THE USE OF PRIOR PROBABILITY DISTRIBUTIONS IN STATISTICAL INFERENCE AND DECISIONS

D. V. LINDLEY
UNIVERSITY OF CAMBRIDGE

1. Introduction

The object of this paper is to discuss some of the general points that arise in common statistical problems when a prior probability distribution is used. Most current statistical thinking does not use such a distribution but we feel that some illumination and understanding can be gained through its use, and, in our more enthusiastic moments, we even feel that only a completely Bayesian attitude towards statistical thinking is coherent and practical. The purpose of this paper is not to present propaganda in favor of the Bayesian attitude but merely to explore the consequences of it.

The first two sections of the paper deal with the large-sample problem where the effect of the prior distribution is small and the inferences and decisions made using a Bayesian approach correspond fairly closely with current practice and are similar to those given by Le Cam [7] using the usual approach. If we emphasize any differences that do exist it is because they are of interest: the most important point is the similarity. In the final sections of the paper the problem of small samples is discussed. Here the prior distribution is important and much of the discussion centers around the choice of it. Some tentative rules are suggested for the choice and comparison is made with the work of Jeffreys. The discussion here is fragmentary and incomplete: but it is hoped that the arguments, unsatisfactory as they are, will stimulate others to produce better ones.

Throughout this paper only the problem of the analysis of an experiment is discussed. Design problems are not considered.

2. Large-sample problems of the estimation type

We first formulate the mathematical model following, apart from the prior distribution, the usual pattern. A *sample space* $\mathfrak{X}$ of points x supports a σ-algebra $\mathfrak{B}$ of sets X. For each point $\boldsymbol{\theta}$ of a *parameter space* Θ there is defined a probability measure over $(\mathfrak{X}, \mathfrak{B})$. It will be assumed that each of these measures is dominated by a σ-finite measure over $(\mathfrak{X}, \mathfrak{B})$; from which it follows that there exist probabil-

Present address: University College of Wales, Aberystwyth.

ity densities $p(x|\boldsymbol{\theta})$ describing the probability measures by integration with respect to the dominating measure. Such an integration will be denoted by $\int dx$ followed by the function to be integrated. These densities considered as functions of $\boldsymbol{\theta}$ for fixed x will be termed *likelihoods*.

In the present paper only the case where the parameter space Θ is a subset of Euclidean space of a finite number, k, of dimensions will be considered. A point of the space can be written in terms of its coordinates; $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$. A prior probability distribution is a probability distribution over this space with the usual Borel σ-algebra. Again the distribution is described by a density $\pi(\boldsymbol{\theta})$ with respect to some σ-finite measure. The problem will be called of *estimation type* if this σ-finite measure is Lebesgue measure, and integration with respect to it will be written $\int d\boldsymbol{\theta}$ followed by the function to be integrated. Such a prior distribution implies that no values of $\boldsymbol{\theta}$ are of a higher order of prior belief than others. This is perhaps not the case with problems of the testing type considered in the next section.

Next there is a *decision space* $\mathfrak{D}$ of points d. (The distinction between inference and decision problems will be discussed briefly in section 4.) The relationship between the decision space and the parameter space is described by means of a *utility function* which is a real-valued function $U(d, \boldsymbol{\theta})$ defined on $\mathfrak{D} \times \Theta$, and is supposed to represent the utility to the experimenter of taking d when the true parameter value which determined the probability distribution of the observed sample point (or observation) x, was $\boldsymbol{\theta}$. Most writers use loss functions: any reader who wishes to do so can use the negative of the utility. A *decision function*, denoted either by δ or $d(\cdot)$, is a function with domain $\mathfrak{X}$ and range in $\mathfrak{D}$.

The problem is to determine the best decision function. If the decision function $\delta{:}d(\cdot)$ is used the expected utility is

$$\int dx \int d\boldsymbol{\theta}\, U[d(x), \boldsymbol{\theta}]p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{2.1}$$

That decision function is judged best which maximizes (2.1). It will be assumed that (2.1) is finite for all decision functions. This will certainly be true if $U(d, \boldsymbol{\theta})$ is bounded and $\pi(\boldsymbol{\theta})$ is a normed probability distribution, that is one for which $\int d\boldsymbol{\theta}\, \pi(\boldsymbol{\theta}) = 1$. The assumption that $U(d, \boldsymbol{\theta})$ be bounded seems desirable in order to avoid paradoxes like the St. Petersburg one (see [3]). We shall have occasion later to use unnormed distributions.

By Fubini's theorem the double integral (2.1) may be written as a repeated integral $\int dx \left\{\int d\boldsymbol{\theta}\, U[d(x), \boldsymbol{\theta}]p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\right\}$ and the maximization may be carried out for each x separately. The optimum decision function is defined by taking for each x that d which maximizes

$$\lambda(d, x) = \int d\boldsymbol{\theta}\, U(d, \boldsymbol{\theta})p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{2.2}$$

Although it is usual to start from (2.1) it is arguable that the optimum decision procedure should be defined as that which maximizes (2.2). For, by Bayes theorem, the posterior probability of $\boldsymbol{\theta}$ has density

$$\pi(\boldsymbol{\theta}|x) = \frac{p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(x)}, \tag{2.3}$$

where

$$p(x) = \int d\boldsymbol{\theta}\, p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{2.4}$$

so that (2.2) may be written

$$\frac{\lambda(d, x)}{p(x)} = \int d\boldsymbol{\theta}\, U(d, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|x), \tag{2.5}$$

the expected utility, given x, of taking decision d. Once x has been observed (2.5) and not (2.1) is the relevant expression to consider. The sample space is only necessary before experimentation, that is before x has been observed, in order to choose between experiments, a topic which will not be discussed in the present paper. In the approach using (2.5) we again assume the utility function bounded and the posterior distribution normed.

