Bayes Estimates for the Linear Model

By D. V. LINDLEY AND A. F. M. SMITH

University College, London

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1971, Mr M. J. R. HEALY in the Chair]

SUMMARY

The usual linear statistical model is reanalyzed using Bayesian methods and the concept of exchangeability. The general method is illustrated by applications to two-factor experimental designs and multiple regression.

Keywords: LINEAR MODEL; LEAST SQUARES; BAYES ESTIMATES; EXCHANGEABILITY; ADMISSIBILITY; TWO-FACTOR EXPERIMENTAL DESIGN; MULTIPLE REGRESSION; RIDGE REGRESSION; MATRIX INVERSION.

INTRODUCTION

ATTENTION is confined in this paper to the linear model, $E(\mathbf{y}) = \mathbf{A}\boldsymbol{\theta}$, where $\mathbf{y}$ is a vector of observations, $\mathbf{A}$ a known design matrix and $\boldsymbol{\theta}$ a vector of unknown parameters. The usual estimate of $\boldsymbol{\theta}$ employed in this situation is that derived by the method of least squares. We argue that it is typically true that there is available prior information about the parameters and that this may be exploited to find improved, and sometimes substantially improved, estimates. In this paper we explore a particular form of prior information based on de Finetti's (1964) important idea of exchangeability.

The argument is entirely within the Bayesian framework. Recently there has been much discussion of the respective merits of Bayesian and non-Bayesian approaches to statistics: we cite, for example, the paper by Cornfield (1969) and its ensuing discussion. We do not feel that it is necessary or desirable to add to this type of literature, and since we know of no reasoned argument against the Bayesian position we have adopted it here. Nevertheless the reader not committed to this approach may like to be reminded that many techniques of the sampling-theory school are basically unsound: see the review by Lindley (1971b). In particular the least-squares estimates are typically unsatisfactory: or, in the language of that school, are inadmissible in dimensions greater than two. This follows since, by a well-known device in least-squares theory (see, for example, Plackett, 1960, p. 59), we may write the linear model after transformation in the form $E(z_i) = \xi_i$ for $i \leqslant p$ and $E(z_i) = 0$ for $i > p$. Here the z's are transforms of the data, and the ξ's of the parameters. Adding the assumption of normality, we can appeal to the results of Brown (1966), generalizing those of Stein (1956), which show that for a very wide class of loss functions the estimate of ξ_i by z_i, for $i \leqslant p$ is inadmissible. In Section 1 of this paper we do comment on the admissibility of the Bayesian estimates and try to show, in a way that might appeal to an adherent of orthodox ideas, that they are likely to be superior, at least in some situations, to the least-squares estimates.

1. EXCHANGEABILITY

We begin with a simple example. Suppose, in the general linear model, that the design matrix is the unit matrix so that $E(y_i) = \theta_i$ for $i = 1, 2, \ldots, n$, and that $y_1, y_2, \ldots, y_n$ are independent, normally distributed with known variance σ^2. Such a simple model might arise if y_i was the observation on the ith variety in a field trial, of average yield θ_i. In considering the prior knowledge of the θ_i it may often be reasonable to assume their distribution *exchangeable*. That is, that it would be unaltered by any permutation of the suffixes: so that, in particular, the prior opinion of θ_7 is the same as that of θ_4, or any other θ_i; and similarly for pairs, triplets and so on. Now one way of obtaining an exchangeable distribution $p(\boldsymbol{\theta})$ is to suppose

$$p(\boldsymbol{\theta}) = \int \prod_{i=1}^{n} p(\theta_i|\mu)\, dQ(\mu),$$

where $p(\theta_i|\mu)$, for each μ, and $Q(\mu)$ describe arbitrary probability distributions. In other words, $p(\boldsymbol{\theta})$ is a *mixture*, by $Q(\mu)$, of independent and identical distributions, given μ. Indeed, Hewitt and Savage (1955), in generalization of de Finetti's original result, have shown that if exchangeability is assumed for *every* n, then a mixture is the *only* way to generate an exchangeable distribution.

In the present paper we study situations where we have exchangeable prior knowledge and assume this exchangeability described by a mixture. In the example this implies $E(\theta_i) = \mu$, say, a common value for each i. In other words there is a linear structure to the *parameters* analogous to the linear structure supposed for the observations $\mathbf{y}$. If we add the premise that the distribution from which the θ_i appear as a random sample is normal, the parallelism between the two stages, for $\mathbf{y}$ and $\boldsymbol{\theta}$, becomes closer. In this paper we study the situation in which the parameters of the general linear model themselves have a general linear structure in terms of other quantities which we call *hyperparameters*.† In this simple example there is just one hyperparameter, μ.

Indeed, we shall find it necessary to go further and let the hyperparameters also have a linear structure. This will be termed a *three-stage model* and is analysed in detail in the next section. There are straightforward extensions to any number of stages.

Returning to the simple example with $E(y_i) = \theta_i$, $E(\theta_i) = \mu$ and respective variances σ^2 and τ^2, say, the situation will be completely specified once a prior distribution has been given for μ. (Effectively this is the third stage just mentioned.) Supposing μ to have a uniform distribution over the real line—a situation usually described by saying there is vague prior knowledge of μ—Lindley (1971a) has obtained the posterior distribution of θ_i and found its mean to be

$$E(\theta_i|\mathbf{y}) = \frac{y_i/\sigma^2 + y_./\tau^2}{1/\sigma^2 + 1/\tau^2}, \tag{1}$$

where $y_. = \sum y_i/n$. The detailed analysis has been given in the reference just cited, so we content ourselves with a brief discussion to serve as an introduction to the general theory in the next section.

The estimates, (1), will be referred to as *Bayes* estimates, and it is these that we propose as substitutes for the usual least-squares estimates. We denote them by θ_i^*,

† We believe we have borrowed this terminology from I. J. Good but are unable to trace the reference.

and reserve the usual notation, $\hat{\theta}_i$, for the ordinary estimates. Notice that θ_i^* is a weighted average of $y_i = \hat{\theta}_i$ and the overall mean, $y_.$, with weights inversely proportional to the variances of y_i and θ_i. Hence the natural estimates are pulled towards a central value $y_.$, the extreme values experiencing most shift. We shall find the weighted average phenomenon will persist even within the general model. Of course the estimate (1) depends on σ^2 and τ^2, which will typically be unknown, but their estimation presents no serious difficulties. If, for each i, there is replication of the y_i then σ^2 may be estimated as the usual within variance. Since we have replication (from the distribution $N(\mu, \tau^2)$ underlying the exchangeability assumption) for the θ_i, τ^2 may be estimated. For example $\sum(\theta_i^* - \theta^*)^2/(n-1)$ might be a reasonable estimate of τ^2, although in fact the reference just cited shows this can be improved upon. These estimates of σ^2 and τ^2 can be used in place of the known values used in (1) and the cycle repeated.

Let us now digress from the Bayesian viewpoint and try to persuade an orthodox statistician that (1) is a sensible estimate for him to consider, and indeed is better than the least-squares estimate. Of course, θ_i^* is a biased estimate of θ_i, so its merit cannot be judged by its variance. We use instead the mean-square error $E(\theta_i^* - \theta_i)^2$. This is just a criterion for judging the merit of one of the n estimates, so let us look at the average mean-square error over the n values. Simple, but tedious, calculations enable this to be found and compared with the corresponding quantity for $\hat{\theta}_i$, namely σ^2. The condition for the average m.s.e. for θ_i^* to be less than that for $\hat{\theta}_i$ is that

$$\sum(\theta_i - \theta)^2/(n-1) < 2\tau^2 + \sigma^2. \tag{2}$$

The m.s.e. for θ_i^* depends on θ_i and hence this condition does also. Consequently the Bayes estimates are not always superior to least-squares. But consider when (2) obtains. The θ_i are, by supposition, given μ, τ^2, a random sample from $N(\mu, \tau^2)$ so that the left-hand side of (2) is the usual estimate of τ^2, had the θ_i been known. Hence the condition is that the estimate of τ^2 be less than $2\tau^2 + \sigma^2$. The distribution of the estimate is a multiple of χ^2 and simple calculations show that the chance—according to the $N(\mu, \tau^2)$ distribution—of (2) being satisfied is high for n as low as 4 and rapidly tends to 1 as n increases. But τ^2, as we have seen, can itself be estimated, so with this in (1) we are almost certain to have a smaller m.s.e. for θ_i^* than for $\hat{\theta}_i$. In particular the expectation (over the θ-distribution) is always in favour of the Bayes estimate.

That argument is heuristic. Our estimates are similar to those proposed by Stein (1956), which he rigorously showed to be superior (in the average m.s.e. sense) to the least-squares estimates. It has been pointed out to us by L. Brown (personal communication) that (1), with known σ^2, τ^2, is an admissible estimate. Essentially this is because the impropriety in our prior distribution is confined to one dimension—in μ. We digress to amplify this statement.

If a *proper* prior distribution (that is, one whose integral over the whole space is unity) and a *bounded* utility function are used, then the estimate obtained by using as an estimate that value which maximizes the expected (over the parameter distribution) utility is always admissible. This is easy to demonstrate since, under the two conditions stated, all the usual mathematical operations, such as reversals of order of integration, are valid. Difficulties arise if either of the italicized conditions above are violated. Quadratic loss, leading to m.s.e. is unbounded, but can conveniently be replaced by

$$1 - \exp\{-(\boldsymbol{\theta} - \mathbf{e})^{\mathrm{T}} \boldsymbol{\Lambda} (\boldsymbol{\theta} - \mathbf{e})\} \tag{3}$$

for estimate $\mathbf{e}$, where $\mathbf{\Lambda}$ is positive semi-definite and, in particular, a unit matrix. The use of vague prior knowledge, with a uniform, and therefore improper, prior distribution does cause difficulties and it is this feature, at least in dimensions higher than two, that gives rise to inadmissible estimates, as Stein was the first to show. In the general theory of the next section all our estimates will be admissible in terms of the bounded loss function (3) provided the prior distribution is proper; we conjecture admissibility if the impropriety is confined to at most two dimensions.

Returning, then, to the inequality (2), we see that there is good reason within the orthodox framework for preferring the new estimates to the old. Further justification may be found in papers by Hoerl and Kennard (1970a, b) who discuss a special case of the estimates that we shall develop in Section 5.3. We do not take these justifications very seriously, feeling that the Bayesian viewpoint is supported by so many general considerations in which criteria, like mean-square error, play little or no part, that the additional validation they provide is of small consequence.

Before proceeding to the general discussion one point must be emphasized. In the example we have assumed an exchangeable prior distribution. The estimates (1) are therefore only suggested when this assumption is practically realistic. It is the greatest strength of the Bayesian argument that it provides a formal system within which any inference or decision problem can be described. In passing from the real-world problem to its mathematical formulation it becomes necessary to make, and to expose, the assumptions. (This applies to any formalism, Euclidean geometry, for example, and not just to Bayesian statistics.) Here exchangeability is one such assumption, and its practical relevance must be assessed before the estimates based on it are used. For example, if, as suggested above, our model described the observed yields of n varieties in an agricultural field trial, the exchangeability assumption would be inappropriate if one or more varieties were controls and the remainder were experimental. However, the assumption might be modified to one of exchangeability within controls and separately within experimental varieties. Similarly with a two-way classification into rows and columns, it might be reasonable to assume separately that the rows and the columns were exchangeable. In any application the particular form of the prior distribution has to be carefully considered.

It should be noted that in assigning a prior distribution to the θ_i of the above form, whilst we are effectively regarding them as a random sample from $N(\mu, \tau^2)$, we are not thereby passing to a Model II, random effects, situation such as has been discussed by Fisk (1967) and Nelder (1968). We are interested in the estimation of the *fixed* effects. One of us (A. F. M. S.) has studied the genuine Model II situation and obtained estimates for μ ($\boldsymbol{\theta}_2$ in the general model below) but this will be reported separately.

We now turn to the general theory. The mathematics is not difficult for someone familiar with matrix algebra, and the main result is stated as a theorem with corollaries. The results in Section 2 all assume *known* variances. The extensions to unknown variances will be described later.

2. General Bayesian Linear Model

The notation $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{D})$ means that the column vector $\mathbf{y}$ has a (multivariate) normal distribution with mean $\boldsymbol{\mu}$, a column vector, and dispersion $\mathbf{D}$, a positive semi-definite matrix.

Lemma. Suppose, given $\boldsymbol{\theta}_1$, a vector of p_1 parameters,

$$\mathbf{y} \sim N(\mathbf{A}_1 \boldsymbol{\theta}_1, \mathbf{C}_1) \tag{4}$$

and that, given $\boldsymbol{\theta}_2$, a vector of p_2 hyperparameters,

$$\boldsymbol{\theta}_1 \sim N(\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_2). \tag{5}$$

Then (a) the marginal distribution of $\mathbf{y}$ is

$$N(\mathbf{A}_1\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}), \tag{6}$$

and (b) the distribution of $\boldsymbol{\theta}_1$, given $\mathbf{y}$, is $N(\mathbf{Bb}, \mathbf{B})$ with

$$\mathbf{B}^{-1} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1} \tag{7}$$

and

$$\mathbf{b} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}+\mathbf{C}_2^{-1}\mathbf{A}_2\boldsymbol{\theta}_2. \tag{8}$$

(Here $\mathbf{y}$ is a vector of n elements and $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{C}_1$ and $\mathbf{C}_2$ are known positive-definite matrices of obvious dimensions.)

The lemma is well known but we prove it here, both for completeness and because the proof has an unexpected byproduct.

To prove (a) we write (4) in the form $\mathbf{y} = \mathbf{A}_1\boldsymbol{\theta}_1+\mathbf{u}$, where $\mathbf{u}\sim N(\mathbf{0}, \mathbf{C}_1)$ and (5) as $\boldsymbol{\theta}_1 = \mathbf{A}_2\boldsymbol{\theta}_2+\mathbf{v}$ where $\mathbf{v}\sim N(\mathbf{0}, \mathbf{C}_2)$. Hence, putting these two equalities together, we have $\mathbf{y} = \mathbf{A}_1\mathbf{A}_2\boldsymbol{\theta}_2+\mathbf{A}_1\mathbf{v}+\mathbf{u}$. But, by the standard properties of normal distributions, $\mathbf{A}_1\mathbf{v}+\mathbf{u}$, a linear function of independent normal random variables, is $N(\mathbf{0}, \mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}})$ and the result follows.

