THE ASYMPTOTICALLY UNBIASED PRIOR DISTRIBUTION[1]

By J. A. Hartigan

Princeton University

1. Summary. In estimation of a real valued parameter θ, using observations from the probability density $f(x \mid \theta)$, and using loss function $L(\theta, \phi)$, the prior density which minimizes asymptotic bias of the associated estimator is shown to be

$$J(\theta) = \mathcal{E}((\partial/\partial\theta) \log f)^2/[(\partial^2/\partial\phi^2)L(\theta, \phi)]^{\frac{1}{2}}_{\phi=\theta}.$$

Results are also given for estimation in higher dimensions.

2. Introduction. In order to find out about a parameter θ, an observation x is made, which is connected to θ by having a probability density $f(x \mid \theta)$ which varies as θ varies. An attractive way of reporting the information about θ in x specifies the conditional probability distribution of θ given x. If our prior knowledge of θ gives rise to a *prior density* $h(\theta)$, then the *posterior density* of θ given x is

$$g(\theta \mid x) = f(x \mid \theta)h(\theta)/\int f(x \mid \theta)h(\theta)\, d\theta,$$

by Bayes' theorem.

We will be concerned with determining reasonable prior densities when there is no prior information; these are called *prior densities on ignorance*. Prior distributions on ignorance are technically important because they offer a reasonably objective route to prior distributions in general—we suppose that our actual prior information is equivalent to an observation y with probability density $g(y \mid \theta)$; starting from a prior density on ignorance we calculate the posterior density of θ given y. This posterior density is now the prior density representing prior information. Some previous suggestions for determining prior distributions on ignorance follow.

(i) *Bayes' prior.* Early followers of Bayes assumed (following Bayes' procedure for the binomial) that ignorance could be represented by a uniform distribution over the parameter space. However, the parameter space is essentially defined only up to 1–1 transformations [its only function is to index the probability distributions on x], and a uniform distribution on different versions of the parameter space yields contradictory posteriors; thus the rule of a uniform prior distribution on ignorance is inadequate.

(ii) *Jeffreys' prior.* Sir Harold Jeffreys [8] in the 1930's began a breakaway from the uniform tradition; in 1946 [9] he suggested a "bootstrap" technique whereby the prior h is based on properties of the family of "genuine" probability densities $f(x \mid \theta)$. It should be noted at this point that it is mathematically con-

Received 21 December 1964.

[1] Work supported partly by the Office of Naval Research and partly by the Atlas Computer Laboratory, Harwell, England.

venient to admit as possible priors, densities with infinite total measure. For θ 1-dimensional, let $f_p = (\partial/\partial\theta)^p \log f(x \mid \theta)$; then Jeffreys' prior is $J_{\frac{1}{2}}(\theta) = I^{\frac{1}{2}}(\theta) = \{\mathcal{E}(f_1^2)\}^{\frac{1}{2}}$ where $I(\theta)$ is Fisher's information (and $\mathcal{E}$ denotes expectation over the sample space). A family of priors similar to Jeffreys' consists of the densities $J_\alpha(\theta)$ where

$$(\partial/\partial\theta) \log J_\alpha(\theta) = [\mathcal{E}(f_1 f_2) + \alpha\mathcal{E}(f_1^3)]/\mathcal{E}(f_1^2).$$

This family, suggested by the author in [7], satisfies the minimum requirements of a prior density on ignorance—invariance, under 1–1 transformations of the sample space and parameter space, under replication of the sample space, and under restriction of the parameter space. Thus for the binomial, $J_\alpha = (pq)^{\alpha-1}$; for normal location, $J_\alpha = 1$; for normal scale, $J_\alpha = \sigma^{-3+4\alpha}$; for scale and location, $J_\alpha = \sigma^{-5+6\alpha}$ [using usual terminology]. Huzurbazar (so Jeffreys reports in the 3rd edition of his book) has suggested $\alpha = 0$ and also $\alpha = 1$ for the exponential family; in general most prior densities seem to be about J_α, $0 \leqq \alpha \leqq 1$.

(iii) *Decision theoretic priors.* More generally let us consider the whole decision theoretic apparatus—densities $f(x \mid \theta)$, *loss function* $L(d, \theta)$ (associated with a space D of *decisions* d), further criteria C. Corresponding to each prior density $h(\theta)$, there is a Bayes' decision procedure which, given the observation x, makes the decision which minimizes the average loss with respect to the posterior distribution [i.e. $\int L(d, \theta)g(\theta \mid x)\, d\theta$]; we now select that prior density (possibly non-unique) whose decision procedure is best by the criteria C. Such a "coat-tail" technique of selection of priors is suggested by Wald's complete class theorem—under regularity conditions, any admissible decision procedure is a Bayes' procedure or a limit of Bayes' procedures. Thus any criteria C which are used to select an admissible decision procedure, also specify associated prior distributions.

Welch and Peers [18] have recently shown that Jeffreys' prior distribution, $J_{\frac{1}{2}}$, generates "good" asymptotic confidence regions—suppose that θ is 1-dimensional, and x is a continuous random variable. Let $\theta(\alpha, \mathbf{x})$ be such that $\theta < \theta(\alpha, \mathbf{x})$ with Bayesian probability α according to the posterior distribution of θ given $x_1, x_2, \cdots x_n$; under regularity conditions, the region $[\theta \mid \theta < \theta(\alpha, \mathbf{x})]$ is of confidence size $\alpha + O(n^{-\frac{1}{2}})$, for any smooth prior density; for Jeffreys' prior $J_{\frac{1}{2}}$, the region is of confidence size $\alpha + O(n^{-1})$. Thus, asymptotically, regions generated from Jeffreys' prior are closer to being confidence regions of size α, than those of any other prior. The result does not hold for discrete distributions as pointed out by Welch in the discussion of Thatcher's paper [16].

Perks [14] first suggested that the prior distribution should be based on the asymptotic volume of confidence regions; if θ_0 is the true value, the efficient confidence region about θ_0 is of volume proportional to $J_{\frac{1}{2}}^{-1}(\theta_0)$. Thus, with prior density $J_{\frac{1}{2}}(\theta_0)$, the prior probability of efficient confidence regions is asymptotically constant, as the true value θ_0 varies. Lindley [13] has used a similar argument to justify $J_{\frac{1}{2}}$ for θ 1-dimensional, based on a definition of information he discussed in [12].

The function of a prior distribution on ignorance, to treat all parameter values equally, is performed in decision theory by invariance and unbiasedness criteria. Invariance theory shows that Bayes' decision procedures are invariant if the prior density $h(\theta)$ is *relatively invariant* under transformations T leaving the decision problem invariant; more precisely, suppose that there is a transformation T mapping x onto Tx and a corresponding transformation (also called T) mapping θ onto $T\theta$, such that the distribution of Tx given $T\theta$ is the same as the distribution of x given θ; then we need that the posterior distribution of $T\theta$ given Tx be the same as the posterior distribution of θ given x, which means $h(T\theta)(dT\theta/d\theta) = C_T h(\theta)$ for all $\theta \varepsilon \Theta$, the parameter space. We say that h is *relatively invariant* under T. (See Hartigan [7].) Barnard first suggested *left* invariant priors in [2] and Fraser showed that *right* invariant priors generated fiducial distributions in [5]. Specific results are $h(\mu) = 1$ for normal location and $h(\mu, \sigma) = \sigma^\alpha$ for location and scale; the method is not generally applicable because there aren't many variant problems.

