International Institute for
Applied Systems Analysis
IIASA www.iiasa.ac.at

Bayesian Statistical Analysis of Experimental Data

Winkler, R.L.

IIASA Research Report
September 1973

RR-73-4

BAYESIAN STATISTICAL ANALYSIS OF EXPERIMENTAL DATA

Robert L. Winkler

September 1973

Bayesian Statistical Analysis of Experimental Data*

Robert L. Winkler**

1. Introduction

Bayesian statistics has received a considerable amount of attention in the past two decades. For instance, statistical journals have included numerous articles on Bayesian methods in recent years. Several books regarding Bayesian statistics have been published recently (some references will be given in Section 5), and many general statistics texts now include one or more chapters on Bayesian statistics. Moreover, the interest in Bayesian methods is not limited to mathematical statisticians. Primarily because of their implications for decision making, these methods have received much attention in business schools. Psychologists concerned with human behavior in inferential and decision-making situations have used Bayesian methods extensively. Economists have used Bayesian methods to compare economic models and to develop models of rational economic behavior. Other interesting applications have involved medicine, law, meteorology, and many additional areas.

Formally, Bayesian statistics consists of a set of statistical procedures that involve the use of Bayes' theorem to revise probabilities as new information is obtained.[1]

---

*This paper will appear in the Proceedings of the Fourth Research Conference on Subjective Probability, Utility, and Decision Making, Rome, 1973. This research was supported in part by the U.S. National Science Foundation under Grants GA-31735 and GA-41232.

**Graduate School of Business, Indiana University, U.S.A.; research scholar at the International Institute for Applied Systems Analysis, Laxenburg, Austria.

The term "Bayesian statistics" covers a wide spectrum of topics, and this paper is concerned with only a portion of that spectrum, the use of Bayesian procedures in the analysis of experimental data. In order to carefully delineate the area of interest, it is first useful to distinguish inferential procedures from decision-making procedures. The motivation for much of the recent interest in Bayesian methods has been decision-theoretic in nature. These methods are adaptive in the sense that they allow for the revision of probabilities on the basis of new information, and thus they provide a useful framework for decision making models. In decision making, the ultimate objective is the choice of an action from a particular set of alternatives. The objective of inferential statistics, on the other hand, is not to choose an action, but simply to make inferences about some population or process on the basis of limited information concerning that population or process. Of course, the distinction between inference and decison making is often somewhat vague. For instance hypothesis-testing procedures can be thought of as inferential in nature, or they can be considered in a decision-making framework. Nevertheless, the main concern in the analysis of experimental data is generally inference, and this paper is thus oriented primarily toward the inferential end of the inference-decision spectrum.

In order to further clarify the approach taken in this paper, it is useful to distinguish between statistical theory and statistical practice. The <u>theory</u> of Bayesian statistics

has received considerable attention in terms of underlying foundational matters, the development of Bayesian procedures, and the comparison of these procedures with most of the commonly encountered inferential procedures of classical statistics. In terms of statistical practice, however, the interest in Bayesian methods has been almost exclusively from a decision-making standpoint. Little actual use has been made of Bayesian inferential procedures except insofar as they are useful in a decision-making context. In this paper, no new theoretical developments are presented; the emphasis is on statistical practice rather than statistical theory.

This paper, then, is concerned with the use of Bayesian procedures in the analysis of experimental data. In Section 2, current statistical practice with regard to the analysis of experimental data is investigated and criticized, and the factors influencing current practice are discussed briefly in Section 3. The analysis of experimental data is approached from a Bayesian standpoint in Section 4, and Section 5 contains a brief concluding discussion.

2. Current Statistical Practice

Although there are certainly some variations caused by different types of problems and approaches in different areas of application, it seems that the general approach to the analysis of experimental data in practice is similar across areas. Therefore, although some specific references are made in this section to a particular field, that of experimental psychology,

the discussion applies not just to that field, but to the analysis of experimental data in general. Experimental psychology is of special interest because it appears that psychologists have been exposed to much more material concerning Bayesian methods than have researchers in other areas where experimental data are frequently collected and analyzed. Beginning with Edwards, Lindman, and Savage [14], numerous articles concerning Bayesian procedures vis-a-vis classical procedures have appeared in psychological journals (e.g. Edwards, [13], Bakan, [1]; and Wilson, Miller, and Lower, [48]), and some psychologists are deeply concerned about the philosophical implications of various statistical procedures (e.g. Meehl, [27]). Moreover, many experimental psychologists are actively involved in research in the area of human behavior in inferential and decision-making situations, and Bayes' theorem is frequently used as a normative standard of comparison (e.g. see Slivac and Lichtenstein, [47]).

Despite the fact that numerous psychologists are familiar with the Bayesian approach, the vast majority of analyses of experimental data in the field of experimental psychology are classical in nature. Of course, some Bayesian analyses can be found (e.g. Beach and Phillips, [2]; Pitz, [32], [33]), but they are the exception rather than the rule. In addition, there are papers where no inferential technique is explicitly stated and where the results are reported by a (hopefully judicious) choice of descriptive statistics; papers of this nature may reflect a trend toward simpler descriptions of data in reporting

experimental results. Of course, this most often occurs when the results are fairly obvious and the choice of inferential procedures is more or less irrelevant.

The area of experimental psychology is not unique. In general, regardless of the area of application, Bayesian analyses of experimental data are seldom encountered in the published literature. Moreover, as viewed from a Bayesian vantage point, the situation is even worse than the lack of Bayesian analyses suggests. When classical procedures are used carefully and appropriately, they can sometimes be given a Bayesian interpretation, and Bayesian procedures often represent an extension of classical procedures instead of a completely unrelated set of procedures (e.g. see Pratt, [36]).[2] It appears, however, that in practice the classical procedures are frequently not used carefully and appropriately and that within the classical school of thought there is a large gap between theory and practice. This gap will be examined here by considering a particular type of procedure, tests of sharp null hypotheses.

