The Likelihood Principle and
the Reliability of Experiments†

Andrew Backe

Department of History and Philosophy of Science
University of Pittsburgh

Abstract. The likelihood principle of Bayesian statistics implies that information about the experimental design from which evidence is collected does not enter into the statistical analysis of the data. In the present paper, I argue that information about the experimental design is nevertheless of value for an assessment of the *reliability of the experiment*, which is a pre-experimental measure of how well a contemplated experiment is expected to discriminate between hypotheses. I show that reliability assessments enter into Bayesian inquiries and eliminate some extreme cases of optional stopping.

Draft of a paper to be presented at the Sixteenth Biennial

Meeting of the Philosophy of Science Association in Kansas City,

Missouri, October 22-24, 1998.

1. Introduction. According to Bayes' 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 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 experimental design do not enter into
the statistical analysis of the evidence. This corollary is often
cited as an advantage of Bayesian statistical analysis over
frequentist statistical analysis, since the latter summarizes
evidence through error probabilities and thus is influenced by
information about the design. My main objective in the following
sections of this paper is to show that the implications of the
likelihood principle are somewhat limited in experimental
practice. After providing a detailed description of the
likelihood principle's consequences for the design of

experiments, I maintain that a major concern in any statistical inquiry is the reliability of one's experiment. Reliability indicates how well an experiment can distinguish a true hypothesis from among alternatives. I argue that reliability assessments will restrict Bayesians' choices of experimental designs and thereby eliminate some extreme cases where a question of optional stopping might otherwise arise.

2. The Likelihood Principle and Experimental Design. An implication of the likelihood principle is that it renders irrelevant certain aspects about the design of the experiment from which the evidence was collected. This consequence is best understood when evaluating the validity of optional stopping plans. Edwards, Lindman, and Savage (1963, 193) 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.

To see the full import of the likelihood principle for optional stopping, 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 independent, 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: $\theta = \theta_0 = \frac{1}{2}$ and H_1: $\theta = \theta_1 = t$, where t is some value greater than $\frac{1}{2}$. 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 calculates the likelihood function after each trial. In this case, the likelihood function is captured by the likelihood ratio $P(x|\theta_1)/P(x|\theta_0)$. Suppose also that the researcher plans to stop the experiment only when $P(x|\theta_1)$ exceeds $P(x|\theta_0)$ by the critical ratio k, at which point the researcher will judge H_1 to be confirmed and report its posterior probability. Otherwise, the researcher will continue to conduct trials. If the researcher observes the desired critical ratio after, say, 50 trials, 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.[2] Because the
evidence is summarized in the likelihood function, 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. 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 is reflected in the identical likelihood
ratios. Consequently, information about the design of the
experiment is of no inferential value.

In contrast to the Bayesian approach, the frequentist
approach to statistical inference does not entail this
consequence. The frequentist approach 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 when a
particular hypothesis has been accepted. In a choice between two
hypotheses H_0 and H_1 about 2 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 " and the latter $.
Any two experiments yielding evidence corresponding to identical
" and $ probabilities will have the same evidential import.

Error probabilities *will* be influenced by the design of the experiment. 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 " level (also called the "significance" level) of .05. Under this circumstance, the " 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 noted that the law of the iterated logarithm shows that, *with probability one*, a significant result will occur if the researcher continues to sample and apply a fixed-sample-size test.

To compensate for this problem of reasoning to a foregone conclusion, procedures have been developed that adjust the " level according to the number of times a test is applied while evidence is accumulating. Armitage, McPherson, and Rowe (1969, 236-237) have outlined a sequential plan for this purpose. They note 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. The Reliability of Experiments. There is a fundamental difference in concerns between Bayesian statistical analysis and frequentist statistical analysis underlying each's respective consequences for the design of experiments. Bayesian statistical analysis is concerned primarily with yielding numerical measures of the degree to which evidence confirms the hypotheses being assessed (see Howson (1997, 268-279)). This probabilistic characterization of the relationship between evidence and hypotheses implies that any bearing that the evidence has on the hypotheses is contained only in the outcome actually observed.[3] The other instances of evidence and the frequencies with which they occur are irrelevant to the import of the obtained evidence and do not enter into the statistical analysis. In the example considered above, the fact that the Bayesian sequential plan will only terminate with evidence confirming hypothesis H_1, and will never terminate with evidence confirming hypothesis H_0, has no bearing on the statistical analysis once evidence in favor of H_1 is observed.

In contrast to Bayesian statistical analysis, frequentist

statistical analysis requires a researcher to average over possible observations when making an inference. Consequently, if a researcher attempts to establish the truth of H_1 through a sequential sampling plan that could never indicate significant evidence for H_0, then such information will enter into the statistical analysis. Mayo (1996, 7) notes that this information permits researchers to perform an assessment of the *reliability* of the experiment being used to collect the evidence. A Reliable experiment is one which, with a high probability, can indicate the true hypothesis among alternatives. The measure of reliability reflects the efficiency of the experiment and is attached to the long-run success of the design chosen by the researcher (see Hacking (1980, 143)).

How well an experiment facilitates choosing the true hypothesis will depend upon the specific design selected by the researcher. When planning an experiment to test hypothesis H_1 against H_0, for example, a researcher interested in assuring with a high probability that the true hypothesis is chosen might apply a design that reduces " to .05 and $ to .05. Only 5% of the time will this method erroneously accept H_1 and only 5% of the time will it erroneously accept H_0. In the restricted, technical sense in which the term "reliable" is being used here, such an experiment would be considered highly reliable, and considered more reliable than one that merely focuses on the probability of

erroneously accepting a particular hypothesis, such as H_1.