We now derive an approximation for $\lambda(d, x)$, equation (2.2), valid for large n, when the sample point is that of n independent observations with the same probability distribution. We write $x^{(n)}$ where we have previously written x and $p^{(n)}$ instead of p, p now denoting the density of a single observation. Let $x^{(n)} = (x_1, x_2, \cdots, x_n)$ and

$$p^{(n)}(x^{(n)}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}). \tag{2.6}$$

$$L(x_i|\boldsymbol{\theta}) = \log p(x_i|\boldsymbol{\theta}),$$

$$L^{(n)}(x^{(n)}|\boldsymbol{\theta}) = \log p^{(n)}(x^{(n)}|\boldsymbol{\theta}) = \sum_{i=1}^{n} L(x_i|\boldsymbol{\theta}). \tag{2.7}$$

Since the utility function and the prior density occur so frequently as a product it is convenient to introduce the notation

$$w(d, \boldsymbol{\theta}) = U(d, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.8}$$

when (2.2) becomes

$$\lambda(d, x^{(n)}) = \int d\boldsymbol{\theta}\, w(d, \boldsymbol{\theta}) \exp [L^{(n)}(x^{(n)}|\boldsymbol{\theta})]. \tag{2.9}$$

Let $\hat{\boldsymbol{\theta}}^{(n)}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$ and suppose sufficient regularity conditions obtain to ensure that this maximum can be obtained by the usual differentiation process. Furthermore write

$$nc_{ij}(\hat{\boldsymbol{\theta}}^{(n)}) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} L^{(n)}(x^{(n)}|\hat{\boldsymbol{\theta}}^{(n)}), \tag{2.10}$$

assuming this to exist. (The left side depends on $x^{(n)}$ also but for notational simplicity this dependence has not been indicated.) Then we may write (2.9) in the form

$$\lambda(d, x^{(n)}) = \int d\boldsymbol{\theta}\, w(d, \boldsymbol{\theta}) \exp\Big[L^{(n)}(x^{(n)}|\hat{\boldsymbol{\theta}}^{(n)}) - \frac{n}{2} \sum_{ij} (\theta_i - \hat{\theta}_i^{(n)})(\theta_j - \hat{\theta}_j^{(n)}) c_{ij}(\hat{\boldsymbol{\theta}}^{(n)})\Big] + o(1), \tag{2.11}$$

the order of the remainder following from the fact that the derivatives of the likelihood function are $O(n)$. Now if $w(d, \boldsymbol{\theta})$ is a "smooth" function of $\boldsymbol{\theta}$ and bounded away from zero in the neighborhood of $\hat{\boldsymbol{\theta}}^{(n)}$, it too may be expanded and, to the order being considered, only the first term $w(d, \hat{\boldsymbol{\theta}}^{(n)})$ is relevant. There is no loss in generality in supposing this, since a constant may be added to the utility function. On practical grounds, the prior probability should not vanish. It therefore follows that

$$\begin{aligned} \lambda(d, x^{(n)}) &\sim w(d, \hat{\boldsymbol{\theta}}^{(n)}) \exp\,[L^{(n)}(x^{(n)}|\hat{\boldsymbol{\theta}}^{(n)})] \\ &\quad \int d\boldsymbol{\theta} \exp\Big[-\frac{n}{2} \sum_{ij} (\theta_i - \hat{\theta}_i^{(n)})(\theta_j - \hat{\theta}_j^{(n)}) c_{ij}(\hat{\boldsymbol{\theta}}^{(n)})\Big] \\ &\sim (2\pi)^{k/2} w(d, \hat{\boldsymbol{\theta}}^{(n)}) p^{(n)}(x^{(n)}|\hat{\boldsymbol{\theta}}^{(n)})\, \Delta_n^{-1/2} \end{aligned} \tag{2.12}$$

where Δ_n is the determinant of the matrix whose typical element is $nc_{ij}(\hat{\boldsymbol{\theta}}^{(n)})$.

The optimum decision function when $x^{(n)}$ is observed is to take that d which maximizes $\lambda(d, x^{(n)})$, and hence a possible approximately optimum decision function is to take that d which maximizes the right side of (2.12). But in that expression d only appears in $w(d, \hat{\boldsymbol{\theta}}^{(n)}) = U(d, \hat{\boldsymbol{\theta}}^{(n)})\pi(\hat{\boldsymbol{\theta}}^{(n)})$ and hence only in the utility function. Consequently the suggested approximately optimum decision function is equivalent to taking the d which maximizes $U(d, \hat{\boldsymbol{\theta}}^{(n)})$. It is easy to see that this approximating function has asymptotically the same expected utility as the optimum decision function; for the expected utility using d is $\lambda(d, x^{(n)})/p(x^{(n)})$, equation (2.5), and an asymptotic form for $p(x^{(n)})$, equation (2.4), can be obtained in the same way as for $\lambda(d, x^{(n)})$ by putting $U = 1$ identically. Hence,

$$p(x^{(n)}) \sim (2\pi)^{k/2} \pi(\hat{\boldsymbol{\theta}}^{(n)}) p^{(n)}(x^{(n)}|\hat{\boldsymbol{\theta}}^{(n)})\, \Delta_n^{-1/2} \tag{2.13}$$

and so the asymptotic expected utility of any decision d is the ratio of (2.12) and (2.13), namely $U(d, \hat{\boldsymbol{\theta}}^{(n)})$, the same as with the approximation.

We have therefore the following large-sample optimum decision function: *If the prior probability has a density with respect to Lebesgue measure and if the likelihood function behaves respectably then choose that decision which maximizes the utility at the maximum likelihood value of the parameter. Or, act as if the maximum likelihood value were the true value.*

Thus in large samples the exact form of the prior distribution is irrelevant and maximum likelihood estimation (in the broad sense considered here) is optimum. In the particular case of point estimation this has already been remarked by Jeffreys (see section 4.0 in [6]) and in the general case such a rule has been advocated by Chernoff [2]. Comparable results have also been given by Le Cam [7]. Any BAN estimate, in the terminology of Neyman [12], could be substituted for $\hat{\boldsymbol{\theta}}^{(n)}$ since such an estimate, by the argument of Fisher [5], must be asymptotically perfectly correlated with the maximum likelihood value and the error committed in replacing one by the other is of smaller order than $\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}$, that is it is $o(1/\sqrt{n})$.