Many (perhaps most) statistical analyses of experimental data that are reported in the published literature involve tests of sharp null hypotheses, and the reporting of these tests is frequently limited to presenting a significance level. Unfortunately, this is exactly the type of procedure that careful analysis reveals to be extremely suspect. Because of the widespread use of such procedures, it is useful to examine them in more detail here, although space limitations prevent a thorough discussion.

First, consider the nature of the hypotheses frequently encountered in practice. The "null" hypothesis is usually an exact hypothesis, such as the hypothesis that a population mean, μ, exactly equals a particular value, μ_0. Very seldom is it reasonable to think that such a hypothesis is exactly true. In general, the experimenter is really interested in whether μ is close to μ_0, where the notion of closeness differs from situation to situation. Thus the question of interest in the experimental situation is not adequately reflected by the choice of a statistical hypothesis. If it is recognized that the statistical hypothesis is only a rough approximation of the real hypothesis of interest, a careful analysis can allow in part for the degree of approximation and can produce results that make some sense in terms of the real hypothesis. Alternatively, the statistical hypothesis can be altered so that it includes an interval of values (e.g. $\mu_1 < \mu < \mu_2$) instead of just a single value. Such hypotheses can be handled within the classical framework, although the analysis requires a little more time and effort on the part of the experimenter than is the case with a sharp null hypothesis. In practice, unfortunately, the sharp null hypothesis is often taken at face value, and the resulting inferences provide answers to the wrong questions.

Next, consider the reporting of tests of sharp null hypotheses in terms of a single significance level. It is common to present just a statistic such as a t-statistic or an F-statistic, along with a statement such as "significant at the .05 level." This approach emphasizes only one of the two types

of errors. For example, consider a very simple situation in which both the null hypothesis and the alternative hypothesis are exact hypotheses. For a given experiment, the probabilities of Type I and Type II errors, α and β, can be computed for each possible choice of a rejection region. The ultimate choice of a rejection region should depend on a trade-off between these two types of errors, but in practice only one type of error is usually taken into consideration. In terms of more complicated situations involving sharp null hypotheses such as $\mu = \mu_0$ and inexact alternative hypotheses such as $\mu \neq \mu_0$, classical statistical theory provides power curves, operating characteristic curves, error curves, and so on, to enable the statistician to take both types of errors into consideration. In practice, these curves are seldom encountered.

The combination of testing a sharp null hypothesis against a two-tailed alternative hypothesis and reporting only a significance level has very unfortunate implications. For example, consider a test of $\mu = \mu_0$ versus $\mu \neq \mu_0$, where μ is the mean of a normally-distributed population with known variance σ^2. The sharp null hypothesis is rejected if the sample mean falls outside of the interval from $\mu_0 - z\sigma/\sqrt{n}$ to $\mu_0 + z\sigma/\sqrt{n}$, where z depends only on the choice of a significance level. But as n increases, the interval becomes narrower and narrower, implying that the rejection region becomes larger and larger. In essence, as the sample size increases, the test becomes more and more sensitive to small deviations of μ from μ_0. In theory this is fine, if such small deviations are of interest. In

most situations, however, the question of interest is whether μ is close to μ_0, not whether μ is equal to μ_0. Thus, ironic as it may seem, a very large sample size provides more precision than is necessary, and blind adherence to a particular significance level means that the null hypothesis is almost certain to be rejected. This procedure, which is very common in practice, provides the right answer to the wrong question; by taking a large enough sample, one can be virtually certain of rejecting a sharp null hypothesis that no one really believed was exactly true in the first place. Moreover, this will be true even if the experimental data strongly support the hypothesis that the parameter of interest is close to the particular value of interest. This general problem was noted over three decades ago by Berkson [3] in the context of tests of goodness-of-fit; for more recent discussions, see Lindley [23] and Jeffreys [20].

Another difficulty with the practice of testing sharp null hypotheses and reporting only significance levels relates to the distinction between a sampling distribution and a likelihood function. If θ is the parameter of interest and y represents the data, then the conditional distribution of y given θ is a sampling distribution. Significance levels in classical hypothesis testing correspond to areas under sampling distributions. But a sampling distribution involves a fixed θ and variable y, whereas a likelihood function involves a fixed y (the observed y from the experiment) and a variable θ. For a likelihood function, the entire distribution of y given a

particular θ is not of interest. Instead, one considers the conditional probability (or density) of y given θ, evaluated at the observed value of y. This particular conditional probability (density) is the likelihood corresponding to the particular θ, and by finding such a probability (density) for all possible values of θ, one generates a likelihood function. The likelihood principle states that the entire evidence of the sample with respect to inferences about θ is contained in the likelihood function (see Birnbaum, [4]). In an investigation of a sharp null hypothesis and an alternative hypothesis, the use of a sampling distribution to determine a significance level completely ignores values of θ other than the value specified by the exact null hypothesis. The likelihood function, on the other hand, considers all possible values of θ and ignores values of y other than the observed value on the grounds that inferences should depend on the observed experimental data, not on data that might have been observed but were not. It must be stressed here that some classical procedures based on the likelihood function have been developed, and once again the problem is in part one of theory versus practice rather than classical statistics versus Bayesian statistics.