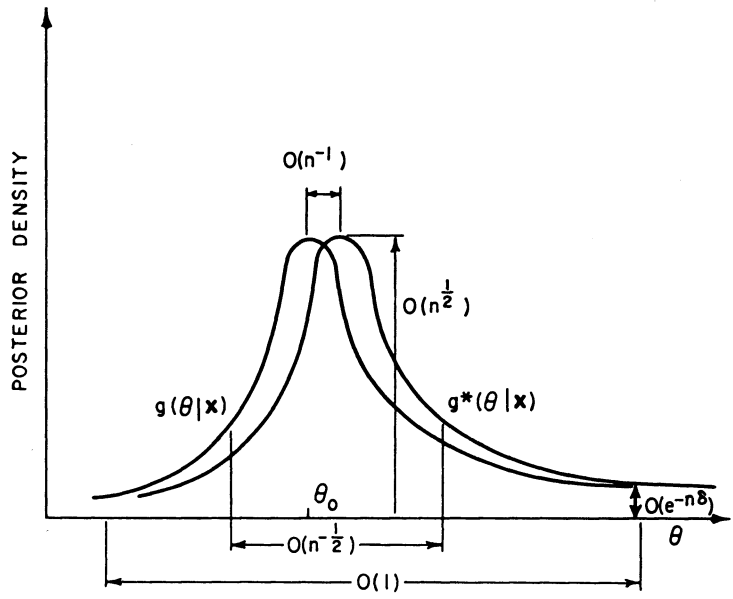
We will consider priors chosen to give *unbiasedness*; if we make the decision $d(\mathbf{x})$ on the basis of the n observations $x_1, \cdots, x_n$, and if the true value of the parameter is θ, the loss is $L(d(\mathbf{x}), \theta)$. Unbiasedness requires that in some average sense (the average being over the possible observations $\mathbf{x}$), $L(d(\mathbf{x}), \theta) > L(d(\mathbf{x}), \theta_0)$ if θ_0 is the true value of the parameter, and θ is some other value. Thus if

$$\mathcal{E}(L(d(\mathbf{x}), \theta) \mid \theta_0) \geqq \mathcal{E}(L(d(\mathbf{x}), \theta_0) \mid \theta_0) \quad \text{all} \quad \theta, \theta_0,$$

we say $d(\mathbf{x})$ is unbiased *in the mean*; we will consider only this type of unbiasedness. Let us consider in particular estimation; here the decision space D coincides with the parameter space Θ; the loss function $L(\phi, \theta)$ is the loss if ϕ is estimated as the true value when θ is the true value; the Bayes estimator associated with the prior distribution h is $\hat{\theta}_h$ minimizing $\int L(\hat{\theta}_h, \theta) g(\theta \mid \mathbf{x})\, d\theta$; we show that for θ 1-dimensional, $\hat{\theta}_h$ is asymptotically unbiased in the mean, if and only if

$$h(\theta) = J_{\frac{1}{2}}^2 / [\partial^2 L(\theta, \phi)/\partial \phi^2]_{\phi=\theta}^{\frac{1}{2}}.$$

Thus, for the squared distance loss function, $h = J_{\frac{1}{2}}^2$, and for Jeffreys' loss function, [9], $h = J_{\frac{1}{2}}$, are the "mean unbiased" prior distributions. The proof of the result requires a host of regularity conditions, of the type used in validating maximum likelihood; principally, we require "boundedness" with respect to the x variable, and smoothness with respect to the θ variable. A good general exposition of asymptotic approximations is contained in Wallace [17]; Cramér [4] contains many of the mathematical details. Investigations of the asymptotic *consistency* and *normality* of posterior distributions have been carried out by LeCam [10], [11], Lindley [13], and Freedman [6]; these are properties which hold for all smooth prior distributions. In order to select prior distributions on the basis of asymptotic behavior of the posterior distribution, more accurate knowledge is required; it is summarized below and in the figure.



ASYMPTOTIC BEHAVIOUR OF POSTERIOR DISTRIBUTIONS

Asymptotic properties of the posterior distribution. First order results, true for any prior density, show that $g(\theta \mid x_1, \cdots x_n)$ is $O(n^{\frac{1}{2}})$ within $O(n^{-\frac{1}{2}})$ of the true value θ_0 and $O(e^{-n\delta})$ outside $O(1)$ of θ_0. If $\hat{\theta}$ is the maximum likelihood estimator, the posterior density of θ given $x_1, \cdots x_n$ is such that θ is asymptotically normal with mean $\hat{\theta} + O(n^{-1})$ and variance $I^{-1}(\theta_0) = [\mathcal{E}(f_1^2)]^{-1} = O(n^{-1})$. Higher order results show that a single observation causes a change of $O(1)$ in g near θ_0; a change in the prior density h also causes a change of $O(1)$ in g; if $h_1 = [(\partial/\partial\theta) \log h(\theta)]_{\theta=\theta_0}$, $\theta + h_1/I$ has a density which deviates from a normal density by terms of $O(1)$, but which changes as the prior changes by terms of $O(n^{-\frac{1}{2}})$ (negligible compared to the effect of one observation). The essential effect of a change in prior distributions is thus a change in location of $O(n^{-1})$. Now suppose $\hat{\theta}_h$ is the Bayesian estimator of θ_0 based on the prior density h; for any other prior h^*, $\hat{\theta}_{h^*} = \hat{\theta}_h + (h_1 - h_1^*)/I + O(n^{-\frac{3}{2}})$; correct choice of h should ensure that $\hat{\theta}_h$ is unbiased (in any reasonable sense we care to specify) to $O(n^{-\frac{3}{2}})$.

3. Terminology and regularity conditions. We start with a sample space $\mathfrak{X}$, a Borel field B of subsets of $\mathfrak{X}$, a family of probability distributions P_θ, $\theta \in \Theta$. We suppose that the family is dominated so that for all $\theta \in \Theta$ P_θ has a *probability density* $f(x \mid \theta)$ with respect to some measure ν, say. The *expectation*, $\mathcal{E}(m(x) \mid \theta)$, of a function $m(x)$ given θ is defined by $\mathcal{E}(m(x) \mid \theta) = \int m(x)\, dP_\theta(x)$. Let μ be some measure on the parameter space Θ; if a measure on the parameter space has density h with respect to μ, we define

$$g(\theta \mid x) = f(x \mid \theta)h(\theta)/\int f(x \mid \theta)h(\theta)\, d\mu(\theta)$$

to be the density (w.r.t. μ) of the *posterior distribution of* θ *given* x, when the *prior density* is h. If θ was a random variable with probability density $h(\theta)$, and $f(x \mid \theta)$ the conditional probability density of x given θ, then $g(\theta \mid x)$ would be the conditional probability density of θ given x. We wish to consider prior measures which are not probability measures, firstly, in the practical sense that no frequency interpretation will be given to them, and secondly, in the mathematical sense that the measures need not assign measure 1 to the whole parameter space, but may even be unbounded. (Posterior measures in contrast, do give measure 1 to the parameter space.) Now suppose that it is required to choose a decision $d \in D$, some *space of decisions*, which is connected to the observation x, through the parameter space Θ, by a real valued *loss function* $L(d, \theta)$ on $D \times \Theta$. $L(d, \theta)$ is the loss if the decision d is made when the true value of the parameter is θ. A *Bayes decision* $d_h(x)$, given the observation x, with respect to a prior density h, minimizes the average loss over the posterior distribution $\int L(d, \theta)g(\theta \mid x)\, d\mu(\theta)$. We may consider more general *decision functions* δ, functions from $\mathfrak{X}$ to D, which state for each possible observation x which decision will be taken; before the observations are made, a decision function might be evaluated by its *average loss*, $L\mathfrak{X}(\delta, \theta) = \mathcal{E}(L(\delta(x), \theta) \mid \theta)$; one criterion for choosing among decision functions is that of *unbiasedness*—a decision function δ is *unbiased* if

$$\mathcal{E}(L(\delta(x), \phi) \mid \theta) \geqq \mathcal{E}(L(\delta(x), \theta) \mid \theta)$$

for all $\theta, \phi \in \Theta$. If $L(d, \theta)$ is interpreted as a measure of incompatibility of d and θ, an unbiased δ is such that, when θ is true, $\delta(x)$ is, on the average, more compatible with θ than with any other value of the parameter. Let us define the *bias of* δ *at* θ by $B(\delta, \theta) = \sup_{\phi \in \Theta} \mathcal{E}\{(L(\delta(x), \theta) - L(\delta(x), \phi)) \mid \theta\}$.

We will be concerned in what follows to investigate the bias of the Bayes decision functions $d_h(x)$; in particular, to ask if a prior distribution h can be found which is of minimum bias for all θ. We shall show that when the decision problem is estimation of a real valued parameter, and under regularity conditions on f, h, and L, there exists a prior distribution of minimum asymptotic bias.

The regularity conditions of f are somewhat stronger than those required to validate maximum likelihood. In the following, all probability statements and expectations are with respect to $f(x \mid \theta_0)$.

$f(x \mid \theta)$, $\theta \in \Theta$, is *regular* at θ_0 if

(F1). Θ is a closed subset of the real line of which θ_0 is an interior point

(F2). $f(x \mid \theta)$, for each $\theta \in \Theta$, $\theta \neq \theta_0$, differs from $f(x \mid \theta_0)$ on more than a set of probability zero

(F3). $\log f(x \mid \theta)$ is continuous in θ uniformly in x

(F4). $\log f(x \mid \theta)$ has finite fourth moments for all $\theta \in \Theta$

(F5). either Θ is bounded, or for some $k > 0$, $\sup_{|\theta| > k} \log (f(x \mid \theta)/f(x \mid \theta_0))$ has negative first moment and finite fourth moment

(F6). $(\partial^r/\partial\theta^r) \log f(x \mid \theta)$ exists and is continuous uniformly in x, for

$1 \leqq r \leqq 4$, in some neighborhood of θ_0; furthermore, $(\partial^r/\partial\theta^r) \log f(x \mid \theta)$ has finite $4/r$th moment and $(\partial/\partial\theta) \log f(x \mid \theta)$ has positive second moment.

