The Likelihood Principle and the Reliability of Experiments

Andrew Backe

Philosophy of Science, Vol. 66, Supplement. Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers. (Sep., 1999), pp. S354-S361.

Stable URL:
http://links.jstor.org/sici?sici=0031-8248%28199909%2966%3CS354%3ATLPATR%3E2.0.CO%3B2-B

Philosophy of Science is currently published by The University of Chicago Press.

The Likelihood Principle and the Reliability of Experiments

Andrew Backe†‡
University of Pittsburgh

The likelihood principle of Bayesian statistics implies that information about the stopping rule used to collect evidence does not enter into the statistical analysis. This consequence confers an apparent advantage on Bayesian statistics over frequentist statistics. In the present paper, I argue that information about the stopping rule is nevertheless of value for an assessment of the reliability of the experiment, which is a pre-experimental measure of how well a contemplated procedure is expected to discriminate between hypotheses. I show that, when reliability assessments enter into inquiries, some stopping rules prescribing optional stopping are unacceptable to both Bayesians and frequentists.

1. Introduction. According to Bayes's theorem, the relevant information from an experiment is contained in the likelihood function $P(x|\Theta)$. This function summarizes the probability of the experimental evidence x occurring under each hypothesis about an unknown parameter value Θ. Summarizing evidence in this way entails that, if any two instances of evidence yield likelihood functions which are the same apart from some constant factor, then the inferences drawn from the experiments should be the same. Stated formally, if $P(x|\Theta) = cP(x'|\Theta)$, where c is some positive constant and x and x' denote instances of evidence from different experiments investigating the same hypotheses about Θ, then

†Department of History and Philosophy of Science, 1017 Cathedral of Learning, University of Pittsburgh, Pittsburgh, PA 15260.

‡This paper developed out of discussions with Deborah Mayo. I thank both her and Merrilee Salmon for commenting on earlier drafts. I also thank Teddy Seidenfeld for his comments and guidance on recent drafts.

Philosophy of Science, 66 (Proceedings) pp. S354–S361. 0031-8248/99/66supp-0027$0.00

the two instances of evidence have identical evidential import. This implication is known as the *likelihood principle*.[1]

A corollary of the likelihood principle is that details about some aspects of the experiment do not enter into the statistical analysis of the evidence. Particularly, rules about when the collection of evidence should stop are rendered irrelevant to the analysis. This corollary has been cited as an advantage of Bayesian statistical analysis over frequentist statistical analysis, since the latter summarizes evidence through significance levels, which are influenced by information about the stopping rule. My main objective in the following sections of this paper is to show that this implication of the likelihood principle is limited by another concern arising in experimental inquiry. After providing a detailed description of the likelihood principle's consequences for stopping rules, I maintain that a major concern in any experimental inquiry is the reliability of one's experiment. In the restricted sense used in this paper, reliability indicates how well an experiment can distinguish a true hypothesis from among alternatives. I argue that reliability assessments will limit Bayesians' choices of experimental procedures and thereby eliminate extreme stopping rules which are supposed to confer an advantage on Bayesian statistical analysis.

2. The Likelihood Principle and Stopping Rules. An implication of the likelihood principle is that it renders irrelevant certain aspects about the experimental procedure used to collect the evidence. This consequence is best understood when evaluating stopping rules. Edwards, Lindman, and Savage note:

> The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience. (1963, 193)

To see the full import of the likelihood principle for stopping rules, consider the following illustration, which is oversimplified from actual practice in order to highlight the theoretical point. Suppose that a researcher is conducting a binomial experiment to investigate whether a new drug is better than a placebo. In the experiment, each trial is a comparative recovery rate on a pair of subjects. Each trial is indepen-

1. See Berger and Wolpert 1984 and Birnbaum 1962 for comprehensive discussions of the likelihood principle. The principle also has been discussed in the work of Edwards, Lindman, and Savage (1963, 237–238), Hacking (1965, 106–109), Mayo (1996, 337–359), and Savage (1962, 17–18), among others.

dent, given the treatment effects, and takes the value of "0" ("unfavorable" to the new drug) or "1" ("favorable" to the new drug). The experiment tests a set of hypotheses H_0 and H_1 regarding the value of the parameter Θ, which indicates the true probability of a "favorable" outcome. Specifically, the experiment tests the hypothesis H_0: $\Theta = \Theta_0 = .5$ and H_1: $\Theta = \Theta_1 = t$, where t is some value greater than .5. The researcher's prior probability distribution over the two hypotheses is uniform (i.e., $P(H_0) = P(H_1) = .5$).