A slight extension of the above argument can provide an approximation to the posterior distribution of $\boldsymbol{\theta}$ in large samples. The posterior distribution is, equation (2.3),

$$\pi^{(n)}(\boldsymbol{\theta}|x^{(n)}) = \frac{p^{(n)}(x^{(n)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(x^{(n)})} \tag{2.14}$$

and an asymptotic form for the denominator has already been obtained, equation (2.13). A series expansion of the numerator on the lines of that in (2.11) shows immediately that

$$\pi^{(n)}(\boldsymbol{\theta}|x^{(n)}) \sim (2\pi)^{-k/2}\,\Delta_n^{-1/2}\exp\left[-\frac{n}{2}\sum_{ij}(\theta_i - \hat{\theta}_i^{(n)})(\theta_j - \hat{\theta}_j^{(n)})c_{ij}(\hat{\boldsymbol{\theta}}^{(n)})\right], \tag{2.15}$$

asymptotically normal with means at the maximum likelihood values and dispersion matrix $||n^{-1}c^{ij}(\hat{\boldsymbol{\theta}}^{(n)})||$, inverse to $||nc_{ij}(\hat{\boldsymbol{\theta}}^{(n)})||$. The result differs only slightly from a similar result which is widely used in circumstances where a Bayesian would use the posterior distribution: namely the result that the maximum likelihood estimates are asymptotically normally distributed about the true values with dispersion matrix the inverse of $||n\bar{c}_{ij}(\boldsymbol{\theta}_0)||$ where $\bar{c}_{ij}(\boldsymbol{\theta}_0)$ is the expectation of $c_{ij}(\boldsymbol{\theta}_0)$ at the true value $\boldsymbol{\theta}_0$. This result is useless as it stands since $\boldsymbol{\theta}_0$ is necessarily unknown, but it may be replaced by $\hat{\boldsymbol{\theta}}^{(n)}$. If a Bayesian makes a similar approximation, and one of the same order of error, and replaces $c_{ij}(\hat{\boldsymbol{\theta}}^{(n)})$ by $\bar{c}_{ij}(\hat{\boldsymbol{\theta}}^{(n)})$, then the two results are practically equivalent. But we shall see below that the latter approximation is unsound in principle.

We conclude this section by showing how better asymptotic approximations may be obtained by an extension of the argument leading to (2.12). I am indebted to Professor H. E. Daniels for helpful discussion on the form of the expansions used here. To avoid complicated notations only the case of a single parameter ($k = 1$) will be discussed and the dependence on anything other than $\theta_1 = \theta$ will not be reflected in the notation. Thus we write the integral in (2.9) as

$$\int d\theta\, w(\theta) \exp\,[L(\theta)]. \tag{2.16}$$

Let $L_i(\hat{\theta})$ denote the ith partial derivative of $L(\theta)$ with respect to θ evaluated at $\hat{\theta}$, and define $w_i(\hat{\theta})$ similarly. Then (2.16) is

$$(2.17)\qquad \int d\theta\, w(\theta) \exp\left[L(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 L_2(\hat{\theta}) + \frac{1}{3!}(\theta - \hat{\theta})^3 L_3(\hat{\theta}) + \cdots\right]$$

$$= \exp L(\hat{\theta}) \int d\theta \left\{w(\hat{\theta}) + (\theta - \hat{\theta})w_1(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 w_2(\hat{\theta}) + \cdots\right\}$$

$$\left\{1 + \frac{1}{3!}(\theta - \hat{\theta})^3 L_3(\hat{\theta}) + \frac{1}{4!}(\theta - \hat{\theta})^4 L_4(\hat{\theta}) + \frac{1}{2}\left[\frac{1}{3!}(\theta - \hat{\theta})^3 L_3(\hat{\theta})\right]^2 + \cdots\right\}$$

$$\exp\left[\frac{1}{2}(\theta - \hat{\theta})^2 L_2(\hat{\theta})\right].$$

The order of the terms needs some care. $(\theta - \hat{\theta}) = O(1/\sqrt{n})$ and $L_i(\hat{\theta}) = O(n)$. In the two sets of braces only the terms up to $O(1/n)$ have been included. Collecting terms of like order together we have for (2.16)

$$(2.18)\qquad e^{L(\hat{\theta})} \int d\theta\, e^{(1/2)(\theta - \hat{\theta})^2 L_2(\hat{\theta})} \left\{w(\hat{\theta}) + \frac{1}{3!}(\theta - \hat{\theta})^3 L_3(\hat{\theta}) w(\hat{\theta})\right.$$

$$+ (\theta - \hat{\theta})w_1(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 w_2(\hat{\theta}) + \frac{1}{3!}(\theta - \hat{\theta})^4 L_3(\hat{\theta}) w_1(\hat{\theta})$$

$$\left. + \frac{1}{4!}(\theta - \hat{\theta})^4 L_4(\hat{\theta}) w(\hat{\theta}) + \frac{1}{2}\left[\frac{1}{3!}(\theta - \hat{\theta})^3 L_3(\hat{\theta})\right]^2 w(\hat{\theta}) + \cdots\right\}$$

$$= e^{L(\hat{\theta})} \left[\frac{-2\pi}{L_2(\hat{\theta})}\right]^{1/2} \left\{w(\hat{\theta}) - \frac{w_2(\hat{\theta})}{2L_2(\hat{\theta})} + \frac{L_3(\hat{\theta}) w_1(\hat{\theta})}{3! L_2(\hat{\theta})^2}\right.$$

$$\left. + \frac{3L_4(\hat{\theta}) w(\hat{\theta})}{4! L_2(\hat{\theta})^2} - \frac{5 \cdot 3 L_3(\hat{\theta})^2 w(\hat{\theta})}{2(3!)^2 L_2(\hat{\theta})^3} + \cdots\right\}.$$

The first term in braces is the one originally obtained in (2.12). The remaining terms in the braces are all $O(1/n)$.

