# BAYESIAN INFERENCE*

Harry V. Roberts, University of Chicago

Bayesian inference or Bayesian statistics is an approach to statistical inference based on the theory of subjective probability. A formal Bayesian analysis leads to probabilistic assessments of the object of uncertainty. For example, a Bayesian inference might be: "The probability is .95 that the mean of a normal distribution lies between 12.1 and 23.7." The number .95 represents a degree of belief, either in the sense of "subjective probability consistent" or "subjective probability rational" (see PROBABILITY: PHILOSOPHY AND INTERPRETATION, which should be read in conjunction with the present article); .95 need not and typically does not correspond to any "objective" long-run relative frequency. Very roughly, a degree of belief of .95 can be interpreted as betting odds of 95 to 5 or 19 to 1. A degree of belief is always potentially a basis for action; for example, it may be combined with utilities by the principle of maximization of expected utility (see STATISTICAL DECISION THEORY).

By contrast, the traditional or "classical" approach to inference leads to probabilistic statements about the method by which a particular inference is obtained. Thus a classical inference might be: "A .95 confidence interval for the mean of a normal distribution extends from 12.1 to 23.7." The number .95 here represents a long-run relative frequency, namely the frequency with which intervals obtained by the method that resulted in the present interval would in fact include the unknown mean. (It is not to be inferred from the fact that we used the same numbers, .95, 12.1, and 23.7, in both illustrations that there will necessarily be a numerical coincidence between the two approaches.)

The term "Bayesian" arises from an elementary theorem of probability theory, named after the Rev. Thomas Bayes, an English clergyman of the 18th century, who first enunciated it and proposed its use in inference. Bayes' theorem is typically used in the process of making Bayesian inferences, as will be explained below. For a number of historical reasons, however, current interest in Bayesian inference is quite recent--dating, say, from the 1950's. Hence the term "neo-Bayesian" is sometimes used instead of "Bayesian."

## An Illustration of Bayesian Inference

For a simple illustration of the Bayesian approach, consider the problem of making inferences about a Bernoulli process with parameter p. A Bernoulli process can be visualized in terms of repeated independent tosses of a not-necessarily "fair" coin. It generates "heads" and "tails" in such a way that the conditional probability of heads on a single trial is always equal to a parameter p regardless of the previous history of heads and tails.

Suppose first that we have no direct sample evidence from the process. Based on experience with similar processes, introspection, general knowledge, etc., we may be willing to translate our judgments about the process into probabilistic terms. For example, we might assess a (subjective) probability distribution for $\tilde{p}$ (the tilde "$\sim$" indicates that we are now thinking of the parameter p as a random variable). Such a distribution is called a prior distribution because it is usually assessed prior to sample evidence. Purely for illustration, suppose that the prior distribution of $\tilde{p}$ is uniform on the interval from 0 to 1: the probability that $\tilde{p}$ lies in any subinterval is that subinterval's length, no matter where the subinterval is located between 0 and 1. Now suppose that we observe heads, heads, and tails on three tosses of a coin. The probability of observing this sample, conditional on p, is

$$p^2(1-p) .$$

If we regard this expression as a function of p, it is called the likelihood function of the sample. Bayes' theorem shows how to use the likelihood function in conjunction with the prior distribution to obtain a revised or posterior distribution of $\tilde{p}$. "Posterior" means after the sample evidence, and the posterior distribution represents a reconciliation of sample evidence and prior judgment. In terms of inferences about $\tilde{p}$, we may write Bayes' theorem in words as

Posterior probability (density) at p, given the observed sample =

$$\frac{\text{Prior probability (density) at } p \times \text{likelihood}}{\text{Prior probability of the observed sample}}$$

Expressed mathematically,

$$f''(p|r,n) = \frac{f'(p)\, p^r(1-p)^{n-r}}{\int_0^1 f'(p)\, p^r(1-p)^{n-r}dp} ,$$

where $f'(p)$ denotes the prior density of $\tilde{p}$, $p^r(1-p)^{n-r}$ denotes the likelihood if r heads are observed in n trials, and $f''(p|r,n)$ denotes the posterior density of $\tilde{p}$ given the sample evidence.