Suppose that the researcher starts conducting trials and determines the likelihood function after each trial. In the present illustration, the likelihood function is related to the likelihood ratio $P(x|\Theta_1)/P(x|\Theta_0)$. Suppose specifically that the researcher intends to stop the experiment only when $P(x|\Theta_1)$ exceeds $P(x|\Theta_0)$ by the critical ratio λ, at which point the researcher will make a terminal decision to accept H_1 and report its posterior probability. Otherwise, the researcher will continue to conduct trials. If the researcher obtains the desired critical ratio after, say, 50 trials, at which point 40 "favorable" outcomes have been observed, then it should make no difference to the interpretation of the result to know that it came from a sequential arrangement that stopped once the researcher observed desirable evidence. The import of the evidence from the sequential design is the same as it would have been had the researcher intended to conduct a fixed-sample-size experiment of 50 trials and observed 40 "favorable" outcomes. The respective likelihood functions of the instances of evidence from the sequential experiment and the fixed-sample-size experiment would be constant multiples of each other, which would be reflected in identical likelihood ratios (see Savage 1962, 72–73). Consequently, information about the stopping rule is of no inferential value.

In contrast to Bayesian statistical analysis, frequentist statistical analysis does not entail this consequence. Frequentist analysis rests primarily on the Neyman and Pearson (1933) theory of hypothesis testing. What researchers seek from an application of a Neyman-Pearson test is the probability that an error has been committed once a particular hypothesis has been accepted. In a choice between two hypotheses H_0 and H_1 about Θ with evidence x, a researcher must consider two possible errors: either a) erroneously concluding that H_1 is true or b) erroneously concluding that H_0 is true. The probability of the former error is denoted α and the latter β. Any two experiments yielding evidence corresponding to identical α and β probabilities will have the same evidential import.

Error probabilities *will* be influenced by the stopping rule. Suppose that a researcher applying a Neyman-Pearson test attempts to establish the truth of H_1 (i.e., reject H_0) by sampling until obtaining evidence

corresponding to an α level, also referred to as the "significance" level, of .05. Under this circumstance, the α level calculated for a fixed-sample-size test will not be appropriate for evaluating the evidence. Because the researcher is continually looking for a "significant" result, there will be a higher probability of finding one purely by chance than if the test were performed merely once, at the final stage of the experiment, as in a fixed-sample-size experiment. In fact, Anscombe (1954, 92–93) has shown that, *with probability one*, a significant result will occur if the researcher continues to sample and apply a fixed-sample-size test. Thus, if the sequential properties of a stopping rule are ignored, persisting in the rule will permit a frequentist to acquire evidence that leads to the rejection of H_0 at any significance level. Such evidence is therefore not really indicative of the truth of the alternative hypothesis.

To compensate for this problem of reasoning to a foregone conclusion, frequentist methods have been developed that adjust the α level according to the number of times a test is applied while evidence is accumulating. Armitage (1975, 27–28) has outlined sequential methods for this purpose. He notes that, to calculate the error probability of rejecting H_0 when a test is applied after each trial, the researcher must determine the probability, given the truth of H_0, that a significant result will occur on or before the given trial. This "overall" significance level will be larger than the nominal significance level of the fixed-sample-size experiment, and, as the sample size increases, the overall probability will become very large and approach one.

3. Reliable Experiments. The preceding discussion highlights a fundamental difference between Bayesian statistical analysis and frequentist statistical analysis. Knowing that an instance of evidence (consisting, for example, of 40 "favorable" outcomes in 50 trials) was collected using a rule prescribing optional stopping rather than fixed stopping will not influence a Bayesian inference, but will influence a frequentist inference. This difference between the two statistical approaches obtains even if the optional stopping rule dictates that the collection of evidence terminate only when and if evidence favorable to one particular hypothesis, such as H_1, is observed.

