DOCUMENT RESUME

ED 134 609 95 TM 006 015

AUTHOR Brumet, Michael E.
TITLE Applying Bayesian Statistics to Educational Evaluation. Theoretical Paper No. 62.
INSTITUTION Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 76
CONTRACT NE-C-00-3-0065
NOTE 32p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Decision Making; Educational Research; *Evaluation; Hypothesis Testing; *Probability; Statistics
IDENTIFIERS Statistical Inference

ABSTRACT

Bayesian statistical inference is unfamiliar to many educational evaluators. While the classical model is useful in educational research, it is not as useful in evaluation because of the need to identify solutions to practical problems based on a wide spectrum of information. The reason Bayesian analysis is effective for decision making is that it defines probability as a measure of opinion or belief, rather than as long-term frequency. Defining probability as a measure of opinion or belief enables the Bayesian investigator to consider a wider range of information than is possible with the traditional model. Personal expertise, logical analysis, and soft data from a wide variety of sources serve to shape opinion about a state of nature, with experimental data providing additional information either for or against the prior opinion of the evaluator. In classical statistics, prior knowledge or opinion is ignored. However, when practical decisions must be made the Bayesian stresses that all knowledge should be brought to bear on the problem rather than just an isolated set of data. Because of the decision-making orientation of the evaluator, the Bayesian model should be considered as an alternative to classical inference. Since the Bayesian model views probability as a measure of opinion rather than as a long-term frequency, the statistical requirements for it are actually greater than for the classical statistician. Use of a wider range of distributions than with classical statistics demands more statistical skills than many evaluators currently possess. However, the questions raised by the Bayesian model are useful even if the model is not totally adopted. (Author/RC)

TM

ED134609

THEORETICAL PAPER NO. 62

applying bayesian statistics to educational evaluation

MAY 1976

WISCONSIN RESEARCH
AND DEVELOPMENT
CENTER FOR
COGNITIVE LEARNING

Theoretical Paper No. 62

APPLYING BAYESIAN STATISTICS TO EDUCATIONAL EVALUATION

by

Michael E. Brumet

Report from the Project on IGE Evaluation

Conrad G. Katzenmeyer
Faculty Associate

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

May 1976

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that agency should be inferred.

Center Contract No. NE-C-00-3-0065

4

WISCONSIN RESEARCH AND DEVELOPMENT CENTER FOR COGNITIVE LEARNING

MISSION

The mission of the Wisconsin Research and Development Center for Cognitive Learning is to help learners develop as rapidly and effectively as possible their potential as human beings and as contributing members of society. The R&D Center is striving to fulfill this goal by

- conducting research to discover more about how children learn
- developing improved instructional strategies, processes and materials for school administrators, teachers, and children, and
- offering assistance to educators and citizens which will help transfer the outcomes of research and development into practice

PROGRAM

The activities of the Wisconsin R&D Center are organized around one unifying theme, Individually Guided Education.

FUNDING

The Wisconsin R&D Center is supported with funds from the National Institute of Education; the Bureau of Education for the Handicapped, U.S. Office of Education; and the University of Wisconsin.

ACKNOWLEDGMENTS

I wish to express my appreciation to Conrad G. Katzenmeyer for providing the opportunity for me to write this paper and for reviewing each draft. My gratitude is also extended to Steven Jurs and Joseph Shaffer who were stimulators for some of the ideas presented in this paper.

6

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Because of the current dominance of classical statistics in the tradition of Fisher, Neyman, and Pearson, an alternative approach known as Bayesian statistical inference is unfamiliar to many educational evaluators. While the classical model is useful in educational research, it is not as useful in evaluation because of the need to identify solutions to practical problems based on a wide spectrum of information.

Business and marketing researchers have utilized the Bayesian model for many years because they need to make practical decisions rather than assertions about some unknown parameter, which is the function of traditional statistics. The reason Bayesian analysis is effective for decision making is that it defines probability as a measure of opinion or belief, rather than as long-term frequency.

Defining probability as a measure of opinion or belief enables the Bayesian investigator to consider a wider range of information than is possible with the traditional model. Personal expertise, logical analysis, and soft data from a wide variety of sources serve to shape opinion about a state of nature, with experimental data providing additional information either for or against the prior opinion of the evaluator. In classical statistics, prior knowledge or opinion is ignored. However, when practical decisions must be made the Bayesian stresses that all knowledge should be brought to bear on the problem rather than just an isolated set of data. Because of the decision-making orientation of the evaluator, the Bayesian model should be considered as an alternative to classical inference.

Since the Bayesian model views probability as a measure of opinion rather than as a long-term frequency, the statistical requirements for it are actually greater than for the classical statistician. Use of a wider range of distributions than with classical statistics demands more statistical skills than many evaluators currently possess. However, the questions raised by the Bayesian model are useful even if the model is not totally adopted.

I

INTRODUCTION

THE BAYESIAN VERSUS CLASSICAL INFERENCE CONTROVERSY

Throughout the sequence of statistical courses digested by most prospective educational evaluators, little notice is usually given to the controversy between so called Neyman-Pearson and Bayesian statisticians. This lack of awareness is no doubt due to the dominant position the Neyman-Pearson (classical) statistician has enjoyed in social science and educational research. However, in some fields where practical decisions must be made on the basis of all available information, the Bayesian statistical model has proven its usefulness. In business and marketing research and to a lesser extent in engineering Bayesian analysis has been effectively utilized to determine the appropriateness of alternative decision choices.

The basic philosophical difference between the two approaches concerns the use of prior information or beliefs. The classical statistician assumes that only specified data gathered after hypothesis formation can be used for inference. The Bayesian statistician contends that the data gathered in an experiment only serve as additional information to be combined with the investigator's prior information or beliefs. This combining of data and opinion is done through the use of Bayes' Theorem. To most educational evaluators the Bayesian approach may seen foreign to all they have learned about the appropriate use of statistical inference.