The discussion in this section has dwelled upon a single type of procedure, the testing of sharp null hypotheses by reporting significance levels. This is admittedly more susceptible to criticism than many other procedures, but it is also the type of analysis that is most frequently encountered in practice. As noted in Section 1, this paper is concerned

more with statistical practice than with statistical theory. Furthermore, the primary interest here goes beyond differences between classical statistics and Bayesian statistics to the more general question of "good statistics" versus "bad statistics." In this regard, it should be emphasized that classical methods do not have a monopoly with respect to the problem of misuse. It is certainly possible for Bayesian methods to be used inappropriately. As will be seen in Section 4, however, there is generally a more direct relationship between the questions of interest in reality and the questions attacked by Bayesian methods than is the case with classical methods. Therefore, it might be hoped that Bayesian methods would be less subject to misuse. Because of the scarcity of Bayesian analyses appearing in the literature, insufficient evidence exists at the present time regarding the extent of the misuse of Bayesian procedures in practice.

As noted earlier in this section, Bayesian procedures sometimes represent an extension of classical procedures instead of a completely unrelated set of procedures. The extension lies in the inclusion of prior information, and arguments concerning the inclusion or exclusion of such information are primarily philosophical in nature. The mathematics of Bayesian procedures are not in dispute; the issues involved are more foundational in nature. The discussion of scientific reporting in Section 4 will touch on a few of these important issues, and more detailed discussions can be found in Savage [41],

[42], Kyburg and Smokler [21], Cornfield [7], and de Finetti [8], [9].

3. Factors Influencing Current Statistical Practice

In the previous section, current statistical practice with regard to the analysis of experimental data was criticized. Statistical theory provides sound techniques for making inferences from experimental data, and some of these techniques will be discussed in Section 4. Why, then, do experimenters often use such weak, poorly-justified techniques to analyze their data? In other words, what causes the apparent gap between theory and practice in statistics?

The theory-practice gap appears to be due to a combination of factors, including tradition, statistical training, lack of availability, computational difficulties, reporting difficulties, and perceived resistence by journal editors. Roberts [40] writes as follows:

> There is no shortage of possible explanations for inadequate reporting: editorial pressure for brevity; the emphasis of much statistical teaching on formalistic analysis and stylized conclusions--such as the ritual of "p-values"; the easy accessibility of packaged computer programs to those who understand little about statistics; and a climate of opinion in which statistics is seldom taken more seriously than any other mechanical prerequisite for publication, such as correct spelling or inclusion of references.

In this section some of these factors will be discussed briefly; for a more complete discussion, see Winkler [50].

Because of the history of controversy between proponents of Bayesian and classical methods, it might be thought that differences relating to philosophical considerations concerning the foundations of statistics might play an important role in the choice of methods of statistical analysis. However, the grounds for the criticism in Section 2 are much more basic than an overly simplified Bayesian-classical dichotomy. Even in terms of classical statistics alone, there is a serious theory-practice gap. Thus, the problems apparently cannot be explained in terms of philosophical considerations alone. Although this paper is written wholeheartedly from the Bayesian approach, the choice of a philosophical approach to statistical inference still seems to be subordinate to the question of whether the approach is used consistently, carefully, and appropriately.

Tradition obviously plays an important role in the choice of inferential procedures. If an experiment is to be conducted in a particular area, it is easy to look at past experiments in the same area and to use a similar type of analysis. In this regard, it might be said that poor statistical practice breeds more poor statistical practice.

The effect of tradition is also felt in the area of statistical training. Most users of statistics are by no means mathematical statisticians; they are specialists in some area of application. While some users may have extensive training in statistics, many have been exposed formally to statistical methods only through one or more basic statistics courses.

Such courses are often taught by instructors who have very little training in statistics themselves and who tend to perpetuate the procedures encountered in practice. Instructors generally use traditional textbooks and teach traditional methods. The stress placed on decision making (rather than inference) by many Bayesians has further slowed the pace of the dissemination of introductory-level material on Bayesian inference. Introductory-level Bayesian textbooks with stress on decision-making have appeared, but books with stress on Bayesian inference at an introductory level are not as common (however, see Section 5). Even after such books become readily available, there will be a lag before they are widely used and the methods are widely applied.

A related problem is caused by the fact that statistical theory has not, in general, been translated into a form that makes it readily accessible to experimenters, most of whom do not (or cannot) read the statistical literature. In other words, Bayesian techniques are not readily available for the average researcher, where availability is to be interpreted in terms of elementary discussions of the procedures, computer programs, appropriate tables, and so on. Thus, at the present time, the Bayesian approach requires a greater commitment of time and effort on the part of the experimenter than do traditional methods that are widely used. Of course, a careful, appropriate classical analysis also requires more time and effort than the simple reporting of a significance level for a test of a sharp null hypothesis.

One of the advantages of Bayesian methods is that the results can be presented in intuitively appealing and easily interpretable forms. For example, it is much more appealing to associate probabilities with hypotheses or with intervals of values of the parameter of interest than to think in terms of significance levels or classical confidence intervals. (With respect to confidence intervals, classical statisticians take great pains to emphasize the appropriate classical interpretation, but this interpretation is so counterintuitive that many users of statistics seem to think of classical interval estimates in terms of the Bayesian interpretation.) Nevertheless, Bayesian procedures are encountered so seldom in analyses of experimental data that their interpretations may not be widely understood. Thus, the experimenter using Bayesian procedures must explain the procedures and the interpretation of the results. A classical t test, for instance, is familiar to virtually all experimenters, whereas the Bayesian counterpart may require a paragraph or two of explanation. Until Bayesian methods are more widely used, applications of such methods will be more difficult to communicate to readers than are applications of standard classical procedures.

