Don't Throw Out the Error Control Baby With the Bad Statistics Bathwater: A Commentary

Deborah G. MAYO

The American Statistical Association is to be credited with opening up a discussion into p-values; now an examination of the foundations of other key statistical concepts is needed.

Statistical significance tests are a small part of a rich set of "techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, p. 1033). These may be called *error statistical methods* (or *sampling theory*). The error statistical methodology supplies what Birnbaum called the "one rock in a shifting scene" (ibid.) in statistical thinking and practice. Misinterpretations and abuses of tests, warned against by the very founders of the tools, shouldn't be the basis for supplanting them with methods unable or less able to assess, control, and alert us to erroneous interpretations of data.

P-value. The significance test arises to test the conformity of the particular data under analysis with H_0 in some respect:

> To do this we find a function $t = t(y)$ of the data, to be called the test statistic, such that
>
> - the larger the value of t the more inconsistent are the data with H_0;
> - the corresponding random variable $T = t(Y)$ has a (numerically) known probability distribution when H_0 is true.
>
> ...[We define the] p-value corresponding to any t as $p = p(t) = P(T \geq t; H_0)$ (Mayo and Cox 2006, p. 81).

Clearly, if even larger differences than t occur fairly frequently under H_0 (p-value is not small), there's scarcely evidence of incompatibility. But even a small p-value doesn't suffice to infer a genuine effect, let alone a scientific conclusion—as the ASA document correctly warns (Principle 3). R. A. Fisher was clear that we need not isolated significant results:

> ...but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher 1947, p. 14)

If such statistically significant effects are produced reliably, as Fisher required, they indicate a genuine effect. This is the essence of statistical falsification in science. The logic differs from inductive updating probabilities of a hypothesis, or a comparison of how much more probable H_1 makes the data than does H_0, as in likelihood ratios. Given the need to use an eclectic toolbox in statistics, it's important to avoid expecting an agreement on numbers from methods evaluating different things. Hence, it's incorrect to claim a p-value is "invalid" for not matching a posterior probability based on one or another prior distribution (whether subjective, empirical, or one of the many conventional measures).

Effect sizes. Acknowledging Principle 5, tests should be accompanied by interpretive tools that avoid the fallacies of rejection and nonrejection. These correctives can be articulated in either Fisherian or Neyman-Pearson terms (Mayo and Cox 2006; Mayo and Spanos 2006). For an example of the former, looking at the p-value distribution under various discrepancies from H_0: $\mu = \mu_0$ allows inferring those that are well or poorly indicated. If you very probably would have observed a more impressive (smaller) p-value than you did, if $\mu > \mu_1$ (where $\mu_1 = \mu_0 + \gamma$), then the data are good evidence that $\mu \leq \mu_1$. This is akin to confidence intervals (which are dual to tests) but we get around their shortcomings: We do not fix a single confidence level, and the evidential warrant for different points in any interval are distinguished. The same reasoning allows ruling out discrepancies when p-values aren't small. This is more meaningful than power analysis, or taking nonsignificant results as uninformative. Most importantly, we obtain an evidential use of error probabilities: to assess how well or *severely tested* claims are. Allegations that frequentist measures, including p-values, must be misinterpreted to be evidentially relevant are scotched.

Biasing selection effects. We often hear it's too easy to obtain small p-values, yet replication attempts find it difficult to get small p-values with preregistered results. This shows the problem isn't p-values but failing to adjust them for cherry picking, multiple testing, post-data subgroups and other *biasing selection effects*. The ASA correctly warns that "[c]onducting multiple analyses of the data and reporting only those with certain p-values" leads to spurious p-values (Principle 4). The *actual* probability of erroneously finding significance with this gambit is not low, but high, so a *reported* small p-value is invalid. However, the same flexibility can occur with likelihood ratios, Bayes factors, and Bayesian updating, with one big difference: The direct grounds to criticize inferences as flouting error statistical control is lost (unless they are supplemented with principles that are not now standard). The reason is that they condition on the actual data; whereas error probabilities take into account other outcomes that could have occurred but did not.

The introduction of prior probabilities—which may also be data dependent—offers further leeway in determining if there has even been replication failure. Notice the problem with biasing selection effects isn't about long-run error rates, it's being unable to say that the *case at hand* has done a good job of avoiding misinterpretations.

Model validation. Many of the "other approaches" rely on statistical models that require "diagnostic checks and tests of fit

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Deborah G. Mayo, Department of Philosophy, Virginia Tech, (Email: mayod@vt.edu).

which, I will argue, require frequentist theory significance tests for their formal justification" (Box 1983, p. 57), leading Box to advocate ecumenism. Echoes of Box may be found among holders of different statistical philosophies. "What we are advocating, then, is what Cox and Hinkley (1974) call 'pure significance testing', in which certain of the model's implications are compared directly to the data…" (Gelman and Shalizi 2013, p. 20).

We should oust recipe-like uses of p-values that have been long lampooned, but without understanding their valuable (if limited) roles, there's a danger of blithely substituting "alternative measures of evidence" that throw out the error control baby with the bad statistics bathwater.

References

Birnbaum, A. (1970), "Statistical Methods in Scientific Inference (letter to the Editor)," *Nature* 225(5237), 1033.

Box, G. (1983), "An Apology for Ecumenism in Statistics," in *Scientific Inference, Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard, and D. F. J. Wu, New York: Academic Press, 51–84.

Cox, D., and Hinkley, D. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Fisher, R. A. (1947), *The Design of Experiments* (4th ed.), Edinburgh: Oliver and Boyd.

Gelman, A., and Shalizi, C. (2013), "Philosophy and the Practice of Bayesian Statistics" and "Rejoinder," *British Journal of Mathematical and Statistical Psychology*, 66(1), 8–38; 76–80.

Mayo, D. G., and Cox, D. R. (2006), "Frequentists Statistics as a Theory of Inductive Inference," in *Optimality: The Second Erich L. Lehmann Symposium*, ed. J. Rojo, Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS), 49, 77–97.

Mayo, D. G., and Spanos, A. (2006), "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction," *British Journal for the Philosophy of Science*, 57(2), 323–357.