As an illustration of the use of this asymptotic expansion for $\lambda(d, x^{(n)})$ consider the case where the decision space is also Euclidean of one dimension and the problem is one of point estimation of the parameter. Suppose that the prior probability is uniform in the neighborhood of $\hat{\theta}$ (that is, the density is constant) and that the utility function is approximately quadratic there, so that $w(d, \theta) \propto U - (d - \theta)^2$ say, where U is the utility of the correct decision, supposed constant. Then in (2.18)

$$(2.19)\qquad \begin{aligned} w(\hat{\theta}) &\propto U - (d - \hat{\theta})^2, \\ w_1(\hat{\theta}) &\propto 2(d - \hat{\theta}), \\ w_2(\hat{\theta}) &\propto -2. \end{aligned}$$

If we attempt to maximize (2.18) over d, only the part in braces is relevant and we have to maximize (writing $L_i(\hat{\theta}) = L_i$)

$$(2.20)\qquad [U - (d - \hat{\theta})^2]\left(1 + \frac{L_4}{8L_2^2} - \frac{5}{24}\frac{L_3^2}{L_2^3}\right) + 2(d - \hat{\theta})\left[\frac{L_3}{6L_2^2}\right] + \frac{1}{L_2}.$$

Differentiation with respect to d gives the optimum d, $\hat{d}$, as the solution of

$$-2(\hat{d} - \hat{\theta})[1 + O(n^{-1})] + \frac{L_3}{3L_2^2} = 0, \tag{2.21}$$

hence,

$$\hat{d} = \hat{\theta} + \frac{L_3(\hat{\theta})}{6L_2^2(\hat{\theta})}. \tag{2.22}$$

In the sense of having smaller mean-square error this estimate is to be preferred in large samples to the maximum likelihood estimate. The correction term takes account of the skewness in the likelihood function. The result may be compared with similar corrections for the same purpose considered by Bartlett [1].

3. Large-sample problems of the hypothesis testing type

There are problems in which it does not seem reasonable to take a prior distribution which is "smooth", or in technical language which is dominated by Lebesgue measure. These problems are those usually referred to as problems of testing a null hypothesis. The null hypothesis is usually regarded as occupying some special position and to a Bayesian this could be reflected in a concentration of prior probability on the null values of the parameters.

A significance test is first formulated in decision-theoretic language. The null hypothesis is a subset of the parameter space and as most significance tests are tests of the null hypothesis that some or all of the parameters take given values, we shall consider only this case. Furthermore, to simplify the notation we take $k = 2$, write $\theta_1 = \theta$, $\theta_2 = \phi$ and suppose the null hypothesis is that $\theta = 0$. The extension to a general number of parameters will be obvious. ϕ is a nuisance parameter. The decision space contains only two elements, d_0 and d_1, which are to be interpreted as decisions to accept and reject the null hypothesis respectively. In order to express this interpretation mathematically some assumptions must be made about the utility function. Clearly these must be that

$$\begin{aligned} &U(d_0; 0, \phi) > U(d_1; 0, \phi), \\ &U(d_0; \theta, \phi) < U(d_1; \theta, \phi) \qquad \text{for} \quad \theta \neq 0. \end{aligned} \tag{3.1}$$

The choice of decision function is determined by the maximum of (2.2) and the optimum function is not affected by subtracting any function of $\boldsymbol{\theta}$ from $U(d, \boldsymbol{\theta})$. In the testing situation we subtract $U(d_1; 0, \phi)$ from the utility when $\theta = 0$ and $U(d_0; \theta, \phi)$ when $\theta \neq 0$. The effect will be to replace the original utility function by another (for which the same notation is used) where the inequalities (3.1) are replaced by

$$\begin{aligned} &U(d_0; \theta, \phi) = 0, \qquad \text{for} \quad \theta \neq 0 \\ &U(d_1; 0, \phi) = 0 \end{aligned} \tag{3.2}$$

and otherwise

$$U(d_i; \theta, \phi) > 0. \tag{3.3}$$

This is equivalent to using "regret" instead of "loss" in the usual treatment. It will not affect the choice of d but it will affect the expected utility by a constant amount, and so could affect the choice of experiment.

In the Neyman-Pearson theory of tests, the probability of error of the first kind is considered on a different level from that of the second kind. This suggests that the null hypothesis is on a different level from the alternative. This can be interpreted by a Bayesian as meaning that there is a concentration of prior probability on the null values of the parameter. (It could also be interpreted as meaning that the utilities for the two decisions are radically different. Because of the way the utilities and prior probabilities combine, equation (2.8), this would amount mathematically to the same results.) We shall therefore follow Jeffreys [6] and introduce such a prior distribution. Specifically: along the line $\theta = 0$ there is a density $\pi_0(\phi)$ with respect to Lebesgue measure on this line, over the remainder of Θ there is a density $\pi_1(\theta, \phi)$ with respect to Lebesgue measure (area) over Θ. The ratio of the integrals

$$\frac{\int d\phi\ \pi_0(\phi)}{\iint d\theta\, d\phi\ \pi_1(\theta, \phi)}, \tag{3.4}$$

over the line $\theta = 0$ and over Θ respectively, gives the prior odds in favor of the null hypothesis. (Necessarily the sum of the two integrals is one; the distribution is normed.)

The expressions for the integrals (2.2) are, for the two decisions,

$$\begin{aligned} \lambda(d_0, x) &= \int d\phi\ U(d_0; 0, \phi)p(x|0, \phi)\pi_0(\phi), \\ \lambda(d_1, x) &= \iint d\theta\, d\phi\ U(d_1; \theta, \phi)p(x|\theta, \phi)\pi_1(\theta, \phi). \end{aligned} \tag{3.5}$$

The former, for example, follows from the fact that $U(d_0; \theta, \phi) = 0$ off the line $\theta = 0$, equation (3.2). Both of these integrals may now be approximated, in the case of independent, identically distributed observations, by the asymptotic results above. If

$$\begin{aligned} w_0(\phi) &= U(d_0; 0, \phi)\pi_0(\phi), \\ w_1(\theta, \phi) &= U(d_1; \theta, \phi)\pi_1(\theta, \phi), \end{aligned} \tag{3.6}$$

we have from (2.12)

$$\lambda(d_0, x^{(n)}) \sim \sqrt{2\pi}\ w_0(\tilde{\phi}^{(n)})p^{(n)}(x^{(n)}|0, \tilde{\phi}^{(n)})[nc_{22}(0, \tilde{\phi}^{(n)})]^{-1/2} \tag{3.7}$$

and

$$\lambda(d_1, x^{(n)}) \sim 2\pi w_1(\hat{\theta}^{(n)}, \hat{\phi}^{(n)})p^{(n)}(x^{(n)}|\hat{\theta}^{(n)}, \hat{\phi}^{(n)})\ \Delta_n^{-1/2}, \tag{3.8}$$

where $\tilde{\phi}^{(n)}$ is the maximum likelihood value of ϕ based on $x^{(n)}$ and assuming $\theta = 0$, and the rest of the notation is as before, remembering θ_1, θ_2 in (2.10) and (2.12) are now θ, ϕ.