4. Reliability and Bayesian Inference. A Bayesian researcher interested in successfully identifying the true hypothesis from among a group of alternatives should be no less concerned than a frequentist researcher with the reliability of the experiment used to collect the evidence. In particular, the Bayesian should choose an experimental plan that will permit an hypothesis to be confirmed if that hypothesis is actually true.

To see how a Bayesian plan 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: $\boldsymbol{\vartheta}$ = $\boldsymbol{\vartheta}_0$ = s against H_1: $\boldsymbol{\vartheta}$ = $\boldsymbol{\vartheta}_1$ = t, where t is some value greater than s. 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 Bayesian experimental plans that the researcher might use. The first plan, E, is a sequential plan that attempts to establish the truth of hypothesis H_1. In this plan, the experiment stops only when, and if, $P(x|\boldsymbol{\vartheta}_1)$ exceeds $P(x|\boldsymbol{\vartheta}_0)$ by the critical ratio $\boldsymbol{8}$, at which point the researcher will judge H_1 to be confirmed and report its posterior probability.

The other experimental plan, E', is also a sequential plan. It is carried out according to the following rules:

(1) If $P(x|\mathbf{2}_1)/P(x|\mathbf{2}_0) \geq \mathbf{8}$, then stop the experiment and

report the posterior probability of H_1 (i.e., confirm H_1).

(2) If $P(x|\mathbf{2}_1)/P(x|\mathbf{2}_0) \leq \mathbf{(}$, then stop the experiment and

report the posterior probability of H_0 (i.e., confirm H_0).

(3) If neither (1) nor (2) obtain, continue sampling.

Unlike plan E, this plan permits an assessment of the truth of H_0

as well as H_1. Furthermore, with probability one, the plan will

terminate.

Suppose that the researcher begins collecting evidence and

that, after, say, the 50th trial, observes a likelihood ratio of

$($. How this evidence is interpreted will depend upon which

sampling plan the researcher adopted at the start of the

experiment. If the researcher adopted plan E, then the inference

at the 50th trial will be to "continue sampling." If, however,

the researcher adopted plan E', then the inference at the 50th

trial will be to "stop the experiment and report the posterior

probability of H_0." The two experimental plans yield different

results even though the evidence collected by the 50th trial is

the same.

In this illustration, plan E is unreliable. The plan will

never permit the researcher to stop the experiment with

confirming evidence for 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 data cannot be used as disconfirming evidence
against H_1. The broader implication here is that the researcher
has no guarantee that the experiment will ever stop. Kadane,
Schervish, and Seidenfeld (1996, 1229) have provided a formula
for determining the positive probability of non-termination of an
experiment such as E. 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 $p(1-q)/q(1-p)$. If the researcher carrying out plan E
requires the critical ratio **8** 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 confirming H_1 when H_0 is actually true.

The practical consequences of adopting plan E are quite
serious. 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 plan E, then approximately half of

the experiments will either yield no conclusion at all or lead

the researcher to confirm 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 plan E will be boundless.[4]

The problems just cited are not restricted to Bayesian

sequential experiments. A Neyman-Pearson experiment that permits

a researcher only to accept H_1 will also be unreliable. The

adjustment of the " probability for such a plan, as Armitage,

McPherson, and Rowe (1969, 236-237) prescribe, does not avoid the

fact that the plan 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 "

probabilities as does Armitage, McPherson, and Rowe's plan.

5. 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 experimental design are of no inferential value. This

consequence pertains to the post-experimental measure of the

degree to which a specific instance of evidence confirms

hypotheses. I have argued in the present paper that in

experimental practice researchers also are concerned with the

pre-experimental measure of the reliability of an experiment. A measure of reliability indicates how good an experimental design 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 particular sequential plans that are often presented to show particular advantages of Bayesian statistical analysis over frequentist statistical analysis. Nonetheless, there remain serious practical cases where the dispute over optional stopping remains a live issue, such as in inverse sampling.

NOTES

† I thank Deborah Mayo and Merrilee Salmon for commenting on an early draft of this paper. I am particularly indebted to Teddy Seidenfeld for his comments and guidance on two recent drafts of this paper.

[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).

[2] Kadane, Schervish, and Seidenfeld (1996) have shown that, with such a sampling plan, the researcher cannot be certain that the plan will stop.

[3] Birnbaum (1962, 271) labels this assumption the *principle of conditionality*. It asserts the "irrelevance of experiments not actually performed." For a detailed discussion of the principle of conditionality, see Berger and Wolpert (1984, 5-18).

[4] This consequence was recognized by Teddy Seidenfeld and shared with me during personal communication.

REFERENCES

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

Armitage, P., McPherson, C. K., and Rowe, B. C. (1975), "Repeated Significance Tests on Accumulating Data", *Journal of the Royal Statistical Society* A 132: 235-244.

Berger, J. O., and Wolpert, R. L. (1984), *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference", *Journal of the American Statistical Association* 57: 269-306.

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70: 193-242.

Hacking, I. (1965), *Logic of Statistical Inference*. Cambridge, MA: Cambridge University Press.

_____. (1980), "The Theory of Probable Inference: Neyman, Peirce and Braithwaite", In D. H. Mellor (ed.), *Science, Belief, and Behaviour: Essays in Honor of R. B. Braithwaite*. Cambridge: Cambridge University Press, pp. 141-160.

Howson, C. (1997), "A Logic of Induction", *Philosophy of Science* 64: 268-290.

Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1996),

"Reasoning to a Foregone Conclusion", *Journal of the American Statistical Association* 91: 1228-1235.

Mayo, D. (1996), *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Neyman, J., and Pearson, E. S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses", *Philosophical Transactions of the Royal Society* (A) 231: 289-337.

Savage, L. (1962), *The Foundations of Statistical Inference*. London: Methuen.

Wald, A. (1947), *Sequential Analysis*. New York: John Wiley & Sons.