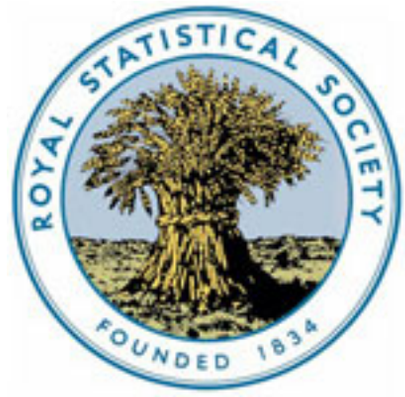

Bayesian Interpretation of Standard Inference Statements

By JOHN W. PRATT

Graduate School of Business Administration, Harvard University

[Read at a RESEARCH METHODS MEETING OF THE SOCIETY, December 2nd, 1964, Professor D. V. LINDLEY in the Chair]

1. INTRODUCTION

THIS paper is an attempt to present in an orderly way various ideas about the interpretation of standard inference statements from the Bayesian point of view. A Bayesian of any variety naturally wants to know what use he can make of standard methods, because they are often used by others and because they have been worked out for so many situations, including some which seem difficult or unrewarding to handle Bayesianly. We shall see that a Bayesian can make considerable use of some standard methods. A non-Bayesian, if he feels there is some element of sense in some Bayesian point of view in some circumstances, may expect a Bayesian lamp to throw some light on his methods. I hope this expectation will be fulfilled for some non-Bayesians, as it is for me.

The use of insufficient statistics will be considered first, in Section 2. It proves easy to assimilate them into the Bayesian framework. An example is given in which this leads to some progress on a problem of Bayesian non-parametric statistics. Estimation and confidence regions are taken up in Sections 3 and 4 respectively, and it is shown that classical properties give approximately, in a certain weak Bayesian sense, corresponding Bayesian properties. Classical anomalies no longer seem disturbing from this point of view. Ideas related to maximum likelihood are postponed to Section 5, and some general remarks concerning the approximation idea of Sections 3 and 4, including the application of the likelihood principle, are postponed to Section 6. Tests of hypotheses are discussed in the next two Sections, significance levels and *P*-values in Section 7 and common uses of tests in Section 8. The general conclusion here is that only certain one-tailed *P*-values are interpretable Bayesianly and that, even when this interpretation is applicable, conventional tests are seldom well articulated to practical problems. Section 9 contains a few final comments.

Reference to "standard" methods is not meant to suggest that any particular methods are definitely standard or that any particular person's practical or theoretical philosophy leads, for instance, to exactly the uses of tests that will be described. Nevertheless, I intend what I will say to apply with only minor modification and qualification to most commonly used methods developed in the "orthodox", "classical", "objective", "frequency" or "Neyman–Pearson" tradition or traditions.

Though I make no formal claim to originality, I will identify those isolatable ideas which I am conscious of having obtained from others without their being so widely known as to need no reference. No attempt will be made, however, to trace the origins of any specific idea, let alone of the present climate of statistics.

2. INSUFFICIENT STATISTICS

The first step in a standard analysis is to choose a statistic, and except in the simplest problems the statistic is not sufficient: for instance, typical *t*, *F* and chi-square

statistics; the sample mean and variance when the population may not be normal; the sample median in a non-parametric situation; multiple R^2; measures of association.

A full Bayesian analysis typically requires a sufficient statistic. In fact, if the prior density is positive everywhere (and if it is not, the parameter space may as well be reduced so that it is), a statistic determines the posterior distribution given all observations if and only if it is sufficient. (This follows immediately from Bayes's formula. See also Raiffa and Schlaifer, 1961, Section 2.2.) Insufficient statistics may be brought into the picture by breaking the full analysis into two steps, of which the second may often seem not worth carrying out: (1) compute the posterior distribution given the statistic; (2) with this as prior and with the likelihood obtained from the conditional distribution of the observations given the statistic, compute the posterior distribution given the observations. Specifically, denote the observation set by x, the statistic by t and the parameter of interest by θ. (In general x, t and θ may be vectors. We do not assume, however, that θ includes all nuisance parameters relevant to the distribution of x or even of t.) The first step is to compute the density of θ given t, which is, in self-explanatory notation,

$$f_1(\theta) = f(\theta \mid t) = \frac{f_0(\theta) f(t \mid \theta)}{f(t)}. \tag{1}$$

The second step, leading to the density of θ given x, would be

$$f_2(\theta) = f(\theta \mid x) = \frac{f_1(\theta) f(x \mid t, \theta)}{f(x \mid t)}. \tag{2}$$

Often, however, this second step is very troublesome to carry out and yet seems unlikely to change the posterior distribution of θ very much, so one might decide to omit it. Thus it is sometimes reasonable for a Bayesian to make inferences conditional on an insufficient statistic. Furthermore, one might expect a standard inference statement based on an insufficient statistic to be more like a posterior probability statement conditional on the statistic than like one conditional on the full data, and we shall see that this is true in a sense explained precisely below (Theorems 3 and 6). Of course nothing said here justifies choosing a statistic (such as chi-square) just because its distribution under some hypothesis is particularly easy to compute; this may lead to an inference about the wrong parameters and discard too much information about the parameters of real interest.

For an example of the Bayesian use of an insufficient statistic, suppose x is a large sample from a population whose mean θ is the parameter of interest. The shape of the population is unknown but its variance σ^2 is known to an adequate approximation. (Perhaps someone told us the sample variance.) Let t be the sample mean. Then t is approximately normal with mean θ and variance σ^2/n. The posterior density of θ given t is, therefore,

$$f(\theta \mid t) \simeq K f_0(\theta) \exp\{-\tfrac{1}{2} n(\theta - t)^2/\sigma^2\}. \tag{3}$$

(Here and later K denotes a normalizing constant which may depend on the statistics but not on the parameters, and $\simeq$ denotes approximate equality.) This calculation requires the prior distribution of θ only. To get the posterior distribution of θ given x, however, we would have to assess the prior joint distribution of θ and the population shape, compute the posterior joint distribution of θ and the shape, and then integrate out the shape. The shape parameter would be multi-dimensional or, under a non-parametric assumption, infinite-dimensional; a responsible assessment of its prior

distribution would involve considerable mental anguish; and the computation might become difficult or expensive. At the same time, the data beyond the sample mean (with σ^2 known) will presumably affect the distribution of θ very little, so an equal effort placed elsewhere might be expected to yield more. Technically, computation of $f(x \mid t, \theta)$ in (2) would involve integration with respect to a distribution of the shape parameter. A similar step would be required for $f(x \mid \theta)$ in a direct calculation

$$f(\theta \mid x) = \frac{f_0(\theta) f(x \mid \theta)}{f(x)}. \tag{4}$$

The observations are not independent given the mean θ alone, but only given θ and the shape.

2.1. *Further Conditioning in Connection with the Mean*

In the situation above, it would be more satisfying to treat σ^2 as unknown and let $t = (\bar{x}, s^2)$, where $\bar{x}$ and s^2 are the sample mean and variance. We shall now consider first this t and then, more generally, $t = (\bar{x}, s^2, t_1, \ldots, t_r)$. One imprecise theorem will be presented for whatever interest it may have in the area of asymptotic Bayesian non-parametric statistics, but we shall not even begin to discuss this area as such. As far as the main topic of this paper is concerned, this entire subsection is a side-road which can and perhaps should be by-passed.

Consider then the situation leading to (3). Conditioning on $t = (\bar{x}, s^2)$ gives the same result as that above with $\sigma^2 = s^2$, provided the joint prior density of θ, σ^2 and the third and fourth central moments of the population varies gently with σ^2. This may be seen as follows. Introduce the quantities

$$u_1 = n^{\frac{1}{2}}(\bar{x} - \theta)/s, \quad u_2 = n^{\frac{1}{2}}(s^2 - \sigma^2)/s^2. \tag{5}$$

Given the population, u_1 and u_2 are approximately joint normal with means 0, variances 1 and λ_4, covariance λ_3, and density

$$(1/2\pi)(\lambda_4 - \lambda_3^2)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\lambda_4 u_1^2 - 2\lambda_3 u_1 u_2 + u_2^2)/(\lambda_4 - \lambda_3^2)\}, \tag{6}$$

where λ_3 and λ_4 depend on the population but not on n. (Specifically, $\lambda_3 \sigma^3$ and $\lambda_4 \sigma^4 + \sigma^4$ are the third and fourth central moments of the population.) Notice that (u_1, u_2) is a one-to-one function of $(\bar{x}, s^2)$ for given (θ, σ^2), and of (θ, σ^2) for given $(\bar{x}, s^2)$. A straightforward argument, explained in more detail in a more general case later, shows that the joint density of $u_1, u_2, \lambda_3, \lambda_4$ given $\bar{x}, s^2$, equals the product of the joint density of u_1, u_2 given λ_3, λ_4, which is approximately (6), and

$$K(1 - n^{-\frac{1}{2}} u_2) f_0(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2 - n^{-\frac{1}{2}} s^2 u_2, \lambda_3, \lambda_4), \tag{7}$$

where $f_0(\theta, \sigma^2, \lambda_3, \lambda_4)$ is the joint prior density of $\theta, \sigma^2, \lambda_3, \lambda_4$. The density of u_1 and hence of θ given $\bar{x}$ and s^2 may be obtained by integrating this product with respect to u_2, λ_3 and λ_4. If $f_0(\theta, \sigma^2, \lambda_3, \lambda_4)$ varies gently with σ^2, then (7) is approximately

$$K f_0(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2, \lambda_3, \lambda_4). \tag{8}$$

Let us integrate the product of (6) and (8) with respect to u_2 first. Since (8) does not depend on u_2, and (6) is a normal density in u_1 and u_2, where u_1 has mean 0 and variance 1, we obtain

$$K f_0(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2, \lambda_3, \lambda_4) e^{-\frac{1}{2}u_1^2}. \tag{9}$$

Now integrating with respect to λ_3 and λ_4, we find that

$$f(u_1 | \bar{x}, s^2) \simeq K f_{\theta,\sigma^2}(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2)\, e^{-\frac{1}{2}u_1^2}, \tag{10}$$

where f_{θ,σ^2} is the prior joint density of θ and σ^2, marginal on λ_3 and λ_4. Dividing by the marginal prior density of σ^2 at s^2, which may be subsumed into the normalizing constant, we may replace f_{θ,σ^2} in (10) by the prior conditional density of θ given σ^2. The density of θ given $\bar{x}$ and s^2 is therefore

$$f(\theta | \bar{x}, s^2) \simeq K f(\theta | \sigma^2 = s^2) \exp\{-\tfrac{1}{2}n(\theta - \bar{x})^2/s^2\} \tag{11}$$

in self-explanatory if inelegant notation. This is the same as the approximate density of θ given $\bar{x}$ $(= t)$ obtained above, at (3), with σ^2 treated as known and equal to s^2.

Note that, as posterior distributions of θ, the t and normal distributions are indistinguishable to the order of approximation used here. It is natural to suppose that an order of approximation sufficient to distinguish them will yield approximately the t distribution with the usual number of degrees of freedom (Raiffa and Schlaifer, 1961, Section 11.5.5) if the shape parameters λ_3 and λ_4 are assumed *a priori*, or appear *a posteriori*, to be approximately equal to their values for normal populations (and an appropriate prior distribution of θ and σ^2 is assumed) but not generally otherwise.

Note also that we can obtain the approximate posterior distribution of θ given $t = (\bar{x}, s^2)$ without actually assessing the joint prior distribution of θ, σ^2, λ_3 and λ_4, the parameters of the sampling distribution of t. What is required is a qualitative statement about the joint distribution, that it varies gently with σ^2, together with an assessment of the prior distribution of the parameter θ of interest. To obtain the posterior distribution of θ given x, one would have to say something about the prior distribution of all the parameters of the distribution of x.

We might go further and condition on $t = (\bar{x}, s^2, t_3, \ldots, t_r)$, where $t_3, \ldots, t_r$ are statistics estimating some further characteristics of the population. Specifically, suppose u_1, u_2 and

$$u_i = n^{\frac{1}{2}}(t_i - \tau_i) \quad (i = 3, \ldots, r) \tag{12}$$

are approximately jointly normal with means 0 and variance matrix $\Lambda = [\lambda_{ij}]$, where $\tau_3, \ldots, \tau_r$ and Λ depend on the population but not on n. (Since u_1 and u_2 are the same as before, $\lambda_{11} = 1$, λ_{12} is the former λ_3 and λ_{22} is the former λ_4.)

Theorem 1. The density of θ given t is

$$f(\theta | t) \simeq K f(\theta | \sigma^2 = s^2, \tau_3 = t_3, \ldots, \tau_r = t_r) \exp\{-\tfrac{1}{2}n(\theta - \bar{x})^2/s^2\} \tag{13}$$

if the foregoing assumption holds and if there exists a set of elements of Λ, which we shall regard as the components of a vector λ, such that:

(a) $\theta, \sigma^2, \tau_3, \ldots, \tau_r$ and the components of λ are functionally independent;

(b) λ_{1j} is a function of λ, θ, and $\tau_{j+1}, \ldots, \tau_r$ for $2 \leqslant j \leqslant r$;

(c) λ_{ij} is a function of λ, θ, and $\tau_{i+1}, \ldots, \tau_r$ for $2 \leqslant i \leqslant j \leqslant r$;

(d) the joint prior density of $\theta, \sigma^2, \tau_3, \ldots, \tau_r$, and λ is approximately constant for variations of order $n^{-\frac{1}{2}}$ in $\sigma^2, \tau_3, \ldots, \tau_r$.

We do not require $\theta, \sigma^2, \tau_3, \ldots, \tau_r$, and all λ_{ij} to be functionally independent because we might want to apply the theorem to a parametric situation or, for instance, to let t_3 and t_4 be estimates of λ_{12} and λ_{22}, so that $\lambda_{12} = \tau_3$ and $\lambda_{22} = \tau_4$.

The theorem sometimes becomes applicable upon replacing $t_3, \ldots, t_r$ by some other statistics $T_3, \ldots, T_R$ such that $(\bar{x}, s^2, T_3, \ldots, T_R)$ is a one-to-one function of $(\bar{x}, s^2, t_3, \ldots, t_r)$. For instance, it is possible to satisfy at least (a)–(c) in this way when $t_3, \ldots, t_r$ is any

finite set of sample moments and fractiles (not depending on n) if the population is permitted to have an arbitrary density which is positive in some neighbourhood of each relevant population fractile.