The purpose of this paper is to show, first, that the differences between educational research and educational evaluation result in the conclusion that statistical techniques appropriate for the former are not necessarily suitable for the latter; and, second, that the Bayesian inferential approach offers an alternative statistical model for the educational evaluator, as he/she is frequently in the position where classical inferential statistics does not allow for utilization of the type of information which he/she possesses.

FOUNDATIONS OF BAYESIAN STATISTICS

The classical (Neyman-Pearson) versus Bayesian controversy can be related to a basic problem in the history of science: the roles of rationalism and empiricism and the interpretation of probability statements (Weber, 1973).

To the Greeks, the laws of science were completely precise and demonstrable through the process of deduction. Fluctuation and variability were considered error and a reflection of lack of knowledge of laws. This deterministic view of knowledge was assumed by early philosophers of science such as Descartes and scientists such as Newton. However, as science matured there was less and less certainty that a rational system could be developed from which any data could be seen as logical consequence. Modern stochastic models such as Maxwell's thermodynamic laws, Mendel's genetic laws, and Einstein's theory of relativity saw science developing probability models of phenomena.

Probability models which have been recognized for a long time were used to cope with ignorance about laws of nature or errors of measurement. However, as stochastic models became more popular, probability models were seen as characteristic of nature itself rather than simply reflecting ignorance. Thus, modern philosophers were forced to reconsider the alternatives that were available to the concept of probability. For the Bayesian, probability means degree of personal belief about some phenomenon. This approach contrasts distinctly with that of the classical statistical school, which considers probability to be long-term relative frequency. The probability of occurrence of an event has been defined as the limit of the relative frequency of its occurrence in some specified reference class of events (Fisher, 1956). With few exceptions, modern statistical textbooks use this relative frequency interpretation of probability.

Bayesians find this view of probability too restrictive. Often statements must be made about nonrepeating events which have a degree of uncertainty. For example, the statement "The probability is greater that a man will land on Mars than that a man will land on Jupiter" makes intuitive sense, yet both events are unique or nonrepeating. By viewing probability as a measure of belief, the concept takes on broader and potentially more useful meaning.

Bayesian statistics stems indirectly from a paper by Thomas Bayes that originally appeared in 1763; however, only in 1961 did the first systematic use of subjective probability and other elements of the Bayesian model emerge with the appearance of Schlaifer's Introduction to Statistics for Business Decisions. Since then, several additional books with a general business application emphasis have appeared, such as Savage (1962), Thompson (1972), and Zellner (1971). However, Edwards, Lindman, and Savage (1963) and McGee (1971) have presented Bayesian approaches to social sciences, and Meyer (1971) presented a general paper on Bayesian statistics at the 1971 American Educational Research Association meeting. This author has not found any additional publications relating the model to educational situations. In this paper, the potential for the use of Bayesian analyses in educational evaluation will be discussed.

II

EDUCATIONAL EVALUATION

DEVELOPMENT OF EDUCATIONAL EVALUATION

Evaluation models were developed in the past decade because of the need to account for and determine the worth of scores of new federally funded educational programs. The first years after the passage of the Elementary and Secondary Education Act (ESEA) of 1965 saw evaluative chaos--every new project was judged by a unique plan. That many of the resulting evaluations would be inadequate was inevitable. Few educators possessed the skills needed to conduct effective evaluation.

A few scholars developed generalizable evaluation plans. Such models were primarily developed by educational researchers, who attempted to compromise the rigor of traditional experimental design with the demands of programmatic educational activities. Although the models did not answer all of the problems of local evaluation, they did provide general guidelines that eliminated some of the deficiencies of early ESEA evaluation reports. These models have been primarily nonquantitative in nature. They consist of general organizational patterns listing the sequence of evaluation activities and key decision points. In view of this statistical vacuum, most evaluators continue to use the traditional statistical inference procedures taught in most graduate schools of education. Such procedures are usually appropriate in educational research; however the differences between research and evaluation are wide enough that a common statistical methodology should not be assumed.

EVALUATION VERSUS RESEARCH

Despite the fact that both evaluation and research can be classified as disciplined inquiries demanding empirical support, logical analysis, and openness to public scrutiny, there are some definite distinctions between these two educational activities. The distinctions are primarily a matter of degree rather than kind. Additionally, the extent and type of differences will depend upon the theoretical models under consideration. However, for the purposes of this paper it will be sufficient to note that "evaluation is the determination of the worth of a thing" whereas "research is the activity aimed at obtaining generalizable knowledge by contriving and testing claims about relationships among variables or describing generalizable phenomena [Worthen & Sanders, 1973, p. 19]." Other distinctions which may be noted between research and evaluation are not necessarily inherent in the two activities but rather reflect the way in which they have been characterized in practice.

The rules of legitimate evidence for evaluation data are broader than for research data. Prior to a research investigation, an experimenter must determine the exact dependent variable which will be measured. This is not to say that other behaviors will be ignored, but their consideration must be secondary to the one specified in the hypothesis. No one would object to the researcher looking elsewhere for support should the specified behavior not support the hypothesis; however, this starts an entirely new investigation with a new hypothesis, new design and new data. Researchers have always looked with scorn at post hoc analysis as something less than scientific.

Evaluation activities are more flexible. With emphasis on the determination of value, the evaluator's duty is to consider evidence from as many angles as possible. For example, the evaluator may be requested to measure the effects of a flexible schedule on a number of independent student projects. If the evaluator notices that such a schedule seems to result in higher student absenteeism, this data may be very important in the evaluation even though it is a side effect.

Research activities generally do not provide for feedback loops. The sequence "problem - hypothesis - sample - data - inference [Hays, 1973, p. 856]" is usually followed until completion. In fact, frequently the researcher may assign most of the activities to rather unsophisticated assistants who must simply follow the research plan formed by the experimenter. After the data are analyzed spinoffs may arise, but the experiment at hand has <u>ended</u>. Even an overview of the major evaluation models will demonstrate the extensive feedback system operating in the models. For example, Stufflebeam (1973) presents a system which is based on continual looping between decisions, activities, and evaluation.