Perceived resistance of journal editors to new approaches may also dissuade researchers from considering improvements in statistical practice. Some researchers have the notion that it is necessary to obtain a very low significance level in order to have a paper accepted for publication. Unfortunately, as observed in the previous section, a very low significance

level for the test of a sharp null hypothesis can be virtually guaranteed by taking a large enough sample. Thus, a perceived association between a low significance level and the probability of acceptance of a paper encourages poor statistical practice. Why should an experimenter invest a great deal of time and effort in a careful, appropriate analysis when it appears that a simple significance level for a test of a sharp null hypothesis will serve the same purpose quite well in terms of yielding publishable results that are acceptable professionally?

4. Bayesian Analysis of Experimental Data

In scientific experiments, statistical methods generally enter into the picture at several stages, including the design of the experiment, the analysis of the data, and the reporting of the experimental results to the general scientific community. These stages are interrelated to a considerable degree, of course; for instance, considerations regarding analysis and reporting must be taken into account during the design stage, and considerations regarding reporting must be taken into account during the analysis stage. In the first part of this section, the question of scientific reporting is considered. In the second part of the section, hypothesis testing is considered once again, and Bayesian alternatives to the procedures criticized in Section 2 are discussed.

Scientific Reporting

The goal of scientific reporting might be stated in an

oversimplified form as "complete disclosure." This implies that the experimenter should report all details concerning the design and carrying out of the experiment, the data that are collected, any assumptions that are made, any analyses that are conducted, and so on. These details enable a reader of the report to understand fully each step taken by the experimenter, to consider alternative assumptions and analyses, and even to replicate the experiment if it is deemed desirable to do so. Complete disclosure is useful for a reader who is intimately interested in the problem that is being studied and who wishes to be able to investigate carefully the experiment and its results.

Of course, not all readers of a scientific report are interested in all of the details. Many readers are only interested in a brief summary of the results of the experiment, with enough information included to enable them to see if the analysis seems to be appropriate and reasonable. Such a reader may not want to "wade through" a complete report, which is obviously the least concise form of report. In most instances of scientific reporting it is necessary to strike a balance between completeness and conciseness, with the point of balance depending upon the details of the particular situation.

To reconstruct an analysis or to consider other analyses "from scratch," it is necessary to have the raw data from an experiment. When the amount of data is not too great, it may be possible to include the data in the report. In many cases, however, reporting the raw data from an experiment requires

too much space to satisfy space limitations imposed by journals. An alternative is to omit the raw data from the report but to make it readily available to any interested parties. This compromise makes the report more concise while still making it possible for interested readers to obtain "complete disclosure."

Even if the data are included in the report, they are not adequate for reporting purposes, since they generally necessitate too much effort on the part of the reader to understand the results of the experiment. Therefore, some summarization is needed, and an obvious choice is to report the likelihood function, since the likelihood principle states that the entire evidence of a sample is contained in the likelihood function.

In most cases where a classical parametric analysis is encountered, enough assumptions are made to allow the researcher to determine the likelihood function. To the extent that different individuals would agree that the assumptions are reasonable, then, the likelihood function might be considered reasonably "public" (i.e. most individuals, given the raw data, would tend to agree with the assumptions and hence with the likelihood function). It must be remembered, however, that choices regarding the acceptance or rejection of various assumptions in building a model of the data-generating process are ultimately subjective choices. Thus, elements of subjectivity enter into the determination of sampling distributions and hence of likelihood functions. Because of frequent reliance

on important mathematical results such as the central limit theorem, some might argue that this element of the analysis is "objective" in nature. Perhaps this is true to a degree, but ultimately the entire model-building process is a subjective process, and it is important in any application to carefully investigate the appropriateness of assumptions such as independence and normality. For reporting purposes, the experimenter should make every effort to justify all assumptions and, insofar as possible, to present enough information to enable the reader to make a personal decision regarding the applicability of the assumptions. Although many statisticians stress the importance of investigating assumptions, it appears that this step is too frequently "glossed over" in practice. Virtually any assumption is an approximation to reality, and the reader has the right to know how "good" the approximation is.

Given the models and assumptions frequently encountered in practice, the likelihood function is usually based on a reasonably simple sufficient statistic. If a tractable sufficient statistic is not available, it may be possible to determine a partial likelihood function based on a nonsufficient statistic. The presentation of a partial likelihood function may even be desirable when a full likelihood function is available if it results in little loss of information and if the partial likelihood function is much simpler and easier to communicate than the full likelihood function.

Knowledge of the likelihood function enables individuals to insert their own prior distributions and to compute the corresponding posterior distributions. In Bayesian inference, the primary inferential statement of concern is the posterior distribution, which summarizes an individual's uncertainty about a parameter after the experimental data have been observed. Except in simple cases, however, the determination of a posterior distribution may require a fair amount of time and effort on the part of the reader. To reduce the computational burden on the reader, the experimenter might assume the burden of performing the application of Bayes' theorem. This could be accomplished by presenting posterior distributions corresponding to a variety of prior distributions, the variety being broad enough to include (at least approximately) the prior distributions, as anticipated by the experimenter, of as many readers as possible. (By way of analogy, note that if there is some question concerning the assumptions underlying the likelihood function, one might perform the analysis under different possible sets of assumptions.) The set of prior distributions may include the experimenter's own prior distribution, but it should not be limited to that distribution.[3]

If the above approach is taken by the experimenter, the problem is to select a set of prior distributions that is not too large or too difficult to work with but is thought to be "representative" of the prior distributions of the audience for which the report is intended. One candidate for inclusion

in the set is a diffuse prior distribution, which is a prior distribution that is relatively "flat" when compared with the likelihood function (see Edwards, Lindman, and Savage, [14]). The use of this distribution invokes Savage's principle of stable estimation and yields a posterior distribution that is approximately proportional to the likelihood function. Therefore, this approach is similar to reporting the likelihood function. Nevertheless, the posterior distribution is a proper probability distribution and probability statements can be made concerning the parameter of interest, so the interpretation is different from that of the likelihood function and easier to understand for the average reader.