*A revised version of this paper will appear in the (forthcoming) International Encyclopedia of the Social Sciences.

In our example, $f'(p) = 1$, $(0 \leq p \leq 1)$, $r = 2$, $n = 3$, and

$$\int_0^1 f'(p)\, p^r(1-p)^{n-r}dp = \int_0^1 p^2(1-p)dp = 1/12 ,$$

so

$$f''(p|r = 2, n = 3) = 12\, p^2(1-p), \quad 0 \leq p \leq 1$$
$$= 0 \quad \text{otherwise.}$$

Thus we emerge from the analysis with an explicit probability distribution for $\tilde{p}$. This distribution characterizes fully our judgments about $\tilde{p}$. It could be applied in a formal decision-theoretic analysis in which utilities of alternative acts are functions of $p$. For example, we might make a Bayesian point estimate of $p$ (each possible point estimate is regarded as an act), and the seriousness of an estimation error ("loss") might be proportional to the square of the error. The best point estimate can then be shown to be the mean of the posterior distribution; in our example, this would be .6. Or, we might wish to describe certain aspects of the posterior distribution for summary purposes; it can be shown, for example, that

$$P(\tilde{p}<.194) = .025 \quad \text{and} \quad P(\tilde{p}>.932) = .025 ,$$

so a .95 "credible interval" for $\tilde{p}$ extends from .194 to .932. Again, it can be shown that $P(\tilde{p}>.5) = .688$: the posterior probability that the coin is "biased" in favor of heads is a little over 2/3.

## The Likelihood Principle

In our example, the effect of the sample evidence was wholly transmitted by the likelihood function. All we needed to know from the sample was $p^r(1-p)^{n-r}$; the actual sequence of individual observations was irrelevant so long as we believed the assumption of a Bernoulli process. In general, a full Bayesian analysis requires as inputs for Bayes' theorem only the likelihood function and the prior distribution. Thus the import of the sample evidence is fully reflected in the likelihood function, a principle known as the likelihood principle (see also LIKELIHOOD). Alternatively, given that the sample is drawn from a Bernoulli process, the import of the sample is fully reflected in the numbers $r$ and $n$, which are called sufficient statistics (see SUFFICIENCY).

The likelihood principle implies certain consequences that do not accord with traditional ideas. Here are examples: (1) Once the data are in, there is no distinction between sequential analysis and analysis for fixed sample size. In the Bernoulli example, successive samples of $n_1$ and $n_2$ with $r_1$ and $r_2$ successes could be analyzed as one pooled sample of $n_1 + n_2$ trials with $r_1 + r_2$ successes. Alternatively, a posterior distribution could be computed after the first sample of $n_1$; this distribution could then serve as a prior distribution for the second sample; finally, a second posterior distribution could be computed after the second sample of $n_2$. By either route the posterior distribution after $n_1 + n_2$ observations would be the same. Under almost any situation that is likely to arise in practice, the "stopping rule" by which sampling is terminated is irrelevant to the analysis of the sample. For example, it would not matter whether $r$ successes in $n$ trials were obtained by fixing $r$ in advance and observing the $r^{th}$ success on the $n^{th}$ trial, or by fixing $n$ in advance and counting $r$ successes in the $n$ trials. (2) For the purpose of statistical reporting, the likelihood function is the important information to be conveyed. If a reader wants to perform his own Bayesian analysis, he needs the likelihood function, not a posterior distribution based on someone else's prior, nor traditional analyses such as significance tests, from which it may be difficult or impossible to recover the likelihood function.

## Vagueness about Prior Probabilities

In our example we assessed the prior distribution of $\tilde{p}$ as a uniform distribution from 0 to 1. It is sometimes thought that such an assessment means that we "know" $\tilde{p}$ is so distributed, and that our claim to knowledge might be verified or refuted in some way. It is indeed possible to imagine situations in which the distribution of $\tilde{p}$ might be known, as when one coin is to be drawn at random from a number of coins, each of which has a "known" $p$ determined by a very large number of tosses. The frequency distribution of these $p$'s would then serve as a prior distribution, and all statisticians would apply Bayes' theorem in analyzing sample evidence. But such an example would be unusual. Typically, in making an inference about $\tilde{p}$ for a particular coin, the prior distribution of $\tilde{p}$ is not a description of some distribution of $p$'s but rather a tool for expressing judgments about $\tilde{p}$ based on evidence other than the evidence of the particular sample to be analyzed.