Parsimony in science requires the avoidance of unnecessary significant effects. The consequences of asserting that a variable has an effect when it doesn't (Type I error) are generally considered worse than saying a variable doesn't have an effect when it does (Type II error). Traditionally, researchers are willing to make Type I errors only five or one percent of the time, while the frequency of Type II errors is rarely considered. The dangers of a Type I error are always in the forefront of an experimenter's mind. Before a variable is given a role in a theory, it is carefully examined to see if it is necessary. Hence a good researcher is characterized as being cautious, conservative, and skeptical.

However, the evaluator must constantly juggle the relative costs of a Type I or Type II error. For example, a university may have devoted years to developing and implementing a competency-based teacher education (CBTE) program. Suppose that an evaluator were to compare such a program with a traditional teacher education program. The null hypothesis would postulate that both types of programs were equally effective. Because of the added expense of a CBTE program, a finding of no difference might result in the university's reverting to a traditional program. However, if the CBTE program were found to be superior, little change in activity would occur; the program would simply continue. In this situation, a Type II error would signify that the lengthy development of the CBTE program was in vain; as a result, the university might decide to switch back to a traditional program. Of course, in other situations a Type I error might still cause the most damage. The point is that in evaluation the consequences of both Type I and Type II errors can vary considerably, depending on the circumstances and costs. The traditional procedure of setting a Type I error probability level at .05 or .01 does not involve consideration of relative costs.

FAILURE OF CLASSICAL STATISTICS IN EDUCATIONAL EVALUATION

Classical Neyman-Pearson statistics provides a good model for typical research. A problem is first recognized, whether from an unexplained variable in a theoretical model or just some question arising out of daily experience. From this problem a researcher postulates a hypothesis about a state of nature. For example, "Does a 15-minute work break in the morning increase total morning output?" To test this hypothesis, the researcher may define his population as all office workers in the city. He will then randomly place 20 such office workers into one of two experimental groups. One group will receive the coffee break and one will not. The total morning output will be determined for both groups. Using classical statistics, it would be possible to test for a difference in the total output of the two groups that is sufficiently large to be unlikely due to chance.

This example illustrates the following points. First, the assumption is made that our researcher has no knowledge of the outcome of the experiment. In reality the experimenter obviously expects certain results or he/she would probably not conduct the experiment. In fact, it is because he/she has some knowledge of expected results that the ironical situation arises of the experimenter's assuming that he/she has no expectancy. Hopefully the research will be public and will not be subject to the demand characteristics of the experimenter. The experiment is to be a microcosm where the effects of variables will be assessed strictly in terms of what occurs in the experiment itself. All outcomes will be based on the likelihood of a sample result, not on any activity which is determined by the prejudices of the experimenter. If the differences between the two groups are "statistically significant," we then conclude that the variable does have an effect and it can be included in our theoretical model. However, if the results are not statistically significant, it cannot be concluded that the variable does not have an effect. All we can say is that there is no evidence to include the variable in our model. This technique is appropriate for the slow process of theory building. However, an evaluator frequently does not have the chance for noncommitment that is given to the researcher. Given that the evaluator is forced to make a decision, failure to reject the null hypothesis will probably lead to the conclusion that the programs were equivocal. However, in doing this the rules of inference are being compromised. Classical procedures were not designed for the purposes of effective decision making, but rather for making assertions about population parameters (Edwards et al., 1963).

Secondly, when an evaluator is hired the employer wants to utilize all of the expertise available at that time. The classical approach does not allow for this since the researcher must be considered ignorant of expected outcomes--only the data can provide information. In reality the evaluator or other staff members should have considerable knowledge about the situation.

Finally, traditional statistics does not lend itself to the feedback system of evaluation. Inferences are strictly limited to the sample data under current consideration. Any additional tests will be based on new sample data. Frequently in evaluation, if conclusions cannot be clearly made on available evidence, more data, perhaps of a slightly different nature, will be examined. With the classical approach, the first set of data would be examined and tested independent of the second set. Thus, if proceeding through a feedback loop in an evaluation model, the test of the new data must be conducted and analyzed independent of the first set.

The conclusion to be drawn is that the classical statistical procedures utilized in research are ineffective in evaluation for the following reasons: (1) The classical approach does not allow for maximum utilization of the expertise available from the evaluator and/or project staff; (2) It can only reject the null hypothesis, never accept it; (3) The classical approach cannot effectively consider relative costs of Type I or Type II errors; (4) It cannot handle feedback loops, which increase the relative store of information.

III

BAYESIAN ANALYSIS

It should be emphasized that all of the mathematical concepts used in traditional statistics are available to Bayesian statisticians. However, some of the interpretations of the classical concepts are broader among Bayesians. Perhaps the most important of these concerns the issue of probability.

DEFINITIONS OF PROBABILITY

At least three separate "types" of probability can be distinguished:

1. A priori probability. This type of probability follows the branch of pure mathematics called the calculus of chances. It is not based on any data that have been gathered, but rather on logical assumptions that are made about the nature of events. It often relies on physical characteristics, such as symmetry. The toss of a coin or a die is a popular example of symmetry.

2. Statistical probability. Probabilities of this nature are estimated by observing the ration of the frequency of occurrence of some event to the total number of opportunities that are available for the event to occur. It is the relative frequency of a given kind of event or phenomena within a class of phenomena, usually called a "population." For example, to say that the probability of a particular child being born male is .52 means that of all known births, .52 of them have been males. This is a case where statistical probability may not correspond to *a priori* probability, since the *a priori* probability of a male being born is .50 according to the principle of indifference. Of course, there may be some explanation for this discrepancy which will change the *a priori* probability.

3. Subjective probability. The third type of probability is distinctly Bayesian and is perhaps the cornerstone of the differences between Bayesian and classical statistics. It is important to note that the Bayesian notion of subjective probability cannot be properly criticized on mathematical grounds, since it employs the same mathematical machinery as other concepts of probability.