To prove (b) we use Bayes's theorem,

$$p(\boldsymbol{\theta}_1|\,\mathbf{y}) \propto p(\mathbf{y}|\,\boldsymbol{\theta}_1)\,p(\boldsymbol{\theta}_1).$$

The product on the right-hand side is $e^{-\frac{1}{2}Q}$ where Q is given by

$$\begin{aligned}(\mathbf{y}-\mathbf{A}_1\boldsymbol{\theta}_1)^{\mathrm{T}}\mathbf{C}_1^{-1}(\mathbf{y}-\mathbf{A}_1\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_1-\mathbf{A}_2\boldsymbol{\theta}_2)^{\mathrm{T}}\mathbf{C}_2^{-1}(\boldsymbol{\theta}_1-\mathbf{A}_2\boldsymbol{\theta}_2)\\ = \boldsymbol{\theta}_1^{\mathrm{T}}\mathbf{B}^{-1}\boldsymbol{\theta}_1-2\mathbf{b}^{\mathrm{T}}\boldsymbol{\theta}_1+\{\mathbf{y}^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}+\boldsymbol{\theta}_2^{\mathrm{T}}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2\boldsymbol{\theta}_2\}\end{aligned}$$

on collecting the quadratic and linear terms in $\boldsymbol{\theta}_1$ together, and using the expressions (7) and (8) for $\mathbf{b}$ and $\mathbf{B}$. Completing the square in $\boldsymbol{\theta}_1$, Q may finally be written

$$(\boldsymbol{\theta}_1-\mathbf{Bb})^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{\theta}_1-\mathbf{Bb})+\{\mathbf{y}^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}+\boldsymbol{\theta}_2^{\mathrm{T}}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2\boldsymbol{\theta}_2-\mathbf{b}^{\mathrm{T}}\mathbf{Bb}\}. \tag{9}$$

The term in braces is a constant as far as the distribution of $\boldsymbol{\theta}_1$ is concerned, and the remainder of the expression demonstrates the truth of (b).

The proof of the lemma is complete, but by combining the separate proofs of (a) and (b) an interesting result can be obtained. On integrating $e^{-\frac{1}{2}Q}$, with Q given by (9), with respect to $\boldsymbol{\theta}_1$, the result is proportional to the density of $\mathbf{y}$, already obtained in (a). The integration does not affect the term in braces in (9) so that, in particular, the quadratic term in $\mathbf{y}$ in (9)—remembering that $\mathbf{b}$ contains $\mathbf{y}$—may be equated to the quadratic term obtained directly from (6), with the result that

$$\mathbf{C}_1^{-1}-\mathbf{C}_1^{-1}\mathbf{A}_1\mathbf{B}\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1} = \{\mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}\}^{-1}.$$

We therefore have the

Matrix lemma. For any matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{C}_1$ and $\mathbf{C}_2$ of appropriate dimensions and for which the inverses stated in the result exist, we have

$$\mathbf{C}_1^{-1}-\mathbf{C}_1^{-1}\mathbf{A}_1(\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1})^{-1}\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1} = \{\mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}\}^{-1}. \tag{10}$$

The result follows from the last equation on inserting the form for $\mathbf{B}$, equation (7). It is, of course, easy to prove the result (10) directly once its truth has been conjectured: furthermore $\mathbf{C}_1$ and $\mathbf{C}_2$ do not have to be positive definite. It suffices to multiply the left-hand side of (10) by $\mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}$ and verify that the result is a unit matrix. The above proof is interesting because it does not require an initial conjecture and because it uses a probabilistic argument to derive a purely algebraic result. The matrix lemma is important to us since it provides simpler forms than would otherwise be available for our estimates. This result has been given by Rao (1965, Exercise 2.9, p. 29).

We next proceed to the main result. As explained in Section 1, we are dealing with the linear model, which is now written in the form $E(\mathbf{y}) = \mathbf{A}_1\boldsymbol{\theta}_1$, the suffixes indicating that this is the first stage in the model. We generalize to an arbitrary dispersion matrix, $\mathbf{C}_1$, for $\mathbf{y}$. The prior distribution of $\boldsymbol{\theta}_1$ is expressed in terms of hyperparameters $\boldsymbol{\theta}_2$ as another linear model, $E(\boldsymbol{\theta}_1) = \mathbf{A}_2\boldsymbol{\theta}_2$ with dispersion matrix $\mathbf{C}_2$. This can proceed for as many stages as one finds convenient: it will be enough for us to go to three, supposing the mean, as well as the dispersion, known at the final stage. For our inferences, and in particular for estimation, we require the posterior distribution of $\boldsymbol{\theta}_1$. This is provided by the following result.

Theorem. Suppose that, given $\boldsymbol{\theta}_1$,

$$\mathbf{y} \sim N(\mathbf{A}_1\boldsymbol{\theta}_1, \mathbf{C}_1), \tag{11.1}$$

given $\boldsymbol{\theta}_2$,

$$\boldsymbol{\theta}_1 \sim N(\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_2) \tag{11.2}$$

and given $\boldsymbol{\theta}_3$,

$$\boldsymbol{\theta}_2 \sim N(\mathbf{A}_3\boldsymbol{\theta}_3, \mathbf{C}_3). \tag{11.3}$$

Then the posterior distribution of $\boldsymbol{\theta}_1$, given $\{\mathbf{A}_i\}$, $\{\mathbf{C}_i\}$, $\boldsymbol{\theta}_3$ and $\mathbf{y}$ is $N(\mathbf{Dd}, \mathbf{D})$ with

$$\mathbf{D}^{-1} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1 + \{\mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}\}^{-1} \tag{12}$$

and

$$\mathbf{d} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y} + \{\mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}\}^{-1}\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3. \tag{13}$$

(Here $\boldsymbol{\theta}_i$ is a vector of p_i elements and the dispersion matrices, $\mathbf{C}_i$, are all supposed non-singular.)

The joint distribution of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is described in (11.2) and (11.3). The use of part (a) of the lemma enables the marginal distribution of $\boldsymbol{\theta}_1$ to be written down as

$$\boldsymbol{\theta}_1 \sim N(\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3, \mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}). \tag{14}$$

(Notice that this is the prior distribution of $\boldsymbol{\theta}_1$ free of the hyperparameters $\boldsymbol{\theta}_2$. We could have expressed the prior in this way but in applications we find the hierarchical form more convenient.)

Then, with (14) as prior, (11.1) as likelihood, part (b) of the lemma shows that the posterior distribution of $\boldsymbol{\theta}_1$ is as stated.

In particular the mean of the posterior distribution may be regarded as a point estimate of $\boldsymbol{\theta}_1$ to replace the usual least-squares estimate. The form of this estimate is a generalization of the form noted in the example of Section 1; namely, it is a weighted average of the least-squares estimate $(\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}$ and the prior mean $\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3$ (equation (14)) with weights equal to the inverses of the corresponding dispersion matrices, $\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1$ for the least-squares values, $\mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}$ for the

prior distribution (14). For the simple example considered in Section 1 we produced an heuristic argument to show that, with respect to our prior distribution, we were confident of satisfying inequality (2) and thus achieving smaller mean square error than with the least-squares estimate. This result can be shown to hold generally for Bayes's estimates derived from hierarchical prior structures, as in (11.1)–(11.3), and will be presented in a future paper.

The matrix lemma enables us to obtain several alternative forms for the term in braces in (12), and hence for the posterior mean and variance, both of which involve this expression. These alternatives look more complicated than those already stated but are often useful in applications. Notice that a computational advantage of the matrix lemma is that its use reduces the order of the matrices to be inverted. The matrix on the right-hand side of (10) is of order n, whereas on the left-hand side, apart from $\mathbf{C}_1$ which is usually of a simple structure (often $\mathbf{C}_1 = \sigma^2\mathbf{I}$), the matrix to be inverted is of order p_1, typically much less than n.

Corollary 1. An alternative expression for $\mathbf{D}^{-1}$ (equation (12)) is

$$\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1}-\mathbf{C}_2^{-1}\mathbf{A}_2(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2+\mathbf{C}_3^{-1})^{-1}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}. \tag{15}$$

This is immediate on applying (10), with the suffixes all increased by one, to the second term in (12).

In most applications of these results the design of the experiment rather naturally suggests the second stage, (11.2), in the hierarchy but at the third stage we find ourselves in a position where the prior knowledge is weak. (Least-squares results apply when the second-stage prior knowledge is weak.) It is natural to express this by supposing the third-stage dispersion matrix $\mathbf{C}_3$ to be large, or to let its inverse, the precision matrix, be zero. In the original form of (12) and (13) it is not easy to see what happens when $\mathbf{C}_3^{-1} = \mathbf{0}$, but (15) enables the form to be seen easily.

Corollary 2. If $\mathbf{C}_3^{-1} = \mathbf{0}$, the posterior distribution of $\boldsymbol{\theta}_1$ is $N(\mathbf{D}_0\mathbf{d}_0, \mathbf{D}_0)$ with

$$\mathbf{D}_0^{-1} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1}-\mathbf{C}_2^{-1}\mathbf{A}_2(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1} \tag{16}$$

and

$$\mathbf{d}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}. \tag{17}$$

The form for $\mathbf{D}_0^{-1}$ follows by direct substitution of $\mathbf{C}_3^{-1} = \mathbf{0}$ in (15). That for $\mathbf{d}_0$ follows by remarking that if the second and third terms in (15) are postmultiplied by $\mathbf{A}_2$ the result is zero, but such postmultiplication takes place in the original expression for $\mathbf{d}$, equation (13).

This corollary is the form we shall most often use in applications.

It is possible to extend the theorem to cases where some or all of the dispersion matrices $\mathbf{C}_i$ are singular. This can be accomplished using generalized inverses and will be the subject of a separate paper. Notice that we have not assumed, as in the usual least-squares theory, that $\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1$ is non-singular. (The case $\mathbf{C}_1 = \sigma^2\mathbf{I}$ will be more familiar.) In the standard exposition it is usual to constrain the individual parameters in the vector $\boldsymbol{\theta}_1$ to preserve identifiability in the likelihood function. Identifiability problems do not arise in the Bayesian formulation since, provided the prior distribution is proper, so is the posterior, whether or not the parameters referred to in these two distributions are identifiable or not in the likelihood function. An example below will help to make this clear.

The situation described in Section 1 has already been discussed in detail by Lindley (1971a), though not within the general framework which was briefly described

in Lindley (1969). The interested reader can easily fit the example into the argument of this section. Corollary 2 is relevant and it is an easy matter to perform the necessary matrix calculations. We proceed to the discussion of other examples.

3. EXAMPLES

3.1. *Two-factor Experimental Designs*

Consider t "treatments" assigned to n experimental units arranged in b "blocks". If the ith treatment is applied within the jth block and yields an observation y_{ij}, the usual model is

$$E(y_{ij}) = \mu + \alpha_i + \beta_j \quad (1 \leqslant i \leqslant t, 1 \leqslant j \leqslant b)$$

with the errors independent $N(0, \sigma^2)$. In the general notation of (11.1)

$$\boldsymbol{\theta}_1^{\mathrm{T}} = (\mu, \alpha_1, \alpha_2, \ldots, \alpha_t, \beta_1, \beta_2, \ldots, \beta_b)$$

and $\mathbf{A}_1$ describes the design used.

For the second stage we argue as follows. It might be reasonable to assume that our prior knowledge of the treatment constants $\{\alpha_i\}$ was exchangeable, and similarly that of the block constants $\{\beta_j\}$, but that these were independent. We emphasize the word "might" in the last sentence. In repetition of the point made in Section 1, we remind the reader that this *assumption* is not always appropriate and our recipes below are not necessarily sensible when this form of exchangeability is unreasonable. For example, it may be known that the treatments are ordered, say $\alpha_1 \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_t$. In this case other forms of prior information are available and alternative estimates are sensible: these will be reported on in a separate paper.

Adding the assumptions of normality we therefore describe the second stage (11.2) by

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad \beta_j \sim N(0, \sigma_\beta^2), \quad \mu \sim N(\omega, \sigma_\mu^2),$$

these distributions being independent. The means of α_i and β_j have been chosen to be zero. Any other value would do since the likelihood provides no information about them, but the choice of zero mean is convenient, since it leads to straightforward comparisons of the Bayes and (constrained) least-squares estimates as deviations from an average level. We shall consider the case where the prior knowledge of μ is vague, so that $\sigma_\mu^2 \to \infty$; ω will then be irrelevant. A third stage is not necessary. We proceed to calculate expressions (12) and (13) for the posterior distribution of $\boldsymbol{\theta}_1$.

The matrix $\mathbf{C}_2$ is diagonal, so the same is true of $\mathbf{C}_2^{-1}$ and its leading diagonal is easily seen to be

$$(\sigma_\mu^{-2}, \sigma_\alpha^{-2}, \ldots, \sigma_\alpha^{-2}, \sigma_\beta^{-2}, \ldots, \sigma_\beta^{-2})$$

and as $\sigma_\mu^2 \to \infty$, the first element tends to zero. $\mathbf{C}_1$ is the unit matrix times σ^2. We can therefore substitute these values into (12) and (13), remembering that $\mathbf{C}_3 = \mathbf{0}$ and $(\mathbf{A}_3 \boldsymbol{\mu})^{\mathrm{T}} = (\omega, 0, \ldots, 0)$ and easily obtain

$$\mathbf{D}^{-1} = \sigma^{-2} \mathbf{A}_1^{\mathrm{T}} \mathbf{A}_1 + \mathbf{C}_2^{-1}$$

and

$$\mathbf{d} = \sigma^{-2} \mathbf{A}_1^{\mathrm{T}} \mathbf{y}.$$

Hence $\boldsymbol{\theta}_1^*$, the Bayes estimate $\mathbf{Dd}$, satisfies the equations

$$(\mathbf{A}_1^{\mathrm{T}} \mathbf{A}_1 + \sigma^2 \mathbf{C}_2^{-1}) \boldsymbol{\theta}_1^* = \mathbf{A}_1^{\mathrm{T}} \mathbf{y}. \tag{18}$$

These differ from the least-squares equations only in the inclusion of the extra term $\sigma^2 \mathbf{C}_2^{-1}$.

In the case of a complete randomized-block design where each treatment occurs exactly once in each block we have, on arranging the elements of $\mathbf{y}$ in lexicographical order,

$$(\mathbf{A}_1^{\mathrm{T}} \mathbf{A}_1 + \sigma^2 \mathbf{C}_2^{-1}) = \begin{pmatrix} bt & b\mathbf{1}_t^{\mathrm{T}} & t\mathbf{1}_b^{\mathrm{T}} \\ b\mathbf{1}_t & (b+\sigma^2/\sigma_\alpha^2)\mathbf{I}_t & \mathbf{J}_{t,b} \\ t\mathbf{1}_b & \mathbf{J}_{b,t} & (t+\sigma^2/\sigma_\beta^2)\mathbf{I}_b \end{pmatrix}, \tag{19}$$

where $\mathbf{1}_m$ is a vector of m 1's, $\mathbf{I}_m$ is the unit matrix of order m and $\mathbf{J}_{m,n}$ is a matrix of order $m \times n$ all of whose elements are 1. As usual

$$(\mathbf{A}_1^{\mathrm{T}} \mathbf{y})^{\mathrm{T}} = (bty_{..}, by_{1.}, \ldots, by_{t.}, ty_{.1}, \ldots, ty_{.b}).$$

Notice that the matrix (19) is non-singular and the solution to (18) is easily seen to be

$$\mu^* = y_{..}, \quad \alpha_i^* = (b\sigma_\alpha^2 + \sigma^2)^{-1} b\sigma_\alpha^2 (y_{i.} - y_{..}), \quad \beta_j^* = (t\sigma_\beta^2 + \sigma^2)^{-1} t\sigma_\beta^2 (y_{.j} - y_{..}). \tag{20}$$

Consequently the estimators of the treatment and block effects (on being measured from the overall mean) are shrunk towards zero by a factor depending on the ratio of σ^2 to σ_α^2 or σ_β^2 respectively. This is in agreement with the result, equation (1), quoted above. Because this is an orthogonal design the magnitude of the "shrinkage" of the treatment effect does not depend on the exchangeability for the blocks, and vice versa. With a non-orthogonal design, such as balanced incomplete blocks, the same remark is not true.