The regularity conditions for h require that it be smooth in a neighborhood of θ_0, and that it will be possible to obtain posterior densities from it. $h(\theta)$ is a density with respect to Lebesgue measure on Θ. We say h is *regular* at θ_0 if

(H1). $\log h(\theta)$, $[(\partial/\partial\theta) \log h(\theta)]$, $[(\partial^2/\partial\theta^2) \log h(\theta)]$ exist and are continuous in a neighborhood of θ_0

(H2). $\int f(x \mid \theta)h(\theta)\, d\theta$ exists and is a bounded function of x.

Finally, we specify the regularity conditions for the loss function $L(d, \theta)$; for estimation of a real valued parameter, $D = \Theta$ is a closed subset of the real line. L is regular at θ_0 if

(L1). $L(d, \theta)$ is bounded

(L2). $L(d, \theta)$ is continuous in θ at θ_0, uniformly in d. $L(d, \theta)$ is continuous in d at θ_0, uniformly in θ

(L3). in some neighborhood of (θ_0, θ_0), the derivatives $(\partial/\partial d)^r(\partial/\partial\theta)^s L(d, \theta)$, $1 \leqq r + s \leqq 4$, exist and are continuous; $L(d, \theta)$, θ fixed, has a unique minimum at $d = \theta$ and $L(d, \theta)$, d fixed, has a unique minimum at $\theta = d$

(L4). $[(\partial^2/\partial\theta^2)L(d, \theta)]_{d=\theta=\theta_0} > 0$.

The asymptotic behaviour of a sequence of random variables $\{y_n\}$ will be described by the following conventions: $y_n = O(B_n)$ with probability $1 - O(\epsilon_n)$ means that $\exists\ K, k, N$, such that $P(|y_n| > KB_n) < k\epsilon_n$ for all $n > N$; $y_n = O(B_n)$ with probability $1 - o(\epsilon_n)$ means that for each $k > 0$, $\exists\ K, N$ such that $P(|y_n| > KB_n) < k\epsilon_n$ for all $n > N$. If $y_n = O(B_n)$ with probability $1 - o(1)$, we will say $y_n = O(B_n)$. Finally $\mathcal{E}_c^*(y_n)$ denotes the average of the truncated variable $y_n^* = y_n$ if $|y_n| < c$, $y_n^* = 0$ if $|y_n| \geqq c$; in particular, we note that

$$\mathcal{E}_{2c}^*(x_n + y_n) = \mathcal{E}_c^*(x_n) + \mathcal{E}_c^*(y_n) + O(P(|x_n| > c)) + O(P(|y_n| > c)).$$

The asymptotic results require contemplation of an infinite sequence of observations $x_1, \cdots x_n, \cdots$, from the same distribution as x. A set of n independent observations has density $f(\mathbf{x}_n \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$; for each set of n observations as n runs from 1 to ∞, we generate a posterior density $g(\theta \mid \mathbf{x}_n)$ corresponding to the prior density h. We will be concerned with the detailed asymptotic behaviour of $g(\theta \mid \mathbf{x}_n)$ and the Bayes estimators $d_h(\mathbf{x}_n)$.

We set $f_r = [(\partial^r/\partial\theta^r) \log f(\mathbf{x}_n \mid \theta)]_{\theta_0}$,

$$g_r^p = \mathcal{E}(f_r^p), \qquad 1 \leqq r \leqq 4, \qquad g_{12} = \mathcal{E}(f_1 f_2).$$

Then for regular f, $g_1 = 0$, $g_1^2 + g_2 = 0$, $g_1^3 + 3g_{12} + g_3 = 0$.

Further we set $h_r = [(\partial^r/\partial\theta^r) \log h(\theta)]_{\theta_0}$,

$$L_{rs} = [(\partial^r/\partial d^r)(\partial^s/\partial\theta^s)L(d, \theta)]_{d=\theta_0, \theta=\theta_0}.$$

For regular L, $L_{10} = L_{01} = 0$, $L_{02} = L_{20} = -L_{11}$, $L_{30} + 2L_{21} + L_{12} = 0$, $L_{21} + 2L_{12} + L_{03} = 0$.

4. Asymptotic behaviour of the posterior density. The posterior density is given by

$$g(\theta \mid \mathbf{x}_n) = f(\mathbf{x}_n \mid \theta)h(\theta)/\int f(\mathbf{x}_n \mid \theta)h(\theta)\, d\theta.$$

Now $f(\mathbf{x}_n \mid \theta) = \exp[\sum_{i=1}^n \log f(x_i \mid \theta)]$; our approach to the asymptotic behaviour of $g(\theta \mid \mathbf{x}_n)$ when the true value of the parameter is θ_0, will principally depend on (well-known) properties of the sums of independent identical random variables, $\sum_{i=1}^n \log f(x_i \mid \theta)$, and on Taylor expansions of these sums around θ_0.

LEMMA 1. *If f and h are regular at θ_0, for any $c > 0$, $\exists\ \delta > 0$ such that $[f(\mathbf{x}_n \mid \theta)/f(\mathbf{x}_n \mid \theta_0)]$ is $O(e^{-n\delta})$, uniformly in θ such that $|\theta - \theta_0| > c$, with probability $1 - O(n^{-2})$.*

PROOF. Let us first consider, for a single value of θ, $f(\mathbf{x}_n \mid \theta)/f(\mathbf{x}_n \mid \theta_0)$. Define $v = \log f(x \mid \theta) - \log f(x \mid \theta_0)$, so that $\log[f(\mathbf{x}_n \mid \theta)/f(\mathbf{x}_n \mid \theta_0)] = \sum_{i=1}^n v_i$. Now because exp is convex $\exp(\mathcal{E}(v)) \leqq \mathcal{E}(\exp v) = 1$, with equality only if v is constant with probability one; since, from (F2), $f(x \mid \theta)$ differs from $f(x \mid \theta_0)$ with probability greater than zero, v is not constant and $\exp \mathcal{E}(v) < 1$ or $\mathcal{E}(v) < 0$.

For any random variable, from the Chebyshev inequality, $P(|y - \mu| > c) \leqq \mu_4/c^4$, where μ is the first moment and μ_4 the fourth moment. Suppose that v has rth (central) moment μ_r; then $\sum v_i$ has first moment $n\mathcal{E}(v)$ and fourth moment $n\mu_4 + 3n(n-1)\mu_2^2$.

Thus $P(\sum v_i > \frac{1}{2}n\mathcal{E}(v)) \leqq k/n^2$; we therefore have $f(\mathbf{x}_n \mid \theta)/f(\mathbf{x}_n \mid \theta_0)$ is $O(e^{-n\delta})$ with probability $1 - O(n^{-2})$. From [F3], which states that $f(x \mid \theta)$ is continuous in θ uniformly in x, the result holds uniformly for θ in any region $K > |\theta - \theta_0| > \epsilon$. From [F5] we have the result holding outside $|\theta| > K$ for some K large enough; thus $f(\mathbf{x}_n \mid \theta)/f(\mathbf{x}_n \mid \theta_0)$ is $O(e^{-n\delta})$ with probability $1 - O(n^{-2})$, uniformly in θ such that $|\theta - \theta_0| > c$, which proves the lemma.

THEOREM 1. *Let f, h be regular at θ_0; the asymptotic behaviour of $g(\theta \mid \mathbf{x}_n)$ may be treated separately in the three regions*

$$\text{(i) } |\theta - \theta_0| \leqq n^{-\frac{1}{2}+\epsilon}, \qquad \text{(ii) } n^{-\frac{1}{2}+\epsilon} \leqq |\theta - \theta_0| \leqq c, \qquad \text{(iii) } |\theta - \theta_0| > c.$$

For c sufficiently small, and for any $\epsilon > 0$, the following hold:

Uniformly in (i),

$$g(\theta \mid \mathbf{x}_n) = (-g_2/2\pi)^{\frac{1}{2}} \exp(\tfrac{1}{2}\xi^2 g_2)\cdot\{1 + \xi[h_1 + f_1(g_2 - f_2)/g_2 + \tfrac{1}{2}f_1^2 g_3/g_2^2] + \tfrac{1}{2}(\xi^2 + g_2^{-1})[f_2 - g_2 - f_1 g_3/g_2] + \tfrac{1}{6}\xi^3 g_3 + O(n^{-1+6\epsilon})\}$$

with probability $1 - o(1)$, where $\xi = \theta - \theta_0 + f_1/g_2$.

