

The p-value, the Bayes/Neyman-Pearson Compromise and the Teaching of Statistical Inference in Introductory Business Statistics

Thomas W. Woolley, Samford University

Abstract

Traditionally the Neyman-Pearson approach to hypothesis testing has been presented in introductory business statistics courses. However, many students as well as researchers find the decisions reached by this approach, i.e., reject/fail-to-reject, inconsistent with their understanding of the scientific process, namely accumulating evidence in support of a hypothesis. The proposed framework provides an easily understood rationale for introducing the student to I.J. Good's Bayes/Neyman-Pearson compromise as represented by Good's standardized p-values. Standardized p-values are a useful and practical tool for the evidentialist interpretation of data within the context of Neyman-Pearson hypothesis testing, something desired by many students and researchers.

INTRODUCTION

Over the years, there has been a debate raging over certain perceived inadequacies in the p-value specifically, and hypothesis testing in general. (e.g., Carver, 1978; Cohen, 1996; Derrick, 1976; Goodman and Royall, 1988; Kirk, 1996; Meehl, 1967; Oakes, 1986; O'Neill, 1995; Pearce, 1990; Rozeboom, 1960; Salsburg, 1990; Sawyer and Peter, 1983; and Walker, 1986). It has been noted that the p-value lacks the capability to summarize a study's data as evidence in support of a hypothesis. It is my contention that it is this issue that rests at the core of researcher frustration with and misuse of the classical frequentist methodology (Goodman and Royall, 1988; Kirk, 1996). Researchers seemingly desire more from a statistical procedure than rules and numbers that lead them to reject or fail to reject a null hypothesis. Such a decision-theoretic approach fails to dovetail with the incremental nature of accumulating knowledge that most researchers, and in fact lay people, utilize in decision making (Schmidt, 1996).

In teaching introductory business statistics courses, using a variety of textbooks (e.g., Anderson, Sweeney, and Williams, 2002), I have observed that this same sense of frustration is found in students. Often students find the Neyman-Pearson methodology to be incongruent with the decision-making tools utilized in other courses. Consequently, I began devoting a single class at the end of the first semester statistics course to the presentation and discussion of a broad overview of various models for statistical hypothesis testing. Then, within the framework developed in class, I. J. Good's Bayes/Neyman-Pearson compromise is presented in the form of a standardized p-value. I have found these classes to be very well received by the students, giving them a better appreciation of both the field of statistics and one alternative index (the standardized p-value) that can be used as a legitimate measure of evidence. And while the compromise, specifically focusing on the standardized p-value, may not be the last word as a measure of evidence it does alleviate some of the frustration students feel with the use of the

strict Neyman-Pearson approach to hypothesis testing (and p-values) in light of the evidentialist summary measure they, and many researchers, desire.

The purpose of this paper, therefore, is to present my general framework for discussing statistical testing and its use as a vehicle for introducing standardized p-values.

MODELS OF STATISTICAL TESTING

The use of statistical tests is well established in all branches of scientific research and universally taught in applied statistics courses. Perhaps the first indication of their use was by Arbuthnot (1710) in what would be called today a retrospective study. Since their modern introduction by Fisher (1925) a number of competing strategies have been proposed which have led to the current frustration with hypothesis testing so prevalent among scientists.

In order to better understand the various views of hypothesis testing, I use a conceptual framework suggested by Mayo (1985) which distinguishes two broad models of statistical testing: behavioral-decision (*behavioralist*) and evidential-strength (*evidentialist*). Similar classifications have been offered by other authors, e.g., Birnbaum (1977) and Savage (1954). The behaviorist philosophy of statistical testing promotes the view that the primary task of statistics is to provide methods that guide decisions which must be made in the face of uncertainty. Mayo (1985, p. 501) observes that this perspective suggests that "statistics should provide, not rules of inductive inference, but rules for making optimal decisions as how to behave with respect to hypotheses." Under this model statistical tests define *optimal* rules for making a decision on when to reject and when not to reject a hypothesis. The traditional Neyman-Pearson scheme is representative of this model. The evidentialist philosophy of statistical testing asserts that "the task of a theory of statistics (in science) is to provide some means of using data to assign hypotheses a measure of evidential strength (support, probability, reliability, degrees of belief, etc.)" (Mayo, 1985, p. 502). Anecdotal evidence suggests that this understanding of statistical testing resonates with many management scientists.