Not only do we rarely "know" the prior distribution of $\tilde{p}$, but we are typically more or less vague when we try to assess it. This vagueness is comparable to the vagueness that

surrounds many decisions in everyday life. For example, a person may decide to offer $21,250 for a house he wishes to buy, even though he may be quite vague about what amount he "should" offer. Similarly, in statistical inference we may assess a prior distribution in the face of a certain amount of vagueness. If we are not willing to do so, we cannot pursue a formal Bayesian analysis and must evaluate sample evidence intuitively, perhaps aided by the tools of descriptive statistics and classical inference.

Vagueness about prior probabilities is not the only kind of vagueness to be faced in statistical analysis, and the other kinds of vagueness are equally troublesome for approaches to statistics that do not use prior probabilities. Vagueness about the likelihood function, that is, the process generating the data, is typically substantial and hard to deal with. Moreover, both classical and Bayesian decision theory bring in the idea of utility, and utilities often are vague.

In assessing prior probabilities, skillful self-interrogation is needed in order to mitigate vagueness. Self-interrogation may be made more systematic and illuminating in several ways. (1) Direct judgmental assessment. In assessing the prior distribution of $\tilde{p}$, for example, we might ask, "For what p would we be indifferent to an even money bet that $\tilde{p}$ is above or below this value?" (Answer is the .50-fractile or median.) Then, "If we were told that $\tilde{p}$ is above the .50-fractile just assessed, but nothing more, for what value of p would we now be indifferent in such a bet?" (Answer is the .75-fractile.) Similarly we might locate other key fractiles, or key relative heights on the density function. (2) Translation to equivalent but hypothetical prior sample evidence. For example, we might feel that our prior opinion about $\tilde{p}$ is roughly what it would have been if we had initially held a uniform prior, and then seen r heads in n hypothetical trials from the process. The implied posterior distribution would serve as the prior. (3) Contemplation of possible sample outcomes. Sometimes we may find it easy to decide directly what our posterior distribution would be if a certain hypothetical sample outcome were to materialize. We can then work backwards to see the prior distribution thereby implied. Of course, this approach is likely to be helpful only if the hypothetical sample outcomes are easy to assimilate. For example, if we make a certain technical assumption about the general shape of the prior distribution (beta distribution), the answers to the following two simply-stated questions imply a prior distribution of $\tilde{p}$: (a) How do we assess the probability of heads on a single trial? (b) If we were to observe a head on a single trial (this is the hypothetical future outcome) how would we assess the probability of heads on a second trial?

These approaches are intended only to be suggestive. If several approaches to self-interrogation lead to substantially different prior distributions, we must either try to remove the internal inconsistency or be content with an intuitive analysis. Actually, from the point of view of "subjective probability consistent," the discovery of internal inconsistency in one's judgments is the only route toward more "rational" decisions. The danger is not that internal inconsistencies will be revealed but that they will be suppressed by self-deception or glossed over by lethargy.

It may happen that vagueness affects only unimportant aspects of the prior distribution: theoretical or empirical analysis may show that the posterior distribution is insensitive to these aspects of the distribution. For example, we may be vague about many aspects of the prior distribution, yet feel that it is nearly uniform over all values of the parameter for which the likelihood function is not essentially zero. This has been called a diffuse, informationless, or locally-uniform prior distribution. These terms are to be interpreted relative to the spread of the likelihood function, which depends on the sample size; a prior that is diffuse relative to a large sample may not be diffuse relative to a small one. If the prior distribution is diffuse, the posterior distribution can be easily approximated from the assumption of a strictly uniform prior distribution. The latter assumption, known historically as Bayes' postulate (not to be confused with Bayes' theorem), is regarded mainly as a device that leads to good approximations in certain circumstances, although supporters of "subjective probability rational" sometimes regard it as more than that in their approach to Bayesian inference. The uniform prior is also useful for statistical reporting, since it leads to posterior distributions from which the likelihood is easily recovered and presents the results in a form readily usable to any reader whose prior distribution is diffuse.