Uniformly in (ii), $g(\theta \mid \mathbf{x}_n) = O(e^{-n\delta})$ *with probability* $1 - o(1)$.

In (iii), $\int_{|\theta-\theta_0|>c} g(\theta \mid \mathbf{x}_n)\, d\theta = O(e^{-n\delta})$ *with probability* $1 - O(n^{-2})$.

PROOF. Define $L_n(\theta) = f(\mathbf{x}_n \mid \theta)h(\theta)/f(\mathbf{x}_n \mid \theta_0)h(\theta_0)$; then $g(\theta \mid \mathbf{x}_n) = L_n(\theta)/\int L_n(\theta)\, d\theta$. To prove (iii), recall that from Lemma 1, $L_n(\theta)$ is $O(e^{-n\delta})$ uniformly in $|\theta - \theta_0| > c$ with probability $1 - O(n^{-2})$; let x_{n+1} be an $(n+1)$st observation; since $\log f(x_{n+1} \mid \theta_0)$ has finite fourth moment, the proof of Lemma 1 may be

modified to show that $L_n(\theta)/f(x_{n+1} \mid \theta_0)$ is $O(e^{-n\delta})$ uniformly in $|\theta - \theta_0| > c$ with probability $1 - O(n^{-2})$. Also, $\int f(x_{n+1} \mid \theta)h(\theta)\, d\theta$ is bounded as x_{n+1} varies from (H2), so $\int_{|\theta-\theta_0|>c} L_{n+1}(\theta)\, d\theta$ is $O(e^{-n\delta})$ with probability $1 - O(n^{-2})$. It remains to check that $\int L_n(\theta)\, d\theta$ is reasonably large; if c' is sufficiently small, $|\log f(x \mid \theta) - \log f(x \mid \theta_0)| < \frac{1}{2}\delta$ for all x, for $|\theta - \theta_0| < c'$. Thus $\int_{|\theta-\theta_0|<c'} L_n(\theta)\, d\theta > e^{-\frac{1}{2}\delta n}c'$, and

$$\int_{|\theta-\theta_0|>c} g(\theta \mid \mathbf{x}_n)\, d\theta = \int_{|\theta-\theta_0|>c} L_n(\theta)\, d\theta / \int L_n(\theta)\, d\theta \quad \text{is} \quad O(e^{-\frac{1}{2}n\delta}),$$

with probability $1 - O(n^{-2})$.

Let us next consider the behaviour of L for $n^{-\frac{1}{2}+\epsilon} < |\theta - \theta_0| < c$.

$$\log L = \log h(\theta) - \log h(\theta_0) + (\theta - \theta_0)f_1 + \tfrac{1}{2}(\theta - \theta_0)^2 f_2 + \tfrac{1}{6}(\theta - \theta_0)^3 f_3 + \tfrac{1}{24}(\theta - \theta_0)^4 f_4(\phi)$$

where $f_4(\phi) = \sum_{i=1}^{n} (\partial^4/\partial\phi^4) \log f(x_i \mid \phi)$, some $|\phi - \theta_0| < c$. From F6, we can choose c so that for some $\lambda < \infty$, $f_4(\phi) < \lambda n$ for all $|\theta - \theta_0| < c$ with probability $1 - o(1)$. Also $f_3 = nk_3 + O(n^{\frac{1}{2}})$, $f_2 = nk_2 + O(n^{\frac{1}{2}})$ where $k_2 < 0$, and $f_1 = O(n^{\frac{1}{2}})$; here $k_i = \mathcal{E}\{(\partial/\partial\theta)^i \log f(x \mid \theta)\}$. For c small enough, there exists a $\mu > 0$ such that

$$\tfrac{1}{2}(\theta - \theta_0)^2 f_2 + \tfrac{1}{6}(\theta - \theta_0)^3 f_3 + \tfrac{1}{24}(\theta - \theta_0)^4 f_4(\phi) < -n\mu(\theta - \theta_0)^2$$

for $|\theta - \theta_0| < c$, with probability $1 - o(1)$. Also in $n^{-\frac{1}{2}+\epsilon} < |\theta - \theta_0| < c$,

$$\log L < (\theta - \theta_0)f_1 - n\mu(\theta - \theta_0)^2 + \log h(\theta) - \log h(\theta_0) < n^{-\frac{1}{2}+\epsilon}O(n^{\frac{1}{2}}) - n\mu(n^{-1+2\epsilon}) + O(1) < -n^{\epsilon}$$

with probability $1 - o(1)$. Thus $L = O(e^{-n^{\epsilon}})$ uniformly in $n^{-\frac{1}{2}+\epsilon} < |\theta - \theta_0| < c$.

Let us now consider the behaviour of L in $0 < |\theta - \theta_0| < n^{-\frac{1}{2}+\epsilon}$.

$$\log L = (\theta - \theta_0)(f_1 + h_1) + \tfrac{1}{2}(\theta - \theta_0)^2 f_2 + \tfrac{1}{6}(\theta - \theta_0)^3 f_3 + O(n^{-1+4\epsilon})$$

$$(1) \qquad L = \exp[(\theta - \theta_0)f_1 + \tfrac{1}{2}(\theta - \theta_0)^2 g_2]\{1 + (\theta - \theta_0)h_1 + \tfrac{1}{2}(\theta - \theta_0)^2 \cdot (f_2 - g_2) + \tfrac{1}{6}(\theta - \theta_0)^3 g_3 + O(n^{-1+6\epsilon})\}$$

uniformly over $|\theta - \theta_0| < n^{-\frac{1}{2}+\epsilon}$. Now $\int_{|\theta-\theta_0|>n^{-\frac{1}{2}+\epsilon}} \exp[(\theta - \theta_0)f_1 + \frac{1}{2}(\theta - \theta_0)^2 g_2] \cdot \{1 + (\theta - \theta_0)h_1 + \frac{1}{2}(\theta - \theta_0)^2(f_2 - g_2) + \frac{1}{6}(\theta - \theta_0)^3 g_3\}\, d\theta$ is $O(e^{-n^{\epsilon}})$; also $\int_{|\theta-\theta_0|>n^{-\frac{1}{2}+\epsilon}} L(\theta)\, d\theta$ is $O(e^{-n^{\epsilon}})$. This permits us to obtain $\int_{\Theta} L\, d\theta$ by integrating the R.H.S. of 1). Thus

$$(2) \qquad \int_{\Theta} L\, d\theta = \exp(-\tfrac{1}{2}f_1^2/g_2)(-2\pi/g_2)^{\frac{1}{2}}[1 - f_1h_1/g_2 + \tfrac{1}{2}(f_1^2 - g_2)(f_2 - g_2)/g_2^2 + (3f_1g_2 - f_1^3)g_3/6g_2^3 + O(n^{-1+6\epsilon})] = O(n^{-\frac{1}{2}}).$$

Dividing (2) into (1), we obtain that (i) holds uniformly in $|\theta - \theta_0| < n^{-\frac{1}{2}+\epsilon}$. For $n^{-\frac{1}{2}+\epsilon} < |\theta - \theta_0| < c$, $g(\theta \mid \mathbf{x}_n) = L(\theta)/\int L(\theta)\, d\theta = O(e^{-n^{\delta}})$ some $\delta > 0$, and so (ii) holds also.

This proves the theorem.