Proof. Let $t = (\bar{x}, s^2, t_3, \ldots, t_r)$ as above, $u = (u_1, \ldots, u_r)$, and $\tau = (\theta, \sigma^2, \tau_3, \ldots, \tau_r)$. Since u is a one-to-one function of t for given τ, it follows that

$$f(t \mid \tau, \lambda) = f(u \mid \tau, \lambda) J(u : t), \tag{14}$$

where $J(:)$ is the absolute value of the Jacobian of the first argument with respect to the second and where u on the right-hand side is to be expressed as a function of τ and t. (The notation is inexact in several ways but adequate for the present argument.) By Bayes's formula, therefore,

$$f(\tau, \lambda \mid t) = K f_0(\tau, \lambda) f(u \mid \tau, \lambda) J(u : t), \tag{15}$$

where $f_0(\tau, \lambda)$ is the prior density of (τ, λ). Since u is a one-to-one function of τ for given t, it follows that

$$f(u, \lambda \mid t) = K f_0(\tau, \lambda) f(u \mid \tau, \lambda) J(u : t) J(\tau : u), \tag{16}$$

where now on the right-hand side τ is to be expressed as a function of u and t.

In the present situation we have the approximation

$$f(u \mid \tau, \lambda) \simeq K \mid \Lambda \mid^{-\frac{1}{2}} e^{-\frac{1}{2} u \Lambda u'}, \tag{17}$$

where, on the right-hand side, u is a row vector and u' its transpose. Furthermore, $J(u : t) = n^{\frac{1}{2}k} \sigma^2 / s^5$ and $J(\tau : u) = s^3 / n^{\frac{1}{2}n}$, so (16) becomes

$$f(u, \lambda \mid t) \simeq K f_0(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2 - n^{-\frac{1}{2}} s^2 u_2, t_3 - n^{-\frac{1}{2}} u_3, \ldots, t_r - n^{-\frac{1}{2}} u_r, \lambda) \times f(u \mid \tau, \lambda)(1 - n^{-\frac{1}{2}} u_2), \tag{18}$$

where $f_0(\theta, \sigma^2, \tau_3, \ldots, \tau_r, \lambda)$ is the prior density of $(\theta, \sigma^2, \tau_3, \ldots, \tau_r, \lambda)$. By assumption (d), (17) and (18), we have the approximation

$$f(u, \lambda \mid t) \simeq K f_0(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2, t_3, \ldots, t_r, \lambda) \mid \Lambda \mid^{-\frac{1}{2}} e^{-\frac{1}{2} u \Lambda u'}, \tag{19}$$

where the elements λ_{ij} of Λ are now to be regarded as functions of λ, u_1 and $u_3, \ldots, u_r$. Integrating (19) with respect to $u_2, \ldots, u_r$ in order gives

$$f(u_1, \lambda \mid t) \simeq K f_0(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2, t_3, \ldots, t_r, \lambda) e^{-\frac{1}{2} u_1^2}, \tag{20}$$

since the final factor at each stage is a joint normal density in the remaining u_i's, and any λ_{ij} depending on u_m will have $2 \leqslant i \leqslant m$ or $2 \leqslant j \leqslant m$ by assumptions (b) and (c) and hence will have disappeared before the integration with respect to u_m. Integrating with respect to λ now gives

$$f(u_1 \mid t) \simeq K f_{\theta, \sigma^2, \tau_3, \ldots, \tau_r}(\bar{x} - n^{-\frac{1}{2}} s u_1, s^2, t_3, \ldots, t_r) e^{-\frac{1}{2} u_1^2}. \tag{21}$$

Making a transformation from u_1 to θ and dividing by the prior density of $(\sigma^2, \tau_3, \ldots, \tau_r)$ at $(s^2, t_3, \ldots, t_r)$ gives (13), and the theorem is established.

3. Estimation

Let ω index the possible distributions of the observation set x. (A nonparametric family is permitted.) A statistic $t = t(x)$ is called an unbiased (or mean unbiased) estimator of a parameter $\tau = \tau(\omega)$ if

$$E(t \mid \omega) = \tau(\omega) \quad \text{for all } \omega. \tag{22}$$

The natural Bayesian counterpart is the posterior mean of τ, $E(\tau|x)$, or perhaps, in view of the previous Section, $E(\tau|t)$. The concepts of unbiased estimator and posterior mean are not equivalent, of course, but there is a sense in which they are approximately the same, albeit a very limited one. Specifically, we have the following theorem.

Theorem 2. If t is an unbiased estimator of $\tau(\omega)$, then

$$E\{t-E(\tau|x)\}=0. \tag{23}$$

(Here and in similar expressions hereafter, all statistics and all parameters are random variables, and expectations and probabilities refer to their joint distribution. Those who like tildes over random variables may insert them mentally over all statistics and parameters operated on by expectation or probability signs.) Theorem 2 says, concerning the unbiased estimator t and the posterior mean of τ, which are both functions of x, that their difference has prior (marginal) expectation zero. (The expectation of the posterior mean of τ is its prior mean, which is also the prior mean of t if t is unbiased.) Accordingly, if the situation appears typical after observing x or evaluating t, we have some reason to hope that an unbiased estimator is approximately a posterior mean, and vice versa. Section 6 includes a few remarks on "appearing typical".

As pointed out by Robert Berk, the same argument would apply to $E\{\tau|y\}$ for any statistic y (of any number of dimensions); that is, $E(\tau|x)$ can be replaced by $E(\tau|y)$ in Theorem 2, and this raises the question which y makes the approximation closest. Once we have t in hand, it is natural to treat t at least as known and therefore to condition on a statistic y of which t is a function. Among such statistics, the choice $y=t$ makes the approximation closest, and in particular closer than the choice $y=x$. The precise sense of "closer" is that the prior expectation of any convex function of the error is smaller.

Theorem 3. If t is a function of y, and ψ is any convex function, then

$$E[\psi\{t-E(\tau|y)\}]\geqslant E[\psi\{t-E(\tau|t)\}]. \tag{24}$$

Proof. By Jensen's inequality and the fact that t is a function of y,

$$E[\psi\{t-E(\tau|y)\}|t]\geqslant\psi[E\{t-E(\tau|y)|t\}]=\psi\{t-E(\tau|t)\}. \tag{25}$$

Taking the expected value of both sides then gives (24).

We have argued that an unbiased estimator t may be regarded as an approximation to the posterior mean $E(\tau|x)$ or $E(\tau|t)$. Section 4 contains a similar argument for confidence regions. General comments on this kind of approximation will be postponed to Section 6. A few more specific remarks will be made here, however.

The virtue of an unbiased estimator t is that the expectation of its error is zero. This is true conditional on ω or, therefore, on τ, but it is not true conditional on x or t. When an estimate (the numerical value of an estimator) is given in connection with some particular body of data x, it is all too easy to misinterpret "unbiased" as meaning that the expectation of the error is zero under the existing circumstances, which of course include knowing x. Even if it is carefully explained that "unbiased" only means that the average error would approach zero in certain kinds of infinite sequences of repetitions, most people will continue to feel that the error of the estimate at hand for the parameter they are interested in at the moment has expectation zero. I think they should not be blamed for doing so in that any other kind of unbiasedness is really irrelevant under their circumstances. A responsible statistician would presumably not call an estimator unbiased if he would regard as unfair a bet made *with knowledge*

of x and with payoff equal to the error of estimate. (The concept of conditional inference and Fisher's concept of a "relevant subsequence" are pertinent here. They are discussed further in Section 6.) The point of this Section is that what people naturally think, is true in a very limited sense, namely that for an unbiased estimator t, the expected error given x, although not necessarily zero, at least had prior expectation zero, that is,

$$E\{E(t-\tau|x)\} = 0, \tag{26}$$

to paraphrase Theorem 2.

Some find it disturbing from the orthodox point of view that unbiased estimators do not "transform properly": if t is an unbiased estimator of τ, then $g(t)$ is usually not an unbiased estimator of $g(\tau)$. From the point of view advanced here, this is very natural: t is an approximation to $E(\tau|x)$, so $g(t)$ might be an approximation to $g\{E(\tau|x)\}$, but $g\{E(\tau|x)\}$ is usually not equal to $E\{g(\tau)|x\}$, so there is no reason why $g(t)$ should be an approximation to $E\{g(\tau)|x\}$. In short, posterior means do not "transform properly", so there is no reason why approximations to them should. Furthermore, from the Bayesian point of view, whether you should use the posterior mean of τ or of $g(\tau)$ or neither depends on the economics of the situation (losses). Thus the optimum estimator inevitably depends on a relevant external factor and is simply not determined by the probability model in itself. A theory of estimation in which this external factor plays no role is bound to be unsatisfactory in some respect.

By the ordinary method of moments, to estimate a k-dimensional parameter, one sets the first k raw sample moments equal to the corresponding population moments and solves for the parameters. Since the raw sample moments are unbiased estimators of the corresponding population moments, this may be regarded as giving approximate posterior means of the raw population moments. The estimators of other parameters then correspond, but have no unusual virtues except convenience.

Maximum-likelihood estimation will be discussed in Section 5.

Median unbiased estimators are confidence bounds at the 50 per cent level and in this guise are discussed in Section 4.

An estimator t_n of a parameter τ is called consistent if it converges in probability to τ, that is,

$$P(|t_n-\tau|>\varepsilon|\omega)\to 0 \quad \text{for all } \varepsilon \text{ and all } \omega \tag{27}$$

as the sample size (or sizes) $n\to\infty$. (The subscript n expresses the fact that t_n depends on the sample size, so that we are really concerned with a sequence of estimators.) The corresponding Bayesian property is not as easy to state, since the posterior probability $P(|t_n-\tau|>\varepsilon|x_n)$ is conditional on a random variable x_n which depends on the sample size. (Here x_n denotes all observations of a sample of size n.) The following theorem states that this posterior probability does approach zero in the sense of convergence in probability if t_n is a consistent estimator of τ. As with Theorem 2, nothing is guaranteed for any particular x_n, but if the situation appears typical after observing x_n and the sample size is large, we have some reason to suppose that the posterior distribution of τ is concentrated near t_n.

Theorem 4. If t_n is a consistent estimator of τ, then

$$P\{P(|t_n-\tau|>\varepsilon|x_n)>\delta\}\to 0 \quad \text{for all } \varepsilon \text{ and } \delta. \tag{28}$$

Proof. By the properties of conditional probability,

$$E\{P(|t_n-\tau|>\varepsilon|x_n)\} = P(|t_n-\tau|>\varepsilon) = E\{P(|t_n-\tau|>\varepsilon|\omega)\}. \tag{29}$$

The right-hand side tends to 0 by (27) and the Lebesgue convergence theorem. But it is easy to see that if (28) is false, the left-hand side of (29) cannot tend to 0. This proves (28).

4. Confidence Regions

Let ω index the possible distributions of the observation set x, as in Section 3. Again a non-parametric family is permitted. A confidence region $R = R(x)$ for a parameter $\tau = \tau(\omega)$ at the level (or conservative level) $1-\alpha$ is defined by the property

$$P(\tau \in R \mid \omega) \geqslant 1-\alpha \quad \text{for all } \omega. \tag{30}$$

If R satisfies (30) with $\geqslant$ replaced by $=$, it is a confidence region for τ at the exact level $1-\alpha$. The natural Bayesian counterpart of the (especially the exact) confidence level is the posterior probability $P(\tau \in R \mid x)$.

For any system of regions $R = R(x)$, we have

$$E\{P(\tau \in R \mid \omega)\} = P(\tau \in R) = E\{P(\tau \in R \mid x)\}. \tag{31}$$

From this the following theorem is immediate (Pratt, 1963).

Theorem 5. If R is a confidence region for τ at the exact (conservative) level $1-\alpha$, then

$$E\{P(\tau \in R \mid x)\} = (\geqslant)\, 1-\alpha. \tag{32}$$

This says that the posterior probability that R covers τ, which is a function of x, has prior expectation exactly (at least) $1-\alpha$. Accordingly, if the situation appears typical after observing x, we have some reason to hope that a confidence region at the exact level $1-\alpha$ has posterior probability approximately $1-\alpha$. (For convenience, we will consider exact levels for now.) As in Section 3, the same argument would apply to $P(\tau \in R \mid y)$ for any y. If y is a trivial (completely uninformative) statistic, then

$$P(\tau \in R \mid y) = P(\tau \in R) = 1-\alpha$$

and the approximation is exact. Once we have R in hand, however, it is natural to treat R as known and therefore to condition on a statistic y of which R is a function. Among such statistics, the choice $y = R$ (or a one-to-one function of R) makes the approximation closest. The precise sense of "closest" is again that the prior expectation of any convex function of the error is smallest.

Theorem 6. If R is a confidence region for τ at the exact level $1-\alpha$, R is a function of y, and ψ is any convex function, then

$$E[\psi\{1-\alpha-P(\tau \in R \mid y)\}] \geqslant E[\psi\{1-\alpha-P(\tau \in R \mid R)\}]. \tag{33}$$

The proof is like that of Theorem 3.

When a particular confidence region $R(x)$, for a parameter τ which is unknown but of interest, has been computed from some known data x, it is all too easy to think the probability $1-\alpha$ still applies. Again I cannot blame anyone for thinking so since no other interpretation relevant to the situation (x known, τ unknown) is psychologically feasible for me. This natural interpretation is justified to a very limited extent by Theorem 5.

Suppose an exact upper (or lower) confidence bound for a one-dimensional parameter τ is available at every level $1-\alpha$. These bounds approximate the fractiles of the posterior distribution of τ. Thus the "confidence cumulative distribution function (cdf)" approximates the posterior cdf.

There are testing problems in which a result is significant at one level but not significant at a less extreme level by the "best" tests at these levels, where "best" means most powerful (against a simple alternative for a composite null hypothesis; see Stein, 1951, Lehmann, 1959, Chapter 3, Problem 29) or Type A (for a simple null hypothesis; see Chernoff, 1951). The corresponding confidence procedures have the property that the region at one level does not contain that at a less extreme level. There are even "best" confidence regions containing all or none of the possible parameter values, where "best" means uniformly most "accurate" unbiased or invariant (Lehmann, 1959, Chapter 5, Example 11, and Section 7.7). These anomalies seem upsetting from the orthodox point of view, and are hard to reconcile with the notion that the confidence property is a fundamental objective of the analysis. On the other hand, if confidence levels are regarded only as approximate posterior probabilities, then one may be annoyed and mistrust the approximation if the anomaly actually arises, but the possibility does not evidence a fundamental flaw in methodology. Confidence procedures have no *fundamental* Bayesian justification, whatever their merits as approximations.

As regards the confidence cdf, the testing situations mentioned suggest that a "best" confidence cdf need not be monotone. The confidence cdf may still be frequently useful as an approximation to the posterior cdf. Compare Edgeworth's series.

