Editorial

JSE has not published editorials in the past, just Notes from the Editor about the current issue of the journal, but I want to change that and introduce occasional editorials. Here is a first installment, which is a position paper on the use of \( p \)-values, statistical inference, terminology, and related ideas. I thank JSE associate editors for their input, particularly Bill Notz who helped edit what follows.

What Should We Say About Statistical Inference?

Most JSE articles do not include any formal statistical inference, but it is not unusual for a paper to include one or more \( p \)-values and other elements of statistical inference. For example, here is a statement from a JSE article published a few years ago: “the difference was not statistically significant \( (p = 0.2638) \)” Here is another: “... a significant predictor of how much the students gain \( (p = 0.2638) \)”.

What is wrong with those sentences? If you say “nothing is wrong” you stand with the overwhelming majority of statisticians. Indeed, some of us have written very similar sentences in textbooks that we have authored. Moreover, JSE published the papers in which they appear and we could have asked the authors to rewrite their statements, but we did not. Instead, we agreed with the assessments of referees that these were perfectly fine as written.

But here are rewordings of those statements: “the difference was not statistically discernible \( (p = 0.3) \)” and “... a discernible predictor of how much the students gain \( (p = 0.002) \)”. In these rewordings we have made two changes: (1) We have reported the \( p \)-value to just one significant digit, rather than the several digits that computer software routinely produces. (2) We have replaced the word “significant” with the word “discernible.” We take up these two changes in order.

How Many Digits Make Sense?

Typically, there are many steps in the process that leads to a \( p \)-value. Data are collected, perhaps with some element of nonresponse. Aside from any nonresponse, the data collection might include measurement error. Once obtained, the data are fed into a model, but we know that all models are wrong (or to put it more positively, all models are at best reasonable approximations). At the end of the process we have a \( p \)-value that likely has been computed as a reasonable approximation of the “true” \( p \)-value (whatever that might mean), but with considerable associated uncertainty. Reporting \( p = 0.2638 \) implies that all of those digits are meaningful digits and that they all matter, neither of which is true in most cases. Reporting \( p = 0.3 \) tells the reader that the \( p \)-value is in the neighborhood of 0.3 while not claiming that we know the \( p \)-value more precisely. Moreover, it almost certainly does not matter whether \( p = 0.26 \) or 0.28 or 0.30. The second digit does not matter in any practical application that we can think of when the \( p \)-value is large.

Even when there is strong evidence against the null hypothesis, we care about the general magnitude of the \( p \)-value but not the precise value. Again, we cannot think of a practical situation in which the consequence of \( p = 0.0017 \) differs from that of \( p = 0.0020 \). The second significant digit perhaps carries mathematical information, but it does not carry statistical information. In our ongoing quest to convince the world that statistics is not mathematics (although statistics uses mathematics), we prefer to report only digits that matter.

There is a borderline case in which readers might want to know the \( p \)-value to two significant digits, namely when \( p \) is near 0.05. For some readers, \( "p = 0.05" \) will be unsatisfactory, as they will see a world of difference between \( p = 0.048 \) and \( p = 0.053 \). We recognize this sad reality, so we would allow an exception to our “one significant digit” rule when the \( p \)-value is near 0.05, or 0.10, or 0.01. But we hope for the day when readers do not cry out for a second digit even in such cases.

What Does “Statistically Significant” Mean?

Secondly, we note that considerable confusion and misuse of statistical inference derives for the unfortunate fact that “significant” has meaning in everyday speech that is quite different from what we statisticians mean when we say that a treatment effect, say, is “statistically significant.” Indeed, in the preceding paragraphs we used the term “significant digit” as it is commonly used in science; we did not have a statistical implication there. When we say that a result is “statistically significant” what we mean by those words is that a statistical test found evidence of an effect, that the sample effect is unlikely to arise by chance alone if the null hypothesis is true. The word “statistically” is intended to emphasize that “significant” refers to “unlikely to have arisen by chance.” Whether or not the effect matters in practice is quite a different question. We know that “significant” does not mean “important,” but our students and others do not.
Our solution to the problem of misunderstanding of "statistically significant" is to call a result "statistically discernible" rather than "statistically significant." If the p-value is small, it means that the test was able to discern a difference (although this might be a Type I error...). Of course, with a large enough sample size, any small treatment effect will lead to a small p-value. The sample size acts like a magnifying glass, giving us the power to find a difference even when that difference is quite modest. So if our sample is large enough, we can discern a difference.

This new terminology may well help with a related problem, one that we see quite frequently even among PhD statisticians, which is this: Sometimes a researcher writes "The test has found that the population means are significantly different" when what they should write is "The test has found that the differences in sample means give significant evidence that the population means are different." It is the sample means that differ by a statistically significant amount, which gives evidence that $\mu_1 \neq \mu_2$. If the statistics community had agreed many decades ago to replace "statistically significant" with "statistically discernible" then today one might find people writing "The test has found that the population means are discernibly different," which would better convey what is happening. The sample means differed by more than chance would predict, so we infer that the population means are different. We discern a difference in population means, based on seeing a large gap between the sample means.

We note that the March 2019 special issue of The American Statistician was devoted to statistical significance and p-values and was a follow-up to the 2016 ASA Statement on p-Values and Statistical Significance. Most of that discussion concerns the use of p-values and the drawing of a bright line between rejecting and retaining a null hypothesis. Our focus is on the phrase "statistically significant" and not on the misuse of p-values and of hypothesis testing in general (although we have opinions on such misuse).

The August 2019 issue of the journal Significance includes an article "What does it all mean?" by Neil Sheldon, who calls for using the word "outlier" in place of "significance"—a suggestion we do not endorse, but which we mention to highlight the widespread dissatisfaction with current terminology. This dissatisfaction stretches back many years. For example, in 2013 Megan Higgs wrote an article "Do we really need the S-word" that appeared in the American Scientist and that included the recommendation to "replace the s-word with words describing what you actually mean by it." We think that "discernible" does that.

It is challenging to change terminology like "statistically significant" or "a significance test" because those phrases have been in use for many years and have become ingrained in textbooks and research papers. However, statistical terminology has evolved over the years. For example, in the past statisticians used the word "assumptions" in reference to hypothesis tests and to modeling, whereas today one is likely to read about "conditions" under which a t-test, say, is valid or the "conditions" underlying the validity of a statistical model. As another example, "exploratory data analysis (EDA)" is common terminology today, but some decades ago one would have read of "descriptive statistics and graphical displays." A third example is "regression" being subsumed into "predictive modeling and curve fitting." Although the term "regression" persists, we usually have to explain to students the origin of the term regression because to non-statisticians that word does not convey the notion of fitting curves to data.

It is beneficial to use terminology that clearly communicates what is meant, and "statistically significant" fails that test. Statistics educators can play a major role in changing terminology. If we change the terms we use in our classes and textbooks, we can influence the next generation of statisticians and researchers.

We are open to considering other terminology. If you have a better replacement for "statistically significant" than our proposed "statistically discernible" please let us know.

In the meantime, we encourage JSE readers to consider the two recommendations we are putting forward. We will not reject a paper submitted to JSE because the author presents a p-value to four decimal places or calls a result statistically significant (although we may suggest changes during the editing process if the paper looks to be publishable), but we do want JSE papers to communicate ideas and results clearly and we think that the changes we are recommending aid in communication.

Jeff Witmer
Editor-in-Chief