## Probabilistic Prediction

A distribution, prior or posterior, of the parameter $\tilde{p}$ of a Bernoulli process implies a probabilistic prediction for any future sample to be drawn from the process, assuming that the stopping rule is given. For example, the denominator in the right hand side of the Bayes' formula for Bernoulli sampling (p. 3) can be interpreted as the probability of obtaining the particular sample actually observed, given the

prior distribution of $\tilde{p}$. While a person's subjective probability distribution of $\tilde{p}$ cannot be said to be "right" or "wrong," there are better and worse subjective distributions, and the test is predictive accuracy. Thus if Mr. A and Mr. B each has a distribution for $\tilde{p}$, and a new sample is then observed, we can calculate the probability of the sample in the light of each prior distribution. The ratio of these probabilities, technically a marginal likelihood ratio, measures the extent to which the data favor A over B or vice-versa. This idea has important consequences for evaluating judgments and selecting statistical models.

In connection with the previous paragraph a separate point is worth making. The posterior distributions of A and B are bound to grow closer together as sample evidence piles up, so long as neither of the priors was dogmatic. An example of a dogmatic prior would be the opinion that $\tilde{p}$ is exactly .5.

## Multivariate Inference and Nuisance Parameters

Thus far we have used one basic example, inferences about a Bernoulli process. To introduce some additional concepts, we now turn to inferences about the mean $\mu$ of a normal distribution with unknown variance $\sigma^2$. In this case we begin with a joint prior distribution for $\tilde{\mu}$ and $\tilde{\sigma}^2$. The likelihood function is now a function of two variables, $\mu$ and $\sigma^2$. An inspection of the likelihood function will show not only that the sequence of observations is irrelevant to inference, but also that the magnitudes are irrelevant except insofar as they help determine the sample mean $\bar{x}$ and variance $s^2$, which, along with the sample size $n$, are the sufficient statistics of this example (see SUFFICIENCY). The prior distribution combines with the likelihood essentially as before except that a double integration (or double summation) is needed instead of a single integration (or summation). The result is a joint posterior distribution of $\tilde{\mu}$ and $\tilde{\sigma}^2$.

If we are interested only in $\tilde{\mu}$, then $\sigma^2$ is said to be a nuisance parameter. In principle it is simple to deal with a nuisance parameter: we "integrate it out" of the posterior distribution. In our example this means that we must find the marginal distribution of $\tilde{\mu}$ from the joint posterior distribution of $\tilde{\mu}$ and $\tilde{\sigma}^2$.

Multivariate problems and nuisance parameters can always be dealt with by the approach just described. The integrations required may demand heavy computation, but the task is straightforward. A more difficult problem is that of assessing multivariate prior distributions, and research is needed to find better techniques for overcoming the problems presented by vagueness in such assessments.

## Design of Experiments and Surveys

So far we have talked only about problems of analysis of samples, without saying anything about what kind of sample evidence, and how much, should be sought. This kind of problem is known as a problem of design. A formal Bayesian solution of a design problem requires that we look beyond the posterior distribution to the ultimate decisions that will be made in the light of this distribution: the best design depends on the purposes to be served by collecting the data. Given the specific purpose and the principle of maximization of expected utility, it is possible to calculate the expected utility of the best act for any particular sample outcome. We can repeat this for each possible sample outcome for a given sample design. Next, we can weight each such utility by the probability of the corresponding outcome in the light of the prior distribution. This gives an overall expected utility for any proposed design. Finally, we pick the sample design with the highest expected utility. For two-action problems--e.g., deciding whether a new medical treatment is better or worse than a standard treatment--this procedure is in no conflict with the traditional approach of selecting designs by comparing operating characteristics, although it formalizes certain things--prior probabilities and utilities--that often are treated intuitively in the traditional approach.

