Controversy Over the Significance Test Controversy

Deborah Mayo

In the face of misinterpretations and proposed bans of statistical significance tests, the American Statistical Association gathered leading statisticians in 2015 to articulate statistical fallacies and galvanize discussion of statistical principles. I discuss the philosophical assumptions lurking in the background of their recommendations, linking also my co-symposiasts. As is common, probability is assumed to accord with one of two statistical philosophies: (1) *probabilism* and (2) (long-run) *performance*. (1) assumes probability should supply degrees of confirmation, support or belief in hypotheses, e.g., Bayesian posteriors, likelihood ratios, and Bayes factors; (2) limits probability to long-run reliability in a series of applications, e.g., a "behavioristic" construal of N-P type 1 and 2 error probabilities; false discovery rates in Big Data.

Assuming probabilism, significance levels are relevant to a particular inference only if misinterpreted as posterior probabilities. Assuming performance, they are criticized as relevant only for quality control, and contexts of repeated applications. Performance is just what's needed in Big Data searching through correlations (Glymour). But for inference, I sketch a third construal: (3) probativeness. In (2) and (3), unlike (1), probability attaches to methods (testing or estimation), not the hypotheses. These "methodological probabilities" report on a method's ability to control the probability of erroneous interpretations of data: *error probabilities*. While significance levels (p-values) are error probabilities, the probing construal in (3) directs their evidentially relevant use.

That a null hypothesis of "no effect" or "no increased risk" is rejected at the .01 level (given adequate assumptions) tells us that 99% of the time, a smaller observed difference would result from expected variability, as under the null hypothesis. If such statistically significant effects are produced reliably, as Fisher required, they indicate a genuine effect. Looking at the entire p-value distribution under various discrepancies from the null allows inferring those that are well or poorly indicated. This is akin to confidence intervals but we do not fix a single confidence level, and we distinguish the warrant for different points in any interval. My construal connects to Birnbaum's *confidence concept*, Popperian *corroboration*, and possibly Fisherian *fiducial* probability. The probativeness interpretation better meets the goals driving current statistical reforms.

Much handwringing stems from hunting for an impressive-looking effect, then inferring a statistically significant finding. The *actual* probability of erroneously finding significance with this gambit is not low, but high, so a *reported* small p-value is invalid. Flexible choices along "forking paths" from data to inference cause the same problem, even if the criticism is informal (Gelman). However, the same flexibility occurs with probabilist reforms, be they likelihood ratios, Bayes factors, highest probability density (HPD) intervals, or lowering the p-value (until the maximal likely alternative gets .95 posterior). But lost are the direct grounds to criticize them as flouting error statistical control. I concur with Gigerenzer's criticisms of ritual uses of p-values, but without understanding their valid (if limited) role, there's a danger of accepting reforms that throw out the error control baby with the "bad statistics" bathwater.