The Annals of Statistics
1992, Vol. 20, No. 1, 490–509

ESTIMATION OF ACCURACY IN TESTING

BY JIUNN TZON HWANG,[1] GEORGE CASELLA,[2] CHRISTIAN ROBERT,[3]
MARTIN T. WELLS[4] AND ROGER H. FARRELL

*Cornell University, Cornell University, Université Paris VI,
Cornell University and Cornell University*

1. Introduction. Approaches to hypothesis testing have usually treated the problem of testing as one of decision-making rather than estimation. More precisely, a formal hypothesis test will result in a conclusion as to whether a hypothesis is true, and not provide a measure of evidence to associate with that conclusion. In this paper we consider hypothesis testing as an estimation problem within a decision-theoretic framework and are able to arrive at some interesting conclusions. In particular, reasonable loss functions result in decision rules that can be regarded as measures of evidence and, under these loss functions, some interesting properties of p-values emerge.

1.1. *Standard approaches*. Classical hypothesis testing is built around the Neyman–Pearson Lemma [Lehmann (1986)] and results in decision rules that are 0–1 rules (except for randomized tests). These formal tests, although optimal in a strict frequentist sense, have been criticized from many different directions. First, there have been many Bayesian criticisms [e.g., DeGroot, (1973); Dickey, (1977); Berger, (1985a, b)] which point out the drawbacks of the stringent conclusion of the Neyman–Pearson approach. Namely, the experimenter is locked into a two-point action space. Secondly, the assessment of accuracy of the test is typically a predata assessment, most often the *size* of the test. This estimate can be quite unreasonable when viewed postdata, a criticism which has also been leveled at Neyman–Pearson theory by conditionalists [Kiefer, (1977); Robinson (1979a, b)]. Alternatives considered by Kiefer include using p-values as an assessment of the likelihood of the null hypothesis. These ideas are in the direction of those proposed here, that the hypothesis test should result in a postdata assessment of evidence. (In fairness to Neyman–Pearson theory, measures of size and power were proposed as predata operating characteristics, not postdata assessments of accuracy, of a testing procedure.)

Received March 1990; revised June 1991.

[1] Research supported by NSF Grant DMS-88-09016.

[2] Research supported by NSF Grant DMS-89-0039.

[3] Research performed while Professor Robert was Visiting Scientist at the Mathematical Sciences Institute, Cornell University.

[4] Research supported by the Mathematical Sciences Institute, Cornell University.

AMS 1980 *subject classifications*. Primary 62F03; secondary 62A99, 62C07, 62C15.

Key words and phrases. Hypothesis testing, decision theory, p-values.