Another possibility is to consider families of conjugate distributions, such as those developed by Raiffa and Schlaifer [39]. Such families provide relatively simple functions relating the parameters of the posterior distribution to the parameters of the prior distribution. Presentation of the functions allows anyone whose prior distribution can be closely approximated by a member of the conjugate family to compute a posterior distribution. Moreover, if the functions are presented graphically, it should be easy for the reader to see how sensitive the posterior distribution is to changes in the prior distribution. In general, the question of the sensitivity of results to changes in the inputs is an important question in any statistical analysis.

In some instances, the bulk of the available prior information is in the form of previously-observed data. In this case,

the prior distribution might be considered to be reasonably "public" in the same sense that "public" likelihood functions were discussed earlier in this section. This might obviate somewhat the need to consider a variety of prior distributions. Of course, as more and more inputs to the analysis are considered "public," the need to worry about alternative inputs and the sensitivity of the results to changes in the inputs is greatly reduced.

Once a posterior distribution (or a set of posterior distributions corresponding to various prior distributions) has been determined, the question of reporting still remains. Of course, one can report the entire posterior distribution, either in graphical form or in functional form, and graphical presentations of distributions are very valuable. In addition, it may be useful to aid the reader's interpretation of the distributions by summarizing them in some way. A few well-chosen summary measures often convey the main results with little loss of information. Some possible summarizations include parameters of the posterior distribution, if it is a well-known distribution; measures of location; measures of dispersion; probabilities of selected intervals of values; and so on. Credible intervals, which are intervals of values accompanied by the corresponding posterior probabilities, are particularly useful summarizations.

The discussion of scientific reporting in this section has been quite brief, as an attempt has been made to cover

important points without going into much detail. For example, problems that arise in multiparameter situations (e.g. the reporting of marginal posterior distributions for individual parameters, the inclusion of nuisance parameters to broaden the model) have not been considered. For more detailed discussions of some of the points covered here, see Edwards, Lindman, and Savage [14], Hildreth [19], and Roberts [40]. As noted at the beginning of the section, it is necessary to strike a balance between the conflicting goals of completeness and conciseness in reporting experimental results. With regard to the Bayesian approach, a report might include posterior distributions and summarizations of posterior distributions corresponding to one or more prior distributions. Alternatively if the burden of applying Bayes' theorem is to be placed on the reader, the experimenter might simply report the likelihood function (or likelihood functions under different sets of assumptions).

Bayesian Hypothesis Testing

Although a full Bayesian report of experimental data requires the presentation of an entire posterior distribution (or a set of distributions corresponding to different prior distributions), simplifications are possible in the case in which the primary interest is in certain hypotheses. The inferential impact of new information with respect to two hypotheses can be adequately summarized by a simple likelihood ratio, and the multiplication of a likelihood ratio by a prior odds ratio yields a posterior odds ratio. The determination

of likelihood ratios for various specifications of hypotheses will be considered in this section, and some brief remarks will be made concerning the inclusion of prior odds ratios and the notion of scientific reporting in the specific case of hypothesis testing. The discussion will be restricted to the case in which only two hypotheses are of interest; the generalization to more than two hypotheses is straightforward.

If the two hypotheses of interest are labelled H_1 and H_2, and y represents the data, then the likelihood ratio of interest is simply

$$LR = \frac{f(y|H_1)}{f(y|H_2)} \quad ,$$

where $f(y|H_i)$ represents the probability (density) of the sample data, conditional upon H_i, evaluated at the observed y. For a very simple example, suppose that it is assumed that the data are generated by a Bernoulli process and that H_1 is the hypothesis that p, the parameter of the Bernoulli process, is equal to .3, whereas H_2 is the hypothesis that p is equal to .4. If two successes are observed in a sample of size ten, then the likelihood ratio is a ratio of binomial probabilities:

$$LR = \frac{\binom{10}{2}(.3)^2(.7)^8}{\binom{10}{2}(.4)^2(.6)^8} = 1.93 \quad .$$

Similarly, in sampling from a population that is assumed to be normally distributed with known variance and unknown mean, if the hypotheses concerning the mean are exact hypotheses,

the likelihood ratio is a ratio of normal densities. In the same situation with the variance unknown, the likelihood ratio is a ratio of student t densities.

The above situations involve exact hypotheses, whereas the hypotheses of interest in experimental situations are frequently inexact. Given a posterior distribution for a parameter, it is possible to determine probabilities for different sets of values of the parameter. Posterior odds ratios are simply ratios of such probabilities. A Bayesian approach to a one-tailed test, then, might simply be to determine a posterior odds ratio of the form $P(\theta \leq \theta_0)/P(\theta > \theta_0)$ directly from the posterior distribution. Of course, the cautions noted in the first part of this section regarding scientific reporting and the choice of prior distributions for reporting purposes still apply when the experimenter's intent is to report the results to the scientific community rather than simply to make private inferences.