Another feature of confidence regions which is disturbing from the orthodox point of view is that one's freedom to choose the confidence level or the region to be considered is severely circumscribed. For instance, one might be interested in the region $\tau(\theta) \leqslant 0$. In the absence of the pathology of the previous paragraph, one could compute the level at which 0 would be an upper confidence bound for τ. This would be a function of the sample, say $1-\alpha(x)$. In the frequency view of probability, however, $\alpha(x)$ has no special meaning. Orthodox theory permits us to say $\tau(\theta) \leqslant 0$ at level $1-\alpha_0$ if we selected the value α_0 in advance and it turned out $\alpha(x) = \alpha_0$, but we may not say $\tau(\theta) \leqslant 0$ at level $1-\alpha_0$ if we selected $\alpha_0 = \alpha(x)$. The region we may look at is imposed upon us by the sample and may not be at all the one we are interested in. All this is further evidence that confidence levels are not exact measures of any fundamental kind of confidence. We need not be fundamentally disturbed, however, if we regard confidence levels merely as approximate posterior probabilities. We are then free to choose the level in any way, though the nature of the approximation, expressed by Theorem 5, suggests perhaps that it will be more accurate on the whole when α is selected in advance.

Two different, entirely legitimate confidence procedures can attach different levels to the same confidence statement. They can also give two confidence statements at the same level, one of which implies the other. Again, these possibilities are hard to reconcile with the notion that the confidence property is fundamental, but are not surprising in approximations.

Discreteness poses two kinds of problems in standard confidence theory. One is exemplified by the usual confidence limits for the parameter p of the binomial distribution: the probability of coverage is a discontinuous, saw-toothed function of p (Stevens, 1950, Fig. 2). A just "conservative" confidence coefficient is usually quoted, but this is non-representative, and to an extent varying with the sample size. Furthermore, just conservative upper and lower confidence limits give an overconservative confidence interval. For example, when $n = 20$, the usual binomial confidence interval at a nominal α of 5 per cent fails to include p with a probability never exceeding

4·2 per cent (and exceeding 3·5 per cent on only about one-fifth of the range of p). If a representative nominal confidence level is used, one cannot claim that the probability of coverage is at least the level stated.

The second kind of problem posed by discreteness is exemplified by the order statistics as confidence limits for the population median (or other fractile): only certain levels are available. In a sample of size $n = 20$, for example, the order statistics are confidence limits for the median at the levels 1·000, 0·999, 0·994, 0·979, 0·942, 0·868, 0·748, 0·588, etc. Both these problems can be avoided by using randomized confidence limits, but almost no one seriously suggests doing so in practice. Stevens (1950, 1957) is an exception to some extent.

In the case of the binomial confidence limits, exactly how great the discreteness problem is can be described in a Bayesian framework. Consider the upper confidence cdf (the confidence cdf arising from just conservative upper confidence limits) and the lower confidence cdf, arising from just conservative lower confidence limits. (A just conservative lower confidence limit at level $1-\alpha$ is also a just anti-conservative upper confidence limit at level α if the latter is defined in the obvious way.) It follows easily from the relation between the binomial and beta distributions that the upper confidence cdf is the posterior cdf under the prior $dp/(1-p)$, and the lower confidence cdf is the posterior cdf under the prior dp/p. Thus the discreteness problem amounts to the difference between the priors $dp/(1-p)$ and dp/p. These priors seem quite different when graphed, but this is an illusion because the difference corresponds to a change of one observation from success to failure. The same change will carry conservative confidence limits into anticonservative ones. The moral (to me) is that the choice of prior is a far more serious problem than discreteness, and if the latter looks serious enough to matter, then moderate differences of prior opinion will have an important effect on the inference and the analysis should bring this out. I presume this moral carries over to discreteness problems of both kinds in most situations.

The following theorem embodies the mathematical content of the previous paragraph and somewhat more.

Theorem 7. Given a binomial observation, the posterior cdf of p equals the upper confidence cdf if the prior density $f(p)$ is the improper density $1/(1-p)$; it equals the lower confidence cdf if $f(p) = 1/p$; and it lies between the two if $(1-p)f(p)$ is non-increasing and $pf(p)$ is non-decreasing. Conversely, if the proper prior density $f(p)$ is differentiable on $0<p<1$ and the posterior cdf of p lies between the upper and lower confidence cdf's for every possible binomial sample size and observation, then $(1-p)f(p)$ is non-increasing and $pf(p)$ is non-decreasing.

Proof. If $f(p) = 1/(1-p)$ and r successes have been observed in n trials, then

$$P(p \leqslant x \mid r) = \int_0^x p^r(1-p)^{n-r-1}\,dp/B(r+1, n-r)$$

$$= \sum_{r+1}^{n} \binom{n}{j} x^j(1-x)^{n-j} \tag{34}$$

by the relation between the binomial and beta cdf's. Therefore x is a just conservative upper confidence limit for p at the level (34). This proves the first statement. The second follows by symmetry or by a similar proof, and the remainder of the theorem by the following lemma.

Lemma. In any one-parameter problem, if F and G are the prior cdf's and F_1 and G_1 the posterior cdf's corresponding to prior densities f and g respectively, and if f/g is non-decreasing, then $F \leqslant G$ and $F_1 \leqslant G_1$. Conversely, in binomial problems, if f/g is differentiable and $F_1 \leqslant G_1$ for every possible binomial sample size and observation, then f/g is non-decreasing.

It will be evident from the proof that the converse could also be generalized.

Proof. If f/g is non-decreasing, then the familiar argument for monotone likelihood ratios gives

$$\frac{F(x)}{1-F(x)} = \frac{\int_{-\infty}^{x} (f/g)\, g}{\int_{x}^{\infty} (f/g)\, g} \leqslant \frac{\{f(x)/g(x)\} \int_{-\infty}^{x} g}{\{f(x)/g(x)\} \int_{x}^{\infty} g} = \frac{G(x)}{1-G(x)}. \tag{35}$$

It follows that $F(x) \leqslant G(x)$. Since the ratio of the posterior densities $f_1/g_1 = f/g$ is also non-decreasing, it also follows that $F_1(x) \leqslant G_1(x)$. To prove the converse, suppose that f/g is not non-decreasing. Then its derivative must be negative somewhere, and therefore there must be an interval (a, b) on which f/g is strictly decreasing. Let x be in (a, b) and let r and n approach infinity in such a way that r/n approaches a point in (a, b). Then the likelihood L becomes negligible outside (a, b) and, by an argument like (35),

$$\frac{F_1(x)}{1-F_1(x)} = \frac{\int_{a}^{x} fL + \text{rem}}{\int_{x}^{b} fL + \text{rem}} > \frac{\int_{a}^{x} gL + \text{rem}}{\int_{x}^{b} gL + \text{rem}} = \frac{G_1(x) + \text{rem}}{1-G_1(x) + \text{rem}}, \tag{36}$$

where the remainders (rem) are small compared to the terms preceding them. This contradicts the assumption that $F_1 \leqslant G_1$ always, proving the converse.

5. Maximum Likelihood

Because the omission of certain points would leave a serious gap, we shall discuss them briefly even though they are well known.

Let the unknown parameters be the elements of a column vector θ. The maximum-likelihood estimate $\hat{\theta}$ is the posterior mode for a prior density which is constant in a neighbourhood of $\hat{\theta}$ and not too large elsewhere. In large samples it is usually nearly the posterior mean.

An asymptotic confidence distribution, which is therefore an asymptotic, approximate posterior distribution, can be derived from the fact that $J^{\frac{1}{2}}(\hat{\theta} - \theta)$ is asymptotically unit (spherical) normal, where $J^{\frac{1}{2}}$ is any square matrix such that $J^{\frac{1}{2}} J^{\frac{1}{2}\prime} = J$ and J may be, for instance, the (Fisherian) information matrix $J(\theta)$, the maximum-likelihood estimate $J(\hat{\theta})$ of $J(\theta)$, or the sample information matrix $\tilde{J}(\hat{\theta})$ defined as the value at $\hat{\theta}$ of the negative of the matrix of second partial derivatives of the log-likelihood with respect to the coordinates of θ. From the orthodox point of view it is perhaps natural to use $J(\theta)$. Often $J(\hat{\theta})$ is substituted because the confidence region $\{\theta : J^{\frac{1}{2}}(\hat{\theta} - \theta) \in R\}$ is much easier to compute when J does not depend on θ. $\tilde{J}(\hat{\theta})$ has the same computational virtue and can be obtained without evaluating any expectations. Still another possibility is to use the approximate normality of the vector of first partial derivatives of the log-likelihood (Bartlett, 1953a, b; 1955); when one also approximates this vector by its linear expansion around $\hat{\theta}$, one obtains

$J = \tilde{J}(\hat{\theta})J^{-1}(\theta)\tilde{J}(\hat{\theta})$. From the Bayesian point of view, one is approximating the log-likelihood by $-\frac{1}{2}(\hat{\theta}-\theta)'J(\hat{\theta}-\theta)$ and the most natural choice of J is $\tilde{J}(\hat{\theta})$ (Jeffreys, 1961, p. 193; Lindley, 1961; Savage, 1961b). The distinction between $\tilde{J}(\hat{\theta})$ and $J(\theta)$ is the distinction between precision obtained and precision expected (Savage, in Savage *et al.*, 1962). Fisher (1925) introduced $\tilde{J}(\hat{\theta})$ as an "ancillary statistic" to be used in combining the present data with other data, calling its function "analogous to providing a true, in place of an approximate, weight for the value of the estimate". It is unclear to me whether an orthodox conditional inference can legitimately be made conditional on $\tilde{J}(\hat{\theta})$, and if so, how. But in any case, the Bayesian point of view leads straightforwardly to the choice $\tilde{J}(\hat{\theta})$ for J.

Similarly, in the absence of any additional considerations, maximum likelihood appears from the Bayesian point of view to be preferable to any other orthodox BAN procedure on the (asymptotically rare) occasions when they differ appreciably (Savage, 1961b, p. 428). Of course the actual posterior mean and median, for instance, are BAN under mild restrictions and are optimum from the Bayesian point of view under appropriate loss structures, but this brings in additional considerations.

6. Remarks on Approximation

We have been interpreting standard inference procedures as approximations in the Bayesian framework. This kind of approximation has aspects in common with mathematical approximations generally. For instance, here as in general there may be several approximations available, and one may or may not have a feeling for which is likely to be better in a particular instance. If not, it is reasonable to use the one which looks best in some overall sense, according to the available indications, even though it may not be best in the situation at hand. One may or may not have a feeling that a given approximation is likely to be fairly good in a particular instance. If not, one may be satisfied if it is fairly good in some overall sense.

As with many approximations, little or no information is ordinarily available concerning the accuracy of these approximations. However nice it would be to have computable and not too conservative bounds on the errors, there are many useful approximations for which such bounds are not available.

One respect in which the present approximations differ from most is that the argument on which they are based is weaker than most, and in fact extremely weak. This is what is really worrisome about orthodox procedures from the Bayesian point of view, rather than the fact that there may be several possible procedures or that their accuracy is unknown.

Sections 3 and 4 referred to using the approximations "if the situation appears typical". In Section 4, for instance, this calls on us to judge whether the posterior (conditional) probability of a certain event ($\tau \in R$), given the sample at hand, equals the prior probability of that event, marginal over all samples. (Both probabilities are marginal over the parameters, of course.) There is no problem if it is feasible actually to obtain the posterior probability. We are interested in the case when this is not feasible, perhaps because of the difficulty of calculation or of organizing prior judgements into a complete prior distribution. Then we want to decide by informal judgement whether any peculiarity of the sample at hand suggests that the posterior probability of the event ($\tau \in R$), for this sample, is very different from its prior probability, marginal over all samples. In this context, the question, "When is a situation typical?" is as hard to answer definitively as two related questions. In an

orthodox inference problem, which reference sequence is appropriate, that is, what should one condition on? Is there a relevant subsequence (in Fisher's terminology) in which the relative frequency is different? However the latter questions may be answered, presumably a situation is typical only when referred to an "appropriate" reference sequence, that is, under "appropriate" conditioning. But there are doubtless many other kinds of consideration which may be relevant. For instance, the order statistics may be so unevenly spaced as to invite smoothing for some purposes. Further possibilities may be suggested by the next few paragraphs.

Whenever the prior is appreciably non-diffuse, the approximations are likely to be poor. If an estimate or confidence interval is located surprisingly, from an *a priori* point of view, then the posterior mean or median or probability interval will ordinarily be located differently, in the direction of agreement with the prior. Even if there is no discrepancy in location, confidence intervals are ordinarily longer than the corresponding posterior probability intervals, and a confidence distribution ordinarily has greater dispersion than the posterior probability distribution. One could choose a confidence procedure reflecting the prior, for instance, one minimizing the marginal (over the prior) expected length (Pratt, 1961). This is hardly standard, but entirely valid from the orthodox point of view. There is some evidence, however (Pratt, 1963), that the discrepancy between confidence levels and posterior probabilities will not be impressively reduced thereby.

Bayesian methods satisfy the likelihood principle or axiom: if, in a given situation, two random variables are observable, and if the value x of the first and the value y of the second give rise to the same likelihood function, then observing the value x of the first and observing the value y of the second are equivalent in the sense that they should give the same inference, analysis, conclusion, decision, action or anything else. Birnbaum (1962) states this more carefully and gives an elegant argument in favour of it, entirely within the orthodox framework and not based on any of the assumptions usually used to justify the Bayesian position. (His argument is indicated below.) Orthodox methods do not satisfy the likelihood principle. As long as we regard them only as approximations, not as fundamental, this need not disturb us any more than, for instance, the possibility of non-monotonicity in Section 4. It only means that we are approximating something which does satisfy the likelihood principle by something which does not.

6.1. *Bringing in the Likelihood Principle*

The likelihood principle, however, opens up the possibility of, and sometimes strongly suggests, using as an approximation not the orthodox procedure for the experiment actually done but instead that for some observation in some different experiment with the same likelihood function. This is by no means a far-fetched possibility. Consider, for instance, sequential stopping rules, limited here to non-randomized rules depending on the data alone, so that the stopping is "non-informative". The likelihood of an outcome in a sequential experiment is the same as if the same outcome had been obtained with a fixed sample-size. Accordingly, in a sequential situation, a Bayesian may just as validly approximate by an orthodox procedure for fixed sample-size as by one for the sequential rule used (and vice versa). In fact, the more peculiar a sequential stopping rule, the more peculiar adjustments would presumably be required by a procedure orthodox for that rule and the more, therefore, a Bayesian might distrust such a procedure and prefer a procedure orthodox for the size of sample actually occurring.

The application of the likelihood principle to sequential stopping is very broad. I am tempted to say universal. That is, the likelihood principle asserts the irrelevance of the stopping rule regardless of how dependent the observations are permitted to be or how non-parametrically or vaguely circumscribed their distribution may be. This can be clarified by recapitulating Birnbaum's (1962) argument in this situation, where it applies particularly beautifully and convincingly.