3.2. *Exchangeability Between Multiple Regression Equations*

The following practical example stimulated our extension from the example of Section 1 to the general model, and we shall report on its use in Section 5.2. The context was educational measurement where variables x and y were related with the usual linear regression structure. However the values of the regression parameters depended on the school the student had attended. Novick (personal communication) suggested to us that improved estimates might be obtained for any one school by combining the data for all schools. This is just what the Bayes estimates do, and would seem to be appropriate whenever exchangeability *between* regressions (schools) is a sensible assumption. The mathematics for p regressor variables goes as follows.

Suppose

$$\mathbf{y}_j \sim N(\mathbf{X}_j \boldsymbol{\beta}_j, \mathbf{I}_{n_j} \sigma_j^2) \tag{21}$$

for $j = 1, 2, \ldots, m$ and $\boldsymbol{\beta}_j$ a vector of p parameters: that is m linear, multiple regressions on p variables. In the notation of the Theorem, $\mathbf{A}_1$, expressed in terms of submatrices, is diagonal with $\mathbf{X}_j$ as the jth diagonal submatrix; $\boldsymbol{\theta}_1^{\mathrm{T}}$ is $(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_m^{\mathrm{T}})$ of mp elements. The exchangeability of the individual $\boldsymbol{\beta}_j$ added to normality gives us the second stage as

$$\boldsymbol{\beta}_j \sim N(\boldsymbol{\xi}, \boldsymbol{\Sigma}) \tag{22}$$

say. Here $\mathbf{A}_2$ is a matrix of order $mp \times p$, all of whose $p \times p$ submatrices are unit matrices, and $\boldsymbol{\theta}_2 = \boldsymbol{\xi}$. We shall suppose vague prior knowledge of $\boldsymbol{\xi}$ and use the special form of Corollary 2.

Simple calculations show that $(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2)^{-1} = m^{-1}\boldsymbol{\Sigma}$ and then that

$$\mathbf{C}_2^{-1}\mathbf{A}_2(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}$$

is a matrix of order mp *all* of whose $p \times p$ submatrices are $m^{-1}\boldsymbol{\Sigma}^{-1}$. In the usual way $\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1$, expressed in terms of submatrices, is diagonal with $\sigma_j^{-2}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j$ as the jth diagonal submatrix. The equations for the Bayes estimates $\boldsymbol{\beta}_j^*$ are then found to be

$$\begin{pmatrix} \sigma_1^{-2}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1+\boldsymbol{\Sigma}^{-1} & \cdots & \mathbf{0} \\ \cdots & \sigma_2^{-2}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2+\boldsymbol{\Sigma}^{-1} & \cdots \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \sigma_m^{-2}\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m+\boldsymbol{\Sigma}^{-1} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \\ \vdots \\ \boldsymbol{\beta}_m^* \end{pmatrix} - \boldsymbol{\Sigma}^{-1}\begin{pmatrix} \boldsymbol{\beta}_\cdot^* \\ \boldsymbol{\beta}_\cdot^* \\ \vdots \\ \boldsymbol{\beta}_\cdot^* \end{pmatrix} = \begin{pmatrix} \sigma_1^{-2}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} \\ \sigma_2^{-2}\mathbf{X}_2^{\mathrm{T}}\mathbf{y} \\ \vdots \\ \sigma_m^{-2}\mathbf{X}_m^{\mathrm{T}}\mathbf{y} \end{pmatrix}, \tag{23}$$

where $\boldsymbol{\beta}_\cdot^* = \sum \boldsymbol{\beta}_i^*/m$. These equations are easily solved for $\boldsymbol{\beta}_\cdot^*$ and then, in terms of $\boldsymbol{\beta}_\cdot^*$, the solution is

$$\boldsymbol{\beta}_j^* = (\sigma_j^{-2}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j+\boldsymbol{\Sigma}^{-1})^{-1}(\sigma_j^{-2}\mathbf{X}_j^{\mathrm{T}}\mathbf{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_\cdot^*), \tag{24}$$

a compromise between the least-squares estimate and an average of the various estimates. The example of Section 1 is a special case with $p = 1$.

Noting that $\mathbf{D}_0^{-1}$, given in Corollary 2 (16), may, for this application, be written in the form,

$$\begin{pmatrix} \sigma_1^{-2}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1+\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \vdots & \vdots \\ \mathbf{0} & \sigma_m^{-2}\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m+\boldsymbol{\Sigma}^{-1} \end{pmatrix} - m^{-1}\begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ \vdots \\ \boldsymbol{\Sigma}^{-1} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{pmatrix} (\boldsymbol{\Sigma}^{-1} \ldots \boldsymbol{\Sigma}^{-1})$$

and thus may be inverted by the matrix Lemma (10), we can obtain an explicit form for $\boldsymbol{\beta}_j^*$. After some algebra we obtain the weighted form of (24) with $\boldsymbol{\beta}_\cdot^*$ replaced by $\sum \mathbf{W}_i\hat{\boldsymbol{\beta}}_i$ where,

$$\mathbf{W}_i = \left[\sum_{j=1}^{m}(\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\sigma_j^{-2}+\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\sigma_j^{-2}\right]^{-1}(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\sigma_i^{-2}+\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\sigma_i^{-2}.$$

This shows explicitly how the information from the ith regression equation is combined with the information from all equations.

3.3. *Exchangeability Within Multiple Regression Equations*

In contrast to the last section suppose that we have a single multiple regression situation

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n\sigma^2). \tag{25}$$

In the educational context, the p regressor variables might be the results of p tests applied to students and the dependent variable, y, a measure of the students' performance after training. We are interested in the case where the individual regression coefficients in $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \beta_2, \ldots, \beta_p)$ are exchangeable. To achieve this it may be necessary to rescale the regressor variables: for example, to write (25) in correlation form in which the diagonal elements of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ are unity and the off-diagonals are the sample correlations. (Again we emphasize the point that this is an assumption and may not be appropriate). If the assumption is sensible then we may fit it into our general model by supposing

$$\beta_j \sim N(\xi, \sigma_\beta^2). \tag{26}$$

There are at least two useful possibilities: (i) to suppose vague prior knowledge for ξ (Corollary 2), (ii) to put $\xi = 0$, reflecting a feeling that the β_i are small.

In (i) simple but tedious calculations analogous to those of Section 3.2 show that

$$\boldsymbol{\beta}^* = \{\mathbf{I}_p + k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{I}_p - p^{-1}\mathbf{J}_p)\}^{-1}\hat{\boldsymbol{\beta}}, \tag{27}$$

where $k = \sigma^2/\sigma_\beta^2$. Similar calculations in (ii), using only a two-stage model, give

$$\boldsymbol{\beta}^* = \{\mathbf{I}_p + k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}^{-1}\hat{\boldsymbol{\beta}}. \tag{28}$$

The estimates (27) and (28) are very similar to those proposed by Hoerl and Kennard (1970a). The main difference is that k in their argument is a constant introduced for various intuitively sensible reasons, whereas here it is a variance ratio. Also the derivation is different: Hoerl and Kennard argue within the orthodox, sampling theory framework, whereas we use the formal theory. We do not attempt to reproduce their most convincing argument against the least-squares estimates and in favour of (27) and (28), merely referring the sampling-theorist to it and saying that we agree with its conclusions with the reservation that we feel that the estimates may not be so sensible if the exchangeability within the regression equation is inappropriate. We return to this example in Section 5.3 where the estimation of k is discussed.

Examples 3.2 and 3.3 may be combined when there is exchangeability between *and* within regressions. We omit the details of this and many other extensions and instead consider how we might remove the major impediment to the application of the general theory, namely the assumption that all the variances are known. In the next section we show that the simple device of replacing the known variances by estimated values in the Bayes estimates is satisfactory.

4. Estimation with Unknown Covariance Structure

For the purpose of the immediate exposition denote by θ the parameters of interest in the general model and by ϕ the nuisance parameters. The latter will include the dispersion matrices $\mathbf{C}_i$ when these are unknown. Consider how the Bayesian treatment proceeds. We first assign a joint prior distribution to θ and ϕ—instead of just to θ—and combine this with the likelihood function to provide the joint posterior distribution $p(\theta, \phi \mid \mathbf{y})$. This distribution then has to be integrated with respect to ϕ, thus removing the nuisance parameters and leaving the posterior for θ. Finally, if we are using quadratic loss or generally one of the forms given by (3), we shall require the mean of this distribution, necessitating another integration. The calculation of the mean will also require the constant of proportionality in Bayes's formula to be evaluated, involving yet another integration. Any reasonable

prior distributions for ϕ that we have considered lead to integrals which cannot all be expressed in closed form and, as a result, the above argument is technically most complex to execute. We therefore consider an approximation to it which is technically much simpler and yet yields the bulk, though not unfortunately all, of the information required for the estimation.

The first approximation consists in using the *mode* of the posterior distribution in place of the *mean*. Secondly, we mostly use the mode of the *joint* distribution rather than that of the *θ-margin*. The modal values satisfy the equations

$$\frac{\partial}{\partial\theta}p(\theta,\phi|\mathbf{y})=\frac{\partial}{\partial\phi}p(\theta,\phi|\mathbf{y})=0.$$

These equations may be re-written in terms of conditional and marginal distributions. In particular that for θ may be expressed as

$$\frac{\partial}{\partial\theta}p(\theta|\phi,\mathbf{y})p(\phi|\mathbf{y})=0$$

or, assuming $p(\phi|\mathbf{y})\neq 0$, as

$$\frac{\partial}{\partial\theta}p(\theta|\phi,\mathbf{y})=0. \tag{29}$$

But the conditional density $p(\theta|\phi,\mathbf{y})$ in (29) is exactly what has been found in the general theory of Section 2, where it was shown to be normal, with mode consequently equal to the mean. Hence we have the result that the θ-value of the posterior mode of the joint distribution of θ and ϕ is equal to the mode of the conditional distribution of θ evaluated at the modal value of ϕ. Consequently all we have to do is to take the estimates derived in Section 2 and replace the unknown values of the nuisance parameters by their modal estimates. For example, the simple estimate (1) is replaced by

$$\frac{y_i/s^2+y_./t^2}{1/s^2+1/t^2},$$

where s^2 and t^2 are respectively modal estimates of σ^2 and τ^2. This approach avoids the integrations referred to above. The modal estimates of ϕ may, analogous to (29), be found by supposing θ known, and then replacing θ in the result by their modes.

It is reasonably clear that the approximations are only likely to be good if the samples are fairly large and the resulting posterior distributions approximately normal. Also the approach does not provide information about the precision of the estimates, such as a standard error (of the posterior, not the sampling-theoretic distribution!) would provide. But as a first step on the way to a satisfactory description of the posterior distribution, it seems to go a long way and has the added merit of being intuitively sensible. In practice we shall find it convenient to proceed as follows. For an assumed ϕ calculate the mode $\theta^{(1)}$, say. Treating $\theta^{(1)}$ as known we can find the mode for ϕ, $\phi^{(1)}$ say. This may be used to find $\theta^{(2)}$, and so on. This sequence of iterations typically converges and only involves equations for the modes of one parameter, knowing the value of the other.

We now proceed to apply these ideas to the situations discussed in Section 3. At the moment we have no general theory to parallel that of Section 2. The reason for this is essentially that we do not have an entirely satisfactory procedure for

estimating the dispersion matrix of a multivariate normal distribution. This might appear an odd statement to make when there are numerous texts on multivariate analysis available that discuss this problem. But just as the usual estimates of the means are inadmissible, so are those of the variances and covariances (Brown, 1968), and are, in any case, obtained from unrealistic priors. We hope to report separately on this problem and defer discussion of a general theory.

5. Examples with Unknown Covariance Structure

5.1. *Two-factor Experimental Designs*

We saw in Section 3.1 that there were three variances in this situation: σ^2 the usual residual variance contributing to the likelihood function, and $\sigma_\alpha^2, \sigma_\beta^2$ being respectively the variances of the treatment and block effects. ($\sigma_\mu^2 \to \infty$ so does not enter.) It is first necessary to specify prior distributions for these and this we do through the appropriate conjugate family, which is here inverse-χ^2, assuming the three variances independent. This conjugate family involves two parameters and is sufficiently flexible for most applications. Specifically we suppose

$$\frac{\nu\lambda}{\sigma^2} \sim \chi_\nu^2, \quad \frac{\nu_\alpha \lambda_\alpha}{\sigma_\alpha^2} \sim \chi_{\nu_\alpha}^2 \quad \text{and} \quad \frac{\nu_\beta \lambda_\beta}{\sigma_\beta^2} \sim \chi_{\nu_\beta}^2. \tag{30}$$

The joint distribution of all quantities involved can then be written down as proportional to

$$\begin{aligned}
&(\sigma^2)^{-\frac{1}{2}(n+\nu+2)} \exp\left[-\frac{1}{2\sigma^2}\{\nu\lambda + S^2(\mu, \alpha, \beta)\}\right] \\
&\quad \times (\sigma_\alpha^2)^{-\frac{1}{2}(t+\nu_\alpha+2)} \exp\left[-\frac{1}{2\sigma_\alpha^2}\{\nu_\alpha \lambda_\alpha + \textstyle\sum \alpha_i^2\}\right] \\
&\quad \times (\sigma_\beta^2)^{-\frac{1}{2}(b+\nu_\beta+2)} \exp\left[-\frac{1}{2\sigma_\beta^2}\{\nu_\beta \lambda_\beta + \textstyle\sum \beta_j^2\}\right],
\end{aligned} \tag{31}$$

where $S^2(\mu, \alpha, \beta)$ is the sum of squares $\sum(y_{ij} - \mu - \alpha_i - \beta_j)^2$.

If σ^2, σ_α^2 and σ_β^2 are known, the mode of this distribution has been found—equation (18), or in the balanced case, equation (20). We have only to substitute the modal estimates of the three variances into these expressions. To find these modal estimates we can, reversing the roles of θ and ϕ in the general argument of the previous paragraph, suppose μ, α and β known. Using the corresponding Roman letters for these modes, we easily see them to be, from (31),

$$\left.\begin{aligned}
s^2 &= \{\nu\lambda + S^2(\mu^*, \alpha^*, \beta^*)\}/(n+\nu+2), \\
s_\alpha^2 &= \{\nu_\alpha \lambda_\alpha + \textstyle\sum \alpha_i^{*2}\}/(t+\nu_\alpha+2), \\
s_\beta^2 &= \{\nu_\beta \lambda_\beta + \textstyle\sum \beta_j^{*2}\}/(b+\nu_\beta+2).
\end{aligned}\right\} \tag{32}$$

These equations, together with (18) (or (20)), can now be solved iteratively. With trial values of σ^2, σ_α^2 and σ_β^2, (18) can be solved for μ^*, α^* and β^*. These values can be inserted into (32) to give revised values for s^2, s_α^2 and s_β^2, which can again be used in (18). The cycle can be repeated until the values converge.