> If Bayesian methods are to be criticized, the criticism should be based on their intuitive reasonableness and appeal, the reality of their assumptions about human behavior, their pragmatic value, their place within empirical science and so on. They are not properly criticized on mathematical grounds. . . . There is no question of the formal validity of the probability statements such methods yield, so long as we play within the formal rules of the game. [Hays, 1973, p. 810].

The notion of subjective probability can be defined as follows. A probability value is a measure of strength of an individual's opinion or belief about the existence of some situation or the occurrence of some event. This is distinctly different from a relative frequency probability. Edwards, Lindman, and Savage (1963) indicated that with personal probability you are saying something about yourself as well as the event you are trying to predict. For example, if I were told that the probability of getting a head in a coin toss is .5, I would understand that according to traditional ideas of probability, with a large number of flips, half of the results would be heads. However, if I were told that the coin to be used had either 2 heads or 2 tails, the traditional meaning of probability wouldn't operate, yet I would probably place bets in line with my former prediction of .5.

BAYES' THEOREM

Bayes' Theorem is essentially an algebraic relationship by which prior probabilities are revised in view of additional data to obtain posterior probabilities. This relationship is most useful in situations involving subjective prior probability distributions.

Let P(H) equal the probability of a hypothesis being true prior to data collection. This is defined as the prior probability. It is a subjective probability although it does not exclude the possibility of elements of a traditional frequency model. The statement "The probability that a person walking in my office is a college senior is .20," could be considered a subjective probability reflecting opinion prior to data collection. This statement may be based on a frequency concept of probability, if perhaps 20 percent of all persons on campus are seniors, or it may be more subjective, being based on the number of seniors that I know, past experience with people walking in my office, or even a nonspecific "gut" feeling.

According to McGee (1971), this kind of probability can also be expressed in terms of odds. The odds on a statement are determined by the probability of a statement divided by the probability of its denial. If you offer three to one odds in a bet on a football game, this is the same as saying you feel that your chance of winning the bet is .75. The odds you offer are usually based on both objective data, such as the frequency of your team winning, and subjective feelings about the physical condition of both teams, where the game is played, and many other factors.

Let P(D/H) equal the probability of the hypothesis being true after data collection. This is usually the most public probability since many modes are available which are widely accepted by researchers. This probability provides the <u>likelihood function</u>. Traditional studies are heavily based upon this type of conditional probability as reflected in the statement "What is the probability of obtaining a sample mean as great as X, given that the hypothetical population mean is u?"

$$\text{Let } P(D/H) = \frac{P(D \cap H)}{P\ (H)}$$

The probability (PD), or "the probability of the data" is of little intuitive interest and primarily servies a standardizing role. It is defined thus:

$$P(D) = P(D/H_1)\,P(H_1)$$

for each alternative hypothesis H_1.

A little algebra now leads to the basic form of Bayes' Theorem:

$$P(H/D) = \frac{P(D/H)\,P(H)}{P(D)}$$

P(H/D) is the probability that the hypothesis is true on the basis of both the initial probability P(H) and the experimental data.

Thus, in using Bayes' Theorem, a prior probability P(H) is formed based on any information available, including logic or intuition. Then data are collected from a sample and a likelihood function is formed [P(D/H)]. This information is then used to refine the prior probability which results in a posterior probability. Bayes' Theorem can be reformulated to apply to continuous parameters. If a parameter η has a prior probability density function $\mu(\eta)$, and if χ is a random variable for which $\nu(\chi/\eta)$ is the density of χ given η, and $\nu(\chi)$ is the density of χ, then the posterior probability density of η given χ is

$$\mu(\eta/\chi) = \frac{\nu(\chi/\eta)\,\mu(\eta)}{\nu(\chi)}.$$

FORMATION OF PRIOR PROBABILITY DISTRIBUTION [P(H)]

In classical statistical testing, the only source of relevant information about a population is the results gained from a sample. Of course certain assumptions are held about the basic population distribution, but nowhere does the prior opinion of the evaluator enter explicitly as it does in Bayesian analysis. Each sample is drawn as though it were the first of its sort ever taken.

If an important decision is to be made, logic would dictate the use of all relevant information. In reality, this is just what most evaluators do. While they collect their "hard" data, their intuition will predominate should the data turn out to be nonsupportive. Through rationalization, internal analysis of the data, or secondary findings, the evaluator will somehow let her/his subjective opinion interact with the statistical results. The Bayesian researcher attempts to let his/her intuition enter the decision making process through the prior probability distribution rather than through the back door.

By definition there are no explicit rules for determining the shape of the prior probability distribution. However, as noted by Edwards et al. (1963) there are times when the shape of the initial distribution will be very important in the final outcome:

1. If small prior probabilities are initially assigned to areas where the data indicate the parameter is located.
2. If a large probability is assigned to a region where the data are nonsupportive.

3. If both experimenter's prior opinion and the data are diffuse.
4. If observations are expensive and relatively few can be made.
5. If major decisions have to be made prior to the collection of much data, such as sample size.

AN EXAMPLE OF BAYES' THEOREM

Consider a situation where an evaluation team needs an estimate of the Wechsler I.Q. for a school district when no test results are readily available. The evaluation team should utilize teacher comments, observational data, achievement test results, and any other information that is available. The team may then meet together and collectively decide the shape of the probability distribution.

In the absence of any information, the best estimate of the mean I.Q. may form a normal distribution with a mean of 100 and a standard deviation around 15. That is, if they were to know nothing about the school district, they would probably assume that it is average. The probability of an average school district having a mean I.Q. above 130 is not very likely, in their opinion perhaps no more than two or three percent.

However, one of the staff members had reading scores that indicated these students were a year behind in reading. Also, several teachers remarked that the students seemed rather "slow." Because of this additional information, the evaluation team may decide that a positively skewed distribution with a mean of 95 best describes the group opinion. Although some information implies these students may be below normal, such information is not totally reliable and in fact the students might be normal. Therefore, even though the group opinion is that the mean I.Q. score is most likely around 95, they may feel the chances are greater that it is above 95 than below 95. A positively skewed distribution would reflect this feeling.