The posterior distribution of θ is thus concentrated in the neighborhood of θ_0; in fact θ is asymptotically normal with mean $\theta_0 - f_1/g_2$ and variance $-1/g_2$ [recalling the maximum likelihood result that $\hat{\theta} - \theta_0 + f_1/g_2 = O(1/n)$, where $\hat{\theta}$ is the maximum likelihood estimate, we can therefore expect that Bayes estimators, whatever the prior density, will differ from maximum likelihood by $O(1/n)$]. The effect of the prior density, represented by $h_1 = [(\partial/\partial\theta) \log h(\theta)]_{\theta_0}$, appears in lower order terms in the distribution; the moments of the truncated variable $\theta^* = \theta$ if $|\theta - \theta_0| < c$, $\theta^* = 0$ if $|\theta - \theta_0| \geqq c$, are given by

$$(3)\quad \mu_1^* = \theta_0 - (f_1/g_2) - (1/g_2)[h_1 + f_1(g_2 - f_2)/g_2 + \tfrac{1}{2}f_1^2 g_3/g_2^2 - \tfrac{1}{2}g_3/g_2] + O(n^{-\frac{3}{2}+\epsilon}),$$

$$(4)\quad \mu_2^* = -1/g_2 + 1/g_2^2[f_2 - g_2 - \tfrac{1}{2}f_1 g_3/g_2] + O(n^{-2+\epsilon}),$$
$$\mu_3^* = -g_3/g_2^3 + O(n^{-\frac{3}{2}+\epsilon}),$$

and generally, the rth cumulant is given by $K_r^* = O\{n^{-\frac{1}{2}(r+1)}\}$ for $r \geqq 3$. $\theta + h_1/g_2$ has a posterior density which is independent of the prior distribution h up to terms in $O(n^{-\frac{3}{2}+\epsilon})$ [compared with a single observation, which changes the posterior density by a term of $O(1)$]. Asymptotically, the prior distribution h therefore affects the location rather than the shape of the posterior distribution. Therefore, we can expect, when we use the prior distribution to generate statistical decision procedures, that bias rather than efficiency will suggest suitable pricrs.

5. Asymptotically unbiased prior distributions. We will be interested in this section in asymptotic properties of the Bayes estimator $d_h(\mathbf{x}_n)$; $d_h(\mathbf{x}_n)$ is that decision which minimizes $\int L(d, \theta)g(\theta \mid \mathbf{x}_n)\, d\theta$.

THEOREM 2. (i) $d_h(\mathbf{x}_n) - \theta_0 = o(1)$ *with probability* $1 - O(n^{-2})$

(ii) $d_h(\mathbf{x}_n) = \mu_1^* - \frac{1}{2}(L_{12}/L_{20})\mu_2^* + O(n^{-\frac{3}{2}})$

where μ_i^* *is the* i*th moment, of the truncated variable:* $\theta^* = \theta$ *if* $|\theta - \theta_0| < c$, $\theta^* = 0$ *if* $|\theta - \theta_0| \geqq c$, *over the posterior distribution.*

PROOF. If $|d_h(\mathbf{x}_n) - \theta_0| > \epsilon$, since $L[d, \theta_0]$ has a unique minimum at $d = \theta_0$, and since $L[d, \theta]$ is continuous at θ_0, $\exists$ c, δ depending only on ϵ such that $(L(d_h(\mathbf{x}_n), \theta) - L(\theta_0, \theta)) > \delta$ whenever $|\theta - \theta_0| < c$. Comparing the decision $d_h(\mathbf{x}_n)$ with the decisions θ_0, $\int (L(d_h(\mathbf{x}_n), \theta) - L(\theta_0, \theta))g(\theta \mid \mathbf{x}_n) < 0$. Hence

$$\delta \int_{|\theta-\theta_0|<c} g(\theta \mid \mathbf{x}_n)\, d\theta < 2 \sup_{d,\theta} L(d, \theta) \int_{|\theta-\theta_0|>c} g(\theta \mid \mathbf{x}_n)\, d\theta.$$

Now, the posterior probability that $|\theta - \theta_0| > c$ is $O(e^{-n\delta})$ with probability $1 - O(n^{-2})$. Hence the event $|d_h(\mathbf{x}_n) - \theta_0| > \epsilon$ occurs with probability $O(n^{-2})$, which proves (i). To prove (ii), we have that for d near θ_0,

$$(5)\quad \int L(d, \theta)g(\theta \mid \mathbf{x}_n)\, d\theta = \int_{|\theta-\theta_0|<c} L(d, \theta)g(\theta \mid \mathbf{x}_n)\, d\theta + \int_{|\theta-\theta_0|\geqq c} L(d, \theta)g(\theta \mid \mathbf{x}_n)\, d\theta.$$

Let us first find d which minimizes the $|\theta - \theta_0| < c$ term; we will later show that the second term can be ignored. We want d as a solution of $\int_{|\theta-\theta_0|<c}$

$(\partial/\partial d)L(d, \theta)g(\theta \mid \mathbf{x}_n)\, d\theta = 0$ i.e. $\int_{|\theta-\theta_0|<c} [L_{20}(d - \theta) + \frac{1}{2}L_{30}(d - \theta_0)^2 + L_{21}(d - \theta_0)(\theta - \theta_0) + \frac{1}{2}L_{12}(\theta - \theta_0)^2 + O(\text{third order terms})]\, g(\theta \mid \mathbf{x}_n)\, d\theta = 0$, where this Taylor expansion about (θ_0, θ_0) is validated by the regularity conditions on L. Hence

$$L_{20}(d - \mu^*) + \tfrac{1}{2}L_{30}(d - \theta_0)^2 + L_{21}(d - \theta_0)(\mu^* - \theta_0) + \tfrac{1}{2}L_{12}[\mu_2^* + (\mu_1^* - \theta_0)^2] + O(n^{-\frac{3}{2}}) = 0.$$

Hence $d = \mu^* - \frac{1}{2}(L_{12}/L_{20})\mu_2^* + O(n^{-\frac{3}{2}})$ (recalling that $L_{30} + 2L_{21} + L_{12} = 0$). It remains to check that the second term in (5), which is $O(e^{-n\delta})$, does not affect this minimum d. Suppose the decision minimizing $\int L(d, \theta)g(\theta \mid \mathbf{x}_n)\, d\theta$ was d'; then $\int_{|\theta-\theta_0|<c} L(d, \theta)g(\theta \mid \mathbf{x}_n)\, d\theta - \int_{|\theta-\theta_0|<c} L(d', \theta)g(\theta \mid \mathbf{x}_n)\, d\theta = O(e^{-n\delta})$. Thus $d - d' = O(e^{-n\delta})$ and the second term in (5) has only a trivial effect.

THEOREM 3. *Let f, h, and L be regular at θ_0. For any prior distribution h, the bias of $d_h(\mathbf{x}_n)$ at θ_0 is $O(n^{-2})$; the bias is $o(n^{-2})$ if and only if* $[(\partial/\partial\theta) \log h(\theta)]_{\theta=\theta_0} = [(\partial/\partial\theta) \log \mathcal{E}(((\partial/\partial\theta) \log f)^2 \mid \theta)]/[(\partial^2/\partial d^2)L(d, \theta)]^{\frac{1}{2}}_{d=\theta})]_{\theta=\theta_0}$. *Such an h will be called asymptotically unbiased at θ_0. h is asymptotically unbiased in an interval if and only if* $h(\theta) = \mathcal{E}(((\partial/\partial\theta) \log f)^2 \mid \theta)/[(\partial^2/\partial d^2)L(d, \theta)]^{\frac{1}{2}}_{d=\theta}$ *in the interval.*

PROOF. The bias of $d_h(\mathbf{x}_n)$ at θ_0 is defined by $B(d_h, \theta_0) = \sup_{\theta\in\Theta} \mathcal{E}[L(d, \theta_0) - L(d, \theta)]$. It will be convenient to split the integration symbolized by $\mathcal{E}$ into two parts, one over $|d - \theta_0| < c$, which we represent by $\mathcal{E}^*$ and one over $|d - \theta_0| \geqq c$ which we represent by $\mathcal{E}^{**}$.