A few points about these solutions are worth noting. Firstly, the value of S^2 that occurs is not the usual residual sum of squares, which is evaluated about

the least-squares value, but the sum about the Bayes estimates. Since the former minimizes the sum of squares, our S^2 is necessarily greater than the residual: s^2 could therefore be larger than the usual estimate. Secondly, whilst it would be perfectly possible to put $\nu = 0$ (referring to σ^2), so avoiding the specification of a value for λ and thereby taking the usual vague prior for a variance, one cannot put ν_α and ν_β zero. If this is done the modal estimates for the treatment and block effects are all zero. The point is discussed in detail in connection with the example of Section 1 in Lindley (1971a). Essentially the estimation of σ_α^2 and σ_β^2 is difficult, in the sense that the data contain little information about them, when they are small in comparison with σ^2: the residual "noise" is too loud. In the contrary case where σ_α^2 and σ_β^2 are large in comparison with σ^2, the actual values of ν_α, λ_α, ν_β and λ_β do not matter much provided the ν's are both small.

5.2. *Exchangeability Between Multiple Regression Equations*

We continue the discussion of Section 3.2 but mainly confine our attention to the homoscedastic case where $\sigma_j^2 = \sigma^2$, say, for all j. It is only necessary to specify prior distributions for σ^2 and $\boldsymbol{\Sigma}$, the dispersion matrix of the regression coefficients (equation (22)). As in the last example we suppose $\nu\lambda/\sigma^2 \sim \chi_\nu^2$. The conjugate distribution for $\boldsymbol{\Sigma}$ is to suppose $\boldsymbol{\Sigma}^{-1}$ has a Wishart distribution with ρ, say, degrees of freedom and matrix $\mathbf{R}$. We are not too happy with this assumption but at least it provides a large-sample solution (see the remarks at the end of Section 4). $\boldsymbol{\Sigma}$ and σ^2 are supposed independent.

The joint distribution of all the quantities is now

$$\begin{aligned}
&(\sigma^2)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)\right\} \\
&\quad\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}m} \exp\left\{-\frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\xi})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\xi})\right\} \\
&\quad\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}(\rho-p-1)} \exp\{-\tfrac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\mathbf{R}\} \\
&\quad\times (\sigma^2)^{-\frac{1}{2}(\nu+2)} \exp\{-\nu\lambda/2\sigma^2\},
\end{aligned} \tag{33}$$

assuming $\boldsymbol{\xi}$ to have a uniform distribution over p-space. (The four lines of (33) come respectively from the likelihood, the distribution of $\boldsymbol{\beta}$, (22), the Wishart distribution for $\boldsymbol{\Sigma}^{-1}$ and the inverse-χ^2 for σ^2.) The integration with respect to $\boldsymbol{\xi}$ is straightforward and effectively results in the usual loss of one degree of freedom. The joint posterior density for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Sigma}^{-1}$ is then proportional to

$$\begin{aligned}
&(\sigma^2)^{-\frac{1}{2}(n+\nu+2)} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{m}\{m^{-1}\nu\lambda + (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)\}\right] \\
&\quad\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}(m+\rho-p-2)} \exp\left[-\tfrac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\left\{\mathbf{R} + \sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)^{\mathrm{T}}\right\}\right],
\end{aligned} \tag{34}$$

where

$$\boldsymbol{\beta}_\cdot = m^{-1}\sum_{j=1}^{m}\boldsymbol{\beta}_j.$$

The modal estimates are then easily obtained. Those for $\boldsymbol{\beta}_j$ are as before, equation (24), with $\boldsymbol{\Sigma}$ and σ^2 replaced by modal values. The latter are seen to satisfy

$$s^2 = \sum_{j=1}^{m} \{m^{-1}\nu\lambda + (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_1^*)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j^*)\}/(n+\nu+2),$$

and

$$\boldsymbol{\Sigma}^* = \left\{\mathbf{R} + \sum_{j=1}^{m}(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_\cdot^*)(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_\cdot^*)^{\mathrm{T}}\right\}\Big/(m+\rho-p-2). \tag{35}$$

It is possible in this case, as in Section 5.1 and 5.3, to proceed a little differently and obtain the posterior distribution of the $\boldsymbol{\beta}_j$'s, free of σ^2 and $\boldsymbol{\Sigma}$ and consider the modes of this. This is because the integration of (34) with respect to $\boldsymbol{\Sigma}^{-1}$ and σ^2 is possible in closed form. The result is

$$\left[\sum_{j=1}^{n}\{m^{-1}\nu\lambda + (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)\}\right]^{-\frac{1}{2}(n+\nu)}$$

$$\times\left|\mathbf{R} + \sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)^{\mathrm{T}}\right|^{-\frac{1}{2}(m+\rho-1)}. \tag{36}$$

The mode of this distribution can be used in place of the modal values for the wider distribution. The differentiation is facilitated by using the result that, with $\mathbf{V}$ equal to the matrix whose determinant appears in (36),

$$\frac{\partial}{\partial\boldsymbol{\beta}_i}\log|\mathbf{V}| = 2\mathbf{V}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_\cdot).$$

It is then possible to verify that the modes for $\boldsymbol{\beta}_j$ satisfy the same equations as before, (24), with $\boldsymbol{\Sigma}$ and σ^2 replaced by values given by (35) except that the divisors on the right-hand sides are $(n+\nu)$ and $(m+\rho-1)$ rather than $(n+\nu+2)$ and $(m+\rho-p-2)$.

It is possible to extend this model significantly by reverting to the heteroscedastic case as originally considered, (21). Here we have to specify a joint distribution for the σ_j^2. A possible device is to suppose that, like the means, the σ_j^2 are exchangeable. A convenient distribution to generate the exchangeability is to suppose $\nu\lambda/\sigma_j^2 \sim \chi_\nu^2$. In the context of several means (Section 1) Lindley (1971a) has shown how the estimates of the variances get pulled towards a central value. The details are so similar here that we do not repeat them.

As explained in Section 3.2, it was Novick's suggestion to consider this problem in an educational context, and we conclude this section by briefly reporting on an application that he, in conjunction with Jackson, Thayer and Cole (1972), have made of these results. We are most grateful to them for permission to include the details here. Their analysis used data from the American College Testing Program on the prediction of grade-point average at 22 colleges from the results of 4 tests; namely, English, Mathematics, Social Studies and Natural Sciences. We therefore have the situation studied in this section with $p = 5$ (one variable corresponding to the mean), $m = 22$, and n_j varying from 105 to 739. They used the heterosecdastic model but the basic equations (24) and (35) are essentially as here described. With the substantial amounts of data available the prior constants, ν, λ, ρ and $\mathbf{R}$ scarcely affect the analysis: the first three were taken to be small and changes of origin of the regressor variables effected to make the prior judgment that $\mathbf{R}$ was diagonal. With ρ small the diagonal elements again play little role. We omit details of how the calculations were performed and refer the interested reader to their paper.

Data were available for 1968 and 1969. The approach was to use the 1968 data to estimate the regressions, to use these estimated equations on the 1969 x-data to estimate the y's, the grade-point averages, and then to compare these predictions with the actual 1969 y-values, using as a criterion of prediction the mean of the squares of the differences. This operation was done twice; once with the full 1968 data, and once with a random 25 per cent sample from each College. The results are summarized in Table 1.

TABLE 1

Comparison of predictive efficiency

	Average mean-square error	
	Least-squares	*Bayes*
All data	0·5596	0·5502
25% sample	0·6208	0·5603

The first row refers to the analysis of the whole data and shows that the Bayesian method only reduces the error by under 2 per cent. With such large samples there is little room for improvement. With the quarter sample, however, in the second row of the table, the reduction is up to 9 per cent and most strikingly the error is almost down to the value reached with the least-squares estimates for all the data. In other words, 25 per cent of the data and Bayes are as good as all the data and least squares: or the Bayesian method provides a possible 75 per cent saving in sample size. They also provide details of the comparisons between the two estimates of the regression coefficients. These tend to be "shrunk" towards a common value (for each regressor variable) and in some cases with the quarter sample the shrinkage is substantial.

It would be dangerous to draw strong conclusions from one numerical study but the analysis should do something to answer the criticism of those who have said that Bayesian methods are not "testable". We favour the method because of its coherence, but the pragmatists may like to extend the method of Novick *et al.* to other data sets, remembering, of course, that we have made an assumption of exchangeability, and the method cannot be expected to work when this is unreasonable.

5.3. *Exchangeability Within Multiple Regression Equations*

In this section we briefly indicate how the analysis of Section 3.3 proceeds when σ^2, the residual regression variance, and σ_β^2, the variance of the regression coefficients, are both unknown. As before, we assume that independently

$$\nu\lambda/\sigma^2 \sim \chi^2_\nu, \quad \nu_\beta \lambda_\beta/\sigma_\beta^2 \sim \chi^2_{\nu_\beta}.$$

As in Section 5.2 the integration with respect to ξ, the mean of the β_j's, may be performed and the result is that the posterior distribution of $\boldsymbol{\beta}$, σ^2 and σ_β^2 is proportional to

$$(\sigma^2)^{-\frac{1}{2}(n+\nu+2)} \exp\left[-\frac{1}{2\sigma^2}\{\nu\lambda+(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}\right]$$

$$\times (\sigma_\beta^2)^{-\frac{1}{2}(p+\nu_\beta+1)} \exp\left[-\frac{1}{2\sigma_\beta^2}\left\{\nu_\beta\lambda_\beta+\sum_{j=1}^{p}(\beta_j-\beta)^2\right\}\right], \qquad (37)$$

where

$$\beta_\cdot = p^{-1}\sum_{j=1}^{p}\beta_j.$$

The modal equations are then easily seen to be (the first coming from (23))

$$\left.\begin{aligned}\boldsymbol{\beta}^* &= \{\mathbf{I}_p + k^*(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{I}_p - p^{-1}\mathbf{J}_p)\}^{-1}\hat{\boldsymbol{\beta}},\\ s^2 &= \{\nu\lambda + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*)^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*)\}/(n+\nu+2),\\ s_\beta^2 &= \left\{\nu_\beta\lambda_\beta + \sum_{j=1}^{p}(\beta_j^*-\beta_\cdot^*)^2\right\}\Big/(p+\nu_\beta+1).\end{aligned}\right\} \tag{38}$$

The value of k^* is of course s^2/s_β^2. The marginal posterior distribution of $\boldsymbol{\beta}$ can be obtained in a manner similar to that described in the last section.

We are now in a position to compare our method with that of Hoerl and Kennard (1970b). We have taken the example of a 10-factor, non-orthogonal multiple regression summarized in Gorman and Toman (1966) and re-analysed by Hoerl and Kennard using their ridge regression method. The results are summarized in Table 2.

TABLE 2

10-factor multiple regression example

Estimate	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	k
Least-squares	−0·185	−0·221	−0·359	−0·105	−0·469	0·813	0·285	0·383	0·092	0·094	0·000
Bayes	−0·256	−0·178	−0·326	−0·086	−0·289	0·592	0·195	0·349	0·117	0·116	0·039
Ridge	−0·295	−0·110	−0·245	−0·050	−0·040	0·325	0·050	0·240	0·125	0·125	0·250

As already explained, the main difference between ridge regression and the Bayes approach lies in the choice of $k(=\sigma^2/\sigma_\beta^2)$ in equation (23). This has the value zero for least-squares, is chosen subjectively in the ridge method by selecting it so large that the regression estimates stabilize, and is estimated from the data in the Bayes method. In applying the Bayes method we started with $k^* = 0$ in (38), obtained estimates $\boldsymbol{\beta}^*$, which were then used in the other equations in (38) to obtain s^2 and s_β^2. It was found that 10 iterations were needed until the cycle converged. The solution is fairly insensitive to changes in the small, positive values of ν and ν_β and these were set to zero.

In the case of non-orthogonal data, the least-squares procedure has a tendency to produce regression estimates which are too large in absolute value, of incorrect sign and unstable with respect to small changes in the data. The ridge method attempts to avoid some of these undesirable features. The Bayesian method reaches the same conclusion but has the added advantage of dispensing with the rather arbitrary choice of k and allows the data to estimate it. It will be seen from Table 2 that except for β_1, β_9 and β_{10}, all the estimates are pulled towards zero, the effect being greater with the ridge method than with Bayes, the latter choosing a considerably larger value of k than the data suggest.

ACKNOWLEDGEMENTS

We are very grateful to Melvin R. Novick who first stimulated our interest in these problems, has continually made fruitful suggestions and, with his colleagues, has allowed us to include the example in Section 5.2. We are indebted to the referees for constructive comment on a first draft of this paper. The second author would like to thank the Science Research Council and the Central Electricity Generating Board for financial support during the course of this research.

REFERENCES

BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.*, **37**, 1087–1136.

—— (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Ann. Math. Statist.*, **39**, 29–48.

CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics*, **25**, 617–657.

DE FINETTI, B. (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability* (H. E. Kyburg, Jr and H. E. Smokler, eds), pp. 93–158. New York: Wiley.

FISK, P. R. (1967). Models of the second kind in regression analysis. *J. R. Statist. Soc.*, B, **29**, 266–281.

GORMAN, J. W. and TOMAN, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics*, **8**, 27–51.

HEWITT, E. and SAVAGE, L. J. (1955). Symmetric measures on cartesian products. *Trans. Amer. Math. Soc.*, **80**, 470–501.

HOERL, A. E. and KENNARD, R. W. (1970a). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–67.

—— (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

LINDLEY, D. V. (1969). Bayesian least squares. *Bull. Inst. Internat. Statist.*, **43**(2), 152–153.

—— (1971a). The estimation of many parameters. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 435–455. Toronto: Holt, Rinehart and Winston.

—— (1971b). *Bayesian Statistics, A Review*. Philadelphia: SIAM.

NELDER, J. A. (1968). Regression, model-building and invariance. *J. R. Statist. Soc.*, A, **131**, 303–315.

NOVICK, M. R., JACKSON, P. H., THAYER, D. T. and COLE, N. S. (1972). Estimating multiple regressions in *m*-groups; a cross-validation study. *Brit. J. Math. Statist. Psychology*, (to appear).

PLACKETT, R. L. (1960). *Principles of Regression Analysis*. Oxford: Clarendon Press.

RAO, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Sympos.*, **1**, 197–206. Berkeley: University of California Press.

DISCUSSION ON THE PAPER BY PROFESSOR LINDLEY AND DR SMITH

Dr J. A. NELDER (Rothamsted Experimental Station): I welcome this paper as shedding interesting new light on an old topic, the linear model with normal errors. The authors develop a thoroughly Bayesian argument, but this does not mean that we have to be Bayesians in order to make use of their ideas, as I shall try to show.