Another Bayesian approach to inexact hypotheses is to specify the hypotheses not in terms of sets of values of the parameter of interest, but in terms of probability distributions over the parameter space. In general, then, H_i can be expressed in terms of a distribution, $f_i(\theta)$. Note that this includes the case of exact hypotheses, for $f_i(\theta)$ can be taken as the degenerate distribution that places a probability of one on a single value of θ. Now the likelihood ratio, $f(y|H_1)/f(y|H_2)$, is a ratio of probabilities (densities) that are conditional on the entire distributions $f_1(\theta)$ and $f_2(\theta)$

rather than on single values of the parameter. Each of these probabilities (densities) can be obtained by considering the predictive distribution of y, which is the marginal distribution of y after θ is integrated out:

$$f(y|H_i) = \int f(y|\theta)\, f_i(\theta)\, d\theta \quad .$$

(If the distribution of θ is discrete, this is a sum rather than an integral.) The likelihood ratio is then of the form

$$LR = \frac{f(y|H_1)}{f(y|H_2)} = \frac{f(y|\theta)\, f_1(\theta)\, d\theta}{f(y|\theta)\, f_2(\theta)\, d\theta} \quad .$$

For an example, suppose that the population of interest is assumed to be normally distributed with known variance σ^2 and unknown mean μ. Moreover, assume that $f_i(\mu)$ is a normal distribution with mean m_i and variance v_i. For a sample of fixed size n, the sample mean, m, is a sufficient statistic. Thus, for the purposes of inference, the sample data, y, can be replaced by m. For hypothesis H_i, the predictive distribution of m is given by

$$f(m|H_i) = \int_{-\infty}^{\infty} f(m|\mu)\, f_i(\mu)\, d\mu \quad .$$

But $f(m|\mu)$ is a normal distribution with mean μ and variance σ^2/n. Carrying out the integration, $f(m|H_i)$ is a normal distribution with mean m_i and variance $v_i + (\sigma^2/n)$. The likelihood ratio is thus a ratio of normal densities determined from the respective predictive distributions, evaluated at the observed value of m.

In the first part of this section, reference was made to the notion of conjugate distributions. In the above example, $f_1(\mu)$ and $f_2(\mu)$ were conjugate distributions. For various data-generating processes, including many of the processes commonly assumed in applications (e.g. the normal process, the Bernoulli process, the Poisson process, the normal regression process, etc.), the form of the predictive distribution has been developed under the assumption that $f_i(\theta)$ is a conjugate distribution (e.g. Raiffa and Schlaifer, [39]). Therefore, if the hypotheses of interest can be expressed in terms of conjugate distributions, the appropriate predictive distribution can be found in the Bayesian literature and the determination of the likelihood ratio is merely a matter of calculating the appropriate probabilities (densities).

Once a likelihood ratio is determined, it can be multiplied by the prior odds ratio to arrive at the posterior odds ratio. For reporting purposes, the experimenter may want to consider various possible prior odds ratios. Of course, if the likelihood ratio is given, it is easy for any reader to insert a prior odds ratio in order to determine a personal posterior odds ratio.

It should be obvious by now that in the Bayesian approach to hypothesis testing, a great deal of care must be taken in the specification of hypotheses. An exact hypothesis can only be entertained if one is willing to place a nonzero prior probability on the single value represented by the exact hypothesis. For instance, a Bayesian generalization of the notion

of testing a sharp null hypothesis is to consider a "spike" of probability at the value specified by the sharp null hypothesis and an alternative hypothesis that is represented by a distribution over the parameter space (e.g. see Jeffreys, [20]). The alternative hypothesis might be taken to be a diffuse distribution, for example. If a "spike" at a single point seems unreasonable, a further generalization is to let both $f_1(\theta)$ and $f_2(\theta)$ be centered at the exact value corresponding to the classical statistician's sharp null hypothesis but to make $f_1(\theta)$ a much tighter distribution than $f_2(\theta)$.

In general, the primary concern in Bayesian inference is the combination of prior information and sample information to form a posterior distribution. In many cases a Bayesian analysis of experimental data need not involve hypothesis testing at all. In this section, however, an attempt has been made to indicate how the Bayesian approach can be structured in terms of hypothesis testing if the experimenter so desires.

5. Discussion

In summary, there is an increasing interest in Bayesian procedures, although much of this interest is decision-oriented rather than inference-oriented and is concerned with development of theory rather than with the actual use of these procedures in practice. In the analysis of experimental results, the main concern is generally inference rather than decision,

and the bulk of current statistical practice in this area leaves much to be desired, as indicated in Section 2. Many factors, including tradition, statistical training, computational difficulties, and reporting difficulties contribute to poor statistical practice. As noted at the end of Section 3, an experimenter has little incentive to invest a great deal of time and effort in a careful, appropriate analysis when it appears that a simple significance level for a test of a sharp null hypothesis will serve the same purpose quite well in terms of yielding publishable results that are acceptable professionally.

How, then, might the weaknesses in current statistical practice be remedied? Within the classical framework, improvements in statistical training that place emphasis on meaning rather than mechanics would be most useful, as would a willingness on the part of journal editors and referees to demand clear, meaningful statistical analyses. The discussion of scientific reporting in Section 4 is relevant here. Furthermore, since this paper is written from the Bayesian standpoint, the view taken here is that the use of Bayesian techniques would lead to great improvements in statistical practice, provided that these techniques are used carefully and appropriately. Bayesian procedures generally provide answers to the questions of interest to the experimenter rather than answers to related but different questions. For example, probability statements can be made directly about the parameters of interest instead of indirectly in terms of probabilities

of sample outcomes conditional upon the parameters.