Consider a situation in which random variables $X_1, X_2, \ldots$ could be observed, and consider two stopping rules (one perhaps having fixed sample-size) under which it would be possible to observe the values $x_1, \ldots, x_n$ and stop. Let outcomes 1 and 2 be the occurrence of $x_1, \ldots, x_n$ under the first and second rules respectively when the rule has been chosen in the ordinary way. Let $1'$ and $2'$ be the same except that the rule has been chosen by flipping a fair coin. It is usually felt that one ought to proceed on the basis of the stopping rule actually used, that is, that $1'$ is equivalent to 1 and $2'$ to 2 for whatever purposes the observations and analysis are being made. At the same time, the information that the stopping rule was chosen by the flip of a coin and that $x_1, \ldots, x_n$ occurred is sufficient, in the sense of sufficient statistics. That is, it is sufficient to know that $1'$ or $2'$ occurred without knowing which. Thus $1'$ and $2'$ are equivalent. But it follows that all four outcomes are equivalent, and in particular the occurrence of $x_1, \ldots, x_n$ under one stopping rule is equivalent to its occurrence under another.

The first step above is a case of the conditionality principle, but note that only this special case is needed. The second step is the sufficiency principle. One justification of the latter here is that, upon learning that $1'$ or $2'$ occurred, but not which, one could flip a coin and proceed as if $1'$ had occurred in case of "heads" and $2'$ in case of "tails". This duplicates the operating characteristics of the procedure one would have used knowing the stopping rule. The point is that, if the rule is chosen by flip of a coin and $x_1, \ldots, x_n$ observed, then the conditional probability that the first rule was used is one-half, whatever may be the joint distribution of $X_1, X_2, \ldots$.

7. Testing Hypotheses: Significance Levels

Before considering more complicated cases, we restate in our context some familiar ideas concerning the use of a real-valued test statistic T to test hypotheses about a one-dimensional parameter θ, which can take on both positive and negative values. Nuisance parameters are permitted: we do not assume that θ alone determines the distribution of the observations or even of T.

Suppose first that the null hypothesis is $\theta \leqslant 0$, the alternative is $\theta > 0$ and we reject when T is too large. Typically the distribution of T for $\theta = 0$ is completely determined and is "stochastically larger" than any possible distribution of T for $\theta < 0$. In this case, if $T = t$ is observed, then the test rejects at any level $\alpha \geqslant P_0(T \geqslant t)$, where $P_0(T \geqslant t)$ is the probability that $T \geqslant t$ when $\theta = 0$. Thus $T = t$ is just significant at the level $P_0(T \geqslant t)$; this probability is often called the P-value. By the relation between confidence bounds and tests, 0 is a lower confidence bound for θ at the level $1 - \alpha = 1 - P_0(T \geqslant t)$. If we now interpret this confidence level as an approximate posterior probability, we may say that the P-value, $P_0(T \geqslant t)$, is an approximation to the posterior probability that $\theta \leqslant 0$. Of course, if one feels *a priori* that θ is likely to be near 0, or unlikely to be negative, as is often the case in hypothesis testing, and if the strength of this feeling is appreciable relative to the sample information, then one expects the approximation to be poor. If θ is a translation parameter determining the distribution of T, so that, in particular, $T - \theta$ and θ are independent from the Bayesian point of view, and if the

prior distribution of θ becomes "diffuse", then $T-\theta$ and T become independent also, and the P-value becomes exactly the conditional probability that $\theta \leqslant 0$ given T.

The interpretation just given of a one-tailed P-value applies even in situations when a confidence procedure for θ based on T has not been defined. If the distribution of T when $\theta \neq 0$ is not completely determined by θ, then defining the confidence procedure may be problematical. Under the conditions stated, however, a lower confidence bound for θ at level $1-\alpha$ would presumably be greater (less) than 0 for α greater (less) than the P-value, so that the P-value approximates the posterior probability that $\theta \leqslant 0$ in the same sense as before. Suppose, for instance, a treatment has been tested against a control in blocks, and $\theta = 0$ corresponds to no treatment effect. Suppose further that if a treatment effect exists, it can be expected to vary from block to block but to have the same sign in every block. However θ is defined, and whatever confidence technique is used, the foregoing interpretation of the P-value will presumably be valid.

In a really difficult case, for instance, when concerned with a non-parametric measure of association, one might be content to use the P-value as an approximation to the posterior probability that $\theta \leqslant 0$, use some estimate of θ as an approximation to the median (say) of the posterior distribution of θ, and either not attempt to pin down the posterior distribution any further or guess at it from these two points.

Suppose next that the null hypothesis is $\theta = 0$, the alternative is $\theta \neq 0$, and we reject when T is too large or too small, specifically when either one-tailed test at level $\frac{1}{2}\alpha$ would reject. Then in situations where the foregoing interpretation of one-tailed P-values is valid, the posterior probability that θ lies on the opposite side of 0 from that suggested by the observations is approximately *one-half* the P-value of the two-tailed test. This corresponds to choosing which one-tailed test to make on the basis of the data, which is utter heresy according to orthodox dogma. On the other hand, the following confidence procedure amounts to the same thing and permits the most orthodox frequentist to make a statement whose direction depends on the data without doubling α:

Say "$\theta \leqslant 0$" if $T \leqslant$ its lower α critical value;
Say "$\theta > 0$" if $T \geqslant$ its upper α critical value;
Say "$-\infty < \theta < \infty$" otherwise.

If $\theta \leqslant 0$, only the second statement is false, and the probability of making it is α or less; similarly, if $\theta > 0$, the probability of a false statement is α or less. The procedure may also be regarded as simultaneously testing two mutually exclusive null hypotheses, each at level α.

The only widely valid relation between a two-tailed P-value and a posterior probability of natural interest seems to be that just given. The P-value of a two-tailed test of the null hypothesis $\theta = 0$ is often approximately the posterior probability that θ is farther from an estimate $\hat{\theta}$ than 0 is, i.e. that θ is not between 0 and $2\hat{\theta}$, but there is no obvious practical reason to look at the probability of this particular interval. It is easy to verify by examples (as in Jeffreys, 1961, pp. 434–435 and Lindley, 1957) that a two-tailed P-value bears no special relation to the posterior probability that the null hypothesis $\theta = 0$ is true, even in those situations where this is considered possible *a priori*. Similarly, it bears no relation to the posterior probability that $|\theta|$ is "small", say $|\theta| \leqslant \delta$, where δ is the threshold of practical significance or is based in some other way on practical considerations. One might hope to learn something about this posterior probability by testing the null hypothesis $|\theta| \leqslant \delta$ (Hodges and Lehmann, 1954) or the null hypothesis $|\theta| \geqslant \delta$, but the P-values of such tests are typically even

less interpretable than ordinary two-tailed P-values. The use of unequal tails would also serve only to cloud the issue.

Now suppose θ is k-dimensional, where $k>1$. The situation is essentially like that of the one-tailed test in the one-dimensional case if the hypothesis is a half-space, or the space on one side of a sufficiently flat $(k-1)$-dimensional surface dividing the θ-space into two regions. Similarly, the situation is essentially like that of the two-tailed test in the one-dimensional case if θ is k-dimensional and the null hypothesis is that θ lies on a $(k-1)$-dimensional hyperplane or a sufficiently flat $(k-1)$-dimensional surface dividing the θ-space into two regions. That is, the posterior probability that θ lies in the opposite half-space or on the opposite side of the $(k-1)$-dimensional surface from that suggested by the observations is approximately equal to the P-value in the former type of situation and one-half the P-value in the latter. These situations are rare in practice except when they can also be expressed naturally in terms of a one-sided or point hypothesis about a one-dimensional parameter in the presence of nuisance parameters. For example, the half-space hypothesis $\theta_1 \leqslant \theta_2$ can be expressed as a one-sided hypothesis $\theta' \leqslant 0$ where $\theta' = \theta_1 - \theta_2$ is one-dimensional, and the hyperplane hypothesis $\theta_1 = \theta_2$ can be expressed as the point hypothesis $\theta' = 0$.

Usually the null hypothesis is that θ lies on a surface of dimension less than $k-1$. Consider, for instance, typical hypotheses of additivity. In such situations the P-value seems entirely unhelpful from the Bayesian point of view, except possibly in the following case (the other extreme). The P-value of a test of the null hypothesis $\theta = 0$ is often approximately the posterior probability that θ is "farther" from an estimate $\hat{\theta}$ than 0 is. Unfortunately, the metric by which "farther" is defined may be quite bizarre; it is determined by the test statistic, which is not ordinarily chosen to make the metric sensible but to have a convenient null distribution. For instance, let θ be the interaction terms in a two-way analysis of variance and note how the non-centrality parameter depends on the cell sample sizes. Even in the one-dimensional case, this interpretation was not too helpful, and here it seems even less so.

A common situation not yet mentioned is that of a null hypothesis $\theta = 0$ concerning a one-dimensional parameter θ which by its nature cannot be negative. Consider, for instance, the null hypothesis that a multiple-R^2 or a non-centrality parameter is 0. As these examples indicate, such situations can usually be restated in terms of a k-dimensional parameter θ and a null hypothesis that θ lies on a surface of dimension less than $k-1$. This is the situation of the previous paragraph, and the P-value is not helpful.

In short, when the null hypothesis $\theta \leqslant 0$ is tested against the alternative $\theta > 0$, where θ is one-dimensional and $\theta < 0$ is possible, the P-value is usually approximately the posterior probability that $\theta \leqslant 0$. Most other situations where the P-value has a helpful interpretation can be recast in this form. Of course, $\theta \leqslant 0$ can be replaced by $\theta \leqslant \theta_0$ or $\theta \geqslant \theta_0$. And while it is convenient to use P-values in the discussion, those who are interested only in whether or not the results are significant at some preselected level will find similar remarks apply. All the statements about the relation of P-values to posterior probabilities, or lack of it, can be seen easily to hold for a univariate or multivariate normal distribution with known variance or variance matrix.

Good (1950, 1958) has given an inequality which should be mentioned here. Suppose H_0 and H_1 are mutually exclusive and T is a statistic, large values of which favour H_1. Then $T \geqslant t$ favours H_1 more than $T = t$ does, so

$$\frac{P(H_0 \mid T = t)}{P(H_1 \mid T = t)} > \frac{P(H_0 \mid T \geqslant t)}{P(H_1 \mid T \geqslant t)} = \frac{P(H_0)\,P(T \geqslant t \mid H_0)}{P(H_1)\,P(T \geqslant t \mid H_1)} > \frac{P(H_0)}{P(H_1)} P(T \geqslant t \mid H_0). \qquad (37)$$

Thus the odds for H_0 against H_1, given $T = t$, are at least the prior odds times $P(T \geqslant t \mid H_0)$. This last factor is the one-tailed P-value if H_0 determines the distribution of T, as is typical of null hypotheses which can be put in the form $\theta = 0$. Unfortunately, in most such cases tests are of doubtful relevance anyway, as we shall see in Section 8.2 and subsequent Sections.

8. Some Common Uses of Tests

We shall now examine from the Bayesian point of view the appropriateness of various common uses of tests. Our conclusions will be stated in Bayesian terminology. In essence, these conclusions seem to me equally valid from the orthodox point of view, but more awkward to state and more difficult to arrive at. Edwards *et al.* (1963) have covered a good part of the same ground rather fully, with many examples. This justifies abbreviating the discussion here.

Of course, one can regard tests as equivalent to two-decision procedures, and the class of admissible tests is then essentially the same as the class of Bayes procedures. Thus classical tests become equivalent to Bayesian two-decision procedures in sufficiently sophisticated hands. We are concerned here, however, with common and not highly sophisticated uses of tests. The categories of common uses which follow are neither mutually exclusive nor collectively exhaustive.

8.1. *Proving and Disproving One-sided Null Hypotheses*

Consider a null hypothesis $\theta \leqslant 0$, where θ is a real parameter and negative values of θ are possible. If this hypothesis is rejected by a test at a fairly small significance level α (say 0·05 or 0·01), then the hypothesis is generally regarded as disproved and $\theta > 0$ as proved. (We use "prove" as an abbreviation for "prove provisionally", or "establish a presumption in favour of", or whatever other phrase the reader prefers.) The evidence against the null hypothesis is usually regarded as stronger the smaller the P-value. We have seen that the P-value is usually approximately the posterior probability of the null hypothesis in this case, putting the Bayesians in approximate agreement with the orthodox conclusion, though not with the orthodox reasoning. At the same time, it should be noted that if it matters how large θ is, as it usually does, then the test is not a sufficient, or often even an important, part of the analysis. See also Section 8.3 below.

Occasionally people act as though they have proved the null hypothesis $\theta \leqslant 0$ when they have carried out a test and the result was not significant at $\alpha = 0{\cdot}05$, say. The Bayesian interpretation of the P-value disagrees with this, as do most orthodox statisticians. A common, though not universal, orthodox view seems to be that a test can disprove the null hypothesis but otherwise leads to no conclusion, so if one hopes to prove $\theta \leqslant 0$ one must test the null hypothesis $\theta > 0$. If this null hypothesis is rejected, $\theta > 0$ is disproved and hence $\theta \leqslant 0$ proved. This puts us in the position of the previous paragraph with the inequalities reversed.

8.2. *Proving and Disproving Point Null Hypotheses*

Consider a null hypothesis $\theta = 0$, where θ is a real parameter and 0 may be either an endpoint or an interior point of its range. If this hypothesis is rejected by a test at a fairly small significance level α, then the hypothesis is generally regarded as disproved and hence the alternative ($\theta < 0$, $\theta > 0$, or $\theta \neq 0$, whichever is appropriate) as proved. This is in discord with the Bayesian point of view. How depends on the situation.

(a) If $\theta = 0$ has positive prior probability (which is rare except as an approximation), then it will have positive posterior probability, but there is no particular relation between the significance level or P-value of a test and the posterior probability that $\theta = 0$, though Good's inequality (37) may apply. For a given P-value, the posterior probability that $\theta = 0$ will generally be larger the larger the sample size when other things are equal, as illustrated for instance by Jeffreys (1961, pp. 434–435) and Lindley (1957).

(b) If $\theta = 0$ has zero prior probability, then it will have zero posterior probability, and a test to disprove $\theta = 0$ is irrelevant and/or superfluous. The point was made sharply by Berkson (1938):

> . . . we may assume that it is practically certain that any series of real observations does not actually follow a normal curve *with absolute exactitude* in all respects, and no matter how small the discrepancy between the normal curve and the true curve of observations, the chi-square P will be small if the sample has a sufficiently large number of observations in it.
>
> If this be so, then we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all!

When people test $\theta = 0$ despite the fact that it cannot be exactly true, their real purpose is usually not simply to prove $\theta \neq 0$. Some possible real purposes are discussed in the following paragraphs and subsequent subsections.