## Comparison of Bayesian and Classical Inference

Certain common statistical practices are subject to criticism either from the point of view of Bayesian or of classical theory: for example, estimation problems are frequently regarded as tests of null hypotheses, and .05 or .01 significance levels are used inflexibly. Bayesian and classical theory are in many respects closer to each other than either is to everyday practice. In comparing the two approaches, therefore, we shall confine the discussion to the level of underlying theory. In one sense the basic difference is the acceptance of subjective probability judgment as a formal component of Bayesian inference. This does not mean that classical theorists would disavow judgment, only that they would apply it informally after the "purely statistical" analysis is finished: judgment is the "second span in the bridge of inference." Building on subjective probability, Bayesian theory is a unified theory, whereas classical theory is diverse and ad hoc. In this sense

Bayesian theory is simpler. But in another sense Bayesian theory is more complex because it incorporates more into the formal analysis. Consider a famous controversy of classical statistics, the problem of comparing the means of two normal distributions with possibly unequal and unknown variances (the so-called "Behrens-Fisher" problem). Conceptually this problem poses major difficulties for some classical theories (not Fisher's fiducial inference; see FIDUCIAL INFERENCE), but none for Bayesian theory. In application, however, the Bayesian approach faces the problem of assessing a prior distribution involving four random variables. Moreover, there may be messy computational work after the prior distribution has been assessed.

In many applications, however, a credible interval emerging from the assumption of a diffuse prior distribution is identical or nearly identical to the corresponding confidence interval. There is a difference of interpretation, illustrated in the opening two paragraphs of this article, but in practice many people interpret the classical result in the Bayesian way. There often are numerical similarities between the results of Bayesian and classical analyses of the same data, but there can also be substantial differences, for example, when the prior distribution is non-diffuse and when a genuine null hypothesis is to be tested.

Often it may happen that the problem of vagueness, discussed at some length above, makes a formal Bayesian analysis seem unwise. In this event Bayesian theory may still be of some value in selecting a descriptive analysis or a classical technique that conforms well to the general Bayesian approach, and perhaps in modifying the classical technique. For example, many of the classical developments in sample surveys and analysis of experiments can be given rough Bayesian interpretations when vagueness about the likelihood (as opposed to prior probabilities) prevents a full Bayesian analysis. Moreover, even an abortive Bayesian analysis may contribute insight into a problem.

Bayesian inference has as yet received much less theoretical study than has classical inference. It is hard at this writing to predict how far Bayesian theory will lead in modification and reinterpretation of classical theory. Before a fully Bayesian replacement is available there is certainly no need to discard those classical techniques that seem roughly compatible with the Bayesian approach; indeed, many classical techniques are, under certain conditions, good approximations to fully Bayesian ones. In the meanwhile, the interaction between the two approaches promises to lead to fruitful developments in statistical inference, and the Bayesian approach promises to illuminate a number of problems--such as allowance for selectivity--that are otherwise hard to cope with.

## A Few Suggestions for Further Reading

The first book-length development of Bayesian inference, which emphasizes heavily the decision-theoretic foundations of the subject, is Robert Schlaifer, Probability and Statistics for Business Decisions (New York: McGraw-Hill Book Company, Inc., 1959). A more technical development of the subject is given by Howard Raiffa and Robert Schlaifer, Applied Statistical Decision Theory (Boston: Division of Research, Graduate School of Business Administration, Harvard University, 1961). An excellent short introduction with an extensive bibliography is Leonard J. Savage, "Bayesian Statistics," in Robert E. Machol and Paul Gray, eds., Recent Developments in Information and Decision Processes (New York: The Macmillan Company, 1962). An interesting application of Bayesian inference is given, along with a penetrating discussion of underlying philosophy and a comparison with the corresponding classical analysis, in Frederick Mosteller and David L. Wallace, "Inference in an Authorship Problem," Journal of the American Statistical Association (Vol. 58, 1963), pp. 275-310. A fuller description of this study will be found in Mosteller and Wallace, Inference and Disputed Authorship: the Federalist Papers. (Reading, Massachusetts: Addison-Wesley Publishing Company, Inc., in press.) This study gives a specific example of how one might cope with vagueness about the likelihood function. Another example is to be found in George E. P. Box and George C. Tiao, "A Further Look at Robustness via Bayes' Theorem," Biometrika (Vol. 49, 1962), pp. 419-432. A thorough development of Bayesian inference from the viewpoint of "subjective probability rational" is to be found in Harold Jeffreys, Theory of Probability (Oxford: Clarendon Press, 3rd edition, 1961).