Savage (1962) has cited this difference between the two approaches as a reason for using likelihood functions rather than significance levels to summarize evidence. He remarks:

> The likelihood principle . . . affirms that the experimenter's intention to persist does not change the import of his experience. The true moral of the facts about optional stopping is that significance

> level is not really a good guide to 'level of significance' in the sense of 'degree of import', for the degree of import does depend on the likelihood alone. . . . (1962, 18)

The apparent advantage of the likelihood principle is that it generally precludes a Bayesian from *reasoning to a foregone conclusion*. Savage (1962, 72–73) and Kerridge (1963, 1109) have demonstrated that, unlike a frequentist, a Bayesian conducting an experiment that will only stop with evidence favorable to one hypothesis does not have to worry about justifying a false conclusion, since, for the Bayesian, the collection of evidence might not terminate. Kadane, Schervish, and Seidenfeld (1996, 1229) recently have provided a general formula for determining the positive probability of non-termination of a Bayesian sequential stopping rule similar to that outlined above. Where p is the "prior" probability of H_1 and q is its "posterior" probability, the bound for the conditional probability of terminating an experiment when H_1 is false is no more than $p(1 - q)/q(1 - p)$. This formula demonstrates that a Bayesian cannot reject a true hypothesis with certainty.[2]

Although the result of non-termination appears to confer an advantage on Bayesian statistics, the kind of stopping rule to which the result has been applied—namely, a rule that can never indicate significant evidence for H_0—is nevertheless problematic. If a researcher has information that a particular hypothesis cannot be accepted in an experiment, then such information should enter into the inquiry at some point. One way that the information should enter is in an assessment of the *reliability* of the experiment. In the restricted sense used here, a *reliable experiment* is one that can indicate the true hypothesis from among the set of alternatives. An experiment that permits either H_0 or H_1 to be accepted is more reliable than one that merely permits a particular hypothesis, such as H_1, to be accepted.

A Bayesian researcher interested in successfully identifying the true hypothesis from among a group of alternatives should perform a preexperimental evaluation of the reliability of the experiment used to collect the evidence. In particular, the Bayesian should choose a stopping rule that will permit an hypothesis to be accepted if that hypothesis is actually true.

To see how a stopping rule can be ill-suited to this goal and, hence, unreliable, consider the following illustration. Suppose that a researcher desires to conduct a binomial experiment with independent trials testing H_0: $\Theta = \Theta_0 = .5$ against H_1: $\Theta = \Theta_1 = \mathrm{t}$, where t is some

2. This result holds under restricted conditions, particularly when probability is countably additive as opposed to finitely additive.

value greater than .5. Assume that the researcher's prior probability distribution over the two hypotheses is uniform (i.e., $P(H_0) = P(H_1) = .5$). Consider two different stopping rules that the researcher might use. The first stopping rule, E, is a sequential rule that attempts to establish the truth of hypothesis H_1. This rule is similar to that considered in section two above. The researcher intends to stop the experiment only when, and if, $P(x|\Theta_1)$ exceeds $P(x|\Theta_0)$ by the critical ratio λ, at which point the researcher will make a terminal decision to accept H_1 and report its posterior probability.

The other stopping rule, E', is also a sequential procedure. It prescribes the following:

(1) If $P(x|\Theta_1)/P(x|\Theta_0) \geq \lambda$, then make a terminal decision to accept H_1 and report its posterior probability.
(2) If $P(x|\Theta_1)/P(x|\Theta_0) \leq \gamma$, then make a terminal decision to accept H_0 and report its posterior probability.
(3) If neither (1) nor (2) obtain, continue sampling.

Unlike procedure E, this procedure permits an assessment of the truth of H_0 as well as H_1. Furthermore, with probability one, the procedure will terminate (see Wald 1947, 37–40).

Suppose that the researcher begins collecting evidence and that, after, say, the 50th trial, observes a likelihood ratio of γ. How this evidence is interpreted will depend upon which stopping rule the researcher adopted at the start of the experiment. If the researcher adopted rule E, then the inference at the 50th trial will be to continue sampling. If, however, the researcher adopted rule E', then the inference at the 50th trial will be to accept H_0 and report its posterior probability. The two stopping rules yield different results even though the evidence collected by the 50th trial is the same.

