The Problems With P-Values are not Just With P-Values

Andrew GELMAN

The ASA's statement on p-values says, "Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted." I agree, but knowledge of how many analyses were conducted etc. is not enough. The whole point of the "garden of forking paths" (Gelman and Loken 2014) is that to compute a valid p-value you need to know what analyses *would have been done* had the data been different. Even if the researchers only did a single analysis of the data at hand, they well could've done other analyses had the data been different. Remember that "analysis" here also includes rules for data coding, data exclusion, etc.

When I was sent an earlier version of the ASA's statement, I suggested changing the sentence to, "Valid p-values cannot be drawn without knowing, not just what was done with the existing data, but what the choices in data coding, exclusion, and analysis would have been, had the data been different. This 'what would have been done under other possible datasets' is central to the definition of p-value." The concern is not just multiple comparisons, it is multiple *potential* comparisons.

Even experienced users of statistics often have the naive belief that if they did not engage in "cherry-picking . . . data dredging, significance chasing, significance questing, selective inference and p-hacking" (to use the words of the ASA's statement), and if they clearly state how many and which analyses were conducted, then they're OK. In practice, though, as Simmons, Nelson, and Simonsohn (2011) noted, researcher degrees of freedom (including data-exclusion rules; decisions of whether to average groups, compare them, or analyze them separately; choices of regression predictors and iteractions; and so on) can be and are performed after seeing the data.

A *scientific* hypothesis in a field such as psychology, economics, or medicine can correspond to any number of *statistical* hypotheses, and if the ASA is going to issue a statement warning about p-values, I think it necessary to emphasize that researcher degrees of freedom—the garden of forking paths—can and does occur even without people realizing what they are doing. A researcher will see the data and make a series of reasonable, theory-respecting choices, ending up with an apparently successful—that is, "statistically significant"—finding, without realizing that the nominal p-value obtained is meaningless.

Ultimately the problem is not with p-values but with null-hypothesis significance testing, that parody of falsificationism in which straw-man null hypothesis A is rejected and this is taken as evidence in favor of preferred alternative B (see Gelman 2014). Whenever this sort of reasoning is being done, the problems discussed above will arise. Confidence intervals, credible intervals, Bayes factors, cross-validation: you name the method, it can and will be twisted, even if inadvertently, to create the appearance of strong evidence where none exists.

What, then, can and should be done? I agree with the ASA statement's final paragraph, which emphasizes the importance of design, understanding, and context—and I would also add measurement to that list.

What went wrong? How is it that we know that design, data collection, and interpretation of results in context are so important—and yet the practice of statistics is so associated with p-values, a typically misused and misunderstood data summary that is problematic even in the rare cases where it can be mathematically interpreted?

I put much of the blame on statistical education, for two reasons.

First, in our courses and textbooks (my own included), we tend to take the "dataset" and even the statistical model as given, reducing statistics to a mathematical or computational problem of inference and encouraging students and practitioners to think of their data as given. Even when we discuss the design of surveys and experiments, we typically focus on the choice of sample size, not on the importance of valid and reliable measurements. The result is often an attitude that any measurement will do, and a blind quest for statistical significance.

Second, it seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an "uncertainty laundering" that begins with data and concludes with success as measured by statistical significance. Again, I do not exempt my own books from this criticism: we present neatly packaged analyses with clear conclusions. This is what is expected—demanded—of subject-matter journals. Just try publishing a result with $p = 0.20$. If researchers have been trained with the expectation that they will get statistical significance if they work hard and play by the rules, if granting agencies demand power analyses in which researchers must claim 80% certainty that they will attain statistical significance, and if that threshold is required for publication, it is no surprise that researchers will routinely satisfy this criterion, and publish, and publish, and publish, even in the absence of any real effects, or in the context of effects that are so variable as to be undetectable in the studies that are being conducted (Gelman and Carline 2014).

In summary, I agree with most of the ASA's statement on p-values but I feel that the problems are deeper, and that the solution is not to reform p-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Andrew Gelman, Columbia University 1016 Social Work Building New York New York 10027 (Email: gelman@stat.columbia.edu).

References

Gelman, A. (2014), "Confirmationist and Falsificationist Paradigms of Science," Statistical Modeling, Causal Inference, and Social Science blog, 5 Sept. *http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/*.

Gelman, A., and Carlin, J. B. (2014), "Beyond Power Calculations: Assessing Type S (sign) and Type M (magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651.

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465.

Simmons, J., Nelson, L., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allow Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366.