In deriving these results it is assumed that $w_0(\phi)$ and $w_1(\theta, \phi)$ are bounded away from zero near the maximum likelihood values. This may not be true for

$w_1(\theta, \phi)$ as $\theta \to 0$. If not then the difficulty may be avoided by returning to the original utility function.

The null hypothesis is accepted iff $\lambda(d_0, x^{(n)}) > \lambda(d_1, x^{(n)})$, that is, iff

$$\frac{p^{(n)}(x^{(n)}|0, \tilde{\phi}^{(n)})}{p^{(n)}(x^{(n)}|\hat{\theta}^{(n)}, \hat{\phi}^{(n)})} > \frac{w_1(\hat{\theta}^{(n)}, \hat{\phi}^{(n)})}{w_0(\tilde{\phi}^{(n)})} \left[\frac{2\pi n c_{22}(0, \tilde{\phi}^{(n)})}{\Delta_n}\right]^{1/2}. \tag{3.9}$$

Write a_n for the right side of this inequality. Then, apart from the way in which a_n is chosen, the test is easily seen to be the usual likelihood ratio test of the hypothesis $\theta = 0$: for the left side is the ratio of the maximum likelihood under the null, to the maximum likelihood under the alternative hypothesis. It is clear that this identification with the likelihood ratio test will persist whatever be the dimensions of the parameter space and null hypothesis.

We have therefore the following large-sample optimum decision function: *If the prior probability has a density of the above form and if the likelihood function behaves respectably, then carry out a likelihood ratio test.*

As in the estimation case we have a procedure which is similar to that advocated by standard statistical practice; for it is known (Wald [16], theorem VIII) that the likelihood ratio test has optimum properties. The difference is apparent when the choice of a_n is considered. Usually a_n is determined (if possible) from the requirement that the probability of rejection of the null hypothesis, when it is true, be some preassigned small number α, for all ϕ. (It is known that this is asymptotically possible from the known asymptotic distribution of the likelihood ratio on the null hypothesis in large samples: Wilks [17].) In the Bayesian treatment a_n is determined from the utilities and prior probabilities at the maximum likelihood values, the information matrices evaluated there and, most important, from the sample size. As in the estimation case, the usual determination of a_n is unsound in principle.

Improvements on the test (3.9) can be obtained by considering the next terms in the asymptotic expansions of $\lambda(d_0, x^{(n)})$ and $\lambda(d_1, x^{(n)})$ on the lines used to derive (2.18). The situation becomes somewhat complex, especially in dealing with $\lambda(d_1, x^{(n)})$ where two parameters are involved, and the details of this approach will not be discussed here.

Instead we propose to study the choice of a_n and the effect this choice has on the significance level, in order, thereby, to relate the Bayesian approach to the usual one. Denote by $\Lambda_n(x^{(n)})$ the likelihood ratio criterion, the left side of (3.9). Then the null hypothesis is accepted iff

$$-2 \log \Lambda_n(x^{(n)}) < -\log \left\{\frac{w_1^2(\hat{\theta}^{(n)}, \hat{\phi}^{(n)})}{w_0^2(\tilde{\phi}^{(n)})} \frac{2\pi c_{22}(0, \tilde{\phi}^{(n)})}{n^{-2}\, \Delta_n}\right\} + \log n. \tag{3.10}$$

Now it is known (Wald [16], theorem IX) that in large samples the left side is distributed as the square of a normal random variable with unit variance and mean

$$\sqrt{n}\, \mu(\theta_0, \phi_0) = \sqrt{n}\, \theta_0 \left[\bar{c}_{11}(\theta_0, \phi_0) - \frac{\bar{c}_{12}^2(\theta_0, \phi_0)}{\bar{c}_{22}(\theta_0, \phi_0)}\right]^{1/2}. \tag{3.11}$$

In particular when the null hypothesis obtains ($\theta_0 = 0$) it is distributed as χ^2 on one degree of freedom, and we consider this case. The left side of (3.10) is therefore $O(1)$ and the expression in braces on the right side tends to a finite limit as $n \rightarrow \infty$ and differs from this limit by terms $O(1/\sqrt{n})$. Hence we may asymptotically replace that expression by its limit, obtained by replacing $\hat{\theta}^{(n)}$ by $\theta_0 = 0$ and $\hat{\phi}^{(n)}$ and $\tilde{\phi}^{(n)}$ by ϕ_0. Here $w_1(0, \phi_0)$ is to be interpreted as $\lim_{\theta \rightarrow 0} w_1(\theta, \phi)$ and is assumed greater than zero and finite. That is, there is a definite gain in utility by taking d_1 instead of d_0 by however little the alternative hypothesis differs from the null hypothesis. The case of a zero limit requires special treatment, the significance level being much smaller, indeed of a different order, from the expression below. Similar remarks apply if the limit is infinite. With this assumption we can write (3.10) as

$$x^2 < A(\phi_0) + \log n \tag{3.12}$$

with $A(\phi_0)$ finite and x a random variable which is $N(0, 1)$.

