P-values Are Not What They're Cracked Up to Be

The ASA is to be congratulated for its "Statement on Statistical Significance and P-values." Much has been written about p-values in the last 50 years. Many authors have been critical, with pointed warnings about misunderstanding and misinterpreting these strange but ubiquitous beasts. The cumulative impact of such criticisms on statistical practice and on empirical research has been minimal to none. Surprisingly, although statisticians can correctly define p-values and they properly struggle to not overestimate the extent of confidence one can have in a confidence interval, most statisticians do not really understand the issues in applied settings.

Recent attacks on p-values and the role of statistical significance in the "crisis of irreproducibility" has highlighted our lack of understanding. Our collective credibility in the science community is at risk. We cannot excuse ourselves by blaming non-statisticians for their failure to understand or heed what we tell them. The fault for widespread ignorance about statistical significance and for the misuses by substantive scientists of measures we promulgate is ours alone. We must communicate better even if we have to scream from the rooftops, which is exactly what the ASA is doing.

More important than the credibility of our discipline is the impact that misuse and misinterpretation of statistical significance and p-values has on science and society. Patients with serious diseases have been harmed. Researchers have chased wild geese, finding too often that statistically significant conclusions could not be reproduced. The economic impacts of faulty statistical conclusions are great.

The effects extend to the public and affect the lay person's understanding and appreciation of science. For example, anybody who leafs through newspapers has seen many studies showing statistically significant health effects of drinking coffee, with each study contradicting many earlier studies. The public learns to not believe "studies." Count me among them.

The ASA's statement will herald a statistical renaissance. Every statistician must take notice. Statisticians who think they know better than the ASA are wrong. And no teacher of an introductory statistics course can pretend to represent modern statistical philosophy and practice without discussing this statement with their students. An effective tack would be to provide them with examples from the research literature, bad examples as well as good—with my warning that it will be hard to find the latter!

Statistics texts define p-values and show how to calculate them in particular examples assuming a particular statistical model. But they fail to address the confusion and mayhem that these measures cause in practical applications. Texts accentuate the positive. They do not consider applied problems with conclusions of statistical significance when the p-value is less than 0.05, say, but have no inferential content, are scientifically meaningless, and cannot be reproduced.

There is little controversy regarding interpreting p-values as summary statistics for a particular set of data. P-values are handy measures of extremity and serve to describe a set of numbers in a way similar to that of Z-scores and confidence intervals. Errors occur when attributing scientific import to a p-value. For instance, researchers may claim that a small p-value is evidence against the null hypothesis that a treatment is ineffective. The standard Bayesian non-informative-prior data-analytic approach is similar to using p-values for inference but is potentially more dangerous because it ostensibly concludes with a posterior probability of truth.

I will expand here on Principles 1 and 4 of the ASA statement.

The statement gives this “informal” definition: “a p-value is the probability under a specified statistical model that a statistical summary of the data … would be equal to or more extreme than its observed value.” Similarly, Principle 1 indicates that p-values “can indicate how incompatible the data are with a specified statistical model.” Yes, but there are subtleties: choice of statistical model; interpretations of “the data;” deciding what is extreme.

Statisticians are trained to analyze numbers. Suppose you provide a statistician with a spreadsheet containing outcomes for a particular experimental treatment vs control and you request a p-value. To set up a model the statistician may ask about the stopping rule, about whether the two samples were independently collected, and about any covariates. After deciding whether to transform the data the statistician calculates a p-value using a test based on some assumed form of the distribution of outcomes or taking a nonparametric approach. What does the p-value mean regarding whether the treatment is better than control? Not much.

Such a p-value is a descriptive summary of a dataset but it has no inferential content. The critical issue is the interpretation of “the data” in the p-value definition. Inferences require a broader interpretation of data than one based on numbers alone. My dictionary says data are “things known or assumed as facts, making the basis of reasoning or calculation.” P-values ignore many aspects of the evidence in the experiment at hand including information that is obviously known. One important piece of data is the simple fact that you gave the statistician the spreadsheet and requested a p-value. Why did you do that? Had you noticed something unusual about the outcomes? Had you requested p-values for the same data from other statisticians and didn't like their answers?