In this illustration, rule E is unreliable. The rule will *never* permit the researcher to stop the experiment with evidence favorable to H_0. In fact, at a particular trial n, the researcher may observe evidence that has a likelihood ratio extremely favorable to H_0, and the researcher may continue to observe such evidence favorable to H_0 as the experiment continues, but such evidence cannot be used to make a terminal decision. The broader implication here is that the researcher has no guarantee that the experiment will ever stop. Kadane, Schervish, and Seidenfeld's (1996, 1229) formula for determining the positive probability of non-termination of an experiment can be used to show the shortcoming of rule E. If the researcher carrying out rule E requires the critical ratio γ to be such that $q = .99$, then (given $p = .5$) the probability of terminating the experiment when H_1 is false will not exceed .01. Any attempt to increase the probability of termination will

be offset by an increased probability of accepting H_1 when H_0 is actually true.

The practical consequences of adopting rule E are quite severe. Suppose that, during the course of a career in which different experimental processes are investigated, a researcher expects H_0 to be true approximately 50% of the time. If the researcher continually applies rule E, then approximately half of the experiments will either yield no conclusion at all or lead the researcher to accept H_1 when in fact H_0 is true. Moreover, if there is any cost at all to experimentation, then the expected costs of adopting rule E will be boundless.[3]

The problems just cited are not restricted to stopping rules in Bayesian sequential experiments. A Neyman-Pearson experiment that permits a researcher only to accept H_1 will also be unreliable. The adjustment of the α probability for such a procedure, as Armitage (1975, 27–28) prescribes, does not avoid the fact that the procedure will never permit the researcher to accept H_0. A more reliable experiment would be Wald's (1947, 37–44) sequential probability ratio plan. This plan incorporates error probabilities, but it also permits hypothesis H_0 to be accepted and, hence, does not require the excessive adjustments of α probabilities as does Armitage's method.

4. Conclusion. According to Bayesian inference, the import of evidence from an experiment depends only on the likelihood function determined by the evidence observed. Some features of the experiment, such as the stopping rule, are of no inferential value. This consequence pertains to the *post-experimental* import of the evidence. I have argued in the present paper that researchers should also be concerned with the *pre-experimental* measure of the reliability of an experiment. A measure of reliability indicates how good an experiment is at distinguishing the true hypothesis. Such information is of value to Bayesian statistical inquiry as well as frequentist statistical inquiry. Consequently, both a Bayesian and a frequentist will refrain from using stopping rules that are sometimes presented to show a particular advantage of Bayesian statistical analysis over frequentist statistical analysis. Nonetheless, there remain serious practical cases where the dispute over stopping rules remains a live issue, such as in inverse sampling.

REFERENCES

Anscombe, F. J. (1954), "Fixed-Sample-Size Analysis of Sequential Observations", *Biometrics* 10: 89–100.

3. This consequence was recognized by Teddy Seidenfeld and shared with me during personal communication.

Armitage, Peter (1975), *Sequential Medical Trials*, 2nd ed. New York: John Wiley & Sons.
Berger, James O. and Robert Wolpert (1984), *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.
Birnbaum, Alan (1962), "On the Foundations of Statistical Inference", *Journal of the American Statistical Association* 57: 269–306.
Cornfield, Jerome (1966), "A Bayesian Test of Some Classical Hypotheses—With Applications To Sequential Clinical Trials", *Journal of the American Statistical Association* 61: 577–594.
Edwards, Ward, Harold Lindman, and Leonard Savage (1963), "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70: 193–242.
Hacking, Ian (1965), *Logic of Statistical Inference*. Cambridge: Cambridge University Press.
Kadane, Joseph B., Mark Schervish, and Teddy Seidenfeld (1996), "Reasoning to a Foregone Conclusion", *Journal of the American Statistical Association* 91: 1228–1235.
Kerridge, D. (1963), "Bounds for the Frequency of Misleading Bayes Inferences", *Annals of Mathematical Statistics* 34: 1109–1110.
Mayo, Deborah G. (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Neyman, Jerzy and Egon Pearson (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses", *Philosophical Transactions of the Royal Society* (A) 231: 289–337.
Savage, Leonard (1962), *The Foundations of Statistical Inference*. London: Methuen.
Wald, Abraham (1947), *Sequential Analysis*. New York: John Wiley & Sons.