The Fisherian, Bayesian, and Likelihood schools of statistical thought include procedures that can be interpreted in accordance with the evidentialist model. Differences among these procedures center on the way data provide evidential support for a hypothesis. Within the Fisherian school, the p-value has been interpreted as a measure of evidential strength. To properly understand how this can be a valid interpretation of the p-value one needs to examine carefully the notion of *significance testing* (Barnett, 1982; Cox and Hinkley, 1974; Kempthorne, 1976; Kempthorne and Folks, 1971; and Neyman, 1969). In significance testing a null hypothesis of interest is put forward and the probability (under the null distribution) of sampling the observed data (or data even further from the hypothesized value of the parameter) is computed. This computed probability is sometimes called the significance probability, critical level, level of significance, observed significance level, or more commonly, the p-value. The size of the p-value reflects the degree of consistency between the data and the null hypothesis: a small p-value provides evidence against the null since the data are less plausible while a large p-value tends to support the hypothesis. Cox and Hinkley (1974) suggest the following interpretation:

Suppose that we were to accept the available data as evidence against H_0 . Then we would be bound to accept all data with a larger value of [the test statistic; i.e., smaller p-value] as even stronger evidence. Hence [the p-value] is the probability that we would mistakenly declare there to be evidence

against H_0 , were we to regard the data under analysis as just decisive against H_0 " (p. 66).

In order for this interpretation to be valid, however, a "nesting condition" is required on departures from the null hypothesis.

What's wrong with the p-value as a measure of evidence? Goodman and Royall (1988) delineate three major problems that have been noted by statisticians for years. First, a huge effect derived from a small sample can result in the same p-value as a small effect derived from a large sample. Second, there is the logical problem of allowing results that have not occurred to enter into the calculation of the p-value (i.e., the probability of getting this data or data even further from the hypothesized value). Jeffreys (1980) has declared

I have always considered the arguments for the use of p absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened (p. 455).

In the Bayesian approach to statistical testing, the *a posteriori* probability provides a natural measure of evidence. This *a posteriori* probability serves as the Bayesian's measure of the evidence provided by the data relative to the null hypothesis. What, if any, is the relationship between the p-value as a measure of evidence and the *a posteriori* probability? There is a large body of literature that explores this question, e.g., Berger and Delampady (1987), Berger and Sellke (1987), Casella and Berger (1987), DeGroot (1973), and Dickey (1977). Under certain circumstances, some researchers have shown that the two measures of evidence are in close agreement, whereas others have demonstrated conditions under which the "actual evidence against a null (as measured, say, by posterior probability...) can differ by an order of magnitude from the p-value" (Casella and Berger, 1987, p. 112). To further compound the situation, the Lindley-Jeffreys-Good paradox illustrates that it is possible to have a p-value as small as desired while the *a posteriori* probability simultaneously approaches 1.0 and the sample size increases, assuming a fixed prior probability distribution on H_0 (Shafer, 1976b). Still, I.J. Good (1981) feels that while "the use of [the p-value] is... logically shaky... it is useful all the same, and can often be given a rough [Bayesian] justification, at least when sample sizes are not very large" (p. 157). Many scientists have been unwilling to accept the subjective nature of the Bayesian approach. The subjective nature of the Bayesian philosophy "requires from its adherents... not that their beliefs (probabilities) should correspond to reality but that they should be consistent" (Oakes, 1986, p. 135). Such a philosophy seems counter to the understanding of many researchers who view their task as one of discovering reality. However, if current thinking in the philosophy of science is accepted, namely that all data are "theory laden" (Polanyi, 1958), then the Bayesian approach provides what many investigators desire.