Assume for the moment that the decision has been made to actually test a sample of children with Wechsler Scale. This information would then be incorporated into the prior probability distribution to form the posterior opinion. The technique used to accomplish this would be Bayes' Theorem.

An example from Hays (1973), will help to solidify what has been presented. Suppose this time the evaluation team is concerned about the academic ability of a single child. There is general agreement among psychologists that I.Q. is best represented as forming a normal distribution with a mean of 100. Consider the following distribution:

I.Q.	PROBABILITY (H)
130 - greater	.02
115 - 129	.14
100 - 114	.34
85 - 99	.34
70 - 84	.14
less - 69	.02

This is the prior probability distribution P(H) for this example. It is a subjective probability distribution but there is widespread agreement about its shape. In this situation, it may be the distribution for the Wechsler.

Suppose that a two-item test is given to a large group of children. The distribution of test scores given each level of I.Q. is obtained. This would be a conditional probability distribution showing the distribution of I.Q. scores for each score on the two-item test. This would provide P(D/H), that is, "What is the probability of an experimental outcome D, given the prior probability distribution H?" To determine this, distribution scores would be needed for both I.Q. and for the two-item test.

I.Q.	PROBABILITY (D/H)		
130 - greater	.05	.15	.80
115 - 129	.15	.35	.50
100 - 114	.30	.30	.40
85 - 99	.40	.30	.30
70 - 84	.50	.35	.15
less - 69	.80	.15	.05
TEST SCORE	0	1	2

Each column represents a conditional distribution of test scores given intelligence level. Suppose a child is brought in and the probability that his I.Q. is over 130 must be quickly determined. Without any information about the child, the probability would be .02, based on our prior information about the distribution of intelligence. Now suppose that the child is given the brief test. If the child scored 2, the probability of an I.Q. over 130 is .80 on the basis of this test information alone. However, a Bayesian would be uncomfortable with a probability this high, since from previous experience very few people have an I.Q. over 130. The score of a single test should be considered, but it should be tempered with prior feelings about the rarity of the event. This is done through the use of Bayes' Theorem, as presented earlier:

$$P(130/2) = \frac{(.80)(.02)}{(.80)(.02)+(.50)(.14)+\ldots+(.05)(.02)} = \frac{.016}{.346} = .046$$

Using the same procedure, other values would result in the following posterior distribution for those children scoring 2 on the test.

I.Q.	PROBABILITY (H/D)
130 - greater	.046
115 - 129	.202
100 - 114	.393
85 - 99	.295
70 - 84	.061
less - 69	.003

By comparing the posterior probability distribution P(H/D) with the prior probability distribution P(H), one can see the effect of the additional information. I.Q. values greater than 100 have a higher probability than with the prior distribution. The probability of an I.Q. score greater than 130 has increased from .02 to .046. However, the change has not been nearly so great as if the data alone were considered. Thus, for a Bayesian, if an event is seen as rare, one set of data will usually not drastically change the probability of that event's occurrence. With the Bayesian model, this posterior probability distribution can be considered a prior probability for another round of data collection. This process can continue, with each set of data being integrated with the prior distribution. Unlike the classical approach, no set of data can stand alone.

HYPOTHESIS TESTING

In addition to the notion of subjective probability, the Bayesian position on hypothesis testing would also elicit great negative response from classical statisticians. The Bayesian position is best reflected in a statement by D.L. Meyer:

> Our teaching must be revolutionized to the point where topics such as confidence intervals and tests of significance are taught almost as an afterthought for those fortunate enough to have a formal model. In discussing the first course in statistics with the staff of a certain University, they told me that descriptive statistics is taught in only two weeks so that "we get to the important stuff--inference--quickly." I am proud to report that at Syracuse, we spend a full semester on the important stuff--descriptive statistics (1971, p. 4).

So rarely do Bayesians have reason to use inferential tests that a discussion of Bayesian statistics could be relatively complete without bringing up the issue. However, because it is so important in classical statistical thought it cannot be avoided entirely.

The most popular notion of a test is a tentative decision between two hypotheses on the basis of data. This usually leads to a choice between two actions, such as whether to include a supplementary reading program in the curriculum or not. With this purpose in mind, the usefulness of the null hypothesis is questionable since the null hypothesis can be expected to be false from the beginning (Edwards, 1963). The differential treatment of two groups, regardless of the nature of the treatment, is usually sufficient to result in some difference in scores on the dependent variable (Edwards et al, 1963). This is probably magnified in educational environments, where placebo effects are frequently significant.

Of course more is at stake than simply choosing between two options, since economic matters will enter in determining what activity is actually followed. One would hardly choose to implement an expensive reading program unless there were clear evidence of the likelihood of substantial improvement. The whole issue of statistical versus educational significance makes clear the hypothesis-testing procedure is taken only half seriously.

This problem is avoided by the Bayesians since hypothesis testing does not hold any particular relevance. That particular probabilities such as .01 or .05 are highlighted for special treatment is without special interest to the Bayesian. In deciding between two alternatives, simply reviewing their posterior probabilities will usually reveal the appropriate decision. Any further analysis is likely to be based upon utility values (in the language of decision theory) rather than on a particular probability value.

For example, if two math programs are being compared and both involve the same economic factors, then the program that seems to be better should be adopted, however slight the superiority. If one program is more expensive than the other, then a payoff function may be applied to account for cost differences. If one program is markedly more expensive than the other, then it will only be adopted if there is clear evidence of its superiority. A traditional significance level of .01 may not be enough to justify the relative difference in cost. Finally, even if the relative costs are the same, if one program appears to have a stronger theoretical foundation, then we would require rather convincing evidence to justify adoption of the alternative program. Bayesian analysis, through the use of the prior probability, can account for this prejudice; traditional inference cannot.

As noted earlier, although Bayesian statistical inference does not require the testing of hypotheses in the traditional sense, testing can occur since the probability laws that apply in classical statistics also apply to Bayesian analyses, only the meaning is different. Therefore, the same testing techniques available to classic inference can be legitimately used by Bayesians. This difference in meaning, however, can lead to situations where a classical statistician may reject a null hypothesis while a Bayesian will see the same data providing support for the null hypothesis. Edwards et al. (1963) developed this argument in detail; it will not be reviewed in its entirety here. However, the intuitive argument can be made more quickly.