Firstly let us show that the value θ_n maximizing $\mathcal{E}(L(d, \theta_0) - L(d, \theta))$ converges to θ_0 as $n \to \infty$. For any $\epsilon > 0$, consider those n for which $|\theta_n - \theta_0| > \epsilon$. From (L3), we can find δ such that $L(\theta_0, \theta_n) \geqq L(\theta_0, \theta_0) + \delta$ for $|\theta_n - \theta_0| > \epsilon$, from (L2) we can find c such that for $|d - \theta_0| < c$, $|L(d, \theta_0) - L(\theta_0, \theta_0)| < \delta/4$, $|L(d, \theta_n) - L(\theta_0, \theta_n)| < \delta/4$. Thus $L(d, \theta_0) - L(d, \theta_n) < -\delta/2$ for $|d - \theta_0| < c$, $|\theta_n - \theta_0| > \epsilon$. Hence for $|\theta_n - \theta_0| > \epsilon$, $\mathcal{E}(L(d, \theta_0) - L(d, \theta_n)) = \mathcal{E}^*(L(d, \theta_0) - L(d, \theta_n)) + \mathcal{E}^{**}(L(d, \theta_0) - L(d, \theta_n)) \leqq -\delta/2 + O(n^{-2})$, since $|d - \theta_0| > c$ with probability $O(n^{-2})$. Hence, for all n large enough $|\theta_n - \theta_0| < \epsilon$; i.e. $\theta_n \to \theta_0$ as $n \to \infty$. Thus, for maximal θ_n, since $L(d, \theta)$ is continuous in θ at θ_0, uniformly in d,

$$\mathcal{E}(L(d, \theta_0) - L(d, \theta_n)) = \mathcal{E}^*(L(d, \theta_0) - L(d, \theta_n)) + o(n^{-2}). \tag{6}$$

Let us now find the value of θ which maximizes $\mathcal{E}^*[L(d, \theta_0) - L(d, \theta)]$; we need to find θ satisfying

$$\mathcal{E}^*[L_{20}(d - \theta) - \tfrac{1}{2}L_{21}(d - \theta_0)^2 - L_{12}(d - \theta_0)(\theta - \theta_0) - \tfrac{1}{2}L_{03}(\theta - \theta_0)^2] = 0.$$

Now $d = \mu^* - \frac{1}{2}L_{12}\mu_2^*/L_{20} + O(n^{-\frac{3}{2}})$ where μ^* and μ_2^* are given by (3) and (4); it follows that $\mathcal{E}^*(d - \theta_0) = O(n^{-1})$, $\mathcal{E}^*(d - \theta_0)^2 = O(n^{-1})$ and $\mathcal{E}^*(d - \theta_0)^3 = O(n^{-\frac{3}{2}})$; also $P(|d - \theta_0| > c) = O(n^{-2})$, which makes $\mathcal{E}^*$ additive, so $\theta = \mathcal{E}^*(d) - \frac{1}{2}(L_{21}/L_{20})\mathcal{E}^*(d - \theta_0)^2 + O(n^{-\frac{3}{2}})$. To find θ_n maximizing $\mathcal{E}(L(d, \theta_0) - L(d, \theta))$ we must take the term of $o(n^{-2})$ into account in Equation (6). A change $\Delta\theta$ in θ produces a change $\Delta\theta\mathcal{E}^*(\partial L/\partial\theta) = \Delta\theta \times o(n^{-1})$ in $\mathcal{E}^*[L(d, \theta_0) - L(d, \theta)]$.

This change must be $o(n^{-2})$, so $\Delta\theta = o(n^{-1})$. Thus the maximizing θ_n satisfies $\theta_n = \mathcal{E}^*(d) - \frac{1}{2}(L_{21}/L_{20})\mathcal{E}^*(d - \theta_0)^2 + o(n^{-1})$. We are now in a position to evaluate the bias of an estimator at θ_0 ;

$$\begin{aligned}
B(d, \theta_0) &= \mathcal{E}^*[L(d, \theta_0) - L(d, \theta_n)] + o(n^{-2}) \\
&= \mathcal{E}^*\{\tfrac{1}{2}L_{20}(\theta_n - \theta_0)(d - \tfrac{1}{2}(\theta_n + \theta_0)) - \tfrac{1}{2}L_{21}(d - \theta_0)^2(\theta_n - \theta_0) \\
&\quad - \tfrac{1}{2}L_{12}(d - \theta_0)(\theta_n - \theta_0)^2 - \tfrac{1}{6}L_{03}(\theta_n - \theta_0)^3 + o(n^{-2}) \\
&= \tfrac{1}{2}L_{20}(\theta_n - \theta_0)\mathcal{E}^*(d - \theta_0) - \tfrac{1}{4}L_{20}(\theta_n - \theta_0)^2 \\
&\quad - \tfrac{1}{2}L_{21}\mathcal{E}^*(d - \theta_0)^2(\theta_n - \theta_0) + o(n^{-2}) \\
&= \tfrac{1}{4}L_{20}(\theta_n - \theta_0)^2 + o(n^{-2}).
\end{aligned}$$

We thus see that the asymptotic bias depends only on the rate of convergence of the maximal θ_n. Now $\theta_n = \theta_0 + \mathcal{E}^*(d - \theta_0) - \frac{1}{2}(L_{21}/L_{20})\mathcal{E}^*(d - \theta_0)^2 + o(n^{-1})$ and $d = \mu_1^* - \frac{1}{2}(L_{12}/L_{20})\mu_2^* + O(n^{-\frac{3}{2}})$ from Theorem 1, where

$$\mu_1^* = \theta_0 - f_1/g_2 - g_2^{-1}[h_1 + f_1(g_2 - f_2)/g_2 + \tfrac{1}{2}f_1^2 g_3/g_2^2 - \tfrac{1}{2}g_3/g_2] + O(n^{-\frac{3}{2}})$$

$$\mu_2^* = -g_2^{-1} + O(n^{-\frac{3}{2}}).$$

We know that $\mathcal{E}(f_1/g_2) = 0$, but must check that $\mathcal{E}^*(f_1/g_2)$ is small. In fact,

$$\begin{aligned}
|\mathcal{E}(f_1 \mid g_2) - \mathcal{E}^*(f_1 \mid g_2)| &\leqq |\mathcal{E}^{**}(f_1 \mid g_2)| \\
&\leqq c^{-3}\mathcal{E}^{**}(f_1^4 \mid g_2^4) \\
&\leqq c^{-3}\mathcal{E}(f_1^4 \mid g_2^4) = O(n^{-2}).
\end{aligned}$$

A similar result holds for other terms in μ_1^* (except the $O(n^{-\frac{3}{2}})$ term). Thus

$$\theta_n = \theta_0 - g_2^{-1}[h_1 - g_{12}/g_2 - \tfrac{1}{2}g_3/g_2 - \tfrac{1}{2}g_3/g_2] + \tfrac{1}{2}L_{12}/L_{20}g_2^{-1} + \tfrac{1}{2}L_{21}/L_{20}g_2^{-1} + o(n^{-1})$$

i.e. $\theta_n = \theta_0 - g_2^{-1}[h_1 - (g_{12} + g_3)/g_2 - \frac{1}{2}(L_{12} + L_{21})/L_{20}] + o(n^{-1})$. Let us now observe that

$$\begin{aligned}
&(g_{12} + g_3)/g_2 \\
&\quad = [\mathcal{E}((\partial/\partial\theta) \log f(\partial^2/\partial\theta^2) \log f) + \mathcal{E}((\partial^3/\partial\theta^3) \log f)]/\mathcal{E}((\partial^2/\partial\theta^2) \log f) \\
&\quad = [(\partial/\partial\theta) \log (-\mathcal{E}((\partial^2/\partial\theta^2) \log f))]_{\theta=\theta_0},
\end{aligned}$$

and since $L_{12} + L_{21} = -(L_{30} + L_{21})$,

$$\begin{aligned}
&(L_{12} + L_{21})/L_{20} \\
&= -\{[(\partial^3/\partial\, d^3)L(d, \theta)]_{d=\theta} + [(\partial^2/\partial\, d^2)(\partial/\partial\theta)L(d, \theta)]_{d=\theta}\}/[(\partial^2/\partial\, d^2)L(d, \theta)]_{d=\theta} \\
&= -[(\partial/\partial\theta)\{\log [(\partial^2/\partial\, d^2)L(d, \theta)]_{d=\theta}\}]_{\theta=\theta_0}.
\end{aligned}$$

Thus $\theta_n = \theta_0 + O(n^{-1})$ for any prior distribution, and $\theta_n = \theta_0 + o(n^{-1})$ if and

only if $(\partial/\partial\theta_0)(\log h - \log \mathcal{E}((\partial/\partial\theta) \log f)^2 + \frac{1}{2}\log [(\partial^2/\partial d^2)L(d, \theta)]_{d=0}) = 0$. This proves the theorem.

6. Applications of one dimensional result. Let us use as a symbol for the asymptotically unbiased prior density JUB. For loss functions of form $L(d, \theta) = \alpha(\theta)(d - \theta)^2$, we have $\text{JUB}(\theta) = I(\theta)/\alpha(\theta)$; as we change $\alpha(\theta)$ we will run over all smooth prior distributions. Thus there is a 1–1 correspondence between asymptotically unbiased prior distributions and loss functions of this type.

For the "intrinsic" loss functions suggested by Jeffreys [9], $L(d, \theta) = \int [f^{\frac{1}{2}}(x \mid d) - f^{\frac{1}{2}}(x \mid \theta)]^2 \nu(dx)$ and $L(d, \theta) = \int \log (f(x \mid d)/f(x \mid \theta))(f(x \mid d) - f(x \mid \theta))\nu(dx)$, $\text{JUB} = I^{\frac{1}{2}} = J_{\frac{1}{2}}$, Jeffreys' density.

