analyst sees what the data say about another hypothetical effect. In contrast, if those same P-values were used with hypothesis testing procedures to provide the basis for decisions regarding hypotheses then claims of 'cherry picking' and 'P-hacking' would usually be correct. A pre-study power analysis is required, and all of the comparisons to be made must be included for the loss function to be correctly calibrated. Thus P-values used within a hypothesis test decision procedure often need adjustment to take the actual experimental design into account lest the statistical support for decisions or actions is weaker than claimed or implied because of a higher than reported risk of false positive outcomes. Exploratory studies should not be misrepresented as planned studies yielding answers to the third question.

There are two types of P-values: P-values that show what the data say, and P-values to be used in decision processes. Analytical manoeuvres that should be derided as 'P-hacking' and 'cherry picking' in a planned study are perfectly appropriate in the setting of a preliminary study. The rights and wrongs of using P-values are context and purpose dependent.

References

Michael J Lew. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. *British journal of pharmacology*, 166(5):1559–1567, June 2012.

Richard M Royall. *Statistical Evidence: a Likelihood Paradigm*, volume 71 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997.

A Strasak, Q Zaman, G Marinell, and K Pfeiffer. The Use of Statistics in Medical Research: A Comparison of The New England Journal of Medicine and Nature Medicine. *The American Statistician*, 61(1):47–55, 2007.