The specifics of data collection and curation and even your intentions and motivation are critical for inference. What have you not told the statistician? Have you deleted some data points or experimental units, possibly because they seemed to be outliers? Are some entries actually the average of two or more measurements made on the same experimental unit? If so, why were there more measurements on some units than on others? Have you conducted other experiments addressing the same or related questions and decided that this was the most relevant experiment to present to the statistician? And on and on and on.

The answers to these questions may be more important for making inferences than the numbers themselves. They set the context for properly interpreting the numerical aspects of the “data.” Viewed alone, p-values calculated from a set of numbers and assuming a statistical model are of limited value and frequently are meaningless.

How can one incorporate the answers to questions such as those above into a statistical analysis? Standard Bayesian data-analytic measures have the same fundamental limitation as p-values. Subjective Bayesian approaches have some hope, but exhibiting a full likelihood function for non-quantifiable data may be difficult or impossible. As a practical matter, when I worry that I don’t know enough about the extra-numerical aspects of the “data” or about the possibility of incorporating this information into a quantitative measure of evidence then I resort to a “black-box warning:”

> “Our study is exploratory and we make no claims for generalizability. Statistical calculations such as p-values and confidence intervals are descriptive only and have no inferential content.”

When is it appropriate to use p-values for inference? An archetype is drug regulation. Drug sponsors must develop a protocol and a statistical analysis plan in advance of an experiment. These explicitly and unambiguously state the primary endpoint and how it will be analyzed. After the experiment a robot could calculate the p-value.

Principle 4 in the ASA statement is that “Proper inference requires full reporting and transparency and multiplicities.” When there is a prospective study protocol and statistical analysis plan then both should be made available at the time of publication along with any deviations from the original plans. In the absence of a protocol and statistical analysis plan the credibility of conclusions is low, despite honest attempts to say what analyses had been planned, whether done or not, and what planned analyses were not done. And adjusting for associated multiplicities may be difficult in this circumstance. A pragmatic approach is to completely describe the multiplicities, keeping a log of what was done, and then giving “unadjusted” p-values, including a black-box warning similar to the one above.

The “p-value dilemma” is entwined with the bigger problems of global multiplicity and irreproducible research. Drug development is an example. Thousands of drugs are being developed worldwide. Each developer conducts clinical trials to decide whether their drug merits further development. The trials may have completely prospective protocols that are followed meticulously. Development continues into the next phase if $Z > 1.65$, say. Some drugs proceed and some do not. For those that proceed, regression to the mean sets in and most drugs fail in the next phase. A Bayesian analysis can accommodate historical information regarding other drugs to suitably regress the results of any particular clinical trial. But speaking as someone who does that, it is not easy to persuade a developer that their next trial is unlikely to be as promising as the present one! And it's not just the developer who is duped. For example, even though they have lots at stake, venture capitalists over-interpret the present trial's p-value and they have trouble understanding that they cannot take the observed data at face value.

Irreproducible research is a huge problem in science and medicine. Statisticians are well positioned to teach other scientists about reproducibility of research, or lack thereof. However, most statisticians are as naïve in this regard as the scientists themselves. Newly minted statisticians tend to regard p-values as relevant scientifically and interpret statistical significance found from processing a spreadsheet of numbers as being reproducible 95% of the time. Only the cold water of experience teaches them otherwise. Again, the remedy is education. We must change the way statisticians are trained. They will, in turn, retrain the rest of the world.

In brief, p-values are not what they're cracked up to be. They serve to describe a dataset of numbers and in that sense they are useful tools. But the vast majority of small p-values do not deserve the label "statistically significant" and they do not imply any other type of scientific relevance. The ASA statement is a bold attempt to right previous misunderstandings in this regard.

Are there better approaches to inference than using p-values, in clinical research say? Absolutely. But that has not been my focus here. It is important to use any tool correctly, especially if we hope to improve it.

Professor Donald A. Berry
Department of Biostatistics
The University of Texas M.D. Anderson Cancer Center
1400 Pressler Street, 4-5062 Pickens Academic Tower
Houston, TX 77030-1402
(Express Mail: 1400 Pressler Street, Box 301402)
Phone: 713-794-4141
Cell: 713-817-5586
Fax: 713-563-4243 or 713-563-4242
e-mail: dberry@mdanderson.org