In order to increase the use of Bayesian inferential procedures in practice, it is necessary to narrow the "theory-practice gap" by making Bayesian procedures more "available" to experimenters. At the most basic level, this effort involves the use of introductory-level, inference-oriented Bayesian texts. Material on Bayesian inference above the elementary introductory level is available in books such as Raiffa and Schlaifer [39], Jeffreys [20], Lindley [24], [25], Pratt, Raiffa, and Schlaifer [37], Good [17], DeGroot [10], LaValle [22], Zellner [51], and Box and Tiao [5]; many of these references also contain material on decision-making procedures. Most introductory texts that are Bayesian in nature are strongly decision-oriented (e.g. Raiffa, [38], Lindley, [26], Moore, [28], and Brown, Kahr, and Peterson, [6]). Some other introductory Bayesian texts contain a mixture of inferential material and decision-theoretic material. For example, Schlaifer [43] was the pioneering introductory-level book in this area (also, see Schlaifer, [44]; Schmitt [46] places some stress on inference; Winkler [49] includes quite a bit of inferential material; a recent book by Phillips [31] is intended to "fill the gap" somewhat in terms of Bayesian inference; and other books may be in preparation (e.g. Pitz, [35]). More books emphasizing Bayesian inference at the introductory level are needed.

Moving from the training level to the level of actual application of the techniques, further effort should be expended

on expressing Bayesian procedures in forms that make them more accessible to users. This involves such steps as expressing the procedures in simplified form (e.g. simplifying formulas for likelihood ratios as much as possible for situations that are widely-encountered) and developing computer programs. Some individuals have worked on the first step (e.g. Pitz, [34]) and on the second step (e.g. Schlaifer, [45], Novick, [30]). Furthermore, at the level of application, perhaps the most useful step in terms of the advancement of Bayesian inference would be the publication of more actual Bayesian analyses of experimental data in journals in the areas of application. An example of a particularly detailed analysis that might be useful for researchers to look at is a disputed-authorship problem studied in Mosteller and Wallace [29]; some applications in the area of medicine are presented in Cornfield [7]; and an application in the area of education is given in Novick [30]. For an interesting (and somewhat controversial) application of Bayesian hypothesis testing, see Good [18] and Efron [15].

Another area of interest is that of scientific reporting. Research in this area might concentrate on the development of different "packages" of items to be reported in different situations and on attempts to simplify these packages without a considerable loss in terms of the information content of the packages. For example, Dickey [11] develops graphical techniques for relating parameters of prior distributions to

parameters of posterior distributions and considers bounds on odds ratios for various situations (also, see Dickey and Freeman, [12]). More work along these lines would be valuable.

In addition to the need to make simple Bayesian procedures more available to users, further theoretical work would be useful. Such work might involve the development of Bayesian procedures for various types of models that have not been studied extensively from the Bayesian standpoint to date and the development of approximations that might simplify a Bayesian analysis. For instance many different situations are reviewed in Lindley [25], and various models have been considered in recent work in Bayesian econometrics (e.g. see Zellner, [51], and Fienberg and Zellner, [16]).

In this paper, some weaknesses in current statistical practice have been discussed, and suggestions for remedying these weaknesses have been presented. The Bayesian approach, which has received much attention in recent years, particularly in terms of decision making, provides a useful framework for the analysis of experimental data. Efforts are needed to make Bayesian procedures more readily available to researchers dealing with experimental data, and some suggestions for the direction of such efforts have been given in this concluding section.

Footnotes

[1]Procedures that do not involve probability revision are frequently included under the heading "Bayesian statistics." In particular, because Bayesian methods use subjective probabilities as inputs, it is often erroneously assumed that "subjective" and Bayesian" are synonymous. Also, because Bayesian methods are frequently used in decision-making problems, it is often erroneously assumed that the adjective "Bayesian" is always used in conjunction with "decision theory."

[2]Under certain circumstances, Bayesian and classical procedures may yield similar numerical results. Even in such instances, however, the interpretations attached to the numerical results by the two schools of thought are quite different.

[3]Statistical analyses are prepared for many different purposes. If the experimenter only wants to use the analysis for personal purposes, then it is appropriate to consider only the experimenter's prior distribution. If the analysis is being prepared for a particular client, then the client's prior distribution would be the relevant distribution to consider. This paper is primarily concerned with reporting to the general scientific community. For this audience, the posterior distribution following the experimenter's personal prior distribution might be of some interest because of the fact that the experimenter presumably has given the problem at hand serious thought. However, others may have different prior distributions, and it is generally inappropriate to confine the analysis to the experimenter's own posterior distribution.

References

[1] Bakan, D. "The Test of Significance in Psychological Research," Psychological Bulletin, 66 (1966), 423-437.

[2] Beach, L.R., and Phillips, L.D. "Subjective Probabilities Inferred from Estimates and Bets," Journal of Experimental Psychology, 75 (1967), 354-359.

[3] Berkson, J. "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," Journal of the American Statistical Association, 33 (1938), 526-536.

[4] Birnbaum, A. "On the Foundations of Statistical Inference," Journal of the American Statistical Association, 57 (1962), 526-536.

[5] Box, G.E.P., and Tiao, G.C. Bayesian Inference in Statistical Analysis. Reading, Mass., Addison-Wesley, 1973.

[6] Brown, R.V., Kahr, A., and Peterson, C.R. Decision Analysis for the Manager. New York, Holt, Rinehart and Winston, 1974.

[7] Cornfield, J. "The Bayesian Outlook and Its Applications," Biometrics, 25 (1969), 617-657 (with discussion).

[8] de Finetti, B. "Subjective or Objective Probability: Is the Dispute Undecidable?" Symposia Mathematica, 9 (1972), 21-36.

[9] de Finetti, B. "Bayesianism: Its Unifying Role for Both the Foundations and the Applications of Statistics," Proceedings of the 39th Sesssion of the International Statistical Institute, in press, 1973.

[10] DeGroot, M.H. Optimal Statistical Decisions. New York, McGraw-Hill, 1970.