If the usual significance level argument had been used (3.12) would have been replaced by $x^2 < A_1$, so that the primary difference between the two methods lies in the addition of the term in $\log n$. The effect of this term is to make the significance level decrease with increasing sample size. This extra term has been noted before by Jeffreys (section 5.2 in [6]) and Lindley [9]. The significance level, using (3.12), is

$$2 \int_{[A(\phi_0) + \log n]^{1/2}}^{\infty} \sqrt{2\pi}\, e^{-t^2/2}\, dt. \tag{3.13}$$

Denote by I_n the integral $\int_{\mu_n}^{\infty} e^{-t^2/2}\, dt$. Integration by parts gives

$$I_n = -\mu_n e^{-\mu_n^2/2} + \int_{\mu_n}^{\infty} e^{-t^2/2} t^2\, dt > -\mu_n e^{-\mu_n^2/2} + \mu_n^2 I_n. \tag{3.14}$$

A different integration by parts gives

$$I_n = \mu_n^{-1} e^{-\mu_n^2/2} - \int_{\mu_n}^{\infty} e^{-t^2/2} t^{-2}\, dt > \mu_n^{-1} e^{-\mu_n^2/2} - \mu_n^{-2} I_n, \tag{3.15}$$

(3.14) and (3.15) combine to give

$$\frac{\mu_n e^{-\mu_n^2/2}}{\mu_n^2 + 1} < I_n < \frac{\mu_n e^{-\mu_n^2/2}}{\mu_n^2 - 1}, \tag{3.16}$$

so that if $\mu_n \rightarrow \infty$, $I_n \sim \mu_n^{-1} e^{-\mu_n^2/2}$. With the value of $\mu_n = [A(\phi_0) + \log n]^{1/2}$, the significance level, (3.13), is asymptotically

$$\left[\frac{\pi e^{A(\phi_0)}}{2} n \log n\right]^{-1/2}, \tag{3.17}$$

whereas without the logarithm term it is constant. The decrease in significance level is achieved at some expense in loss of power. Without the logarithm term the power reaches a constant value, say 1/2, if $\mu(\theta_0, \phi_0)$ is of the order $1/\sqrt{n}$, whereas, using (3.10), $\mu(\theta_0, \phi_0)$ will have to be of order $(\log n)^{1/2}/\sqrt{n}$ to achieve the same power.

The arguments of this section extend, without difficulty, to the case of a finite number of decisions, such as occur in many analyses of variance. The value of $\lambda(d_i, x^{(n)})$ for the different d_i may be replaced by the approximations.

4. General remarks on the small-sample problem

The large-sample theory has been outlined and seen, naturally enough, to be insensitive to the form of the prior distribution. For small samples the same situation does not obtain and the Bayesian argument is usually criticized on the grounds of this dependence on the prior distribution. We delay until section 5 a discussion of the form of the prior distribution and here make some remarks about the general character of Bayesian inferences and decisions which are relevant in small samples and yet are independent of the precise form of the prior distribution.

There seem to me to be two processes at work: one is that by which a prior distribution is changed by observation into a posterior distribution, and this process I call *inference*, and it is statistical if the observations are statistical. The other is the use made of the posterior distribution to come to decisions, which is called *decision theory*. It seems useful to distinguish these two because many decisions can be made on the basis of a single set of observations and the common part to all these decision processes is the posterior distribution. One need only glance at almost any statistical textbook to see an example of inference problems (though not usually tackled by Bayesian arguments) in which the purpose of the example is to make some concise summary of what the data have to say about some parameter or parameters. One might, for example, make an interval estimate of the effect of some drug without discussing any decision as to whether or not the drug should be used. On the other hand the decision problem uses only the posterior distribution, and once this is known, can be solved without further reference to the observations.

Thus in either inference or decision problems it is the posterior distribution which plays the central role. Inferences, like interval estimates, can be made in terms of it. Decisions can be made by maximizing the expected utility with respect to it [equation (2.5)]. It therefore follows that the inferences or decisions made depend only on the observations through the likelihood function $p(x|\boldsymbol{\theta})$, a function of $\boldsymbol{\theta}$ for fixed x. Consequently the other observations in the sample space are irrelevant. If different observations have the same likelihood function then the inferences or decisions made should be the same. This point has been made before, perhaps first by Jeffreys, see page 360 of [6], and an example in the binomial sampling situation is given in [11]. But it is worth repeating because, simple though the point is, it has important consequences which are at variance with current statistical practice.

Consider first the concept of a significance test. The significance level plays a fundamental role and is obtained by integration over part of sample space and therefore involves other possible observations than the one obtained. The use of the significance level therefore violates the principle that inferences or decisions should depend only on the likelihood function, and is at variance with the Bayesian approach. The test in the Bayesian approach, as we have seen in section 3, determines the critical value quite differently. So far as I am aware no one has explained why the fact that on a hypothesis H_0 the probability of the given observation and others more extreme than it is small, should lead one to reject H_0. Some explanation can be given if it is known that on H_1, some other hypothesis, the probability of the same event is large. But how often is the power calculated in making a test? In any case it is not reasonable to consider such probabilities because they depend on the sample space and the sampling rule used. It matters whether the observations were come by sequentially or by a fixed sample size in ordinary testing, but to a Bayesian such information is irrelevant.

Furthermore there are situations in which it is not clear what the sample space is. The clearest example is due to Cox [4]. A penny is tossed: if it comes down heads an observation x is obtained which is $N(\mu, \sigma_1^2)$; if tails an observation x is obtained which is $N(\mu, \sigma_2^2)$. The result of the toss, x, σ_1^2 and σ_2^2 are known, μ is not and $\sigma_1^2 \gg \sigma_2^2$. Does the sample space consist of two real lines or of the one real line corresponding to the result of the toss of the coin? The former gives the better test in the usual sense but Cox argues in favor of the latter. More familiar but more complicated examples are provided by regression problems (should the sample space be confined to samples having the same values of the independent random variable?), contingency tables (should the margins be held fixed?) and certain confidence interval situations. In the last example I have in mind cases where the interval covers the whole line. To a Bayesian only the likelihood function is relevant, though, as we shall see in the next section, whether or not only part of the likelihood is relevant is an open question. In the classical treatment many of these problems have been solved by Lehmann and Scheffé [8] with the concept of completeness.