(c) One might be interested in whether θ is near 0, say $|\theta| \leqslant \delta$, where δ is chosen on the basis of practical considerations. The P-value of a test of $\theta = 0$ cannot have any special relation to the posterior probability that $|\theta| \leqslant \delta$, since it does not even depend on δ. More sophisticated orthodox approaches would be to test $|\theta| \leqslant \delta$ (Hodges and Lehmann, 1954) or $|\theta| \geqslant \delta$ (rejection of which could justify the conclusion $|\theta| < \delta$, according to a common orthodox view). From the Bayesian point of view, such tests do not have naturally interpretable P-values (except in so far as they become one-tailed in large samples) and by no means solve the problem.

(d) One might be interested in whether θ has the sign suggested by the observations. Here typically the P-value is twice the relevant posterior probability, as pointed out in Section 7.

8.3. *Deciding Between Two Acts: Break-even Values*

The previous two Sections apply particularly to research hypotheses and situations in which terminal action is far in the future and/or nebulously related to the matter under test. They also apply to situations requiring an immediate choice between two terminal actions, but then there is much more to say. This Section and the next say more about two such situations, briefly because the material is far from new.

Consider first the problem of whether to switch from a process with known yield θ_0 to a new process with unknown yield θ. Here it is tempting from the orthodox point of view to test the null hypothesis $\theta \leqslant \theta_0$ and switch if it is rejected, setting α low enough to allow for the fact that switching would be expensive, which makes rejecting when $\theta \leqslant \theta_0$ the more serious of the two kinds of error. The Bayesian procedure, if the utilities are linear in θ, is to find the break-even value θ_b, depending on the economics of the situation, and switch if the posterior mean of θ exceeds θ_b but not otherwise.

Classical counterparts of this Bayesian procedure, which would probably lead to the same action in most practical problems, would be to switch if an estimate $\hat{\theta} > \theta_b$ but not otherwise, or to switch if a test of the null hypothesis $\theta \leqslant \theta_b$ at level $\alpha = 0{\cdot}50$ (not $\theta \leqslant \theta_0$, $\alpha = 0{\cdot}05$) leads to rejection but not otherwise. At least in this problem, using the "natural" null hypothesis $\theta \leqslant \theta_0$ (not depending on change-over costs) and allowing for the greater seriousness of one kind of error by setting a conventionally small α like 0·05, will not lead to a good procedure except by remarkable coincidence or remarkable intuition in the choice of α. Comparing an estimate with a break-even value seems a more natural procedure.

8.4. *Deciding between Two Acts when there are Two States*

Now consider a problem involving two acts and two simple hypotheses H_0 and H_1. (Classification problems are sometimes of this type, and all problems of this type can be restated as classification problems.) Any procedure in such a problem may be regarded as a test of the null hypothesis H_0; then α is the probability, given H_0, of taking the less desirable action. Let β be the probability, given H_1, of taking the less desirable action. The problem can be viewed as a choice among the available (α, β)-points. If the loss attributable to taking the wrong action is the same for both kinds of error, then the orthodox framework rather suggests choosing that test with $\alpha = \beta$ among the admissible (here essentially also most powerful and likelihood-ratio) tests. This will coincide with the Bayesian procedure if H_0 and H_1 are equally likely *a priori* and are suitably symmetric with respect to one another so that the admissible (α, β)-curve is symmetric in α and β. If H_0 and H_1 are equally likely *a priori* but lack this symmetry, then the Bayesian procedure selects that point on the admissible (α, β)-curve where an infinitesimal decrease in either α or β would require an equal increase in the other. This is not generally the point where $\alpha = \beta$, but it always corresponds to a critical likelihood ratio of 1. If the prior probabilities or losses are not equal, the Bayesian procedure minimizes $k\alpha + \beta$, where k equals the prior odds times the ratio of losses. This selects the point where the admissible curve has slope $-k$. It corresponds to a critical likelihood ratio of k. The Bayesian point of view focuses on a critical rate of exchange between α and β, and on the likelihood ratio itself rather than tail probabilities thereof (Savage, 1961a, b). The orthodox point of view, with its different focus, suggests different decision rules (Lehmann, 1958), though as far as the orthodox theory goes formally, it is not in disagreement. In fact, Lindley and Savage have given an argument in the orthodox tradition leading to the Bayesian procedure and essentially the Bayesian view in this problem, or any finite problem, and hence any problem for practical purposes. The argument is reported by Savage (1961a, b).

One further contrast is this. If a decision rule makes α and β both small, say $\alpha = \beta = 0{\cdot}01$, then it is tempting to think each decision is almost surely right. This is true marginally, but given certain observations the probability of error will be large, though these observations must be rare if α and β are small. Suppose, for instance, two known normal populations with the same variance are equally likely *a priori*, and consider classifying an observation according to which population mean it falls nearer to. If the means are far enough apart, α and β will be very small, but given an observation midway between the means, the conditional probability of error will be $\frac{1}{2}$. In this kind of problem, the P-value is not approximately the posterior probability of H_0. *A priori*, α and β are relevant, but after one sees the observations, they are not. What is relevant then is the likelihood ratio.

8.5. *Deciding Whether to Sample Further*

Consider a situation involving two terminal acts, one better when $\theta < 0$, the other when $\theta > 0$. Schlaifer (1959) has emphasized that to decide whether the information at hand is enough or whether further sampling should be done, a common practice is to make a two-tailed test of the null hypothesis $\theta = 0$ and sample further unless it leads to rejection at some level α (presumably 0·05). Two questions arise: (a) What assurance will one have that θ is on the side of 0 it appears to be on when sampling is stopped? (b) Is this a sensible sampling rule?

As regards (a), if each test is made in the ordinary way on all observations to date, then the posterior probability that θ has the opposite sign from that suggested by the data will typically be about half the *P*-value at the time of stopping. A conscientious orthodox statistician might be inclined to compute the level allowing for the stopping rule, but this would presumably give $\alpha = 1$ if the procedure stops with probability 1 when $\theta = 0$.

As regards (b), the sampling rule under discussion must often be poor because it takes no account of how far θ appears to be from 0 or how the loss of a wrong terminal act varies with θ. If many observations have been taken without obtaining significance, then the posterior distribution of θ will be concentrated near 0 so that in typical problems the expected loss of an immediate decision will be small, yet the rule calls for continuing sampling just as much as it does after a few observations have been taken without obtaining significance, when the expected loss of an immediate decision is much larger and the probability is much greater that a small number of additional observations will lead to a different terminal act.

8.6. *Deciding on Sample Size and Decision Rule*

Consider a situation involving two terminal acts, one better when θ is near 0, the other when θ is far from 0, and suppose a terminal act is to be chosen after a single round of sampling. A suggestion of Pearson (1962), oversimplified here, is this: choose a level α, an alternative θ_1, and a type II error probability β, and make the decision according to a two-tailed test at level α after taking a sample just large enough so that such a test will have power $1-\beta$ at θ_1. A similar procedure could be used in one-sided problems; in the situation of Section 8.3, for instance, the null hypothesis would presumably be $\theta = \theta_0$, but θ_1 might be the break-even value and $\beta = \frac{1}{2}$. The argument in favour of procedures like these seems to be that it is sometimes troublesome to think out the prior probabilities and costs (losses, utilities) needed for a full Bayesian analysis of the problem. The question then becomes whether it generally gives better results to select some "reasonable" prior probabilities and costs and use a Bayes strategy or to select some "reasonable" α, θ_1 and β, keeping in the back of one's mind the rough situation as regards prior probabilities and costs, but not attempting to make them explicit. The latter procedure would make me very uneasy because I have no quantitative intuitive feeling about, for instance, how to reflect the cost of sampling in the choice of α and β. I have not even an order of magnitude feeling about how rapidly α and β should decrease as the cost of sampling decreases.

8.7. *To Simplify Description*

This and the next two Sections concern various kinds of "preliminary tests", that is, tests carried out to decide which of two methods to use in the remainder of the analysis, rather than to reach a research conclusion or decide upon an action.

We consider first preliminary tests whose purpose is to simplify description if possible. For definiteness, imagine a two-factor experiment with yield μ_{ij} at the ith level of factor 1 and the jth level of factor 2. Description and comprehension of the situation is much easier if additivity holds than otherwise. One might carry out a test of the null hypothesis of additivity and treat the situation as additive unless the test leads to rejection. However, Berkson's comment holds here. One usually knows the situation is not exactly additive, so what is the point of testing? More relevant would be to estimate some measure of non-additivity and treat the situation as additive unless the estimate is above the value at which the non-additivity would just be practically important. It is sometimes argued that a test of additivity can perfectly well be regarded as a test of approximate rather than exact additivity. This is true as long as the test is weak in the region of approximate additivity. It is not a sufficient defence of the test, however, because the extent of non-additivity that the test is likely to tolerate depends on the sample size, whereas it should depend on the extent of non-additivity that would be practically important.

8.8. *To Improve Analysis*

Sometimes preliminary tests are used in hopes of improving the analysis fundamentally, not just by simplifying it. Consider, for example, the two-factor situation of the previous Section. If one is going to estimate the μ_{ij} by orthodox methods, one will get one set of estimates assuming additivity and another set otherwise. The first procedure is better than the second even when there is non-additivity if the extent of non-additivity is small. Rather than choose in advance one procedure or the other, it may be better to decide between them on the basis of a preliminary test of additivity.

Bayesianly the situation is this (under the usual normality assumptions). The first procedure gives the posterior means for a prior under which the "interactions" are 0 with probability 1 and the remaining parameters have a "diffuse" distribution. The second procedure gives the posterior means for a prior under which everything is "diffusely" distributed. If, under our real prior distribution, there is substantial probability that the interactions are small but some probability that they are quite large, then the procedure using the preliminary test may indeed give a better approximation to the posterior means than either of the two original procedures. This is not to say that it is a really good procedure, however. Against it, among other things, is the fact that the estimators resulting from the preliminary test will be discontinuous functions of the observations, since there is an abrupt jump from one estimating function to another as the test statistic passes from one side of the critical value to the other. Posterior means do not have this unpleasant feature even if there is positive prior probability that the interactions are 0.

Other situations in which it may be hoped that a preliminary test will improve the analysis include those in which there is an "outlying" observation which may be due to a gross error and those in which several variances, regression slopes or what-have-you, may be equal so that their estimates could be "pooled". Orthodox statisticians have already pointed out that what matters in all these situations is the effect on the overall analysis of the preliminary test, so that customary significance levels may not be at all appropriate. (Also, if the subsequent analysis includes another test, the significance level of the overall procedure will not be that of the later test by itself.) As is clear from the current state of such problems, an orthodox statistician must do very difficult calculations to analyse such situations properly from his point of view. A fully Bayesian posterior analysis is often fairly simple. The reason is that here

analysis is much easier to do in "extensive" form (using posterior distributions) than in "normal" form (using decision rules).

8.9. *To Justify the Model*

Most statistical analysis, orthodox and Bayesian, is based on the assumption that a certain sampling model is adequate. One would often like to check this assumption, and sometimes a preliminary test is a tempting tool to use.

Consider once more the two-factor situation discussed above. Suppose one were inclined to assume additivity but wanted to check on this assumption. A preliminary test might be used. However, this is not really a good procedure; the situation is basically the same as that in the last Section, and the remarks made there apply. This typifies situations where one knows what analysis he would make if doubtful assumptions were dropped.

One step away from this is the situation when one knows what alternative model he would use, but does not know what prior probabilities he would use in each of the possible models. For instance, one might want to avoid putting a prior distribution on the interactions if additivity does not hold, or on the "skewness" and "kurtosis" of a population suspected of non-normality. A way to handle this sometimes is by sensitivity analysis: do the analysis under the whole range of models which are plausible in the light of the data for any prior within very broad limits. If the essential conclusions are little affected by the choice among these models, then one need not make up one's mind. On the other hand, if the choice among these models matters greatly, then one will have to be less vague about one's prior or else fall back on some form of superstition.

When one has no particular alternative model in mind, the situation is even more problematical. The general-purpose tests of distribution assumptions are very weak in small samples, while trivial unimportant deviations from the assumptions will lead to almost certain rejection in sufficiently large samples. Again Berkson's comment applies. The null hypothesis is surely not exactly true, so why test it? The present situation has much in common with that of Section 8.7. As there, testing is not to the point, and it would be more to the point to estimate something; if the estimate is large enough to suggest a practically important deviation from the assumption, one could then look at the data from various angles to see if any alternative assumption is suggested. Unfortunately it is much less clear here than in Section 8.7 what to estimate or what would constitute a practically important value of the quantity estimated. The situation is not hopeless. For instance, the chi-square goodness-of-fit statistic does estimate a certain strange quantity, and some examples might give an idea what value of this quantity could be practically important. I would expect to find a procedure developed along these lines preferable to comparing the chi-square statistic with the 0·95 fractile of its distribution under the null hypothesis.

8.10. *Concluding Remarks on Tests*

In summary, conventional tests seem to a Bayesian to have a useful interpretation only under special circumstances. These circumstances are not generally present in practice, and even when they are, conventional tests are not generally well articulated to the real problems of the situation. The situations discussed in the previous Sections may be mostly ones for which tests were not intended by their inventors, and some common uses of tests, proper or improper, may not have been touched on. I do feel, however, that the situations discussed represent many where tests are commonly used

in practice, and that conventional tests are seldom satisfactory tools of analysis. (They are theoretically important for their relation to confidence regions.)

It could be argued that tests are a general-purpose tool and cannot be expected to be ideal for any particular job. I would say that they are entirely inadequate for many jobs, and that it would usually be better to see what the real job is and use a tool fashioned, even though crudely, for this particular job, rather than to use a tool, however easy-to-hand and shiny, which accomplishes an entirely different job. Tests seem to me more like a hammer than a boy-scout knife. Even an amateur does not attempt to do many jobs with a hammer, and a professional carpenter who did would not be allowed in the union.

9. Miscellaneous Remarks

Some will no doubt feel that the Bayesian justification of standard estimation and confidence procedures presented in this paper is very weak. I do too, but consider this an inevitable defect of the theory behind these procedures. The Bayesian and orthodox justifications of these procedures are much alike: both represent the situation for given data by an average of which this situation forms but one component. The Bayesian justification averages over a prior distribution, the orthodox over an infinite sequence.

There has been little or no mention of many important Bayesian topics, for instance, "preposterior" theory (see especially Raiffa and Schlaifer, 1961) and "Bayesian tests" (see especially Jeffreys, 1961). This is because no standard procedures correspond at all closely to Bayesian procedures in these areas, and our concern has been to look at standard procedures with a Bayesian eye, not to present Bayesian procedures.

It should perhaps be pointed out that Sections 3 and 4 have a reverse side. For instance, (31) implies that a region of posterior probability $1-\alpha$ is an approximate confidence region in the sense that its probability of coverage has weighted average $1-\alpha$, where the weighting is given by the prior. Similarly, a posterior mean is approximately unbiased in the sense that its bias has weighted average zero. Because the weighting has no orthodox meaning, this argument may be even less comforting to the orthodox than the reverse argument is to a Bayesian.

