The ASA Guidelines and Null Bias in Current Teaching and Practice

Sander Greenland (corresponding author)
Department of Epidemiology and Department of Statistics
University of California, Los Angeles, CA, U.S.A.
lesdomes@ucla.edu

The ASA statement is a step forward for the profession that I am happy to endorse. Nonetheless, as a compromise among many conflicting views it is bound to leave many readers dissatisfied in some respects. I should like to express my chief concern via this quote from Neyman (1977, p. 106; emphasis added) discussing a hypothetical example of a possibly carcinogenic chemical A:

> “Here we come to the subjectivity of judging importance. From the point of view of the manufacturer the error in asserting the carcinogenicity of A is (or may be) more important to avoid than the error in asserting that A is harmless. Thus, for the manufacturer of A, the 'hypothesis tested' may well be: 'A is not carcinogenic'. **On the other hand, for the prospective user of chemical A the hypothesis tested will be unambiguously: 'A is carcinogenic'.** In fact, this user is likely to hope that the probability of error in rejecting this hypothesis be reduced to a very small value!”

With this passage, Neyman makes clear that testing the “no effect” hypothesis favors only the manufacturer, precisely because it assumes asymmetrically that the cost of erroneously concluding there *is* an effect (a false positive) is higher than the cost of erroneously assuming there is *no* effect (a false negative). Note also that Neyman uses the phrase 'the hypothesis tested' (as opposed to “null hypothesis”) making clear that the tested hypothesis could be that there *is* an effect.

Like most discussions of statistical testing I have seen, the ASA statement fails to appreciate Neyman’s basic points. Against my objections, the statement maintained use of the term “null hypothesis” to designate *any* hypothesis being subject to a statistical test, whether that hypothesis is that there is no effect or that there is an effect. This is in flat contradiction of ordinary English, in which “null” means “nothing,” as in “no association” (e.g., Merriam-Webster 2016, Oxford 2016a) – just as in mathematics it refers to the additive identity element – and “null hypothesis” refers to no difference (Oxford 2016b). But in the field of statistics, many statisticians use “null” to mean “the hypothesis one is attempting to nullify” or refute. Thus the term “null hypothesis” joins “significance” and “interaction” as a term whose meaning in statistical jargon deviates from common usage, resulting in profoundly confused views of inference among users and even among statistics professors and textbooks.

This ingrained (and not always inadvertent) null bias in standard statistical expositions leads to a profound violation of scientific neutrality in disputes, as illustrated in Neyman’s example. An extreme form of this bias is seen in testing *only* the no-effect hypothesis, then claiming the evidence supports no effect because the test was “nonsignificant” (P>0.05) (see Greenland 2004, 2011, 2012 and Greenland & Poole 2011 for examples of this practice among statistics and epidemiology professors when acting as expert witnesses) – even though tests of other effect sizes would show the same statistical evidence even better supports many nonzero effects.

The falsificationist ideal that (quite independently of Popper) inspired Fisher and even more so Neyman was that tests do no more than help us see what our data cannot refute or reject under an assumed data-generation model. A failure to reject may thus allow us to proceed *as if* an unrejected effect size is correct; but it does **not** and cannot tell us which effect size to proceed with among the many that would remain unrejected by the same testing procedure. Testing only the no-effect hypothesis simply assumes, without grounds, that erroneously defaulting to no effect is the least costly error, and in this sense is a methodologic bias toward the null.

Likelihoodists and Bayesians were among the earliest to recognize the problem of focusing on single hypotheses. As Edwards (1972, P. 180) wrote of significance testing,

> “What used to be called judgement is now called prejudice, and what used to be called prejudice is now called a null hypothesis. In the social sciences, particularly, it is dangerous nonsense (dressed up as ‘the scientific method’), and will cause much trouble before it is widely appreciated as such.”