Confidence intervals and unbiased estimates are open to the same criticism, because they both involve an integration over sample space. Maximum likelihood estimation is free from this criticism because it obviously uses only the likelihood function, but the use of the inverse of the expectation of minus the second derivative to obtain the variance of the estimate uses integration over sample space. But here the criticism is little more than a quibble because the difference between the expectation and the actual value of the derivative is small. In small samples, if maximum likelihood could there be justified, the criticism could have more content.

We repeat that the remarks on the irrelevance of the sample space do not, of course, apply to the *design* of experiments. Before x has been observed the average utility (2.1) is relevant, after x has been observed the average utility, conditional on that x, (2.5), is relevant.

Not much of statistics seems to remain free from this Bayesian criticism, and it might appear that most of modern statistics is unsound. But fortunately this is not so. Most of the common statistical methods are unaffected by changing from the usual to the Bayesian approach. An example will illustrate the point. Student's t-distribution gives the distribution of $(\bar{x} - \mu)\sqrt{n}/s$ (the notation is surely well established) in random samples of size n from a normal distribution and provides a basis for the confidence interval statement

$$p\left(\bar{x} - \frac{t_\alpha s}{\sqrt{n}} < \mu < \bar{x} + \frac{t_\alpha s}{\sqrt{n}}\middle|\mu\right) = 1 - \alpha \tag{4.1}$$

for all μ, the probability referring to $\bar{x}$ and s. Under reasonable assumptions on the prior distribution of μ and σ the posterior distribution of μ is such that $(\bar{x} - \mu)\sqrt{n}/s$ has Student's t-distribution on $n - 1$ degrees of freedom (note that μ is the random variable here, $\bar{x}$ and s are fixed). Consequently

$$\pi\left(\bar{x} - \frac{t_\alpha s}{\sqrt{n}} < \mu < \bar{x} + \frac{t_\alpha s}{\sqrt{n}}\middle|\bar{x}, s\right) = 1 - \alpha. \tag{4.2}$$

To a practical man (4.1) and (4.2) probably mean the same thing. Thus a Bayesian will agree with standard practice. The similarity arises because of the intimate relationships between operations in sample space and parameter space that arise when dealing with the normal law. A treatment of many statistical situations using Bayesian methods is given by Jeffreys [6]. Interestingly the Fisher-Behrens test and the exact test of Fisher's for the 2 × 2 contingency table can be justified by such arguments, though the latter involves a slight approximation.

5. Choice of the prior distribution

Because they avoid integrations over sample space and the resulting distributional problems, Bayesian methods are often easier to apply than the usual methods, but they do involve the serious difficulty of the choice of the prior distribution. Some Bayesian writers, for example Savage [15], advocate a subjective point of view and consider that the prior distribution reflects only the beliefs of the person making the inference or decision, and that therefore no prior distribution is more correct than any other. The value of an objective theory would be so great that it seems to the present writer well worth trying to see whether some objective and natural choice of a prior distribution is possible. To do this we consider some indirect and direct attacks on the problem.

Some clues on the nature of a prior distribution can be obtained from statements that are commonly made. We deal with just one such type of statement here, namely where it is said that such and such an observation gives no information about a parameter. For example it is often said that a sample of size one from a normal distribution of unknown mean and variance gives no information about the variance or that the margins of a 2 × 2 table give no information about association in the table. (We might note in passing that if a Bayesian could give a meaning to a parameter being unknown he would have the required objective prior distribution, that of ignorance.) To a Bayesian this presumably means that the prior and posterior distributions are identical. Now if there is just one real parameter θ it is easy to see that this can only happen if the likelihood function does not involve θ: a trivial case. The situation is more interesting however if there are two parameters θ and ϕ. In the notation of sections 2 and 3 the posterior distribution of θ, say, is

$$\int d\phi\, \pi(\theta, \phi|x) = \int d\phi \frac{p(x|\theta, \phi)\pi(\theta, \phi)}{p(x)} \tag{5.1}$$

and the prior distribution of θ is

$$\pi(\theta) = \int d\phi\, \pi(\theta, \phi). \tag{5.2}$$

So we can say that x gives no information about θ if (5.1) and (5.2) are equal; that is if

$$\int d\phi\, p(x|\theta, \phi)\pi(\phi|\theta), \tag{5.3}$$

where $\pi(\phi|\theta)$ is the conditional prior distribution of ϕ given θ, is a function of x only, namely $p(x)$. Hence the prior distribution should be chosen so that (5.3) is satisfied. I have been unable to make any substantial progress with this equation, but it is interesting to note that it gives a condition on the distribution of ϕ for fixed θ, only, and not on the distribution of θ. In the normal case a single observation gives no information if the mean is uniformly distributed over the whole real line independently of the variance. Such a distribution is unnormed but can be treated by the methods of Rényi [13].

If the observations can be written (x, y), so that $p(x, y|\theta, \phi) = p(y|x; \theta, \phi)p(x|\theta, \phi)$ and if $p(x|\theta, \phi)$ and $\pi(\phi|\theta)$ satisfy (5.3), then instead of the likelihood $p(x, y|\theta, \phi)$ we may use $p(y|x; \theta, \phi)$. For example, in the 2 × 2 contingency table in the form of two binomials where we are interested in testing the equality of the two proportions, if y refers to the interior of the table and x to the margins, and if

$$\theta = \frac{p_1q_1}{p_2q_2}, \qquad \phi = \frac{p_1p_2}{q_1q_2} \tag{5.4}$$

(so that we wish to test $\theta = 1$), then (5.3) can be satisfied, at least for small enough tables. In this case we also have the important fact that $p(y|x; \theta, \phi)$ does not involve ϕ. Hence provided only that $\pi(\phi|\theta)$ satisfies (5.3) the explicit form of it need not be considered and the posterior distribution of θ can be obtained from $p(y|x; \theta)$ without considering the nuisance parameter ϕ. Similar remarks generalize to densities of the exponential family.