Acknowledgements

I am most grateful to Harry V. Roberts, Bruce M. Hill, I. Richard Savage and Arthur P. Dempster for comments on the first draft of this paper. The many others to whom I am also indebted in various ways it would be impracticable to acknowledge individually. The research of this paper was supported by the National Science Foundation under grant NSF-G24035.

References

Bartlett, M. S. (1953a), "Approximate confidence intervals, I", *Biometrika*, **40**, 12–19.

—— (1953b), "Approximate confidence intervals, II", *Biometrika*, **40**, 306–317.

—— (1955), "Approximate confidence intervals, III", *Biometrika*, **42**, 201–204.

Berkson, J. (1938), "Some difficulties of interpretation encountered in the application of the chi-square test", *J. Amer. statist. Ass.*, **33**, 526–536.

Birnbaum, A. (1962), "On the foundations of statistical inference", *J. Amer. statist. Ass.*, **57**, 269–306.

Chernoff, H. (1951), "A property of some type A regions", *Ann. math. Statist.*, **22**, 472–474.

Edwards, W., Lindman, H. and Savage, L. J. (1963), "Bayesian statistical inference for psychological research", *Psychol. Rev.*, **70**, 193–242.

Fisher, R. A. (1925), "Theory of statistical estimation", *Proc. Camb. phil. Soc.*, **22**, 700–725.

GOOD, I. J. (1950), *Probability and the Weighing of Evidence*. London: Griffin.
—— (1958), "Significance tests in parallel and in series", *J. Amer. statist. Ass.*, **53**, 799–813.
HODGES, J. L. and LEHMANN, E. L. (1954), "Testing the approximate validity of statistical hypotheses", *J. R. statist. Soc.* B, **16**, 261–268.
JEFFREYS, H. (1961), *Theory of Probability*, 3rd ed. Oxford University Press.
LEHMANN, E. L. (1958), "Significance, level and power", *Ann. math. Statist.*, **29**, 1167–1176.
—— (1959), *Testing Statistical Hypotheses*. New York: Wiley.
LINDLEY, D. V. (1957), "A statistical paradox", *Biometrika*, **44**, 187–192.
—— (1961), "The use of prior probability distributions in statistical inference and decisions", *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1**, 453–468.
PEARSON, E. S. (1962), "Some thoughts on statistical inference", *Ann. math. Statist.*, **33**, 394–403.
PRATT, J. W. (1961), "Length of confidence intervals", *J. Amer. statist. Ass.*, **56**, 549–567.
—— (1963), "Shorter confidence intervals for the mean of a normal distribution with known variance", *Ann. math. Statist.*, **34**, 574–586.
RAIFFA, H. and SCHLAIFER, R. (1961), *Applied Statistical Decision Theory*. Boston: Graduate School of Business Administration, Harvard University.
SAVAGE, L. J. (1961a), "The foundations of statistics reconsidered", *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1**, 575–586.
—— (1961b), *The Subjective Basis of Statistical Practice* (report), Department of Mathematics, The University of Michigan, Ann Arbor, Michigan.
—— *et al.* (1962), *The Foundations of Statistical Inference*. London: Methuen.
SCHLAIFER, R. (1959), *Probability and Statistics for Business Decisions*. New York: McGraw-Hill.
STEIN, C. M. (1951), "A property of some tests of composite hypotheses", *Ann. math. Statist.*, **22**, 475–476.
STEVENS, W. L. (1950), "Fiducial limits of the parameter of a discontinuous distribution", *Biometrika*, **37**, 117–129.
—— (1957), "Shorter intervals for the parameter of the binomial and Poisson distributions", *Biometrika*, **44**, 436–440.

DISCUSSION ON PROFESSOR PRATT'S PAPER

Professor D. V. LINDLEY: The paper that has been presented to us to-night is important. The climate of opinion within this Society towards the Bayesian approach has changed from one almost of derision, to one where we can contemplate enquiring whether any of the standard material is relevant to a Bayesian. It is standard statistics that is being judged: it does not come off too badly. In fact I think it does rather better than our speaker makes out. For much of standard statistics revolves around the normal distribution; and for that distribution there is an exact connection between Bayesian and standard arguments, and the approximations mentioned by the speaker are excellent. We may cite the simple case of the normal distribution with unknown mean and known, say, unit, variance. The standard inference statements are based on $p(\bar{x} \mid \mu)$ being $N(\mu, n^{-\frac{1}{2}})$. The Bayesian substitutes $p(\mu \mid \bar{x})$ being $N(\bar{x}, n^{-\frac{1}{2}})$ for a smooth prior. Both statements may be incorporated into $\bar{x}-\mu$ being $N(0, n^{-\frac{1}{2}})$ (at least if you do not put tildes on your random variables) and this is the pivot of fiducial theory. The argument extends to the case of unknown variance, where Jeffreys has shown that the posterior distribution of μ is the Student's t of standard statistics. It extends even to linear hypotheses and least squares: a topic which embraces a very substantial part of our subject. Consequently the bulk of standard statistics means a lot to a Bayesian. Indeed, I would say that a contributory factor in the success of the Bayesian approach has been the failure of the standard methods when used outside the normal family. Standard statistics has worked because it is largely Bayesian.

Now to turn to points more of detail. That part of our subject which the speaker commends the least is significance tests, yet, in my view, more can be said in their favour. I would like to put forward the view that a significance test is a meaningful expression of a posterior probability, though whether it is the most useful expression is another matter. Consider a posterior distribution of a parameter described by means of its density $p(\theta)$. Initially suppose θ to be one-dimensional. Then, such a distribution may be summarized in many ways, for example through its mean and standard deviation. But a possible way

is by means of its fractiles. A particular fractile that we might consider is the lower 5 per cent one, and we could say that the posterior probability of being less than it is 5 per cent. Similarly the upper one can be used, and the probability of θ lying between them is 90 per cent. For a typical distribution the values of θ included in this set have higher density than the remainder. If not we could consider that set of parameter values having probability 90 per cent and being more probable than those outside the set. It is not unreasonable to think of those in the set as being the plausible values for the parameter, and those outside as being less plausible. The 90 per cent is a measure of the plausibility. If a null value of the parameter of particular interest lies outside the set then it is implausible (or, dare I say it, significant) at a 10 per cent level. This, I contend, is not an unfair interpretation of a significance test. A null value is judged significant at level α if it does not belong to the more likely values included in a set of probability $1-\alpha$.

The extension to many parameters is not difficult. The joint distribution of θ (now vector-valued) has contours of constant probability density. Typically values within such a contour are more probable than those outside. If we choose a contour such that the included set has probability $1-\alpha$ and the null value lies outside this set then it is significant at level α. If this interpretation is adopted then most significance tests with the normal distribution, and many others, conform with standard practice: the F-test in the analysis of variance is an important example. This is not to say that a significance test is the proper way in which to summarize a posterior distribution. But the use of fractiles is sensible and the major modification in significance tests I would suggest (apart from the interpretation) is the use of higher levels than are customary. The use of 50 per cent, on the lines of the old probable error, is a serious possibility.

A standard test has one apparent advantage over Bayesian methods. Its validity (though not its optimality) depends only on the null hypothesis. Alternatives do not have to be mentioned explicitly. This is useful in the many situations where the null is precise and the alternatives are vague; and here the Bayesian has a problem because of the difficulty in describing a prior distribution over a wide class of alternatives. It is perhaps worth pointing out that in one important situation the Bayesian can handle the situation and provide (in the above sense) the usual significance test. The case is where the distributions (of the observed random variables) are grouped, and consequently we are considering multinomial distributions. The null hypothesis is a set (possibly having a single member) of multinomials: the alternatives are all other multinomials. This entire family is finite-parametered and a smooth prior distribution is easily specified. Using the Bayesian interpretation briefly described above the Bayesian significance test is asymptotically the usual Pearson χ^2 with the usual modifications to the degrees of freedom according to the complexity of the null hypothesis. We thus have a Bayesian interpretation of historically the first systematic significance test. It can be argued that significance tests need not involve alternatives, a view to which I cannot subscribe; but doubtless other speakers will have something to say about this.

It has been a privilege to have you with us to-night, Professor Pratt, and to listen to you read your excellent paper. I have much pleasure in proposing the vote of thanks.

Professor G. A. BARNARD: Before I come to the usual task of the Seconder and list criticisms, I would like to make points which would be regarded more as comments, in connection with the section in Professor Pratt's paper concerning the use of insufficient statistics. He suggests a way of using the sample mean to get an approximation to the posterior distribution of the population mean, which is robust with respect to variations in the shape of the population. I look at this from the likelihood point of view, which, with the assumption of a smooth prior (which has been made by both the preceding speakers to-night) need not be distinguished mathematically from the Bayesian point of view. In the situations considered I feel it is better to set out the whole of the likelihood function along the lines that were indicated a long time ago by Fisher. One can, for example, take

the frequency function to be of such a form that its logarithm is a polynomial of some degree k:

$$\log \phi(x) = \alpha_0 + \alpha_1 x + \ldots + \alpha_k x^k$$

(if we take $k = 2$ this corresponds to the assumption of normality). Then the log-likelihood is simply expressible in terms of the sample moments

$$(1/n) \log L = \alpha_0 m_0 + \alpha_1 m_1 + \ldots + \alpha_k m_k,$$

m_k being the sum of the kth powers of observations divided by the sample number. By taking k large enough one often can make this accommodate most of the distributions one might wish to contemplate, and hence reasonably simply look at the way in which the posterior distribution of the parameters depends on various assumptions one might make. In fact we would in practice not use these coefficients α as the parameters of interest, but rather express the density in terms of a quantity $u(x-\theta)/\sigma$, copying as far as possible Professor Pratt's notation, and then one would want to use the moments, not from the origin, but centred at the sample mean. It might then appear, on calculating the higher moments of the sample, the third and fourth moments for example, that with the particular sample one had, the posterior distribution of θ did not in fact depend markedly on the higher power terms and in that case one would be able to say something about the posterior distribution of θ without paying attention to the possible non-normality of the distribution. But if on the other hand one had an unlucky sample and it turned out otherwise, then it would be most misleading, I think, to suggest that one could make statements about the population mean in the light only of the sample mean, without regard to the values of the non-normality parameters. In this connection the need for an analysis of the specific features of each sample must be borne in mind.

We are now in the habit of paying attention to the need for robustness in the way we answer the questions that are put to us; we want our procedures to be robust with respect to departures from normality, for example. But we need also to bear in mind that the questions we are asking need to be robust, and there is such a thing as a robust answer to a highly non-robust question. This can be illustrated by a distribution formed of a mixture of 99·9 per cent $N(0, 1)$, with 0·1 per cent lumped at $x = 10{,}000$. Such a distribution would have mean 10, but for many practical purposes (such as finding 99 per cent limits for the value of x) the mean would be zero. Most samples of moderate size from this distribution would suggest that it was $N(0, 1)$, and would, for many purposes, not thereby be misleading.

Now to come to the more philosophical parts of the paper. Of course a great deal of it, as I already indicated, is interpretable in terms of the theory of likelihood. Anything one says about posterior distributions relative to smooth priors is interpretable either in terms of likelihood itself, or in terms of mean likelihood. And so with a great deal of the paper one finds oneself in considerable agreement, and indeed with the emphasis on the value of the likelihood principle. However, I find myself at odds with Professor Pratt and Professor Lindley, in that it seems to me that the likelihood principle, and other considerations of the sort which have been advanced here to-night, apply just to those situations where we are concerned with choices between relatively well-specified alternatives, but not to the situations where the alternatives are *not specified.* It is not here, as I see it, a question of the alternatives not being *precisely* specified. May I refer to a point I mentioned some weeks ago, a situation really existing in nuclear physics at the moment. Observations have been obtained of nuclear processes which appear to be inconsistent with any interpretation that makes the processes involved invariant under time reversal. Now these observations cast a doubt, it seems to me, on the hypothesis of time reversal, but in themselves they provide little clue as to any alternative hypothesis. Anyone able to suggest an alternative would be in line for a Nobel Prize within a relatively short time. The plain fact of the matter is that in the presence of the observations now available the hypothesis of time reversal would require a belief in miracles. We are disinclined to believe in miracles. And

the disinclination to believe in miracles lies at the root of a wide class of applications of statistics. We would lose a great deal if we gave up the use of this principle.

If on the other hand we ask ourselves, Is the elaborate procedure which has sometimes been called a "test of a hypothesis" necessary, of course I do not think it is. The elaborate procedure I am referring to is one in which one is supposed to set up the hypothesis being tested, and to set up the class of all admissible alternatives, and then to consider what is the sample space, and then consider whether this is to be a best critical region of type A, B, C or any other letter, and so on. One is here assuming that the class of alternatives is well defined, so that a likelihood approach is possible and appropriate.

I may perhaps try to convey my attitude to tests of significance by an oversimplified picture of what I conceive as "a world view". Our multifarious experience can for this purpose be parcelled up into a large number of relatively independent sets of data, which we may imagine laid out in squares on a table, like an enormous checker-board, one set of data to a square. Along with each parcel of data we put an hypothesis, which we think of as "accounting for" (or a "model", thought of as generating) the data. To each hypothesis–data pair there will be a measure of discrepancy attached, as being in our judgement appropriate to the pair in question. Then it will follow that in each square of our checker-board will be a number, the significance level associated with the data, on the hypothesis belonging to the square in question, using the appropriate measure of discrepancy. We shall feel happy with our world-view (ignoring discontinuities) provided the numbers thus attached to the squares are reasonably uniformly distributed between 0 and 1. If we find too many numbers altogether too near to 0, we will have to change some of our hypotheses. But *some* numbers should always be near 0. The reserve with which we treat an hypothesis–data pair which gives us a low number arises from our knowledge, not that such things are impossible, but that they are rare. We cannot admit too many to the checker-board, or we shall find our world-view unsatisfactory. Somewhat as we can, and perhaps should, afford to be extravagant in spending money, every now and then, but we must not be so too often or we shall run into trouble.

The attitude I have tried to describe cannot be satisfactorily explained in terms of categorical attitudes of "acceptance" or "rejection" of hypotheses nor, I think, in terms of irrevocable "decisions". It is a relatively simple-minded attitude which is, none the less, one of the sheet anchors of statistical method.

To conclude, I would like to say that I have for a long time regarded many of the Bayesian criticisms of what are here referred to as the "orthodox procedures" to be justified in many respects. It is certainly true that these "orthodox procedures" ignore the frequently occurring situations where what is wanted is a routine procedure, in industry, or in bioassay, or some other situation, where a repeated set of situations will arise in which it is perfectly valid to make estimates of the relative frequency with which various situations will arise, and where with such estimates it is only natural to use a Bayesian procedure based on those relative frequencies. The fact that such situations tend to get ignored in so-called orthodox statistics is a pity and the Bayesian criticism of it is valid.