Another approach that has been suggested for statistical hypothesis testing is the likelihood function (Goodman and Royall, 1988). This methodology avoids the need to interpret probability as a measure of one's beliefs. Though Edwards (1972) provides the most complete recent account of the likelihood approach to scientific inference, its roots go back to Barnard (1949), Barnard, Jenkins, and Winsten (1962), Birnbaum (1962, 1968), and Fisher (1934, 1959). Barnett (1982) notes that likelihood "has met with some considerable support in different fields of application, notably physics and genetics" and that "there is an obvious lay appeal in the... principle" (p. 285). Goodman and Royall (1988) do a nice job of illustrating the application of the likelihood approach to statistical testing in their field as well as pointing out that the resulting

likelihood ratio cannot be interpreted probabilistically. The likelihood approach, however, is not without its critics.

Oakes (1986) believes that likelihood "is arguably the least ambitious of the approaches" to statistical testing while Bayesians criticize its "reliance on a relative frequency definition of probability and its comparative failure to come to grips with prior knowledge" (p. 144). Both Bayesian and classical (Neyman-Pearson) statisticians are uncomfortable "with what is seen by many to be an ill-defined, or uninterpretable, assignment of numerical measures of relative weight of evidence to alternative models, hypotheses or parameter values" (Barnett, 1982, p. 287). In fact, Plackett (1966) argued that there "are other objections to placing too much emphasis on the likelihood function since, having no definite scale, it allows only the comparison of hypotheses and needs to be supplemented by significance tests and sampling properties." However, the most damaging assessments concerning the proposition that the likelihood ratio can be interpreted as a quantitative measure of the weight of evidence for or against hypotheses, have come from within the likelihood camp itself (Birnbaum, 1968, 1977; Giere, 1977; and Shafer, 1976a, 1976b). In reviewing the book by Edwards (1972), Hacking (1975) states

Allan Birnbaum and myself are very favorably reported in this book for things we have said about likelihood, but Birnbaum has given it up and I have become pretty dubious (p. 137).

So, although the likelihood approach engenders a sincere and sympathetic response from researchers concerned with the appropriateness of traditional statistical approaches to scientific inference, it is clearly not a panacea.

BAYES/NEYMAN-PEARSON COMPROMISE PROPOSAL

Having discussed this material in class, students often feel that there is no appropriate method for interpreting data evidentially. To help students overcome this hurdle I present I.J. Good's compromise proposal (1982). Clearly, the meaningfulness of the p-value diminishes as the sample size increases. This is, in fact, one of the criticisms leveled at the interpretability and generalizability of the p-value and hypothesis testing. I.J. Good (1982) proposed the use of a Bayes factor, that is, a gauge of the weight of evidence against a hypothesis, to develop a standardized tail-area probability (p-value) or what he refers to as the *q-value*. Good bases his proposal on Jeffreys (1939, p. 356) work that showed the Bayes factor against the null for a given p-value is inversely proportional to the square root of sample size. He makes the case that introducing the q-value would "bring p values into closer relationship with weights of evidence while also preserving the appearance of objectivity" (Good, 1992).

Good (1982) points out that a p-value for a sample size of n would convert to a Bayes factor of roughly equal to $p\sqrt{n/100}$ for a sample of 100 observations. Hence, by "standardizing" p-values from different sources to a sample size of 100 using this method, it is possible to interpret p-values comparatively and evidentially. In other words, sample size has been controlled; as such, the only variable in the q-value is the effect size. Good (1982) interprets the standardized p-value by saying that "the evidence against the null hypothesis is (about) the same as if a tail-area probability of [*standardized p-value*] had been obtained from a sample of size 100" (p. 165). More formally, the standardized p-value is defined as $q = \min(0.5, p\sqrt{n/100})$ where the value of 0.5 is somewhat arbitrary, though its purpose is to avoid q-values of greater than 1.