Ironically though, the same null bias found in significance tests is tightly integrated into the confirmationist version of “objective Bayesian” testing, which assigns a spike (point prior mass) to test a point null hypothesis against a continuous composite alternative (Jeffreys, 1961; Berger & Sellke, 1987). That isolated spike identifies the null as the point with special added cost to falsely reject, and the mass of the spike is a surrogate for that added cost. Again Neyman’s example should be borne in mind, for in such cases the spike serves only the parties invested in maintaining the null.

As Neyman's example made clear, defaulting to "no effect" as the test hypothesis (encouraged by describing tests as concerning only "null hypothesis", as in the ASA statement) usurps the vital role of the context in determining loss, and the rights of stakeholders to use their actual loss functions. Those who benefit from this default (either directly or through their clients) have gone so far as to claim assuming "no effect" until proven otherwise is an integral part of the scientific method. It is not; when analyzed carefully such claims hinge on assuming that the cost of false positives is always higher than the cost of false negatives, and are thus circular.

Yes, in many settings (such as genomic scans) false positives are indeed considered most costly by all research participants, usually because everyone expects few effects among those tested will be worth pursuing. But treating these settings as if scientifically universal does violence to other contexts in which the costs of false negatives may exceed the costs of false positives (such as side effects of polypharmacy), or in which the loss functions or priors vary dramatically across stakeholders (as in legal and regulatory settings).

Those who dismiss the above issues as mere semantics or legal distortions are evading a fundamental responsibility of the statistics profession to promote proper use and understanding of methods. So far, the profession has failed abjectly in this regard, especially for methods as notoriously contorted and unnatural in *correct* interpretation as statistical tests. It has long been argued that much of harm done by this miseducation and misuse could be alleviated by suppression of testing in favor of estimation (Yates, 1951, p. 32-33; Rothman, 1978). I agree, although we must recognize that loss functions also enter into estimation, for example via the default of 95% for confidence or credibility intervals, and in the default to unbiased instead of shrinkage estimation. Nonetheless, interval estimates at least help convey a picture of where each possible effect size falls under the same testing criterion, thus providing a more fair assessment of competing hypotheses, and making it easier for research consumers to apply their own cost considerations to reported results.

In summary, automatically defaulting to the no-effect hypothesis is no less mindless of context and costs than is defaulting to a 0.05 rejection threshold (which is widely recognized as inappropriate for many applications). Basic statistics education should thus explain the integral role of loss functions in statistical methodology, how these functions are hidden in standard methods, and how these methods can be extended to deal with settings in which loss functions vary or costs of false negatives are large.

References

Berger, J. O., and Sellke, T. (1987), "Testing a point null hypothesis: the irreconcilability of p-values and evidence" (with discussion), *Journal of the American Statistical Association* 82, 112-122.

Edwards, A.W.F. (1972). Likelihood. Cambridge: Cambridge Univ. Press.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press.

Greenland, S. (2004). “The need for critical appraisal of expert witnesses in epidemiology and statistics,” *Wake Forest Law Review* 39, 291–310.

_______ (2011). “Null misinterpretation in statistical testing and its impact on health risk assessment,” *Preventive Medicine* 53, 225–228.

_______ (2012). “Nonsignificance plus high power does not imply support for the null over the alternative,” *Annals of Epidemiology* 22, 364–368.

Greenland, S., and Poole C. (2011). “Problems in common interpretations of statistics in scientific articles, expert reports, and testimony,” *Jurimetrics* 51, 113–129.

Merriam Webster Dictionary online (2016), accessed 24 Feb., http://www.merriam-webster.com/dictionary/null

Neyman, J. (1977). Frequentist probability and frequentist statistics. Synthese 36, 97–131.

Oxford English Dictionary online (2016a), accessed 24 Feb., http://www.oxforddictionaries.com/us/definition/american_english/null

Oxford English Dictionary online (2016b), accessed 24 Feb., http://www.oxforddictionaries.com/us/definition/american_english/null-hypothesis

Rothman, K.J. (1978). “A show of confidence,” *NEJM* 299, 1362−1363.

Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *Journal of the American Statistical Association* 46, 19–34.