Another valid Bayesian criticism concerns the irrelevance for most purposes of "unbiasedness" and such properties of estimates. I am not sure whether it is a reflection of my own middle age, but I feel now that we are in the hands perhaps of the Committee of Public Safety at this point in the revolution in statistics. I have always felt sympathy with the Marquis de Condorcet who lost his head as a result of changes which perhaps he helped to start to some extent. There is a danger that the revolution in statistics may go too far, and it is a very serious danger. Above all, the point should be borne in mind that nowhere in this paper is the situation really contemplated—and this is a situation which really arises in science—where we simply do not know, and cannot know, what the alternatives are. Everywhere in the paper the assumption is made that if θ does not have one value, then it has another. And this is simply not true in real life. Having made my criticisms, Mr Chairman, I have much pleasure in seconding the vote of thanks.

The vote of thanks was put to the meeting and carried unanimously.

Dr I. J. Good: There is so much meat in Professor Pratt's paper, that, taken in a single helping, it is liable to provoke mental indigestion. Personally I have been able to take in only a small part of the banquet. But, in relation to the whole of it, I cannot resist quoting Doog's *bon mot* "To the Bayesian all things are Bayesian". A statistical technique or inference, and in fact any inference whatever, is reasonable if and only if it is approximately Bayesian. But the degree of approximation ought to be checked in each application. This applies to maximum likelihood, confidence intervals, tail-area probabilities, scientific induction and anything else you like to mention.

My comments will be mainly historical. In the first place Theorem 7 was largely anticipated recently by Thatcher (1964) and by Perks (1947).

The author has referred to an inequality I gave on p. 94 of *Probability and the Weighing of Evidence*, connected with the interpretation of the tail-area probability of a statistic when one is testing a null hypothesis of the form $\theta = 0$. He went on to say that significance tests are of doubtful relevance in such cases. I hope that his remark will not stop any statistician from making frequent use of significance tests. I agree with Professor Barnard on this point, although I believe that a tail-area probability can always be given a Bayesian interpretation. Surely the chi-squared test is very useful in spite of all the objections to it. An approximate Bayesian justification for its use in some circumstances was given in my book.

The author's comment about the doubtful relevance of significance tests of the null hypothesis, coming where it does in his text, might give the impression that I was a dyed-in-the-wool tail-area pundit. Accordingly I should like to state that, in my book, various precautions were mentioned that should be held in mind when using significance tests. Such precautions were especially worth pointing out at a time when Sir Ronald Fisher's influence was at its zenith. For example, several of the following precautions were mentioned.

(i) The null hypothesis is rarely precisely true, and can therefore be rejected on sufficiently large samples. (I did not know this had been anticipated by Berkson.)

(ii) A test, nominally for the hypothesis $\theta = 0$, with sample size n, is usually a test for whether $|\theta|$ is large enough to be detected from a sample of size n.

(iii) One sometimes can and should test the hypothesis $|\theta| < \delta$, where δ is small, instead of testing $\theta = 0$. But *if n is not large*, this comes to much the same as testing $\theta = 0$.

(iv) When n is large, the Bayes factor against the null hypothesis is not as large, for a given tail-area probability, as it is for small n. The handicap is liable to be about $\sqrt{n}$, a fact previously pointed by Jeffreys. (A Bayes factor in favour of a hypothesis, H, provided by evidence, E, is defined by $F(H:E) = O(H/E)/O(H)$, and is equal to a likelihood ratio if H and $\bar{H}$ (not-H) are both simple statistical hypotheses. Jeffreys usually took the initial odds of the null hypothesis as 1, so that, in his book, the Bayes factor became equal to the final odds.)

(v) The Bayes factor against the null hypothesis $\theta = 0$ is $1/(\gamma P)$, where P is the tail-area probability and γ is often in the range (10/3, 30). When $P > 0{\cdot}01$, γ is often about 4 or 5, and when $P < 0{\cdot}001$, γ is often about 10. (See, for example, Good, 1958.) If γ is outside the range mentioned, which is surprising, the Bayesian should reconsider his assumptions carefully.

(vi) If the non-null hypothesis is increased in precision, then the Bayes factor can be very much greater than $1/P$, where P is the tail-area probability based on a statistic appropriate for the vaguer non-null hypothesis.

(vii) A technique for selecting a statistic whose tail-area probability makes sense, is first to set up a Bayesian model that does not seem too absurd, to calculate the Bayes factor for this model, but to interpret it merely as a statistic and *not* as a Bayes factor. I call this the *Bayes/non-Bayes compromise*. (See, for example, Good 1957, especially p. 863.)

(viii) A blind use of tail-area probabilities allows the statistician to cheat, by claiming at a suitable point in a sequential experiment that he has a train to catch. This must have been known to Khintchine when he proved in 1924 that, in sequential binomial sampling,

a "sigmage" of nearly $\sqrt{(2 \log \log n)}$ is reached infinitely often, with probability 1. (Weaker results had been proved earlier by other mathematicians.) But note that the iterated logarithm increases with fabulous slowness, so that this particular objection to the use of tail-area probabilities is theoretical rather than practical. To be reasonably sure of getting 3σ one would need to go sampling for billions of years, by which time there might not be any trains to catch. I did not notice this theoretical objection to the use of tail-area probabilities until 1950.

I was puzzled by the speaker's comment, in the second paragraph of the Introduction, that only single-tailed tail-area probabilities are interpretable Bayesianwise. This seems to me to be false when testing the hypothesis $p = \frac{1}{2}$ in binomial sampling. If, given the non-null hypothesis, the initial distribution of the physical probability p is symmetrical about $p = \frac{1}{2}$, then a double-tail-area probability of 0·01 has roughly the same Bayesian significance that a single tail of 0·01 would have if the initial probability density were folded across $p = \frac{1}{2}$. On the other hand, in a chi-squared test for a multinomial distribution, the left and right tails have very different significance. A value of χ^2 very near 0 suggests either sampling without replacements, or, more probably, an error of the third kind in the calculation of χ^2.

Professor M. S. BARTLETT: I am pleased to add my welcome to Professor Pratt during his visit to this country; if I have occasion in a moment to cross swords with him, I hope he will not take it personally in any way. I am glad indeed to add also my thanks to him for his careful paper, which will be of permanent value in the difficult task of getting Bayesians and non-Bayesians, if not exactly agreeing, at least finding some common meeting ground. Perhaps wisely, Professor Pratt does not discuss broad issues, only specific ones. However, I think a reader of his paper has in consequence to be rather critical of Professor Pratt's notation and terminology. It is all very well mathematically to say (as is said after equation (23)) that "all statistics and all parameters are random variables", but what is this really supposed to mean? After all, even within the Bayesian camp their creed is not so unique as all that; there is the epistemological approach of Jeffreys, the personal or individualistic approach of Savage, and the frequency interpretation of prior probabilities attributable, I believe, to Karl Pearson. I am not being altogether facetious in suggesting that, while non-Bayesians should make it clear in their writings whether they are *non-Bayesian Orthodox* or *non-Bayesian Fisherian*, Bayesians should also take care to distinguish their various denominations of *Bayesian Epistemologists*, *Bayesian Orthodox* and *Bayesian Savages*. (In fairness to Dr Good, I could alternatively have referred to *Bayesian Goods*; but, oddly enough, this did not sound so good.) There would then be less danger in criticism that we are querying the wrong creed. I would have thought, and would concede some points in its favour in such a context, that Professor Pratt's membership of a Business School would have earmarked him as a *Bayesian Savage*, but I cannot tell from his paper. How precisely, for example, do we interpret "reasonable" in his phrase in Section 8.6: "reasonable prior probabilities"? The wording which precedes it is: "The question then becomes whether it *generally* gives better results . . ." (my italics). Has Professor Pratt here become *Bayesian Orthodox*? If so, I do not know how we can answer his question in the abstract, even if his own paper remains on that plane. We would need to consider some concrete examples, with perhaps a post-mortem or follow-up investigation (dare I say statistical investigation?) to see how alternative possible procedures would have fared.

Coming to one or two more specific points, I was interested to see the Bayesian handling of sufficient and insufficient statistics in Section 2, but the author's suggestion of using the sample mean whether sufficient or not would seem to me to need some further *caveat* about the shape of the distribution. Professor Pratt remarks: "the data beyond the sample mean . . . will presumably affect the distribution of θ very little". We know, however, that if the distribution were, say, like a double exponential half the information would be thrown away by such a procedure, and if like a t distribution with no more than two degrees of freedom, all of it.

On the question of stopping rules, I do not accept the likelihood principle as a postulate, or the argument that a statistic T, say, is sufficient (in the orthodox sense at least) without knowing which stopping rule was employed. The distribution of any other statistic, given T, has to be independent of θ, *for all* T, not just a particular T. In the case, say, of binomial or inverse binomial sampling, which experiment was carried out is, in Fisher's terminology, ancillary information; in particular, the information is different for the two types of sampling. With regard to the two primary statistics, r (the number of occurrences) and n (the number of trials), which one is an ancillary statistic depends on the type of sampling. The essential point here Professor Pratt himself notes at the end of his paper. The *non-Bayesian Orthodox* statistician (to avoid irrelevant argument I will exclude *Fisherians*) always views the sample as one of a class, and not as unique. He does not deny the interest and possible value of anything which is unique, but he cannot deal with it as a statistician; he may, like me, doubt whether he can deal with it as a scientist. If a Bayesian claims to surmount this difficulty by averaging over a prior distribution, he is either carrying out a purely conventional averaging as an *Epistemologist*, a unique personal averaging as a *Savage*, or an *Orthodox* statistical averaging which, being based on only vague prior information, is almost certainly wrong. *Bayesian Savages* sometimes claim polemically a degree of unanimity over their respective and unique assessments, and the *non-Bayesian Orthodox* statistician will always be happy to add to his investigations and investigate this claim for them also. If the Bayesians cavil by drawing attention to the fictitiousness of the non-Bayesian statistical probability concept, the non-Bayesian can retort that the fictitiousness of the velocity concept does not prevent a Bayesian being fined for exceeding the speed limit, or the fictitiousness of the mass concept prevent his being charged on excess baggage.

Mr J. Aitchison: I found this a most stimulating paper and would like to add my thanks to those of previous speakers.

Two of the central results (Theorems 2 and 5) of Professor Pratt's paper are direct consequences of the possibility of computing expected utilities in two equivalent ways:

$$E_{p(\theta)}\, E_{p(x|\theta)}\, U\{d(x), \theta\} = E_{p(x)}\, E_{p(\theta|x)}\, U\{d(x), \theta\}, \qquad \text{(A)}$$

d being the conclusion or decision function. Thus Theorem 2 takes $d(x)$ as an estimate and $U\{d(x), \theta\} = d(x) - \theta$, sets $E_{p(x|\theta)}\, U\{d(x), \theta\} = 0$ and concludes that the right-hand side of (A) is 0. Again Theorem 5 takes $d(x)$ as a confidence region with

$$U\{d(x), \theta\} = \begin{cases} 1 & \text{if} \quad \theta \in d(x), \\ 0 & \text{if} \quad \theta \notin d(x), \end{cases}$$

sets $E_{p(x|\theta)}\, U\{d(x), \theta\} = 1 - \alpha$ and concludes that the right-hand side of (A) is $1 - \alpha$. Corresponding results could be obtained by similar arguments for other standard inference procedures, such as tolerance regions. My purpose in introducing (A) is that I believe examination of it, within the framework of the clearer concepts of decision theory, throws some light on an underlying reason why many statisticians feel dissatisfied with standard inference procedures. The problem of choosing d to maximize (A) through its left-hand side is, of course, frequentist or normal decision analysis and through its right-hand side is Bayesian or extensive decision analysis, as defined by Raiffa and Schlaifer (1961). It is typical of standard inference procedures that they choose utility functions which are so simple that to attempt to find a d which maximizes $E_{p(x|\theta)}\, U\{d(x), \theta\}$ leads to triviality. For example, the best confidence region is the parameter space. Rather than introduce a more realistic utility specification which places higher values on "smaller" confidence regions, orthodoxy escapes by setting $E_{p(x|\theta)}\, U\{d(x), \theta\}$ equal to some predetermined constant, e.g. $1 - \alpha$ in the confidence-region case, below the attainable maximum. From the class of d's which satisfy such a restriction an attempt is made to find one which satisfies some other optimum property. The choice of the predetermined constant, which is widely recognized as an arbitrary act with little real meaning except in situations where the

informative experiment is being repeated, is given pride of place. This relegation of optimization to a very secondary role is, I believe, one of the causes of anomalies and paradoxes in standard inference procedures. We cannot expect to get a sensible answer every time we ask a silly question.

I would like to see a much more serious attempt on the part of statisticians to encourage their clients to realize that hard thinking about utility functions is a necessary part of the inference or decision process. As a small contribution towards such a move I would like to mention a method of specifying utility functions which I have found useful in practice. Often the consequences of a conclusion or decision $d(x)$ are most readily reckoned in terms of repetitions of some future experiment. For example, $d(x)$ may be the conclusion that one drug or variety is better than others and the basic future experiment would be the recording of the effect y of the drug or the yield y of the variety. Suppose that the density describing this experiment is $p\{y \mid d(x), \theta\}$ and that the value or y-utility of using $d(x)$ and of observing y in the future is $V\{y, d(x)\}$. Then the appropriate θ-utility for the decision problem is the average value per future experiment:

$$U\{d(x), \theta\} = E_{p\{y \mid d(x), \theta\}} V\{y, d(x)\}.$$

Often V is more directly assessable by an experimenter than U.

Mr W. PERKS: The author mentioned that unbiased statistics do not transform "properly", nor does the mean of a posterior distribution. While I appreciate that the mode has something to be said for it as an estimate of the parameter although it does not transform "properly", the median of the posterior distribution, and the other fractiles, not only transform "properly" but have the property of invariance if the prior indifference rule used is invariant under transformation of the parameter. The mean implicitly uses linear utilities but the median and the other fractiles are completely free of utilities. There is something to be said for the statistician supplying estimates completely free from his own prior beliefs and from any arbitrary scale of utilities.

Mr KERRIDGE: Professor Pratt is to be congratulated on his defence of frequentist methods from a Bayesian point of view. As we have been asked to declare our affiliation, I should explain that I am a frequency Bayesian roughly in the Karl Pearson or Von Mises tradition. As such, the agreement between Bayesian and, for example, confidence arguments is of vital interest to me. I would be more inclined to justify Bayesian methods, in the absence of definable frequencies, by confidence arguments, but both ways of looking at the problem are relevant.

To complement what Professor Pratt has said, I might mention a result which is in a sense the converse of his Theorem 5. That is, if we start with a Bayesian statement we can show that in large samples an approximate confidence property will appear. For simplicity, the notation of discrete hypotheses and sample space is used, but the generalization to other cases is easily carried out.