The authors have laid great stress on their estimates being *Bayesian* estimates, and have compared their estimates with least-squares estimates which they interpret only within a sampling-theoretic framework. Essentially what they are doing is incorporating in their model extra information about a set of (say) population means to the effect that they can be taken as a random sample from a "hyper-population". Such situations undoubtedly occur: e.g. the authors' own example about a random subset of the many lines produced in a breeding program. The same notion underlies designed experiments in incomplete blocks, where the blocks of the design are allocated at random to the finite population in the field. Here the Lindley-Smith estimates of the treatment effects when the block effects are assumed exchangeable but nothing is assumed about the treatments are the familiar

ones obtained from the standard analysis using both inter-block and intra-block information. They have replaced the traditional one-stage model

$$y = \beta_i + t_j + e_{ij},$$

where

$$\beta_i \sim N(\mu, \sigma'^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma^2),$$

with a two-stage model

$$y_{ij} \sim N(\beta_i + t_j, \sigma^2),$$

where

$$\beta_i \sim N(\mu, \sigma'^2).$$

Here the Bayesian analysis of the two models gives identical estimates of t_j and these are the same as the estimates first derived by Yates using maximum likelihood with the one-stage model. The interesting question now arises: what is the corresponding maximum-likelihood procedure for the two-stage model? I want to take a point of view neither Bayesian nor sample-theoretic and look at inferences from the likelihood function. How do we cope with two-faced parameters like the β_i which are both random and fixed at the same time? The basic idea for incorporating prior information, due to Edwards (1969), is that we should express it by means of *prior likelihoods*, rather than by prior probabilities. The distinction is non-trivial; the use of prior likelihoods has the property that we do not close the universe of possible models, because we specify only the *relative* weights to be attached to the alternatives. Further alternatives can always be added. In the simple model with

$$y_i \sim N(\theta_i, \sigma^2),$$
$$\theta_i \sim N(\mu, \tau^2)$$

the log likelihood relevant to the location parameters is just

$$-\tfrac{1}{2}[\textstyle\sum (y_i - \theta_i)^2/\sigma^2 + \sum (\theta_i - \mu)^2/\tau^2],$$

where the second term is the prior likelihood expressing the fact that the θ_i's are regarded as a random sample from a normal population of variance τ^2. The ML estimators are just

$$\hat{\mu} = y_.,$$
$$\hat{\theta}_i = (y_i/\sigma^2 + y_./\tau^2)/(\sigma^{-2} + \tau^{-2}),$$

i.e. the same as the authors'. Applying this procedure to the incomplete-block example produces the same t_j estimates as before and β_i estimates equal to those of the authors.

This parallelism can be carried through a large part of the paper, and I conclude that though the sampling-theory school may be caught in the toils of inadmissibility, and perhaps deserve the authors' strictures, this does not mean that we have to be Bayesians to make use of their results.

I want to say something about ridge regression. The authors say of Hoerl and Kennard's papers that "they do not attempt to reproduce their most convincing argument against the least-squares estimates . . .". This is a pity because I cannot find any convincing arguments in those papers; they measure the deviation of $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{\beta}$ in terms of the Euclidean distance, but it is hard to see why they should unless there is some underlying idea of exchangeability. When there is, the procedures of this paper can be used and provide a justification for their procedure; when there is not we have to accept the least-square estimates *with their associated information surface*. It is true that they may be very sensitive to small changes in the data, but this is not an indictment of least squares, rather it is an indictment of the *data* used for fitting that particular model. The data are generating information in what is almost a subspace of the parameter space, and it is

important that this should be recognized by the experimenter and the necessary action taken. As for estimates with the wrong sign, this usually implies inadequate scaling of the parameters in the model. It is interesting in this context that Marquardt, who developed the equivalent of the ridge regression method in optimization algorithms to cope with near-singular surfaces, has recently concluded (Marquardt, 1970) that the use of generalized inverses, i.e. of acceptance of a subspace of estimability, is equally effective. As Marquardt says, "the generalized inverse method confines the solution to a linear subspace containing the origin, whereas the ridge method confines the regression vector to a sphere about the origin".

Finally a comment on the predictive efficiency of the new estimates as given in Table 1. It is hard to assess their success without some more knowledge of the internal consistency of the data in this respect, but taking them at their face value I would interpret them to mean that a better model gave better predictions. To paraphrase Dr Lehrer, mathematician and singer of songs at the piano, "a model is like a sewer, what you get out of it depends on what you put into it".

It will be clear, I hope, that I have very much enjoyed this paper, and so have real pleasure in proposing the vote of thanks.

Dr V. D. BARNETT (University of Newcastle upon Tyne): It is a pleasure to second the vote of thanks to the authors, on an interesting and important paper. The linear model features as a central one for the application of statistical method, in unifying a vast range of practical problems including typically the various analysis of variance and regression situations. Much of what the authors might term our "orthodox" heritage of statistical methodology stems from applying the principle of least-squares to the linear model, or more particularly from linear hypothesis theory with an assumed normally distributed error structure. Anyone concerned with the comparative aspects of statistical inference, or simply wishing to explore the variety of methods which might be brought to bear to draw meaningful inferences in this area, must surely welcome this evening's detailed study of estimation for the linear model from the Bayesian viewpoint. It casts a new light on an old and important problem.

The authors themselves make it quite clear that they do not regard this work as the final word on the Bayesian analysis of estimation for the linear model. It provides a basic framework for extension in various respects, some of which are to be considered in promised further work. At the same time the basic assumptions underlying the present work are rather particular ones and demand careful scrutiny with regard both to practical application and formal justification. I should like to address a few remarks to these questions of basic premises and implementation.

The principle of least squares for parameter estimation has a rather unique universality, in making no distributional assumptions about the error structure. At this basic level its sample-specific nature may not be entirely satisfying, and we seek consolation in any asymptotic optimality properties that accrue when the error structure is normal. We have then a *general* principle of estimation; with inherited characteristics in a special case. We might ask to what extent it is possible to parallel this from the Bayesian viewpoint, in promoting a *general* principle of estimation reflected against the hierarchical normal structure? In initiating their study with a discussion of de Finetti's concept of *exchange ability*, it was tempting to believe that the authors would be placing crucial practical emphasis on what for long has seemed a vital idea. I must confess to being somewhat confused, however, in the later parts of the paper as to just how central this concept is in their work. To what extent, if any, can we attribute importance to exchangeability, *per se*, as an ingredient in Bayesian estimation? It figures in delimiting the normal hierarchical structure later used—and is offered as a *natural* expression of a certain form of prior information in certain practical situations—but it is in no way specific to the normal structure. Are any desirable features of the estimates to be credited to exchangeability, or to normality? In this respect some of the claims of Section 1 seem rather generous: in

promising study of exchangeability as expressed by mixtures applied to linear models where the parameters themselves have linear structures in relation to hyperparameters. It does seem that only a very special form of this has been presented. Can we go further; for example without the normality assumption?

The bulk of the paper considers the known variance situation. Whilst much of the general theory of Section 2 has already appeared in Lindley (1971a) its fuller treatment is illuminating, and clearly illustrated in the market-place applications of Section 3. But in real-life applications we will not know the variance structure, and it seems that the major feature of this paper, in terms both of its novelty and applicability, must be its consideration of the case of unknown variance structure. The modal invariance property is a fascinating one, and a useful aid to Bayesian analysis of such multi-parameter problems. The suggestion, however, that estimation of unknown variances "presents no serious difficulties" seems optimistic and hardly fully substantiated; specifically the claim at the end of Section 3 that "the simple device of replacing the unknown variances by estimated values in the Bayes estimates is satisfactory". Even in the simple example of Section 1 the proposed iterative estimation of τ^2 is not entirely convincing—even less so for the more complex situations described later. Two issues are involved which cause concern and surely need further comment or study. The first is the question of the *stability* of such iterative procedures. What do we really know about this in the context of the Bayesian estimation problems of this paper? Experience of iterative methods applied to "orthodox" multiparameter problems suggests that we should hardly be surprised to encounter anomalous behaviour. The second matter that seems to require attention is the question of how far we should proceed through the normal hierarchy before (that is, at what "stage") we call a halt to hyperparameterization in the unknown variances case.

A simple example (somewhat similar to that in Section 1) seems to highlight this. Suppose $\mathbf{y}$ is an independent random sample of size n from $N(\theta_0, \sigma^2)$ where σ^2 is known. In the spirit of this paper we might take a two-stage model where θ_0 has a prior $N(\mu, \tau^2)$ distribution. If μ, τ^2 are known, then θ_0 has posterior distribution

$$N(\Delta\{(n\bar{y}/\sigma^2)+(\mu/\tau^2)\}, \Delta), \quad \text{with} \quad \Delta = (\tau^{-2}+n\sigma^{-2})^{-1}.$$

But suppose μ, τ^2 are not known and instead we have k previous "true" values of θ; $\theta_1, \ldots, \theta_k$ (a sample from $N(\mu, \tau^2)$). This could arise in industrial batch production problems, where batches of components are used in assembly and as a result precise values of the means for k earlier batches are elicited. Here we do not even need to *estimate* $\theta_1, \ldots, \theta_k$. Following the authors we might approximate the posterior distribution of θ_0 by substituting $t^2 = \sum(\theta_i - \theta_\cdot)^2/(k-1)$ for τ^2 and $\theta_\cdot$ for μ. But t^2 is only an *estimate* of τ^2 and exhibits sampling fluctuations which are not taken into account in any way. When k is small these can be important (if k is large, less so); the estimation of θ_0 surely needs to reflect this. Similar difficulties exist for the example of Section 1 but in more extreme form since here we do not know the θ_i but must estimate them iteratively. It might therefore seem more appropriate to introduce a third stage with a prior distribution for (μ, τ^2) proportional to τ^{-2} (expressing prior ignorance) and to "up date" this by $\theta_1 \ldots \theta_k$ to yield a prior distribution for θ_0 which in turn is augmented by $\mathbf{y}$ to produce the posterior distribution

$$\pi(\theta_0 \mid \mathbf{y}, \boldsymbol{\theta}) \quad \propto \quad \exp\left\{-\frac{n}{2\sigma^2}(\theta_0-\bar{y})^2\right\} \Big/ \left\{1+\frac{k(\theta_0-\theta_\cdot)^2}{(k^2-1)\,t^2}\right\}^{k/2}.$$

These two posterior distributions are quite different in form (though coincident when $k \to \infty$) and in certain cases, particularly when k or t^2 are small, *resulting inferences differ greatly*; both modal estimates and Bayesian confidence intervals. (Modal estimates are surprisingly similar, however, unless k or t^2 are very small.) How are we to choose between the two approaches? In a sense the three-stage model has more appeal. Why not always proceed to the stage where we can introduce prior ignorance assumptions, and avoid the iterative estimation?

There is one small point I should like to raise. On various occasions reference is made to the use of the *mean* of the posterior distribution as a natural estimator, and the *mode* is later proposed as an "approximation". Leaving aside their coincidence in normal distributions, is not the *mode* the more natural Bayesian estimator (as the "most likely" value) and the *mean* only supported if we superimpose a special form of loss structure?

Finally, I should like to comment on comparisons and pre-requisites. The authors firmly state at the outset their resolve to avoid further proliferation of comparative discussion of the relative merits of Bayesian and non-Bayesian methods. And yet they immediately appear to abandon this attitude, and the "orthodox" statistician finds himself frequently enlightened or instructed. I do not want to dwell on the "new lamps for old" arguments, but I cannot help feeling that the attributed or adopted criteria are not always entirely fair. I should not be surprised to find later discussants taking issue on some of these points. However, there is one constructive aspect of this matter on which I should appreciate the authors' comments. No statistician can afford to avoid recognizing the basic assumptions of his model or of the techniques he uses. Pragmatic assumptions may be necessary on occasions and concepts and criteria must have an element of arbitrariness about them. This is true of orthodox methods as well as Bayesian methods. Unfortunately, what is "unreasonable" to one person may be "obviously true" to another. But in so far as someone may reject outright a Bayesian approach there is undoubtedly an element of "preferring the devil they know"—in the sense of having built up a practical feeling for the dangers which may arise from the uncertain, or dubious, elements of the orthodox approach. Inevitably there is less experience of this from the Bayesian standpoint—and the authors would provide a valuable service if they could tell us something of the real nature of their "devils". They warn us to seriously beware of applying the methods of this paper if exchangeability, or normality, is unreasonable. But how do we assess this? In their examples even, I find it most difficult to assess the propriety of exchangeability or normal prior distributions on parameters or hyperparameters. What guidance have we on these matters?

In conclusion may I say how much I have enjoyed this paper. It provides much food for thought through cogent argument and compelling illustrations. Undoubtedly it is a vital contribution to the Bayesian study of an important topic and should stimulate useful further work in this area. I am most pleased to second the vote of thanks.

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

Professor CEDRIC A. B. SMITH (Galton Laboratory, University College London): Imagine that we are collecting information about the heights of men in various African tribes. In tribe T_i the expected mean height is θ_i, the expected variance v_i, the observed mean $\bar{y}_i$ in a sample of m_i individuals. We are interested in the values of the θ_i. We simplify the discussion by assuming all m_i reasonably large and equal (with common value m), all distributions within tribes to be nearly Gaussian (as will be in any case true for the $\bar{y}_i$ by the central limit theorem), and (less realistically) that all variances v_i are equal, with common value v (which is known or estimated with sufficient accuracy). We can concentrate attention on the $\bar{y}_i$ which have variance $\sigma^2 = v/m$.

The θ_i themselves in the various tribes will have a distribution, strictly speaking discontinuous but representable nearly enough as $\phi(\theta)\,d\theta$. In the authors' treatment this is assumed Gaussian. But this seems less than realistic in many practical cases.

The question arises, can we regard the θ_i in the tribes actually examined as a random sample from the distribution ϕ? Strictly speaking, this is almost never the case. We do not select tribes by formal randomization procedure. But we may feel justified in analysing the data *as if* it were a random sample. This may be called a "random effects model" or "exchangeability". Despite the authors' protestations, the distinction between these terms seems rather fine, but note that it involves an act of judgment, i.e. subjective and not objective probabilities, even in the random effects model.

At first sight the "pulling-in" of the estimates appears to contradict common sense; one would think it obvious that in each tribe the "best" estimate of θ_i is $\bar{y}_i$. If one could assume a very nearly flat prior distribution for ϕ, this would indeed be so, and it is tempting to think that the "pulling-in" found by the authors is mainly due to a rather severe assumption of a Gaussian form for ϕ. However, the following argument shows that this is not so. Note that if we have only a small number n of tribes, we will not (in general) know the exact form of ϕ, and we will have to proceed as the authors do by using a mixture of various plausible forms ϕ_j for ϕ, each with its own prior probability (or density). However, for the sake of discussion, let us assume that we have examined a large number n of tribes, so that we can identify ϕ with sufficient accuracy; also that we have a large number of observations in each tribe, so that $\sigma^2 = v/m$ is small. The posterior probability density for θ_i is then of the form

$$\text{const} \times \sigma^{-1} \exp\{-(\bar{y}_i - \theta_i)^2/2\sigma^2\}\, \phi(\theta_i).$$

Suppose that, near $\bar{y}_i$, we have $\ln \phi(\theta) = A_i + B_i(\theta - \bar{y}_i) +$ negligible terms, where

$$B_i = \left\{\frac{d \ln \phi(\theta)}{d\theta}\right\}_{\theta=\bar{y}_i} \simeq \frac{\phi'\theta_i)}{\phi(\theta_i)}.$$

Then, on substituting this into the expression above, we see that to this order of approximation the posterior distribution is Gaussian with mean (and mode)

$$\theta_i^* = \bar{y}_i + \sigma^2\, \phi'(\theta_i)/\phi(\theta_i)$$

and with variance σ^2. The second term here represents the "pulling-in" effect.

