We're using a common statistical test all wrong. Statisticians want to fix that.

Retraction Watch Tracking retractions as a window into the scientific process

After reading too many papers that either are not reproducible or contain statistical errors (or both), the American Statistical Association (ASA) has been roused to action. Today the group released six principles for the use and interpretation of p values. P-values are used to search for differences between groups or treatments, to evaluate relationships between variables of interest, and for many other purposes. But the ASA says they are widely misused. Here are the six principles from the ASA statement:

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.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

We spoke with Ron Wasserstein, ASA's executive director, about the new principles.

Retraction Watch: Why release these "six principles" now? What about this moment in research history made this a particularly pertinent problem?

Ron Wasserstein: We were inspired to act because of the growing recognition of a reproducibility crisis in science (see, for example, the National Academy of Sciences recent report) and a tendency to blame statistical methods for the problem. The fact that editors of a scholarly journal – *Basic and Applied Social Psychology* — were so frustrated with research that misused and misinterpreted p-values that they decided to ban them in 2015 confirmed that a crisis of confidence was at hand, and we could no longer stand idly by.

Retraction Watch: Some of the principles seem straightforward, but I was curious about #2 – I often hear people describe the purpose of a p value as a way to estimate the probability the data were produced by random chance alone. Why is that a false belief?

Ron Wasserstein: Let's think about what that statement would mean for a simplistic example. Suppose a new treatment for a serious disease is alleged to work better than the current treatment. We test the claim by matching 5 pairs of similarly ill patients and randomly assigning one to the current and one to the new treatment in each pair. The null hypothesis is that the new treatment and the old each have a 50-50 chance of producing the better outcome for any pair. If that's true, the probability the new treatment will win for all five pairs is $(½)^5 = 1/32$, or about 0.03. If the data show that the new treatment does produce a better outcome for all 5 pairs, the p-value is 0.03. It represents the probability of that result, *under the assumption that the new and old treatments are equally likely to win.* It is not the probability the new treatment and the old treatment are equally likely to win.

This is perhaps subtle, but it is not quibbling. It is a most basic logical fallacy to conclude something is true that you had to assume to be true in order to reach that conclusion. If you fall for that fallacy, then you will conclude there is only a 3% chance that the treatments are equally likely to produce the better outcome, and assign a 97% chance that the new treatment is better. You will have committed, as Vizzini says in "The Princess Bride," a classic (and serious) blunder.

Retraction Watch: What are the biggest mistakes you see researchers make when using and interpreting p values?

Ron Wasserstein: There are several misinterpretations that are prevalent and problematic. The one I just mentioned is common. Another frequent misinterpretation is concluding that a null hypothesis is true because a computed p-value is large. There are other common misinterpretations as well. However, what concerns us even more are the misuses, particularly the misuse of statistical significance as an arbiter of scientific validity. Such misuse contributes to poor decision making and lack of reproducibility, and ultimately erodes not only the advance of science but also public confidence in science.

Retraction Watch: Do some fields publish more mistakes than others?

Ron Wasserstein: As far as I know, that question hasn't been studied. My sense is that all scientific fields have glaring examples of mistakes, and all fields have beautiful examples of statistics done well. However, in general, the fields in which it is easiest to misuse p-values and statistical significance are those which have a lot of studies with multiple measurements on each participant or experimental unit. Such research presents the opportunity to p-hack your way to findings that likely have no scientific merit.

Retraction Watch: Can you elaborate on #4: "Proper inference requires full reporting and transparency"?

Ron Wasserstein: There is a lot to this, of course, but in short, from a statistical standpoint this means to keep track of and report all the decisions you made about your data, including the design and execution of the data collection and everything you did with that data during the data analysis process. Did you average across groups or combine groups in some way? Did you use the data to determine which variables to examine or control, or which data to include or exclude in the final analysis? How are missing observations handled? Did you add and drop variables until your regression models and coefficients passed a bright-line level of significance? Those decisions, and any other decisions you made about statistical analysis based on the data itself, need to be accounted for.

Retraction Watch: You note in a press release accompanying the ASA statement that you're hoping research moves into a "post $p<0.05$" era – what do you mean by that? And if we don't use p values, what do we use instead?

Ron Wasserstein: In the post $p<0.05$ era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy. (As a start to that thinking, if p-values are reported, we would see their numeric value rather than an inequality ($p=.0168$ rather than $p<0.05$)). All of the assumptions made that contribute information to inference should be examined, including the choices made regarding which data is analyzed and how. In the post $p<0.05$ era, sound statistical analysis will still be important, but no single numerical value, and certainly not the p-value, will substitute for thoughtful statistical and scientific reasoning.

Retraction Watch: Anything else you'd like to add?

Ron Wasserstein: If the statement succeeds in its purpose, we will know it because journals will stop using statistical significance to determine whether to accept an article. Instead, journals will be accepting papers based on clear and detailed description of the study design, execution, and analysis, having conclusions that are based on valid statistical interpretations and scientific arguments, and reported transparently and thoroughly enough to be rigorously scrutinized by others. I think this is what journal editors want to do, and some already do, but others are captivated by the seeming simplicity of statistical significance.

Source: http://retractionwatch.com/2016/03/07/were-using-a-common-statistical-test-all-wrong-statisticians-want-to-fix-that/