Suppose that we have hypotheses $H_1, H_2, \ldots, H_i, \ldots$ and outcomes of an experiment $X_1, X_2, \ldots, X_j, \ldots$, the frequencies $F(X_j \mid H_i)$ being known, but the frequencies $F(H_i)$ being unknown or non-existent. (It is convenient to use the F notation for frequencies, reserving the more usual P notation for non-frequency probabilities.)

A rule of assertion A is constructed as follows. Assuming prior probabilities $P(H_i)$, find any assertion A_j corresponding to each X_j, such that $P(A_j \mid X_j)$, the conditional probability that A_j is true $= \alpha$ for each j. Exact equality may not be possible because of the discrete nature of the problem, but approximate equality is sufficient for the argument. The assertion may be of the confidence type, i.e. that some subset of hypotheses contains the true hypothesis, or of more general character, for example that some future sample will have a specified property. A is then the complete rule of assertion, "assert A_j when X_j is observed". Then

$$\sum_i P(H_i)\, F(A \mid H_i) = P(A) = \sum_j P(A_j\, X_j) = \sum_j P(A_j \mid X_j)\, P(X_j) \simeq \alpha.$$

Suppose we maintain the rule of assertion, but change the prior probabilities to $P'(H_i)$. In large samples there will be no appreciable difference in the posterior probabilities, and so arguing as before, we must have

$$\sum_i P'(H_i)\, F(A \mid H_i) \simeq \alpha.$$

But the $P'(H_i)$ may be any probabilities, so that we must have

$$F(A \mid H_i) \simeq \alpha \quad \text{for each } i.$$

Hence the rule of assertion generated by a Bayesian method leads to a correct assertion in a proportion of cases almost equal to the posterior frequency used in generating it.

This is in some ways stronger than Professor Pratt's result, which deals only with expectations, but, on the other hand, we cannot say how large a sample is required to make the approximation good. A particular example of a predictive type of assertion has been considered by Thatcher (1964) in connection with Laplace's Rule of Succession.

It would be interesting to have converses in the above sense to other results in this very useful paper: it seems very likely that they will be true. It is to be hoped that the two-way traffic between Bayesian and orthodox ideas will become an established and fruitful feature of statistics.

The following written contribution was received after the meeting:

Miss V. R. Cane: The statement that Bayesian methods satisfy the likelihood principle seems to involve the implicit assumption that the experiment chosen by the experimenter is independent of his prior beliefs about the parameter he is investigating. The corresponding assumption made by Birnbaum is that, if we have two experiments E and E', we are allowed to make up a mixture experiment from them; it is not the same as his principle of conditionality, which takes the mixture as already existing. Consider, however, the following example: trials of a new drug have been performed on animals which are thought to be physiologically similar to human beings (for instance, animals of this species may in the past have responded to other drugs in the same way as human beings do) and the drug has been found to be effective for 99 per cent of the animals tested. An experimenter in possession of this information has to set up an experiment to test the drug on human patients. It seems unlikely that he would decide on a sequential stopping rule of the form "go on until there have been 100 failures" although he might accept "go on until there have been 2 failures" or "go on until there have been 100 successes"; even if he were to decide on an experiment with fixed sample size, he would not choose this size by consulting a random number table.

His decision whether or not to do a particular experiment can, it seems, only be based on considerations of efficiency. It seems easiest to discuss this in terms of orthodox statistical methods, using at the same time the idea that an experiment is a form of communication between the experimenter and an observer, who may be the same person. Suppose that the experimenter believes that the probability of success is $\geqslant \theta_1$, and knows that the observer will test the result against a null hypothesis $\theta = \theta_0 < \theta_1$; if he chooses an experiment of fixed size n_0, with s_0 and n_0 determined by

$$\sum_{r=s_0}^{n_0} b(r;\, n_0 \mid \theta_0) = \alpha, \quad \sum_{r=s_0}^{n_0} b(r;\, n_0 \mid \theta_1) = 1-\beta,$$

then the observer will reject $\theta = \theta_0$ in favour of $\theta > \theta_0$ at significance level α if there are s_0 or more successes and the probability that he will do so if $\theta \geqslant \theta_1$ is at least $1-\beta$. The equivalent experiment with the rule of stopping after f_1 failures and rejecting $\theta = \theta_0$ if the number of trials is n_1 or more requires

$$\sum_{r=n_1-f_1}^{n_1-1} b(r;\, n_1-1 \mid \theta_0) = \alpha, \quad \sum_{r=n-f_1}^{n_1-1} b(r;\, n_1-1 \mid \theta_1) = 1-\beta,$$

that is, $n_1 = n_0+1$, $f_1 = (n_0-s_0)+1$; if the rule is to stop after s_2 successes, and rejection of $\theta = \theta_0$ occurs if the number of trials is $\leqslant n_2$, the equivalent experiment requires $s_2 = s_0, n_2 = n_0$. Consideration of these three cases will show that they all lead to the same rule: he ought to go on until there are s_0 successes or until n_0 patients have been treated, whichever occurs sooner; in effect, he will be using his previous knowledge, which suggests that success is likely, to make the experiment short, while guarding against doing an experiment of indefinite length if in fact the probability of success is low. If he thinks that failure is likely (i.e. $\theta \leqslant \theta_1 < \theta_0$) he should work with the number of failures rather than the number of successes. Thus whether he chooses to count successes or failures will indicate his prior beliefs.

An experiment should be acceptable to an experimenter; it seems clear that there are at least some cases in which the experimenter will use prior information to choose which experiment to perform, and, if he thinks he has chosen the best experiment for his purpose, he is unlikely to accept some arbitrary experiment or mixture of experiments proposed to him.

Professor PRATT replied briefly at the meeting and subsequently in writing as follows:

I am most grateful to all the contributors to the discussion for the courtesy, carefulness and substance of their remarks. I am also most grateful to all those who helped make this trip not only possible but also extremely pleasant and stimulating.

With Professor Lindley's reasoning and his conclusion that the bulk of standard statistics *means a lot* to a Bayesian, I agree. But I still feel (perhaps he does too) that standard tests as standardly applied are usually not *appropriate*, and reliance on them at some stage seriously weakens all too many analyses of real data, including those using normal theory. His interpretation of significance levels seems to me to make this all the more evident. The posterior probability of the region inside (or outside) a contour of constant probability density is largely irrelevant to the purposes of most applications of tests. (See Section 8 of the paper.) Incidentally, in the case of a uniform prior distribution, such a region is "snug" in the terminology of Hildreth (1963). This very interesting paper I should certainly have referred to earlier, especially in connection with insufficient statistics.

An interesting question arises in connection with the exponential family introduced by Professor Barnard: Are there one-to-one functions of the first k sample moments satisfying the conditions of Theorem 1 under this family? (Note that here the conditional distribution of θ given all the observations is the same as that given the first k sample moments, since the latter form a sufficient statistic.) If the answer to the question is yes, then for large samples, Professor Barnard's recommended analysis of the specific features of each sample has high probability of giving approximately the same distribution for the population mean as that based on only the sample mean and variance, even when the population distribution is a highly non-normal member of the family. It might be worth carrying out this analysis nevertheless, depending on the importance of the problem relative to the cost and effort of computation (and perhaps of assessing prior distributions). And of course the situation changes if one would become interested in a different parameter (perhaps asking a more robust question) upon learning that the population was highly non-normal. The questions we ask do not always need to be robust, however—there may be good reason for an interest in the population mean even when it is unfortunately not very robust. Sometimes we should face difficulties head on rather than change our problems to ones with nicer answers.

The important matter on which Professor Barnard and I disagree is tests of significance in the situation "where we simply do not know, and cannot know, what the alternatives are". This situation is contemplated in the paper (last paragraph of Section 8.9), but very briefly because I have little that is constructive to say about it. What I did say is not relevant to the case of particular concern to Professor Barnard where there is a scientific (and statistical) null hypothesis which really might be almost exactly true but no specified

scientific alternative. The more I think about tests of significance in such situations, the more dissatisfied I become. The reason is this. The odds ratio (Bayes factor) provided by the data is $f_0(x)/f_1(x)$, where $f_0(x)$ is the density of the observed x under H_0 (marginal on any parameters of H_0 if H_0 is composite) and $f_1(x)$ is the density of the observed x under the hypothesis that H_0 is false. If "H_0 false" were a parametric alternative, then $f_1(x)$ could be obtained by assessing a distribution for the parameters and integrating them out. Here this is impossible. In principle, $f_1(x)$ can be assessed directly. Unfortunately any one person's judgement of $f_1(x)$ is liable to be very vague and different people's judgements are liable to be very different. And in practice people do feel very uncertain and disagree greatly with one another about how far the data discredit H_0 in the kind of situation under discussion. Computing a P-value runs the danger of hiding this real uncertainty and legitimate disagreement behind a screen of irrelevant precision. The relevant quantity is $f_0(x)/f_1(x)$, however uncertain and disagreeing we may be about its value. The only way in which I have been able to convince myself that P-values are relevant is through Good's inequality (37) when it is applicable.

I shall comment on only a few of the remarks in the banquet provided by Dr Good. I am very grateful for his references. I like to think I would have taken note of Thatcher's paper in due course, but the extremely interesting paper by Perks (1947) seems to be unknown in the United States.

It never occurred to me that anyone could conceivably think Dr Good either "dyed-in-the-wool" or a "tail-area pundit", but I apologize indeed if I conveyed that impression.

The information Dr Good gives in his precaution (v) on the relation between the Bayes factor (odds) and the P-value is sufficient to justify taking the precaution he suggests. It also would be very pleasant often to be able to estimate the Bayes factor from the P-value without actually calculating the former. For this purpose, I hope Dr Good will sometime publish a comprehensive collection of real-life examples, giving for each the P-value of a test of $\theta = 0$ and the Bayes factor for the null hypothesis of real interest (which is presumably of the form $|\theta| \leqslant \delta$ usually).

In reply to Dr Good's final paragraph, the second paragraph of Section 1 was not intended to suggest (and a word has now been added to avoid the suggestion) that single-tailed tail-area probabilities are interpretable Bayesian-wise in all cases. In particular they are not when the null hypothesis is a single point. (Cf. Section 7.)

Professor Bartlett is correct in classifying me a Bayesian Savage, though I might take exception to his word order. On the whole, I would rather be called a Savage Bayesian than a Bayesian Savage. Of course I can quite see that Professor Bartlett might not want to admit the possibility of a Good Bayesian.

In Section 8.6 I used what I took to be the language of the non-Bayesian Orthodox in an attempt to convince them that, even from their own point of view, choosing some α, β and θ_1 is not automatically a better way to determine a procedure than choosing some prior distribution and costs and then optimizing. I would rather try to define "reasonable" (which I did put in quotes) for prior probabilities than for α, β and θ_1.

As regards distributions like a double exponential, I do think that in those situations in which the population mean is really the parameter of interest (not some more robust parameter), I would very rarely be willing to assume enough about the population (such as exact symmetry) to want to use something in place of the sample mean. Of course, if a double exponential distribution is assumed, the conditions of Theorem 1 are not satisfied for t_3 = the sample median, for instance.

As regards sufficient statistics, I believe I am using the same definition as Professor Bartlett. The next-to-last paragraph of Section 6 is perhaps too informally worded. It would have been more precise to say that $(X_1, X_2, \ldots, X_N)$ is a sufficient statistic, where N is the number of observations actually made. Consider, for instance, a Bernoulli process with parameter p, and suppose the experiment is to do either binomial sampling with $n = 10$ or inverse binomial sampling with $r = 4$, as determined by the flip of a fair coin (independent of the process). Then the actual outcomes $(X_1, X_2, \ldots, X_N)$ form a sufficient

statistic in an almost degenerate way: they determine everything observed except that they do not determine the side of the coin observed in cases when $N = 10$, $\Sigma_1^{10} X_i = 4$, $X_{10} = 1$; but in any such case the conditional probability of each side of the coin is $\frac{1}{2}$ no matter what the value of p. According to the sufficiency principle, then, we need not care which stopping rule was used if the actual outcomes do not determine it. This is all that is needed for this part of Birnbaum's argument.

Mr Aitchison's method of proof is illuminating, and the rest of his remarks I can only say I agree with.

Mr Perks was certainly succinct. I am not sure of his exact meaning in one or two places, and I apologize if the following comments reflect a misunderstanding. The fact (I hope I am correct) that Mr Perks is a Bayesian Epistemologist would explain his statement, with which a Bayesian Savage of course disagrees, that the median of the posterior distribution is free from prior beliefs. But what I do not see is the sense in which the median is free of utilities. In an estimation problem, for instance, the median is optimum if the loss is proportional to absolute error, while the mean of the posterior distribution is optimum if the loss is proportional to squared error. The mean also determines which act to take in a two-action problem with linear utilities for each act (by whether it is larger or smaller than the "break-even" value). Different functionals of the posterior distribution are relevant to different decision problems. Do the Bayesian Epistemologists have special ways of determining their decision problems as well as their prior distributions?

Mr Kerridge's strong converse of Theorem 5 is interesting. It is perhaps worth mentioning that no correspondingly stronger version of Theorem 5 itself will hold without considerable added restrictions. For a trivial example suggesting less trivial ones, in any ordinary problem, as the number of observations approaches infinity, the posterior probability of a confidence region based on half the observations will not converge to the confidence level in any usual sense of convergence.

The problem raised by Miss Cane I shall skirt somewhat. I agree that the experimenter's prior beliefs influence his choice of experiment, and his choice of experiment may well influence someone else's beliefs. If so, the latter's prior distribution for analysing the experiment should summarize his beliefs after learning what experiment was chosen. Of course, the choice of experiment is largely determined in practice by factors other than the experimenter's prior beliefs. In any case, the non-Bayesian Orthodox handling of stopping rules, for instance, is not an attempt to reflect their influence on prior opinions, and I remain unconvinced of the need for reflecting the influence of prior opinions on stopping rules.

Miss Cane's remarks do point to the need for care in stating the likelihood principle, Birnbaum's argument for it, and the fact that Bayesian methods satisfy it, but they do not seem to me to weaken them. I shall not attempt to make the more careful statements here but merely point out that if the experiment is completely determined by circumstances not under the experimenter's control and we accordingly regard the actual stopping rule and all hypothetical stopping rules as dictated by external fiat, Miss Cane's problem does not arise and non-Bayesian Orthodox practice is still strikingly discordant with the next-to-last paragraph of Section 6.

References in the Discussion

Good, I. J. (1957), "Saddle-point methods for the multinomial distribution", *Ann. math. Statist.*, **28**, 861–881.

Hildreth, C. (1963), "Bayesian statisticians and remote clients", *Econometrica*, **31**, 422–438.

Perks, W. (1947), "Some observations on inverse probability including a new indifference rule", *J. Inst. Actu.*, **73**, 285–334.

Thatcher, A. R. (1964), "Relationships between Bayesian and confidence limits for predictions", *J. R. statist. Soc.* B, **26**, 176–192.