If a true null hypothesis is being tested under a one-tailed t-test with a large sample, a t-value between 1.68 and 1.96 will occur two percent of the time. Of course, if the null hypothesis is false, the frequency of this outcome will depend on the alternative hypothesis. However, given the condition of no prior knowledge as specified with classical inference, a uniform distribution of alternative hypotheses with t-values between 0 and 20 is not unreasonable. Under such a condition, and assuming the null hypothesis is false t will fall in the range from 1.68 to 1.96 at most 1.40% of the time (i.e., 1.96-1.68/20 = 1.40%). Thus a t-value would occur in this interval more frequently under the condition of a true null hypothesis than under a diffuse alternative hypothesis, despite the fact that the classical statistician would reject the null hypothesis. Obviously, any answer to this argument depends upon the shape of the prior distribution of the alternative hypothesis. The classical statistician cannot handle this problem since he does not consider the shape of such a distribution.

To continue with the current discussion would lead to theoretical issues beyond the scope of this paper. However, the point must be made that although Bayesian statistical analyses do allow for the use of classic testing procedures, sometimes the outcomes will not coincide. Since Bayesians usually hold that traditional testing procedures are simply a nominal statistical exercise, the question of how Bayesians make decisions remains.

DECISION MAKING

Upon reflection, the classical approach of rejecting a sharp null hypothesis doesn't really provide much information. As anyone familiar with information theory is aware, the most knowledge is gained by reducing the number of options in half. Given the stochastic nature of the world, one can obviously not eliminate 50 percent of all alternatives with certainty. However, one should be able to do better than simply reach a probabilistic conclusion that the real value of some parameter does not appear to be at one point along an infinite range of possible values.

This no doubt was the feeling of McGee (1971) when he noted:

> The author had for some years felt the need to consider alternatives to setting up a null hypothesis in order to reject it and, having considered the position of the Bayesians, was a ready convert at an intensive summer session in information theory. . . . The idea of an experimenter approaching a statistician with the request "I want to reject the null hypothesis" was totally out of place [p. 277].

An example of the Bayesian approach to estimating the probability that an I.Q. value was greater than 130 was presented earlier. Now consider a situation where a decision between two alternative math programs must be made. The traditional approach would be to set up a null hypothesis of no difference on some test score. The alternative hypothesis would be that there is a difference between the two programs. A two-tail test is likely to be employed. This is a typical kind of problem found in educational evaluation.

An obvious difficulty can arise in this kind of situation. Suppose a nonsignificant _t_-value were obtained. There would then be no particular reason for favoring one program over the other and a toss of a coin would be an appropriate decision maker. Obviously, the evaluator is given license to draw any conclusion. She/he could simply choose the program which resulted in the slightly higher mean test score and use this score as partial justification. This is often done when one hears the report that "means were in the right direction" or "results were just short of significance." According to the classical rules, these statements carry as much weight as saying the coin almost landed tails. Clearly, with nonsignificant results, the evaluator cannot reach any other conclusion than to assume no difference for the moment. Suppose she/he plays the game and doesn't draw conclusions on the basis of the data. She/he is then likely to appeal to other items, such as cost effectiveness, theoretical soundness, or teacher preference.

When the evaluator begins to look at this kind of evidence, she/he is stepping outside of the public system of inference and becomes open to criticism. While heterogeneous and frequently qualitative data are useful, an evaluator trained in classical statistics has difficulty in dealing with such information in a systematic way. Because she/he has difficulty explaining the procedure she/he used to make decisions, the evaluator is frequently criticized for being biased or subjective. By making public the process whereby she/he used these other sources of information, the Bayesian evaluator reduces the chances that she/he will be criticized for personal bias, fuzzy thinking, or irrational decisions. Others may disagree with her or his prior probability distribution, but the way that the research data and subjective opinion were brought together cannot be disputed, assuming acceptance of the Bayesian paradigm.

Bayesian Analysis for an Evaluation Problem--An Example

Math Program Y has recently been developed to replace the standard Math Program X. School District No. 32 wants to consider adoption of Math Program Y becuase it appeals to a community which considers itself progressive and innovative. Not much money is available for data collection, so the evaluation will have to utilize a small sample. These evaluators, who have considerable knowledge of math education programs, are not enthusiastic about Math program Y. They feel the program has serious flaws and has not been adequately developed.

The evaluators, along with school district personnel, have agreed to use the scores on the Standard Math Achievement Test to help with the judgment. A lot of data have been collected in the past from schools using Program X. However, the data are old, since the program has not been heavily used in recent years. The national norms for schools using Program X are available on the Standard Math Achievement Test and are as follows:

Score	Distribution
15	.00
14	.10
13	.10
12	.10
11	.25
10	.25
9	.10
8	.10
7	.00
6	.00
5	.00
4	.00
3	.00
2	.00
1	.00

In addition to the old data, the following information is also available:

1. School District No. 32 is located in an upper middle class area where the parents of the students are engaged in skilled or professional occupations.
2. Math Program X has been recently modified, and according to some sources, probably would result in better test performance.

On the basis of all this information, the task is to determine where the value of u_x, the mean score on the Standard Achievement Math Test, lies for School District No. 32 using Program X. The prior probability distribution for Program X is presented in Table 1 along with the rest of the data needed for this example.

TABLE 1

PRIOR, DATA, AND POSTERIOR PROBABILITY DISTRIBUTIONS
FOR MATH PROGRAM EXAMPLE

Mean Test Score	Prior Distributions X	Prior Distributions Y	Data Distributions X	Data Distributions Y	Posterior Distributions X	Posterior Distributions Y
15	.20					
14	.25					
13	.25			.05		
12	.15	.10		.90		.95
11	.10	.10	.05	.05	.263	.05
10	.01	.20	.90		.474	
9	.01	.20	.05		.263	
8	.01	.10				
7	.01	.10				
6	.01	.10				
5		.10				
4						
3						
2						
1						

The prior probability distribution for Program Y is almost totally subjective since little data are available. The lack of enthusiasm for the program by the evaluation team is revealed by the prior probability distribution.