Suppose that we have a density $p(x|\theta)$ dependent on a single real parameter θ. The following considerations suggest a way of obtaining an objective prior distribution. In scientific work we progress from a situation where θ is only vaguely known, by experimentation to one where it is fairly precisely known: but it is rarely exactly known. Usually there comes a point where it is known to within an amount $\delta(\theta)$, which may depend on θ. For example it may be known to six decimal places, $\delta(\theta) = 10^{-6}$, or to six significant figures, $\delta(\theta) = 10^{-6}\theta$. The function $\delta(\theta)$ is often known objectively or may be found from some decision problem. Hence we may effectively replace the continuous range of θ by discrete values of θ at mesh $\delta(\theta)$. Knowledge of θ means that it is known which point of the mesh obtains: therefore a possible meaning for ignorance is a distribution of probability which assigns equal probability to all points of the mesh: or returning to the continuous case a prior density, $\pi(\theta)$, proportional to $\delta^{-1}(\theta)$. In a sense such a density would correspond to ignorance of θ and could be used as such.

Apart, however, from location parameters (or parameters transformable thereto, such as scale parameters) there seems no reason for any particular $\delta(\theta)$ outside of a specific decision problem, and an argument for inference problems would be useful. Jeffreys has given one in section 3.9 of [6] and we now give an alternative justification for his prior distribution. Essentially we obtain the function $\delta(\theta)$ in terms of the density function $p(x|\theta)$. Knowledge of θ means knowing the form of the density for x and the amount by which one value of θ differs from $\theta + \delta(\theta)$ can be measured by how much $p(x|\theta)$ differs from $p[x|\theta + \delta(\theta)]$. The difference can be measured by how much information is available in the experiment for distinguishing between θ and $\theta + \delta(\theta)$, the more information the further they are apart. Information can be measured in a way suggested by the author [10]; namely the expected information is

$$(5.5)\qquad \int dx \left\{\int d\theta\, \pi(\theta|x) \log \pi(\theta|x)\right\} p(x) - \int d\theta\, \pi(\theta) \log \pi(\theta)$$
$$= \iint dx\, d\theta\, p(x|\theta)\pi(\theta) \log \left[\frac{p(x|\theta)}{p(x)}\right].$$

The justification for this has been given in the paper referred to: further discussion of the use of Shannon's measure has been provided by Rényi [14]. Now if we consider only θ and $\theta + \delta(\theta)$ and they have equal prior probabilities (representing ignorance), namely 1/2, 1/2 then as $\delta(\theta) \to 0$ it is easy to establish that (5.5) is

$$(5.6)\qquad 2\delta^2(\theta)\mathcal{E}\left\{\left[\frac{\partial \log p(\theta)}{\partial\theta}\right]^2\right\} = 2\delta^2(\theta)I(\theta),$$

to order $\delta^2(\theta)$, where $I(\theta)$ is Fisher's information function. Hence arguing as before a possible prior distribution is

$$(5.7)\qquad \pi(\theta) \propto I(\theta)^{1/2}$$

which is the one suggested by Jeffreys. He gives several examples of the use of it. Notice that in (2.12) $\Delta_n^{-1/2}$, in the case of a single parameter, is approximately $I(\hat{\theta}^{(n)})^{-1/2}$ and therefore if this prior distribution were used (2.12) could be written

$$(5.8)\qquad \lambda(d, x^{(n)}) \sim \sqrt{2\pi}\, U(d, \hat{\theta}^{(n)})p^{(n)}(x^{(n)}|\hat{\theta}^{(n)}).$$

When more than one parameter is involved the above argument seems less satisfactory because no natural mesh suggests itself and Jeffreys extension, using the determinant of Fisher's information function in place of $I(\theta)$, produces ridiculous answers, for example with the t-distribution. A way out of the difficulty when there are two parameters θ and ϕ, and ϕ is a nuisance parameter is to derive $\pi(\phi|\theta)$ for each θ by the above reasoning. This then enables one to calculate $p(x|\theta) = \int d\phi\, p(x|\theta, \phi)\pi(\phi|\theta)$ and then find $\pi(\theta)$ in the same way from $p(x|\theta)$. This removes the difficulty in connection with the t-distribution, but it remains to investigate it in more detail.

REFERENCES

[1] M. S. Bartlett, "Approximate confidence intervals," *Biometrika*, Vol. 40 (1953), pp. 12–19, 306–317; Vol. 42 (1955), pp. 201–204.

[2] H. Chernoff, "Sequential design of experiments," *Ann. Math. Statist.*, Vol. 30 (1959), pp. 755–770.

[3] H. Chernoff and L. E. Moses, *Elementary Decision Theory*, New York, Wiley, 1959.

[4] D. R. Cox, "Some problems connected with statistical inference," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 357–372.

[5] R. A. Fisher, "Theory of statistical estimation," *Proc. Cambridge Philos. Soc.*, Vol. 22 (1925), pp. 700–725.

[6] H. Jeffreys, *Theory of Probability*, Oxford, Clarendon Press, 1948 (2nd ed.).

[7] L. Le Cam, "On the asymptotic theory of estimation and testing hypotheses," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1956, Vol. 1, pp. 129–156.

[8] E. L. Lehmann and H. Scheffé, "Completeness, similar regions and unbiased estimation. Part II," *Sankhyā*, Vol. 15 (1955), pp. 219–236.

[9] D. V. Lindley, "Statistical inference," *J. Roy. Statist. Soc., Ser. B*, Vol. 15 (1953), pp. 30–76.

[10] ———, "On a measure of the information provided by an experiment," *Ann. Math. Statist.*, Vol. 27 (1956), pp. 986–1005.

[11] ———, "A survey of the foundations of statistics," *Appl. Statist.*, Vol. 7 (1958), pp. 186–198.

[12] J. Neyman, "Contribution to the theory of the χ^2-test," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1949, pp. 239–273.

[13] A. Rényi, "On a new axiomatic theory of probability," *Acta Math. Acad. Sci. Hungar.*, Vol. 6 (1955), pp. 285–335.

[14] ———, "On the notion of entropy and its role in probability theory," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. 1, pp. 547–561.

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

[16] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Amer. Math. Soc.*, Vol. 54 (1943), pp. 426–482.

[17] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, Vol. 9 (1938), pp. 60-62.