EXAMPLE

Suppose that three separate studies are conducted each of which looks at the effect of organizational behavior modification on task performance. Table 1 summarizes the pertinent information from each study:

TABLE 1

Examples of the Standardized p-value (q-value) for Varying Sample Sizes

	<u>Study 1</u>	<u>Study 2</u>	<u>Study 3</u>
p-value	0.04	0.04	0.04
n	10	100	1,000
q-value	0.01	0.04	0.13

In this example, the p-values are meaningless, uninterpretable with respect to offering any evidence relative to the null. Obviously all three studies yielded different treatment effects with Study 1 showing the greatest effect and Study 3 the smallest. But all three studies would lead to a rejection of the null hypothesis of no effect if standard Neyman-Pearson methods were applied (assuming the studies were designed with $\alpha = 0.05$). The q-values, on the other hand, provide the additional information that Study 3 offers less evidence against the null than either Study 1 or Study 2.

CONCLUSION

The Neyman-Pearson approach to hypothesis testing is almost universally taught in introductory business statistics courses. It has been noted that many students find decisions yielded by this procedure in conflict with the scientific process of accumulating evidence in support of a hypothesis. As suggested by Kirk (1996), the routine reporting of effect sizes in research reports would contribute to putting the p-value into its proper perspective by drawing more attention to weight of evidence in support of the null hypothesis. However, effect sizes are neither being taught in our business statistics classes nor being reported by our management scientists in their published research reports.

This paper has presented a framework that I have found useful in presenting an overview of the predominant philosophies of statistical testing. This framework provides an easily understood rationale for introducing the student to I. J. Good's Bayes/Neyman-Pearson compromise as conceived of through Good's standardized p-value concept. My experience with the standardized p-value suggests that it provides a useful and practical tool for the evidentialist interpretation of data desired by many students and researchers without the difficulty of computing or interpreting effect size.

REFERENCES

- Anderson, D.R., Sweeney, D.J., and Williams, T.A. 2002. *Statistics for business and economics*, eighth edition. Cincinnati, OH: South-Western College Publishing.
- Arbuthnot, J. 1710. An argument for divine providence, taken from the constant regularity

- observed in the births of both sexes. *Philosophical Transactions*, 27: 186-190.
- Barnard, G.A. 1949. Statistical inference (with discussion). *Journal of the Royal Statistical Society*, B11: 115-149.
- Barnard, G.A., Jenkins, G.M., & Winsten, C.B. 1962. Likelihood inference and time series (with discussion). *Journal of the Royal Statistical Society*, A125: 351-352.
- Barnett, V. 1982. *Comparative statistical inference*, second edition. New York: Wiley.
- Berger, J.O. & Delampady, M. 1987. Testing precise hypotheses. *Statistical Science*, 2: 317-352.
- Berger, J.O. & Sellke, T. 1987. Testing a point null hypothesis: The irreconcilability of p-values and evidence (with discussion). *Journal of the American Statistical Association*, 82: 112-133, 135-139.
- Birnbaum, A. 1962. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57: 269-306.
- Birnbaum, A. 1968. Likelihood, *International encyclopedia of the social sciences*, volume 9. New York: Macmillan and the Free Press.
- Birnbaum, A. 1977. The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese*, 36: 19-49.
- Carver, R.P. 1978. The case against statistical significance testing. *Harvard Educational Review*, 48: 378-399.
- Casella, G. & Berger, R.L. 1987. Reconciling Bayesian and frequentist evidence in one-sided testing problem (with discussion). *Journal of the American Statistical Association*, 82: 106-111, 123-135.
- Cohen, J. 1996. The earth is round ($p < .05$). *American Psychologist*, 49: 997-1003.
- Cox, D.R. & Hinkley, D.V. 1974. *Theoretical statistics*. London: Chapman and Hall.
- DeGroot, M.H. 1973. Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, 68: 966-969.
- Derrick, T. 1976. The criticism of inferential statistics. *Educational Research*, 19: 35-40.
- Dickey, J.M. 1977. Is the tail area useful as an approximate Bayes factor? *Journal of the American Statistical Association*, 72: 138-142.
- Edwards, A.W.F. 1972. *Likelihood*. Cambridge: Cambridge University Press.
- Fisher, R.A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Fisher, R.A. 1934. Two new properties of mathematical likelihood. *Proceedings of the Royal Statistical Society*, A144: 285-307.
- Fisher, R.A. 1959. *Statistical methods and scientific inference*, second edition. Edinburgh: Oliver and Boyd.
- Giere, R.N. 1977. Allan Birnbaum's conception of statistical evidence. *Synthese*, 36: 5-13.
- Good, I.J. 1981. Some logic and history of hypothesis tests. In *Philosophy in economics*, University of Western Ontario Series on the Philosophy of Science (J.C. Pitt, ed.). Dordrecht: Reidel.
- Good, I.J. 1982. Standardized tail-area probabilities. *Journal of Computation and Simulation*, 16: 65-66.
- Good, I.J. (1992). The Bayes/Non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87: 597-606.
- Goodman, S.N. & Royall, R. 1988. Evidence and scientific research (Commentary). *American*