We might ask for cases where there exists a prior distribution giving an exactly unbiased estimator. Such a case is that of the exponential family, where $f(x \mid \theta) = \exp (a(\theta)b(x) + l(\theta) + m(x))$, $a(\theta)$ monotone, $a(\theta)$ and $l(\theta)$ differentiable; we will suppose further that $f(x \mid \theta) = 0$ on the boundary of Θ. It is a well known by-product of the Cramer-Rao inequality that $b(x)$ is a minimum variance unbiased estimator of $\mathcal{E}_\theta(b(x)) = -l'(\theta)/a'(\theta)$. Let us then take $L(d, \theta) = [l'(\theta)/a'(\theta) - l'(d)/a'(d)]^2$ and find that $I(\theta) = -l''(\theta) + a''(\theta)l'(\theta)/a'(\theta)$ and that $\text{JUB} = |a'(\theta)|$. The JUB is such that $a(\theta)$ is uniformly distributed over Θ; i.e. $\text{JUB} = J_0$ in the notation of the introduction. For this prior distribution the Bayes estimator is $b(x)$ which is exactly unbiased; this then is a case where an exactly unbiased prior distribution exists.

7. Extensions to more than one-dimension. Suppose that Θ is a closed subset of r-dimensional euclidean space, and that θ_0 is an interior point of Θ. Very similar results to the one dimensional case may be obtained for the asymptotic form of the posterior distribution and for specifying the prior distribution which gives asymptotically unbiased estimators; we state them here without proof.

Let us change our notation so that

$$f_i = [(\partial/\partial\theta_i) \log f(\mathbf{x}_n \mid \theta]_{\theta_0}, \qquad h_i = [(\partial/\partial\theta_i) \log h(\theta)]_{\theta_0}$$

$$f_{ij} = [(\partial/\partial\theta_i)(\partial/\partial\theta_j) \log f(\mathbf{x}_n \mid \theta)]_{\theta_0},$$

$$f_{ijk} = [(\partial/\partial\theta_i)(\partial/\partial\theta_j)(\partial/\partial\theta_k) \log f(\mathbf{x}_n \mid \theta)]_{\theta_0},$$

$$g_i = \mathcal{E}(f_i), \qquad g_{ij} = \mathcal{E}(f_{ij}), \qquad g_{ijk} = \mathcal{E}(f_{ijk}), \qquad g_{ij\cdot k} = \mathcal{E}(f_{ij}f_k);$$

let g^{ij} denote the ijth element of the inverse of the $\{g_{ij}\}$ matrix.

Let

$$L_{ij} = [(\partial/\partial\, d_i)(\partial/\partial\, d_j)L(d, \theta)]_{d=\theta=\theta_0}$$

$$L^{12}_{ijk} = [(\partial/\partial\, d_i)(\partial/\partial\theta_j)(\partial/\partial\theta_k)\, L(d, \theta)]_{\theta_0}$$

$$L^{21}_{ijk} = [(\partial/\partial\, d_i)(\partial/\partial\, d_j)(\partial/\partial\theta_k)L(d, \theta)]_{\theta_0}.$$

Finally we will use the tensor summation convention that $a_{ij}b_j = \sum_{j=1}^{r} a_{ij}b_j$, etc.

Then under regularity conditions, the asymptotic behaviour of the posterior

distribution near θ_0 is given by

$$g(\theta \mid \mathbf{x}_n) = (2\pi)^{-\frac{1}{2}r}|-g_{ij}|^{\frac{1}{2}} \exp\ (\tfrac{1}{2}\xi_i\xi_j g_{ij})\{1 + \xi_i(h_i - (f_{ij} - g_{ij})g^{jk}f_k$$
$$+ \tfrac{1}{2}g^{jl}f_l g^{km}f_m g_{ijk}) + \tfrac{1}{2}(\xi_i\xi_j + g^{ij})(f_{ij} - g_{ij} - g^{kl}f_l g_{ijk}) + \tfrac{1}{6}\xi_i\xi_j\xi_k g_{ijk} + O(n^{-1})\},$$

where $\xi_i = \theta_i - \theta_{i0} + g^{ij}f_j$. The moments are given by

$$\mu_i = \theta_{i0} - g^{ij}f_j - g^{ij}[h_j - (f_{jl} - g_{jl})g^{lk}f_k + \tfrac{1}{2}g^{pl}f_l g^{km}f_m g_{jpk} - \tfrac{1}{2}g^{lk}g_{jlk}] + O(n^{-\frac{3}{2}})$$
$$\mu_{ij} = -g^{ij} + O(n^{-\frac{3}{2}})$$
$$\mu_{ijk} = O(n^{-2})$$

The Bayes estimator with respect to h is asymptotically unbiased if the r equations in the derivatives of $\log h$,

$$h_i = g_{ij\cdot k}g^{jk} + g_{ijk}g^{jk} + \tfrac{1}{2}L^{ml}(L^{12}_{ljk} + L^{21}_{jkl})g^{jk}g_{im}, \tag{7}$$

are satisfied.

A solution to these equations may not exist for all θ, and so there may be no prior distribution which is asymptotically unbiased for all θ; a solution does exist when the loss function used is Jeffreys'

$$L(d, \theta) = \int \{\log f(x \mid d) - \log f(x \mid \theta)\}[f(x \mid d) - f(x \mid \theta)]\nu(dx).$$

In that case JUB $= J_{\frac{1}{2}} = |I|^{\frac{1}{2}}$, where I is the information matrix with ijth element $-g_{ij}$; this is the generalization of Jeffreys' density to r-dimensional parameters, as given by Jeffreys in [9]. Another simple loss function is the squared distance $L(d, \theta) = \sum_{i=1}^{r}(d_i - \theta_i)^2$; in this case Equation (7) reduces to $h_i = g_{ij\cdot k}g^{jk} + g_{ijk}g^{jk}$, but, again, this equation may have no solution for all θ.

8. Applications of many-dimensional result. For the exponential family $f(x \mid \theta) = \exp\ [\sum_{i=1}^{r} a_i(\theta)b_i(x) + l(\theta) + m(x))]$, where θ is r dimensional, $\theta \to (a_1(\theta), \cdots a_r(\theta))$ is a 1–1 differentiable transformation of θ and $l(\theta)$ is differentiable, we can find Bayes estimators which are exactly unbiased, similarly to the one-dimensional case. We assume that Θ is such that $f(x \mid \theta)$ vanishes on the boundary of Θ. Recalling that $\mathbf{b}(x) = (b_1(x), \cdots b_r(x))$ is a minimum variance unbiased estimator of $\mathcal{E}(\mathbf{b}(x) \mid \theta)$, we use the loss function

$$L(d, \theta) = \sum_{i=1}^{r}[\mathcal{E}(b_i(x) \mid d) - \mathcal{E}(b_i(x) \mid \theta)]^2;$$

the exactly unbiased prior distribution is given by requiring $[a_1(\theta), \cdots, a_r(\theta)]$ to be uniformly distributed. The associated Bayes estimator is $\mathbf{b}(x)$.

Let us now consider in detail unbiased estimation in the regression problem

$$\mathbf{y} = X\mathbf{u} + \sigma\boldsymbol{\xi},$$

where $\mathbf{y}$ is an observation vector of dimension n, $\mathbf{u}$ is a vector of dimension m, X is an $n \times m$ matrix of rank m, σ is a scalar, and $\boldsymbol{\xi}$, the error variable, is distributed as an n-dimensional spherical normal. The parameter here is $\theta = (\mathbf{u}, \sigma)$ which is $(m + 1)$ dimensional. We note that for prior densities of form $h(\mathbf{u}, \sigma) = \sigma^{p-1}$, the

posterior distribution of μ, σ given $\mathbf{y}$ is given by

$$\mu \mid \mathbf{y}, \sigma \sim N(\hat{\mu}, (X^TX)^{-1}\sigma^2)$$
$$\sigma^{-2} \sim \|\mathbf{y} - X\hat{\mu}\|^{-1}\chi^2_{n-m-p}$$

where $\|\mathbf{u}\| = \sum_{i=1}^n u_i^2$ and $\hat{\mu}$, the least squares estimate, is the value of μ minimizing $\|\mathbf{y} - X\mu\|$.

A dizzying array of possible loss functions now confronts us.