[11] Dickey, J.M. "Scientific Reporting and Personal Probabilities: Student's Hypothesis," Journal of the Royal Statistical Society B, 35 (1973).

[12] Dickey, J.M., and Freeman, P. "Populations of Personal Probabilities and Bayes' Theorem," London, University College, unpublished paper, 1973.

[13] Edwards, W. "Tactical Note on the Relation Between Scientific and Statistical Hypotheses," Psychological Bulletin, 63 (1965), 400-402.

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

[15] Efron, B. "Does an Observed Sequence of Numbers Follow a Simple Rule? (Another Look at Bode's Law)," Journal of the American Statistical Association, 66 (1971), 552-568 (with discussion).

[16] Fienberg, S.E., and Zellner, A., Eds. Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage. Amsterdam, North Holland, 1974.

[17] Good, I.J. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. Cambridge, Mass, The M.I.T. Press, 1965.

[18] Good, I.J. "A Subjective Evaluation of Bode's Law and an 'Objective' Test for Approximate Numerical Rationality," Journal of the American Statistical Association, 64 (1969), 23-66 (with discussion).

[19] Hildreth, C. "Bayesian Statisticians and Remote Clients," Econometrica, 31 (1963), 422-438.

[20] Jeffreys, H. Theory of Probability, 3rd. Ed. Oxford, Clarendon Press, 1961.

[21] Kyburg, H.E., and Smokler, H.E., Eds. Studies in Subjective Probability. New York, Wiley, 1964.

[22] LaValle, I.H. An Introduction to Probability, Decision, and Inference. New York, Holt, Rinehart and Winston, 1970.

[23] Lindley, D.V. "A Statistical Paradox," Biometrika, 44 (1957), 187-192.

[24] Lindley, D.V. Introduction to Probability and Statistics from a Bayesian Viewpoint, 2 Vols. Cambridge, Cambridge University Press, 1965.

[25] Lindley, D.V. Bayesian Statistics: A Review. Philadelphia, Pa., Society for Industrial and Applied Mathematics, 1971.

[26] Lindley, D.V. Making Decisions. New York, Wiley, 1971

[27] Meehl, P.E. "Theory-Testing in Psychology and Physics: A Methodological Paradox," Philosophy of Science, 34 (1967), 103-115.

[28] Moore, P.G. Risk in Business Decision. New York, Wiley, 1973.

[29] Mosteller, F., and Wallace, D.L. Inference and Disputed Authorship: The Federalist. Reading, Mass., Addison-Wesley, 1964.

[30] Novick, M.R. "High School Attainment: An Example of a Computer-Assisted Bayesian Approach to Data Analysis," International Statistical Review, 41 (1973), 264-271.

[31] Phillips, L.D. Bayesian Statistics for Social Scientists. London, Nelson, 1973.

[32] Pitz, G.F. "An Example of Bayesian Hypothesis Testing: The Perception of Rotary Motion in Depth," Psychological Bulletin, 70 (1968), 252-255.

[33] Pitz, G.F. "An Inertia Effect (Resistance to Change) in the Revision of Opinion," Canadian Journal of Psychology, 23 (1969), 24-33.

[34] Pitz, G.F. "Quick and Dirty Data Analysis with a Bayesian Flavor," Carbondale, Ill., Southern Illinois University, unpublished manuscript, 1972.

[35] Pitz, G.F. "Bayesian Data Analysis and Statistical Inference in Psychology," Carbondale, Ill., Southern Illinois University, unpublished manuscript, 1974.

[36] Pratt, J.W. "Bayesian Interpretation of Standard Inference Statements," Journal of the Royal Statistical Society B, 27 (1965), 169-203 (with discussion).

[37] Pratt, J.W., Raiffa, H., and Schlaifer, R. Introduction to Statistical Decision Theory, preliminary ed. New York, McGraw-Hill, 1965.

[38] Raiffa, H. Decision Analysis. Reading, Mass., Addison-Wesley, 1968.

[39] Raiffa, H., and Schlaifer, R. Applied Statistical Decision Theory. Boston, Graduate School of Business Administratio Harvard University, 1961.

[40] Roberts, H.V. "Reporting of Bayesian Studies," in Fienberg and Zellner, 1974.

[41] Savage, L.J. The Foundations of Statistics. New York, Wiley, 1954.

[42] Savage, L.J., et al. The Foundations of Statistical Inference. London, Methuen, 1962.

[43] Schlaifer, R. Probability and Statistics for Business Decisions. New York, McGraw-Hill, 1959.

[44] Schlaifer, R. Analysis of Decisions Under Uncertainty. New York, McGraw-Hill, 1969.

[45] Schlaifer, R. Computer Programs for Elementary Decision Analysis. Boston, Graduate School of Business Administration, Harvard University, 1971.

[46] Schmitt, S.A. Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics. Reading, Mass., Addison-Wesley, 1969.

[47] Slovic, P., and Lichtenstein, S. "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment," Organizational Behavior and Human Performance, 6 (1971), 651-730.

[48] Wilson, W., Miller, H.L., and Lower, J.S. "Much Ado About the Null Hypothesis," Psychological Bulletin, 67 (1967), 188-196.

[49] Winkler, R.L. An Introduction to Bayesian Inference and Decision. New York, Holt, Rinehart and Winston, 1972.

[50] Winkler, R.L. "Statistical Analysis: Theory Versus Practice," Theory and Practice of Measuring Subjective Probability. Dordrecht, D. Reidel, 1974, in press.

[51] Zellner, A. An Introduction to Bayesian Inference in Econometrics. New York, Wiley, 1971.