Test data are obtained for both Program X and Program Y. The mean score for Program X is 10 and for Program Y is 12. However, since there may be measurement error in the test data, some credibility is given to the scores next to the obtained mean (.05 for each score). The posterior distribution has been determined by Bayes' Theorem. The results are somewhat similar to those obtained with traditional statistics, which would rely on the data alone. However, confidence in the true value of the mean for Program X is weak. The evaluators are only 47 percent certain that the value is 10, while they are quite certain (95 percent) about the true value for Program Y. This is because the data tend to go considerably counter to their prior feelings about the true value for Program X. Perhaps more data for Program X would be appropriate.

These findings illustrate the generalization that when the prior probability distribution and the data distribution are dissimilar, the posterior probability distribution will be diffuse. This is logical since data that go against common sense will usually result in greater uncertainty about the real nature of events than data that support prior beliefs. In classical statistics, conclusions should be unaffected by prior opinion.

DATA COLLECTION STOPPING RULE

A final distinction between the two inference techniques lies in their different treatment of data stopping rules. Classical procedures require the experimenter to specify in advance how much data she/he will collect. This principle is outlined in most texts on general statistics (e.g., Hays, 1963). Such a requirement is part of an overall rule by classical statisticians to specify all data collection and data analysis activities prior to any actual data gathering, with the exception of certain post hoc tests. This principle is extremely difficult for an educational evaluator to follow, since his working environment is very fluid; initial specification of all data-gathering procedures is usually impossible. The classical statistician is so specific on this point because of the ease with which a null hypothesis can be rejected. In fact, with repeated cycles of data gathering and testing, an experimenter could be certain of rejecting the null hypothesis even if it were true (Hays, 1963).

However, as Edwards et al. (1962) noted, this is not a problem for a Bayesian who does not use the null hypothesis testing procedure.

> In contrast, if you set out to collect data until your posterior probability for a hypothesis which unknown to you is true has been reduced to .01, then 99 times out of 100 you will never make it, no matter how many data you, or your children after you, may collect [p. 239].

The cornerstone of this difference between classical and Bayesian data collection procedures is known as the "likelihood principle." This principle flows directly from Bayes' Theorem and the concept of subjective probability. It is in operation when two different experimental outcomes (x and y) have the same bearing on opinion about a parameter. That is, if $P(x/\lambda)$ and $P(y/\lambda)$ are proportional functions of λ, then each of the two data x and y have exactly the same thing to say about values of λ.

In the discrete case, if $P(D'/H_1) = kP(D/H_1)$ for some positive constant k, then the likelihood principle operates. For example, in a coin toss, 10 heads out of 20 throws means the same thing as 20 heads out of 40 throws. This simple principle was discussed by classical statisticians such as Fisher (1956). However, in classical testing the principle is lost, according to Savage (1962):

> The likelihood principle is in conflict with many historically important concepts of statistics. For example, whether a test is unbiased depends not on the likelihood alone, but rather on $Pr(x/\lambda)$ considered as a function of x as well as a finction of λ. Similarly with the concepts of significance or confidence level. For instance, it has been widely believed that the import of such a datum as 6 red-eyed flies out of 100 depends on whether the experiment was designed to observe 100 flies or designed to observe 6 red-eyed flies. An estimate unbiased for either of these experiments is biased for the other [p. 17].

Since stopping rules are irrelevant to a Bayesian, greater objectivity actually results than with the traditional model. Once data are collected, the original intentions of the experimenter are irrelevant. The experimenter can collect data until he has proven his point or exhausted all his funds, time, or patience.

Such freedom should be appealing to an educational evaluator who is frequently under pressures that interfere with a predetermined plan of data collection. Frequently a shortage of funds, uncooperative teachers, or pressures of time prevent data collection from being completed. Left with incomplete data, most evaluators continue to grind out inferential tests despite gross violations of principles of classical inference. Frequently these violations go unnoticed or else are rationalized as being necessary to meet the demands of the real world. Violations of assumptions do not necessarily reflect negatively on the practicing evaluator, for statistical models should reflect reality, rather than force reality to reflect the statistical model. If the stopping rule principle is violated so frequently, then the Bayesian model, which disregards the stopping rule, may be more appropriate for evaluation situations.

IV

CONCLUSION

At this time, a textbook describing Bayesian analyses for educational problems does not exist, although there are several texts slanted toward other disciplines which should prove useful to educators (e.g., Morgan, 1968; Zellner, 1971).

One should not assume that less mathematical rigor is required in Bayesian analysis than in classical statistics; in reality, the opposite is true. Classical statistics, with its emphasis on the normal underlying distribution, has been documented so well that students with little mathematical sophistication can perform adequate statistical analyses. The Bayesian model, however, requires a good understanding of distribution theory in order to adequately describe the prior probability distribution. Some simple analyses using discrete distributions may be within the reach of almost anyone; however the full richness of the Bayesian approach may not be appreciated without some mathematical sophistication.

Most evaluators can take some steps to begin to utilize a more Bayesian approach in their daily work. Even if the Bayesian model is not completely accepted, the questions raised by the Bayesians should increase the vigilance of the evaluator to avoid gross violations of the classical model. Frequently, when a model in a given area is widely used, the assumptions underlying the model are taken for granted. (One needs only to look at the 19th century Newtonian physicists or early 20th century behaviorists to see this problem.)

The following recommendations are made and should be easy to implement:

1. Don't parade statistical procedures in an attempt to add respectability to a subjective process. When an evaluator or any scientist attempts to generalize beyond his data, he is engaging in a subjective process (Edwards et al., 1962). The mathematical models may be useful, but they do not automatically objectify any inferential process.