What is the distribution of the θ_i^*, taken over all i? Its expected mean is

$$E(\theta_i^*) = E(\bar{y}_i) + \sigma^2\, E\{\phi'(\theta_i)/\phi(\theta_i)\}.$$

But $E(\bar{y}_i) = E(\theta_i) = \mu$, the mean of the distribution ϕ. And

$$E\left(\frac{\phi'}{\phi}\right) = \int_{-\infty}^{\infty} \frac{\phi'(\theta)}{\phi(\theta)}\, \phi(\theta)\, d\theta = [\phi(\theta)]_{-\infty}^{\infty} = 0.$$

So $E(\theta_i^*) = \mu$, i.e. the "pulling-in" does not affect the mean. It simplifies the calculation of the variance to take measurements from the mean μ, so that we may assume $E(\theta_i^*) = 0$. Then, neglecting terms in σ^4, and setting var $\theta_i = \tau^2$,

$$\text{var}\, \theta_i^* = E(\theta_i^{*2}) \simeq E(\bar{y}_i^2) + 2\sigma^2\, E\{\bar{y}_i\, \phi'(\theta_i)/\phi(\theta_i)\}.$$

But $E(\bar{y}_i^2) = \text{var}\, \bar{y}_i^2 = \tau^2 + \sigma^2$. And, since $\bar{y}_i \simeq \theta_i$,

$$E\left\{\frac{\bar{y}_i\, \phi'(\theta_i)}{\phi(\theta_i)}\right\} \simeq \int_{-\infty}^{\infty} \theta \frac{\phi'(\theta)}{\phi(\theta)}\, \phi(\theta)\, d\theta = \int_{-\infty}^{\infty} \theta \phi'(\theta)\, d\theta.$$

Suppose that $\theta\phi(\theta) \to 0$ as $\theta \to \pm\infty$, as is true apart from very unusual distributions. Then on integrating by parts we find

$$\int_{-\infty}^{\infty} \theta \phi'(\theta)\, d\theta = 0 - \int_{-\infty}^{\infty} \phi(\theta)\, d\theta = -1.$$

Hence var $\theta_i^* = \tau^2 - \sigma^2$. Since σ^2 is not being neglected, this means that the variance of the θ_i^* is appreciably smaller than that of the $\bar{y}_i$, whatever the form of the distribution ϕ. If we regard the θ_i^* as sensible estimates for the θ_i, this means that for any form of the distribution ϕ there will be an appreciable "pulling-in" from the straightforward sample means $\bar{y}_i$; that is, this phenomenon cannot be attributed to the special Gaussian form used by Lindley and Smith.

I would like to add my agreement to Dr Nelder's suggestion that Lindley and Smith's

Mr T. LEONARD (University College London): I would like to make a few remarks about the possible extension of the excellent ideas expressed in this paper to situations where the exchangeable parameters cannot be considered to be normally distributed. In such circumstances, a good procedure is usually to transform the parameters in such a way that the normality assumption is more realistic for the new parameters.

A simple example occurs when the data x_i are independent and binomially distributed, with parameters θ_i and indices n_i $(i = 1, ..., m)$. It is convenient to transform to the log-odds α_i, where

$$\alpha_i = \log\{\theta_i/(1-\theta_i)\}.$$

Under the exchangeability assumption, it is reasonable to suppose that, given μ and σ^2, the α_i are independent and normally distributed with common mean μ and variance σ^2, where μ is uniformly distributed, and σ^2 possesses an inverse χ^2 distribution.

The prior to posterior analysis is straightforward, and when σ^2 is known, it is easily shown that the posterior modes of the α_i are given by the solutions to the equations

$$\frac{e^{\alpha_i}}{1+e^{\alpha_i}} = \frac{x_i}{n_i} - \frac{\alpha_i - \alpha_.}{n_i \sigma^2} \quad (i = 1, ..., m);$$

when σ^2 is unknown it is replaced by its modal estimate.

Unless x_i is close to zero or n_i, approximations are given by

$$\alpha_i = \frac{\sigma^2 l_i + v_i \alpha_.}{\sigma^2 + v_i},$$

where

$$l_i = \log\{x_i/(n_i - x_i)\},$$
$$v_i = x_i^{-1} + (n_i - x_i)^{-1}.$$

These provide similar weighted average forms to those described in the paper.

Logarithmic transformations and exchangeable normal prior distributions may also be employed in the analysis of Poisson or multinomial data. They may also be applied to the situation described on p. 3 in the main paper, where there are normal data with exchangeable variances. In all these cases, simple equations may be found for the exact modes, and weighted average forms may be approximated.

Professor M. R. NOVICK (University of Iowa): Within the time and space limitations imposed on them, Lindley and Smith have done a commendable job of laying out in proper mathematical form a model of intriguing complexity. It is a pity that we cannot carry on into the early hours of the morning, examining the remarkable theoretical implications of this model and, even more excitingly, the applications that can be made of these mathematical results. Such discussions are available, however, and I commend to the attention of this audience a series of reports on these topics. Some of these have or will shortly appear (Novick and Jackson, 1970; Jackson *et al.*, 1971; Novick *et al.*, 1971; Novick *et al.*, 1972). These papers contain elaborate praise for the remarkable work presented here this evening so that it seems appropriate that I concentrate now on whatever faults I can invent for tonight's paper.

First, Lindley and Smith have conveniently concealed the difficulties in actually getting out the required estimates. In the applications we have so far studied, modal estimates can be obtained by the iterative solution of what we have called *Lindley equations*. There are, however, definitely problems of bimodality. These can generally be handled, but only by giving careful attention to the prior distribution of the hyperparameters. It is not, however, a trivial exercise as Lindley and Smith have implied.

Secondly, I disagree with the choice of terminology as to whether all of this is Model I or Model II. I do not think that there is necessarily a right answer to this, but exchangeability does imply the equivalence of a random sampling assumption for parameters and this I (and Box and Tiao, 1968) would call the Random Effects Model.

Lindley and Smith have warned that we must have exchangeability in our prior beliefs to apply these methods. I would elaborate this by saying that we should have an informed exchangeability and not one implied by ignorance. In the cited empirical study, I personally went to great lengths to attain an informed exchangeability, with gratifying results. Later, for a very simple model, I was able to obtain an explicit formula showing that the increase in precision using this method was a function of the between group homogeneity of the parameters.

Let me further remark that this assumption of exchangeability is a very questionable one when imposed across variables rather than across groups. I think that application of the work discussed in Section 5.3, both Bayesian and classical, must be done very carefully. My own feeling is that further theoretical work needs to be done in this area, possibly with a principal components or orthogonal factor analytic model, in a *reduced* variable space. Here it may be more reasonable to make a valid assumption of exchangeability. If so, we may finally be able to do a really good job of estimating a single covariance matrix.

As a final remark, let me chide Lindley and Smith for not having referenced the work of Truman Kelley (1927) who, more than forty years ago used, in an educational measurement context, methods that are very similar to those later adopted by Robbins, by Stein (1966) and by those of us now working within a Bayesian framework with exchangeable prior distributions.

Professor D. R. Cox (Imperial College): Prior information enters regression analysis in the choice of explanatory variables and of the form of relation to be fitted. Once these are provisionally fixed, prior information may affect the estimation of parameter values, some of the main types of such information being (a) knowledge of the sign of one or more coefficients; (b) a relation with analogous parameters in other sets of data, e.g. that if β_p and β_p^* are the current and previous values respectively, then $\beta_p \geqslant \beta_p^*$ or $|\beta_p - \beta_p^*| \leqslant d_p$ or $\beta_p - \beta_p^*$ is a random variable of known distribution; (c) statements about the interrelations of the explanatory variables, such as used in path analysis; (d) the assumption that certain parameters are linked by having been generated by a common physical random mechanism; (e) the postulation of subjective prior distributions for parameters. Points (d) and (e) are linked mathematically, but conceptually are quite different, of course.

It seems to me highly desirable that if any such assumptions are introduced they should as far as possible be tested from the data, either graphically or by significance tests. Failure to do so may mean overlooking inconsistency arising from biased data, from the omission of important explanatory variables or from the prior assumption being misconceived.

In sampling theory, possibility (d) can often be represented as a fully parametrized empirical Bayes problem. That is, the vector observation $\mathbf{Y}$ has density $f_{\mathbf{Y}|\Theta}(\mathbf{y} \mid \theta; \phi)$ depending on unknown parameters ϕ and on parameters Θ with a density $f_\Theta(\theta; \psi)$ depending on further parameters ψ. We can thus find $f_{\mathbf{Y}}(\mathbf{y}; \phi, \psi)$ and obtain estimates $\hat{\phi}, \hat{\psi}$, for example by maximum likelihood. Now if we are interested in some components Θ_1 of Θ we can, for given ϕ, ψ, obtain the posterior density $f_{\Theta_1|\mathbf{Y}}(\theta_1 \mid \mathbf{y}; \phi, \psi)$. This suggests the estimate

$$\hat{f}(\theta_1) = f_{\Theta_1|\mathbf{Y}}(\theta_1 \mid \mathbf{y}; \hat{\phi}, \hat{\psi})$$

or of improved estimates based on refining this. The authors in a fully Bayesian approach have been able to integrate out ϕ, ψ instead of having to substitute point estimates.

Professor R. L. Plackett (University of Newcastle-upon-Tyne): I would like to emphasize the atmosphere of unity which has been such a welcome feature of this evening's discussion by deriving the authors' results from sampling theory without the use of prior distributions or prior likelihoods.

If we take the two-stage linear model with which they begin, then equations (7) and (8) can be obtained by combining the original set of observations $\mathbf{y} \sim N(\mathbf{A}_1 \boldsymbol{\theta}_1, \mathbf{C}_1)$ with an additional set $\mathbf{A}_2 \boldsymbol{\theta}_2 \sim N(\boldsymbol{\theta}_1, \mathbf{C}_2)$. It is no coincidence that equation (10), in the special case when $\mathbf{C}_2 = \mathbf{I}$, appears in the section of Plackett (1950) which is concerned with adjustments to the least-squares estimators, their dispersion matrix and the sum of squared residuals, due to the appearance of additional observations. If we extend this argument to the three-stage model, then we have to combine the last two stages. In that case, we are combining $\mathbf{y} \sim N(\mathbf{A}_1 \boldsymbol{\theta}_1, \mathbf{C}_1)$ with $\mathbf{A}_2 \mathbf{A}_3 \boldsymbol{\theta}_3 \sim N(\boldsymbol{\theta}_1, \mathbf{C}_2 + \mathbf{A}_2 \mathbf{C}_3 \mathbf{A}_2^{\mathrm{T}})$. When the authors come to apply the results of Section 2, their use of exchangeability implies that the additional data $\mathbf{A}_2 \boldsymbol{\theta}_2$ or $\mathbf{A}_3 \boldsymbol{\theta}_3$ have the form $\xi \mathbf{1}$. These relationships explain why the estimators take a weighted form throughout the paper, and why they are closer together than the least-squares estimators.

An attempt to estimate how much data is actually introduced in this way can be made from an analysis of Table 1. Consider the first column. Suppose that the variance of an individual observation is σ^2, and the variance of a predicted value from all the data is $a\sigma^2$. Then

$$\sigma^2(1+a) = 0{\cdot}5596 \quad \text{and} \quad \sigma^2(1+4a) = 0{\cdot}6208,$$

whence $a = 0{\cdot}04$, equivalent to 25 observations, and $\sigma^2 = 0{\cdot}54$. Let Bayes and all the data predict with variance $b\sigma^2$. Then $b = 0{\cdot}02$, equivalent to 50 observations. The conclusion from this posterior analysis is that the effect of using a Bayesian model is to double the amount of information in the data.

Professor P. SPRENT (University of Dundee): The tendency to pull towards the central value gives the estimate in equation (1) a clear advantage over the least-squares estimate. As contributors to this discussion have already pointed out the least-squares estimates are too dispersed. This is evident if we write $y_i = \mu + \delta_i + \epsilon_i$, where $\theta_i = \mu + \delta_i$ and the ϵ_i are $N(0, \sigma^2)$ while the θ_i are $N(\mu, \tau^2)$; assuming independence, the y_i are $N(\mu, \sigma^2 + \tau^2)$, and thus have greater dispersion than the θ_i. The more extreme y_i correspond to cases where δ_i and ϵ_i have the same sign and thus overestimate the magnitude of δ_i.

If we take $\mu = 0$ for convenience and assume n is large (but there is only one replicate for each δ_i) then if $\tau^2 = \sigma^2 = 1$ and $\delta_i = 1$, the probability is approximately 0·95 that y_i (i.e. $\hat{\theta}_i$) will lie between -1 and 3. Assuming $y_\cdot = 0$; for large n when $\mu = 0$, the Bayes estimate θ_i^* will lie between $-0{\cdot}5$ and 1·5 with the same probability. In this case, despite its bias, θ_i^* seems preferable to $\hat{\theta}_i$. If now $\tau^2 = 1$, $\sigma^2 = 9$ and $\delta_i = 1$, then with probability approximately 0·95, $\hat{\theta}_i$ will lie between -5 and 7 while θ_i^* will lie between $-0{\cdot}5$ and 0·7. It is clear that with a probability very close to unity θ_i^* will in this case be less than the true value. The practical man would, one hopes, faced with these values of τ^2 and σ^2, realize the futility of trying to find out anything useful about δ_i from a single replicate corresponding to each δ_i. The ratio $\lambda = \sigma^2/\tau^2$ is a useful guide both to this futility and the seriousness of bias in the Bayes estimate. Is it too frivolous to call it a *coefficient of stupidity*? If $\tau^2 = 9$ and $\sigma^2 = 1$, λ is small; taking $\delta_i = 1$ we then have with probability approximately 0·95 that $\hat{\theta}_i$ lies between -1 and 3 while θ_i^* lies between $-0{\cdot}9$ and 2·7; the Bayes estimate again is reasonable.

Clearly when $\tau^2 = 0$ all $\delta_i = 0$ and $y_\cdot$ is the appropriate estimate of μ in either case. Similarly when $\sigma^2 = 0$, both estimates reduce to y_i, the exact value of $\mu + \delta_i$.

I suggest that when bias in the Bayes estimate is large it cannot be dismissed lightly, but reflects a high value of the coefficient of stupidity, λ. If there are r replicates for each y_i in (1), having mean $y_{i\cdot}$, we may conjecture that σ^2 in (1) should be replaced by σ^2/r, giving $\lambda = \sigma^2/(r\tau^2)$, and the Bayes estimates would approach more closely the least-squares estimates as r increased. A study of degree of bias might prove illuminating in the more sophisticated examples in this paper.