(i) *Jeffreys loss function.* $L(d, \theta) = \int (\log f(x \mid d) - \log f(x \mid \theta))(f(x \mid d) - f(x \mid \theta))\nu(dx)$. For $\theta = (\mu, \sigma)$, $d = (\mathbf{m}, s)$, $L(d, \theta) = \frac{1}{2}\|X\mu - X\mathbf{m}\|(\sigma^{-2} + s^{-2}) + \frac{1}{2}n(\sigma/s - s/\sigma)^2$; the unbiased prior distribution is $h(\mu, \sigma) = \sigma^{-(m+1)}$ (Jeffreys' prior) which leads to estimates $\mathbf{m} = \hat{\mu}$, $s^2 = \|\mathbf{y} - X\hat{\mu}\|/\{n(n-2)/(1 + m/n)\}^{\frac{1}{2}}$. This assignment of prior distribution has been a subject of some controversy because it generates the posterior $\sigma^{-2} \sim \|\mathbf{y} - X\hat{\mu}\|^{-1}\chi_n^2$, and thus the precision of our knowledge of σ, represented by the number of degrees of freedom of the χ^2, is unaffected by the number m of parameters fitted in the model. Jeffreys, in his key 1946 paper [9], questioned the plausibility of his prior $J_{\frac{1}{2}}$ in this situation. Nevertheless, for unbiased estimates according to Jeffreys' loss function, $J_{\frac{1}{2}}$ is the uniquely appropriate prior density.

(ii) *Unbiased estimation of* σ^2. If we wish to estimate σ^2 and μ unbiasedly, $L(d, \theta) = (\sigma^2 - s^2)^2 + \sum(\mu_i - m_i)^2$, the unbiased prior distribution is $h(\mu, \sigma) = \sigma^{-3}$, and the associated Bayes estimates, $\mathbf{m} = \hat{\mu}$, $s^2 = \|\mathbf{y} - X\hat{\mu}\|/(n - m)$, are exactly unbiased.

(iii) *Unbiased estimation of* σ^{-2}. Here $L(d, \theta) = (\sigma^{-2} - s^{-2})^2 + \sum(\mu_i - m_i)^2$, the unbiased prior distribution is $h(\mu, \sigma) = \sigma$, and the associated Bayes estimates, $\mathbf{m} = \hat{\mu}$, $s^2 = \|\mathbf{y} - X\hat{\mu}\|/(n - m - 2)$; are exactly unbiased.

(iv) *Minimum variance unbiased estimators.* The minimum variance unbiased estimators are $\|\mathbf{y}\|$ and $X^T\mathbf{y}$ which have expectations $n\sigma^2 + \mu^TX^TX\mu$ and $X^TX\mu$. In order to estimate $n\sigma^2 + \mu^TX^TX\mu$ and $X^TX\mu$ unbiasedly, we must use the prior distribution $h(\mu, \sigma) = \sigma^{-(2m+3)}$; the associated estimates of μ and σ^2 are $\mathbf{m} = \hat{\mu}$ and $s^2 = \|\mathbf{y} - X\hat{\mu}\|/n$. The posterior distribution of σ is given by $\sigma^{-2} \sim \|\mathbf{y} - X\hat{\mu}\|\chi^2_{n+m+2}$, so that this prior distribution is even more extreme than Jeffreys' in *increasing* the number of degrees of freedom as parameters are added to the model. We must conclude that unbiased prior distributions may give reasonable estimates, but simultaneously give unsatisfactory general purpose posteriors.

(v) *The usual prior distribution.* The usual prior is $h(\mu, \sigma) = \sigma^{-1}$, which generates the posterior

$$\mu - \hat{\mu} \sim N(O, (X^TX)^{-1}\sigma^2),$$
$$\|\mathbf{y} - X\hat{\mu}\|/\sigma^2 \sim \chi^2_{n-m}.$$

The first statement is true given σ, $\hat{\mu}$ and is also true given σ, μ; the second statement is true given σ, and is also true given $\mathbf{y}$. The fiducial and posterior distribu-

tions for σ therefore coincide. This is one way of justifying priors; there is no obvious unbiasedness justification for $h(\mu, \sigma) = \sigma^{-1}$.

9. Outstanding problems.

1. Unbiasedness should determine priors in other types of decision problems; in particular, in testing $\theta = \theta_0$ against $\theta \neq \theta_0$, one Bayesian decision procedure rejects $\theta = \theta_0$ if $g(\theta_0 \mid x) < c$; tests of various levels of significance are obtained by varying c. The test will be unbiased if $P(g(\theta_0 \mid x) < c \mid \theta) \geqq P(g(\theta_0 \mid x) < c \mid \theta_0)$ for all $\theta \in \Theta$. Which prior distribution makes the test "asymptotically unbiased" for all c? Jeffreys' prior would be the first candidate.

2. For a loss function $L(d, \theta) = |d - \theta|$, the Bayes estimator is the median of the posterior distribution; under regularity conditions, it is unbiased for Jeffreys' prior. The same result appears to hold in higher dimensions; does it?

3. In "interval" estimation, suppose we describe the loss in deciding upon the region d when θ is true, by

$$L(d, \theta) = L(d) - c(d, \theta)$$

where $c(d, \theta) = 1$ if $\theta \in d$, $c(d, \theta) = 0$ if $\theta \notin d$, and L is a positive measure on Θ, with density $l(\theta)$ say. The Bayes decision is $\{\theta \mid g(\theta \mid x) > l(\theta)\}$; we may now ask for two types of asymptotic impartiality properties, *confidence* and *unbiasedness*. The region function $x \to \{\theta \mid g(\theta \mid x) > l(\theta)\}$ will be a confidence region (function) if $P(g(\theta \mid x) > l(\theta) \mid \theta)$ is independent of θ; it will be an unbiased region function if $P[g(\theta \mid x) > l(\theta) \mid \theta_0] \leqq P(g(\theta_0 \mid x) > l(\theta_0) \mid \theta_0)$ for all $\theta \in \Theta$. Borges [3] considers some confidence regions of this type. First order confidence properties are guaranteed if and only if $l(\theta) \propto J_{\frac{1}{2}}(\theta)$; higher order results depending on the prior distribution h remain open.

4. Anscombe [1] suggests transforming the parameter space to make the likelihood function look like a normal likelihood function; possibly a uniform distribution in the transformed space would give a reasonable prior. In fact, it gives the prior $J_{\frac{1}{2}}$ mentioned in the introduction, in the one-dimensional case. One would hope for an extension to a higher number of dimensions.

5. Rao [15] defines a general measure of loss $L(T, \theta)$ when an estimator T is used in estimating θ and it may be possible to specify prior distributions h by requiring the estimators T_h (minimizing $\int L(T, \theta)h(\theta)\, d\theta$) to have "good" properties.

REFERENCES

[1] ANSCOMBE, F. J. (1964). Normal likelihood functions. *Ann. Inst. Statist. Math.* **16** 1–20.

[2] BARNARD, G. A. (1952). The frequency justification of certain sequential tests. *Biometrika* **39** 144–150.

[3] BORGES, R. (1962). Subjektivtrennscharfe Konfidenzbereiche. *Z. fur Washrscheinlichkeitstheorie* **1** 47–69.

[4] CRAMÉR, H. (1937). *Random Variables and Probability Distributions*. University Press, Cambridge.

[5] FRASER, D. A. S. (1961). The fiducial method and invariance. *Biometrika* **48** 261–280.

[6] FREEDMAN, D. A. (1963). On the asymptotic behaviour of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403.
[7] HARTIGAN, J. A. (1964). Invariant prior distributions. *Ann. Math. Statist.* **35** 836–845.
[8] JEFFREYS, H. (1961). *Theory of Probability*. 3rd ed. Oxford Univ. Press, London.
[9] JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A.* **186** 453–461.
[10] LE CAM, L. (1954). On the asymptotic theory of estimation and hypothesis testing. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 129–156.
[11] LE CAM, L. (1958). The asymptotic properties of Bayes solutions. *Publ. Inst. Statist. Univ. Paris.* **2** 17–35.
[12] LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005.
[13] LINDLEY, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 453–468.
[14] PERKS, F. J. A. (1947). Some observations on inverse probability, including a new indifference rule. *J. Inst. Actuaries.* **73** 285–334.
[15] RAO, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 531–554.
[16] THATCHER, A. R. (1964). Relationships between Bayesian and confidence limits for predictions. *J. Roy. Statist. Soc. Ser. B* **26** 176–191.
[17] WALLACE, D. L. (1958). Asymptotic approximations to distributions. *Ann. Math. Statist.* **29** 635–654.
[18] WELCH, B. L. and PEERS, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **25** 318–329.