- Journal of Public Health*, 78: 1568-1574.
- Hacking, I. 1975. *The emergence of probability*. Cambridge: Cambridge University Press.
- Jeffreys, H. 1939. *Theory of probability*, first edition. Oxford: University Press.
- Jeffreys, H. 1980. Some general points in probability theory, in *Bayesian analysis in econometrics and statistics*, ed. A. Zellner. Amsterdam: North-Holland.
- Kempthorne, O. 1976. Of what use are tests of significance and tests of hypothesis. *Communications in Statistics*, A5: 763-777.
- Kempthorne, O. & Folks, J.L. 1971. *Probability, statistics, and data analysis*. Ames, Iowa: Iowa State University Press.
- Kirk, R.E. 1996. Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56: 746-759.
- Mayo, D.G. 1985. Behavioristic, evidentialist, and learning models of statistical testing. *Philosophy of Science*, 52: 493-516.
- Meehl, P.E. 1967. Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34: 103-115.
- Neyman, J. 1969. Statistical problems in science: The symmetric test of a composite hypothesis. *Journal of the American Statistical Association*, 64: 1154-1171.
- Oakes, M.L. 1986. *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
- O'Neill, B. 1995. Weak models, nil hypotheses, and decorative statistics: Is there really no hope? *Journal of Conflict Resolution*, 39: 731-748.
- Pearce, N. 1990. White swans, black ravens, and lame ducks: Necessary and sufficient causes in epidemiology. *Epidemiology*, 1: 47-50.
- Plackett, R.L. 1966. Current trends in statistical inference. *Journal of the Royal Statistical Society*, A29: 249-267.
- Polanyi, M. 1958. *Personal knowledge: Towards a post-critical philosophy*. Chicago: University of Chicago Press.
- Rozeboom, W.W. 1960. The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57: 416-428.
- Salsburg, D. 1990. Hypothesis versus significance testing for controlled clinical trials: A dialogue. *Statistics in Medicine*, 9: 201-211.
- Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
- Sawyer, A.G. & Peter, J.P. 1983. The significance of statistical significance tests in marketing research. *Journal of Marketing Research*, 20: 122-133.
- Schmidt, F.L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1: 115-129.
- Shafer, G. 1976a. *A mathematical theory of evidence*. Princeton: Princeton University Press.
- Shafer, G. 1976b. A theory of statistical evidence. In Harper, W.L. and Hooker, C.A. (eds.), *Foundations of probability theory, statistical inference, and statistical theories of science*, II. Dordrecht: Reidel.
- Walker, A.M. 1986. Reporting the results of epidemiologic studies. *American Journal of Public Health*, 76: 556-558.