Dr J. B. COPAS (University of Essex): I am not sure what the "orthodox statistician" is supposed to be assuming when he is being convinced that the estimate labelled (1) is better than the least-squares estimate. Surely no one would disagree that if this special model of the random θ's is true, which implies amongst other things that the observations are positively correlated, then the estimates should be made to depend on each other in the way indicated. When the assumption of such a prior structure is absent, on the other hand, the situation is simply one in which an experiment is repeated n times with different and unknown parameter values: in other words, a compound decision problem in the sense of Robbins. I attempted to show at one of our previous meetings (Copas, 1969) that it is very doubtful that compound decision rules, of which (1) is typical, are in fact better than the classical results, unless a relationship of some sort is assumed between the components, such a exchangeability. For one, only the *average* mean squared error over i from 1 to n is considered, whereas the risk of particular components may be unacceptably high, and in any case there would presumably be no particular reason for supposing that inequality (2) was satisfied unless the θ's were assumed to be some sort of random sample.

As Professor Cox has mentioned, the mathematical models described in this paper are almost identical to those dealt with in the completely non-Bayesian approach known as empirical Bayes. Thus this work could be described as giving Bayes solutions to the empirical Bayes problem. The main difference lies in the meaning of the mixing distribution Q of Section 1. In empirical Bayes, Q is simply a frequency model for a continuing series of repetitions of an experiment. In principle, any assumption in this model can be tested. If I understand correctly, we now see this distribution as a measure of prior beliefs operating through the assumption of exchangeability. Is it not remarkable that one's prior beliefs about treatment i would be identical *in every way* to one's prior beliefs about treatment j? I hope that the elegant mathematics of this paper will not tempt us to rest assured in our ignorance and refrain from enquiring of our experimenter the differences which must inevitably exist between the various parts of his data.

Finally, two brief questions. What is meant by the statement that exchangeability is assumed for *every* n? Does this entail a hypothetical continuation of the number of components in the problem? Secondly, it is easy to show that the correlation between any pair of θ's in the mixture model must be non-negative. Do the authors have a simple representation of exchangeable priors with negative correlations? I have in mind a randomization approach to field trials in which the true plot yields are, in a sense, negatively associated.

Dr D. V. HINKLEY (Imperial College): I have two points to make about this very interesting and useful paper. Both have to do with related unpublished work which I think can augment the solutions proposed by the authors.

The first point has to do with outliers—in the parameters, not the observations. Take the simple problem of estimating $\theta_1, \ldots, \theta_n$ when observations $y_i \sim N(\theta_i, 1)$ are available, and the θ_i are independently $N(0, r^2)$ with r^2 known; this is a simple version of the problem in Section 1. The Bayes estimate for quadratic loss is $\boldsymbol{\theta}^B = \hat{\boldsymbol{\theta}}\{r^2/(r^2+1)\}$, where $\hat{\boldsymbol{\theta}}$ is the vector of least-squares estimates, i.e. y in this case. The Bayes estimate minimizes the Bayes risk, but the risk for any given θ_i is unbounded. Thus the Bayes estimate is particularly vulnerable to atypical parameter values, that is outliers, which will be shrunk too much. Efron and Morris, in a series of unpublished papers at Stanford University and Rand, devise an estimate which limits risk with little relative increase in Bayes risk. For our simple problem of estimating $\theta_1, \ldots, \theta_n$ they propose estimates such as

$$\tilde{\theta}_i = \begin{cases} \hat{\theta}_i + M, & \hat{\theta}_i < -M(r^2+1) \\ \theta_i^B = \hat{\theta}_i\left(\dfrac{r^2}{r^2+1}\right), & -M(r^2+1) \leqslant \hat{\theta}_i \leqslant M(r^2+1), \\ \hat{\theta}_i - M, & M(r^2+1) < \hat{\theta}_i. \end{cases}$$

As an example, suppose that the Bayes risk of $\hat{\boldsymbol{\theta}}$ is normalized to be 1·0, and that $r^2 = 1$. Then the Bayes risk of $\boldsymbol{\theta}^B$ is 0·5. If $\tilde{\theta}_i$ are used with $M = 1$, the Bayes risk is only 0·525 and furthermore the maximum risk is 2 for any θ_i. This approach can be generalized to the cases studied by the authors. Incidentally one suspects from Table 2 that β_6 is an outlier, particularly from a normal plot of the least-squares estimates; the Bayes estimate of β_6 is made normal.

My second point is of a somewhat mathematical nature, and has to do with use of variance estimates in Bayes estimates such as (1). Suppose now that we are estimating multivariate means $\boldsymbol{\theta}_i$, using independent observations $\mathbf{y}_i$ from $N(\boldsymbol{\theta}_i, \mathbf{I})$ where the $\boldsymbol{\theta}_i$ are independent samples from $N(\mathbf{0}, \boldsymbol{\Sigma})$. If we write $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, then the Stein-type estimate for $\boldsymbol{\theta}_i$ is of the form

$$\boldsymbol{\theta}_i^* = \{\mathbf{I} - k\mathbf{S}^{-1}\}\,\hat{\boldsymbol{\theta}}_i,$$

where $\hat{\boldsymbol{\theta}}_i$ is the maximum-likelihood estimate and $\mathbf{S} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}$. In a forthcoming paper by Efron and Morris (1972), reference is made to the fact that Stein has established a uniformly superior replacement for $k\mathbf{S}^{-1}$ in $\hat{\theta}_i^*$, namely $a\mathbf{S}^{-1} + b\mathbf{I}/\mathrm{tr}\,(\mathbf{S})$ for suitable a and b. Does this help in the authors' search for a way of coping with their unknown variances?

Mr E. F. HARDING (Cambridge University): In estimating three means μ_1, μ_2, μ_3 by normal samples with the loss function

$$(\hat{\mu}_1 - \mu_1)^2 + (\hat{\mu}_2 - \mu_2)^2 + (\hat{\mu}_3 - \mu_3)^2 \tag{1}$$

the usual estimate $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ is, by Stein's so-called paradox, not admissible. To use an example I heard from Professor Barnard, if μ_1 refers to butterflies in Brazil, μ_2 to ball-bearings in Birmingham and μ_3 to Brussels sprouts in Belgium, then an admissible estimator will cause the estimates of these three quite unrelated things to be related to each other, the largest $\hat{\mu}$ being on the whole pulled down and the smallest pulled up. Now, if you do not like that sort of thing, and in general I do not, then so much the worse for the loss function; one can, after all, always work a "paradox" backwards.

Coming now to exchangeability as such, take another instance in which the three μ's are the tax-allowable expenses of Professor Lindley, Dr Smith and myself, normally distributed data about these being available. The Tax Commissioners might, I suppose, regard the three of us as exchangeable *a priori*, and use the loss function (1); in which case Professor Lindley could find himself paying too much tax because one of the others of us was paying too little—fair to the Exchequer, perhaps, but not to him—and he could reasonably plead that they should use the estimator fairest *to him*: because he, after all, can hardly think himself exchangeable with, say, me. And this is the point: exchangeability is relative to the one who wants to do the exchanging.

Viewing statistical decision theory as "a game between the statistician and Nature" perhaps encourages the habit of imperiously acting according to principles of exchangeability, etc., since one hardly regards the "states of Nature" as having ethical rights; but when, in particular instances, it becomes a "game" between the statistician and other people, then one may expect one's hypotheses to demand respect.

I would not harp on this (the authors do, here and there, indicate that one ought, strictly, to *verify* exchangeability) had they not passed rather lightly over the matter of verification. Indeed, a first reading suggests that assuming exchangeability is a good thing because it leads to procedures with better statistical qualities. True: but they are better *relative to the loss function* (1), and if you dislike the ethical consequences then, again, so much the worse for the loss function. Of course, if μ_i were a sequence of batch means in a continuing industrial process and the profit was expressible in a form like (1), then exchangeability would lead to wholly desirable consequences. That would be a clear case, unlike, however, the authors' blithe and unqualified application of exchangeability to a

case which is, I think, very like the imaginary tax example: their final example of estimating the merits of the different American colleges. If Federal funds were to be distributed according to merits, then estimates derived from exchangeability could well lead to the better colleges getting less, and the worse more, than their fair shares. In such cases I would prefer to solve a sequence of decision problems, in which the loss function for the ith mean would be $(\hat{\mu}_i - \mu_i)^2$, and I think that a special case has to be made out for doing otherwise, to show that the "unfairness' which must result is either justified or irrelevant. One might even be able to incorporate all this into a complicated loss function!

Mr A. P. DAWID (University College London): I should like to make two points and pose a problem.

Firstly, there seem to be close connections between the approach of tonight's paper and the work of Ericson (1969) on finite sampling theory, in which exchangeability also played a leading role. One might view Ericson's structure as a three-stage model with a first stage that is singular, because in finite sampling one usually observes the first-stage parameters, that is to say the values of the sampled units, without error. Ericson was interested in prediction for the unsampled units, and this raises the question of how we should deal with prediction in general for multi-stage models.

Secondly, let me take a fundamentalist approach, according to my understanding of de Finetti's ideas. Given the possibility of taking a set of observations, a subjectivist should be putting his prior distribution jointly on the set of possible (as yet unobserved) *outcomes*. If one chooses to work with "unknown parameters" it is not, generally, because of any physical structure of the problem, but because, given various assumptions of exchangeability and conditional exchangeability, such parameters are conjured into existence by invoking de Finetti's theorem. What, then, is the status of the first-stage parameter $\boldsymbol{\theta}_1$, and what assumptions about the distribution of $\mathbf{y}$ are we making when we express opinions of exchangeability among certain components of $\boldsymbol{\theta}_1$?

I am also interested to know the true mathematical expression of two-way exchangeability, as considered in the authors' first example, the two-factor design. We are interested in some first-stage parameters $\{\theta_{ij}\}$ measuring somehow a mean response for cell (i, j). Lindley and Smith assume a singular second stage of the interaction-free form: $\theta_{ij} = \mu + \alpha_i + \beta_j$; and then take the α's and β's as independently exchangeable. What I should like to know is this: suppose we have observations $\{y_{ij}\}$ which we are willing to regard as exchangeable both within rows and within columns (but not simultaneously). Is the above structure as general as possible? What about an interaction term: $\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$? If $y_{ij} = \theta_{ij} + \epsilon_{ij}$, this is identical with the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \eta_{ij},$$

where $\eta_{ij} = \gamma_{ij} + \epsilon_{ij}$, so without replicated y's we cannot get much information about the γ's. But even in this case there is a qualitative difference between the two models, and I am now asking for guidance on how to parametrize such a situation when such arbitrariness exists.

Finally, a problem: we have seen the application of multi-stage models to the standard normal theory linear set-up, and in the discussion Mr Leonard has mentioned approaches to similar problems with binomial and multinomial models. However, many situations have a more complicated structure, but the ideas presented tonight would be attractive to apply if they could be made tractable. Let me outline a problem I have in mind and ask whether Lindley and Smith can give me any help with it.

The situation is that a number of patients are asked questions a number of times by a number of doctors. When doctor k asks patient i the jth question for the rth time, he elicits a binary answer S_{ijkr}. We suppose that there is a true answer T_{ij} which the doctors are trying to uncover, but some may be better at doing so than others. There are various exchangeabilities I should like to impose on this problem: firstly, $\{T_{ij}\}$ might be exchangeable over i, and perhaps in a restricted form over j. Then one could look at the conditiona

distribution of the S's given the T's, where we could introduce more symmetries. Perhaps we could make these conditional distributions depend on further exchangeable parameters. I cannot quite see what the multi-stage expression of such structure might be in general, but one can simplify at a number of points. However, the technical difficulties of inference in even the simplified models seem rather daunting. Does the normal theory analysis give any insights into how to tackle this problem?

The following contributions were received in writing, after the meeting.

Dr C. CHATFIELD (University of Bath): I had intended to attend this meeting, but finally decided that the discussion would probably involve yet another round in the endless controversy between Bayesians (subjectivists), Frequentists, Likelihoodlums, etc., which would probably not be profitable for an applied statistician like myself. Arguments about the foundations of statistics are, of course, extremely interesting from an academic point of view, but, as far as one can ascertain, the different approaches seem to give more or less the same answer in the vast majority of cases.

A major difficulty for applied statisticians is that the arguments involved tend to be extremely difficult to follow and centre around rather artificial examples. A further difficulty is that participants in these discussions tend to show scant respect for one another's position—for example, see the authors' statement that they "know of no reasoned argument against the Bayesian position". For my own part I agree with H. O. Hartley in the discussion of Cornfield's paper (1969) that no single theory of inference is entirely free from deficiencies and so, like most applied statisticians, I adopt the standard sampling theory approach for mainly practical reasons. (Of course, on occasions I also find it useful to look at likelihood surfaces or use Bayes theorem.)

Turning from these general matters to the paper, I wonder how many other people find the basic idea of exchangeability hard to swallow except possibly in a few special situations. I can imagine the reaction this assumption would receive from many scientists, and also the likely reaction to the Lindley-Smith estimates in their equation (1). Perhaps their estimates do have a smaller mean-square error than the ordinary least-squares estimates provided that the priors really are exchangeable and provided that one knows both σ^2 and τ^2. However, in practice these variances will almost certainly be unknown and I am very doubtful about using the estimates (1) with estimates of σ^2 and τ^2.

It is nice to see that the authors have actually performed a numerical study in which Bayes estimates do come out marginally superior to the ordinary least-squares estimates. However, as the authors say, one cannot draw strong conclusions from a single study. Moreover I am doubtful if the improvement they achieve justifies the extra effort involved. Scientists have trouble enough with straightforward regression without the complication of accepting the idea of subjective exchangeable priors.

Perhaps I am being slightly unfair in concentrating on the practical aspects of what is really a Research Section paper and which is clearly an important paper from a theoretical point of view. However, the meeting was billed as an ordinary meeting (organized by the Research Section), and I feel that ordinary meetings should be of broader interest and of more practical use than appears to be the case here.

Of course this last remark also applies to many other Research Section meetings, and so perhaps it would be a good idea to discontinue the recently introduced practice of calling all Research Section meetings "ordinary meetings".

Dr STEPHEN E. FIENBERG (University of Chicago): It is a great pleasure for me to have the opportunity to discuss this paper by Professor Lindley and Dr Smith. I regret that I was unable to hear its presentation. The problems discussed here are ones on which I myself have expended considerable energies. It is more than faint praise for me to remark that I wish I had written this paper.

The presentation of the general theory in Sections 1 and 2 is lucid, and perhaps to the surprise of many of us, quite straightforward. Work on the Bayesian analysis of linear models abounds, but the use of a multi-stage Bayesian analysis with successive reductions in the number of parameters at each stage has considerable intuitive appeal.

I was especially glad to see Sections 3.1 and 5.1 dealing with the complete two-way layout, since I once worked my way through this example (Fienberg, 1967) to arrive at the simple generalization of expression (20) for the two-way layout with replications, i.e.