2. Realize the ease with which the null hypothesis is rejected. Just because one has been able to reject a null hypothesis at the .01 or .05 level does not necessarily mean that something of educational significance has been found. As noted by Savage (1962), null hypotheses are frequently rejected inappropriately, and even if appropriately used little of any practical significance can be concluded from the rejection of single null hypothesis.

3. Report probability levels when possible. Instead of using the magical .05 or .01 levels of null hypothesis rejection, the actual probability levels for the alternative hypothesis should be reported. Most evaluators are familiar with power and power functions, but they are rarely discussed beyond a first course in statistics.

4. Specify prior opinion. Since most activities in evaluation involve hypotheses where the evaluator is not neutral, prior opinion and the reasons for this opinion are legitimate information. Instead of using this information covertly when drawing conclusions, openly expressing the initial bias may be more appropriate. If expressed in probability terms, Bayes Theorem could be applied to revise such opinion. Conventional inferential analyses could still be performed if desired.

5. Remember the data stopping rule. If the evaluator is determined to test a sharp null hypothesis, the size of the sample must be specified in advance and sequential testing of data must be avoided. Also, the significance level should be determined prior to data collection.

V

SUMMARY

Numerous arguments have been presented in support of the appropriateness of the Bayesian model for the educational evaluator. For the most part the evaluator is engaging in appropriate practices, but the classical statistical model does not describe how she/he really works. By changing models, the evaluator will be able to continue doing what she/he already does, yet she/he will be better able to explain to others how it is done.

This does not mean that the use of Bayesian model will not require any changes in practice. Such a model requires greater specificity in situations where no rules existed, such as expression of bias towards one program or the other prior to data collection. However, such changes should result in a new sense of freedom since the evaluator can admit that numerous types and sources of data resulted in her/his decisions and at the same time she/he can stay within the limits of a credible statistical model.

REFERENCES

Bayes, T. Essay towards solving a problem in the doctrine of chances. The Philosophical Transaction, 1963, 53, 370-418. (Reprinted from Biometrika, 1958, 45, 293-315.

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. Psychological Review, 1963, 70(3), 193-242.

Fisher, R. A. Statistical methods and scientific inference. Edinburgh: Oliver and Boyd, 1956.

Hays, W. L. Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

McGee, V. E. Principles of statistics: Traditional and Bayesian. New York: Appleton-Century-Crofts, 1971

Meyer, D. L. Bayesian statistics. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971. (ERIC No. ED 051 261).

Morgan, B. W. An introduction to Bayesian statistical decision processes. Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Savage, L. J. The foundations of statistical inference. London: Methuen and Co., 1962.

Schlaifer, R. Introduction to statistics for business decisions. New York: McGraw-Hill, 1961.

Stufflebeam, D. L. An introduction to the PDK book: Educational evaluation and decision-making. In B. R. Worthen and J. R. Sanders (Eds.), Educational evaluation: Theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

Thompson, G. E. Statistics for decisions: An elementary introduction. Boston: Little, Brown, & Co., 1972.

Weber, J. D. Historical aspects of the Bayesian controversy. Tucson, Ariz.: University of Arizona Press, 1973.

Worthen, B. R., & Sanders, J. R. Educational evaluation: Theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

Zellner, A. An introduction to Bayesian inference in econometrics. New York: John Wiley and Sons, 1971.

National Evaluation Committee

Francis S. Chase, Chairman
Emeritus Professor
University of Chicago

Helen Bain
Past President
National Education Association

Lyle Bourne
Professor
University of Colorado

Sue Buel
National Evaluation Committee

Roald F. Campbell
Emeritus Professor
The Ohio State University

George E. Dickson
Dean, College of Education
University of Toledo

Larry R. Goulet
Professor
University of Illinois

Chester W. Harris
Professor
University of California - Santa Barbara

William G. Katzenmeyer
Professor
Duke University

Barbara Thompson
Superintendent of Public Instruction
State of Wisconsin

Joanna Williams
Professor
Teachers College
Columbia University

University Advisory Committee

John R. Palmer, Chairman
Dean School of Education

William R. Bush
Deputy Director
R & D Center

David E. Cronon
Dean
College of Letters and Science

Diane H. Eich
Specialist
R & D Center

Evelyn L. Hoekenga
Coordinator
R & D Center

Dale D. Johnson
Associate Professor
Curriculum and Instruction

Herbert J. Klausmeier
Member of the Associated Faculty
R & D Center

James M. Lipham
Member of the Associated Faculty
R & D Center

Wayne R. Otto
Associate Director
R & D Center

Richard A. Rossmiller
Director
R & D Center

Elizabeth J. Simpson
Dean
School of Family Resources
and Consumer Sciences

Len Van Ess
Associate Vice Chancellor
University of Wisconsin - Madison

Associated Faculty

Vernon L. Allen
Professor
Psychology

B. Dean Bowles
Professor
Educational Administration

Thomas P. Carpenter
Assistant Professor
Curriculum and Instruction

Marvin J. Fruth
Professor
Educational Administration

John G. Harvey
Professor
Mathematics
Curriculum and Instruction

Frank H. Hooper
Professor
Child Development

Herbert J. Klausmeier
V.A.C. Henmon Professor
Educational Psychology

Joseph T. Lawton
Assistant Professor
Child Development

Joel R. Levin
Professor
Educational Psychology

L. Joseph Lins
Professor
Institutional Studies

James M. Lipham
Professor
Educational Administration

Donald N. McIsaac
Professor
Educational Administration

Gerald Nadler
Professor
Industrial Engineering

Wayne R. Otto
Professor
Curriculum and Instruction

Robert G. Petzold
Professor
Music
Curriculum and Instruction

Thomas S. Popkewitz
Assistant Professor
Curriculum and Instruction

Thomas A. Romberg
Professor
Curriculum and Instruction

Richard A. Rossmiller
Professor
Educational Administration

Dennis W. Spuck
Assistant Professor
Educational Administration

Michael J. Subkoviak
Assistant Professor
Educational Psychology

Richard L. Venezky
Professor
Computer Sciences

J. Fred Weaver
Professor
Curriculum and Instruction

Larry M. Wilder
Assistant Professor
Child Development