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (1 \leqslant i \leqslant t,\ 1 \leqslant j \leqslant b,\ 1 \leqslant k \leqslant n),$$

where the second stage is

$$\gamma_{ij} \sim N(0, \sigma_\gamma^2), \quad \alpha_i \sim N(0, \sigma_\alpha^2),$$
$$\beta_j \sim N(0, \sigma_\beta^2), \quad \mu \sim N(\omega, \sigma_\mu^2).$$

In this somewhat more general case, taking the posterior mean and letting $\sigma_\mu^2 \to \infty$, one gets $\mu^* = y...$,

$$\alpha_i^* = (y_{i..} - y...)\frac{bn\sigma_\alpha^2}{bn\sigma_\alpha^2 + n\sigma_\gamma^2 + \sigma^2},$$

$$\beta_j^* = (y_{.j.} - y...)\frac{tn\sigma_\beta^2}{tn\sigma_\beta^2 + n\sigma_\gamma^2 + \sigma^2}$$

and

$$\gamma_{ij}^* = (y_{ij.} - \mu^* - \alpha_i^* - \beta_j^*)\frac{n\sigma_\gamma^2}{n\sigma_\gamma^2 + \sigma^2}.$$

Letting $n = 1$ and $\sigma_\gamma^2 = 0$ we get expression (20). The theorem and the simple form of expression (18) greatly simplify the amount of algebraic manipulation I was forced to carry out.

Going on to the case where the second-stage variances are unknown, it is interesting to note that, using the estimates

$$s^2 = \sum_{ijk}(y_{ijk} - y_{ij.})^2/\{2 + bt(n-1)\},$$

$$s_\gamma^2 = n\sum_{ij}(y_{ij.} - y_{i..} - y_{.j.} + y...)^2/\{(b-1)(t-1) - 2\},$$

$$s_\beta^2 = nt\sum_{j}(y_{.j.} - y...)^2/(b-3),$$

$$s_\alpha^2 = nb\sum_{i}(y_{i..} - y...)^2/(t-3)$$

for σ^2, $(\sigma^2 + n\sigma_\gamma^2)$, $(\sigma^2 + n\sigma_\gamma^2 + nt\sigma_\beta^2)$ and $(\sigma^2 + n\sigma_\gamma^2 + nb\sigma_\alpha^2)$ respectively, in α_i^*, β_j^* and γ_{ij}^* above, one arrives at the estimator of Stein (1966), provided negative estimates of variances are replaced by zeros, in keeping with the classical practice. (Here t and b are assumed to be $\geqslant 3$.) This estimator is of course similar to the one suggested by Lindley and Smith although no iteration is necessary. Stein points out that the anticipated contraction of the main effects is likely to be small relative to that for the interaction effects.

Rather than using point estimates for the prior parameters, Dempster (1971) suggests laying out intervals for them which are acceptably consistent with the data in the sense of passing simple significance tests, or equivalently of belonging in certain confidence regions. These intervals can then be translated into intervals for the posterior means, and in effect yield upper and lower bounds

I am somewhat troubled by the use of modal approximations proposed in Section 4 of this paper, and there are three items of concern. First, integrating out the nuisance parameters from the posterior distribution is not really as technically complex as implied by Lindley and Smith. One often gets posterior distributions of the θ's, which are products

of multivariate t-distributions. Then, one can use an identity such as that given by Dickey (1968) to reduce the integral for the posterior mean to an integral of dimension one less than the number of t-like factors. For the one-way ANOVA model discussed in Section 1 this reduction yields a one-dimensional integral that can be handled easily by simple numerical techniques.

Second, the use of the mode of the posterior distribution in place of the mean is fine when the θ posterior margin is nearly normal. When the posterior distribution is unimodal but asymmetric, the use of the mode may lead to relatively large errors. Mosteller and Wallace (1964) suggest taking modes of transformed parameters relative to a weighted measure, so as to make the resulting mode and mean much closer than the mode and mean of the original θ-margin.

Finally, the use of the mode of the joint posterior distribution in place of that of the θ-margin is one which requires far more justification than that of computational expediency.

A problem not considered in the present paper is that of estimating parameters for linear models, when some or all of the observation vectors are incomplete. In a discussion of a paper by Hartley and Hocking (see Fienberg, 1971), I proposed a crude Stein-like improvement for a simple k-means problem where the maximum-likelihood estimator has a regression form. Extensions of the techniques presented by Lindley and Smith should yield substantial improvement over my estimators, which are, in turn, uniformly superior to the maximum-likelihood estimators according to the criterion of expected quadratic loss. Have the authors undertaken an investigation of such extensions?

Professor Bruce M. Hill (University of Michigan): Professor Lindley and Dr Smith have presented an elegant and informative Bayesian treatment of an important class of problems. Although they apparently would not agree, I think they are dealing with a genuine Model II situation. The distinction they make in terms of where one is *interested* in fixed effects or variance components does not seem an important one to me. Nor do I think that exchangeability is really the appropriate basis for their results, but rather normality. Exchangeability of the θ_i is indeed a natural way to characterize subjective opinion in certain contexts, but there is no more reason (and in fact, I believe, much less) to expect robustness of normal theory prior distributions of the type they assume for the θ_i than to expect robustness of normal theory distributions for errors. Thus I expect their analysis is appropriate, even as an approximation, only for a small subclass of the exchangeable prior distributions for the θ_i. This is not to argue against normal theory, which is often suggestive, as well as being important for more traditional reasons (i.e. central limit theorem mystique), but rather to emphasize that Bayesian inference need not be bound up with normality, either for likelihood or prior distributions, any more than classical theory as, for example, is shown in my paper (Hill, 1969).

Dr R. Thompson (Unit of Statistics, Edinburgh): The authors may care to comment on the relationship between the exchangeability model and the random effects Model II. Consider, for example, the two-factor experimental design discussed in Sections 3.1 and 5.1, when the variances σ^2, σ_α^2 and σ_β^2 are known. Then equation (20) gives the same estimates of θ as the random effects model with the α's and β's normally distributed with variances σ_α^2 and σ_β^2. But estimation of the variances σ^2, σ_α^2 and σ_β^2, by maximum likelihood, assuming a random model, does not in general give the same estimates as equations (32) for the exchangeable model. It seems that, when σ_α^2 and σ_β^2 are large compared with σ^2, the estimates of the α's and β's are pulled more to the general mean on a random effect model than on an exchangeable model.

Lindley and Smith assert in Section 5.1 that when there is vague prior knowledge of σ^2, σ_α^2 and σ_β^2 the modal estimates of the treatment and block effects are zero (i.e. when $\nu = \nu_\alpha = \nu_\beta = 0$). Consideration of the following example suggests that this is not so.

Let

$$b = t = 4, \quad n = 16,$$

$$\sum_i (y_{i.} - y_{..})^2 = \sum_j (y_{.j} - y_{..})^2 = 40{\cdot}5,$$

$$\sum_i \sum_j (y_{ij} - y_{i.} - y_{.j} + y_{..})^2 = 60.$$

Then $s^2 = 4$, $s_\alpha^2 = s_\beta^2 = 8$ satisfy equations (32) (with the minor modification that if there is vague prior knowledge on σ^2, σ_α^2 and σ_β^2 then presumably the 2 in the denominator of equations (32) should be deleted).

It would seem that a stronger form of the matrix lemma, (10), is needed for the within regression example (Section 3.3). For then the so-called inverse of $(\mathbf{C}_2 + \mathbf{A}_2 \mathbf{C}_3 \mathbf{A}_2^{\mathrm{T}})$, found by application of (10), reduces to $(\mathbf{I} - p^{-1} \mathbf{J}_p)$ when $\mathbf{C}_3^{-1} = \mathbf{0}$, and this is a singular matrix.

Professor B. DE FINETTI (University of Rome): I think that the main point to stress about this interesting and important paper is its significance for the philosophical questions underlying the acceptance of the Bayesian standpoint as the true foundation for inductive reasoning, and in particular for statistical inference. So far as I can remember, the present paper is the first to emphasize the role of the Bayesian standpoint as a logical framework for the analysis of intricate statistical situations. So often in such situations the mathematical technicalities play an overwhelming role (usually suppressing any attempt at intellectual reflection), so that one is reduced to an empirical choice amongst various *ad hoc* procedures culled from standard texts.

The present paper not only rebuilds the usual procedures with the innovations of the Bayesian approach, but explains why such innovations are logically and practically necessary. These explanations are given both a Bayesian and "objectivistic" interpretation: for example, the digression that tries "to persuade an orthodox statistician that (1) is a sensible estimate *for him* to consider, and indeed is better than the least-squares estimate". Personally, I am particularly pleased to see the notion of exchangeability introduced into statistical techniques, for example, in between—and within—exchangeability. In this way the concept is not only applied but, through such different forms, its meaning and role should become clearer and more familiar to statisticians.

I would like to express my warmest congratulations to my friend Lindley and his valiant collaborator, Smith.

Professor OSCAR KEMPTHORNE (Statistical Laboratory, Iowa State University): The authors should be congratulated on their presentation. It will be informative to many. I have no detailed questions or remarks about the formal development, I wish mainly to comment on philosophical issues that underly the whole matter under discussion.

It seems *still* to need to be said that there is not, nor ever has been, a controversy about one type of use of Bayes's theorem. This is merely a statement of conditional probability and if the probabilities that enter into the computation are frequency probabilities, then the resultant conditional probability is a frequency probability. Any such conditional probability is verifiable by repetitions of the whole of the sampling process in just the same way that an ordinary probability is verifiable. While I have considerable difficulty understanding the idea of verifiability and literature search has not aided me, I assert that underlying all scientific knowledge is some concept of verifiability that is fairly generally accepted even if not well formalized, and that verification in experimental sciences is a matter of agreement of results among repetitions of an experiment, the repetitions being defined by the protocol of the experiment and done by different people. So, if, for instance, $\mathbf{y} = \mathbf{A}_1 \boldsymbol{\theta} + \mathbf{e}$ conditionally on $\boldsymbol{\theta}$, $\boldsymbol{\theta} = \mathbf{A}_2 \boldsymbol{\phi} + \mathbf{f}$ conditionally on $\boldsymbol{\phi}$ and so on, in which the parameters of the ultimate linear model are known, there is no problem

in applying ideas of conditional probability to obtain the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$. It is useful that the authors have written out exactly how the mathematical computations proceed and the solution with particular priors.

It is worth noting also that the essential aspects of the Bayesian argument have been used widely in psychological testing theory for perhaps sixty years and in the adjustment of animal breeding records at a more complicated level for perhaps forty years (see, for example, Lush, 1937). I have drawn attention to this before (Kempthorne, discussion of Lindley, 1971a) but repetition seems necessary. I have some knowledge of the genetic case, and shall write what is done for a simple case. The situation is that we have records on cows, say, which we denote by y_{ij}, i indexing the cow and j the record within cow. The model used is

$$y_{ij} = \mu + c_i + e_{ij}, \quad j = 1, 2, \ldots, n_i,$$

with μ known—the breed average, say. The terms e_{ij} arise from measurement error and are assumed to be uncorrelated with mean zero and variance σ_e^2. The terms c_j arise from the Mendelian process and are assumed to be uncorrelated with mean zero and variance σ_c^2. Then it is completely standard knowledge (see, for example, Kempthorne, 1957) in these areas that the best mean-square error predictor of c_i which is of the form $K_i(y_{i.} - \mu)$ is obtained with

$$K_i = \frac{\sigma_c^2}{\{\sigma_c^2 + (\sigma_e^2/n_i)\}} = \frac{n_i \rho}{1 + (n_i - 1)\rho},$$

where $\rho = \sigma_c^2/(\sigma_c^2 + \sigma_e^2)$. The "best" estimate then of $\mu + c_i$ is given by

$$\frac{\{(\sigma_e^2/n_i)\mu\} + \sigma_c^2 y_{i.}}{\{\sigma_c^2 + (\sigma_e^2/n_i)\}} = \frac{(\mu/\sigma_c^2) + \{y_{i.}/(\sigma_e^2/n_i)\}}{1/\sigma_c^2 + 1/(\sigma_e^2/n_i)}.$$

That this is a weighted mean of two "estimates" μ and $y_{i.}$ weighting inversely as variance is clear and was mentioned by Alan Robertson in a short note (Robertson, 1955). The next stage of estimating μ, σ_c^2, σ_e^2 by some way or other has been used *very widely*. There is a short description of this written by C. R. Henderson (an animal breeder) (Henderson *et al.*, 1959). It is worth noting that Henderson states and proves (using his notation)

$$(\mathbf{R} + \mathbf{ZDZ}')^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}.$$

This result grew out of a Bayesian process which I call legitimate because the model is based on prior knowledge and *not on lack of knowledge*. It is worth noting that Henderson advocates from the viewpoint of computation the maximization with regard to the θ_i of a sort of likelihood equal to $p(\theta_i)\, p(y \mid \theta_i)$, which gives the mode of the posterior distribution.

Let me turn to the philosophical issues.

First, the authors regard the result of Stein (1956) as compelling. I confess that I have failed to see the force of this result. The problem of treating a multivariate parameter estimation from the naive decision viewpoint is that even if one envisages a vector loss function one will eventually have to reduce the problem to one dimension by some norming in some general space. The loss function is usually a real-valued function and the result that one obtains depends critically on it. In the simple case of a random sample with a real unbounded parameter, the loss function $(\hat{\mu} - \mu)^2$ leads to the sample mean and the loss function $|\hat{\mu} - \mu|$ to the median. So the sample mean is inadmissible—and hence unsatisfactory—and so is the sample median. I may seem to be making a mere play on words here, because the authors know these facts as well as, and probably better than, I, but they say that because the multivariate mean $\boldsymbol{\mu}$ is inadmissible with regard to $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ it is unsatisfactory. Is it not the fact that with a reasonable use of language, any estimate is "typically unsatisfactory" because it will at best be admissible only for a particular loss function or a small class of such functions? The Stein result for the case $\mathbf{y} \sim N(\boldsymbol{\mu}, 1)$, where $\mathbf{y}$ and $\boldsymbol{\mu}$ are p-vectors, uses the loss function $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$, and implies

that if this is your loss function you should not use $\hat{\boldsymbol{\mu}} = \mathbf{y}$. This, it appears, was a great surprise to the "sampling-theory school" to use the authors' phrase. It was so, I am told, because it had seemed previously that best invariant estimates were thought to be admissible. The informed reaction to the question of what is the "best" $\hat{\mu}$ in terms of a particular loss function surely has to be that one does not know because the mathematics is so difficult. But, in any case, even if $\hat{\boldsymbol{\mu}} = \mathbf{y}$ were the best, I would not regard the result as of great interest or binding practical relevance. I am of the opinion that the mathematical workers of the "sampling-theory school" are not in touch with real problems of learning from investigation. The investigator who runs an experiment on k treatments, say $A, B, C, \ldots, K$ does not know what use he will make of the results. He may note, for instance, that another investigator has done an experiment with treatments B, C and E and others unlike those in $A, B, C, \ldots, K$. He will wish to have an estimate of μ_B, μ_C, μ_E to apply in his comparison. The loss function $(\hat{\mu}-\mu)'(\hat{\mu}-\mu)$ is simply irrelevant for that purpose. Furthermore, his use of $\hat{\mu}_B, \hat{\mu}_C, \hat{\mu}_E$ as given by a pulled-in estimator of $\boldsymbol{\mu}' = (\mu_A, \mu_B, \mu_C, \ldots, \mu_K)$ is not necessarily admissible for (μ_B, μ_C, μ_E). The present authors indicate rather definitely that "inadmissible" implies "unsatisfactory". I seem to get the lesson from the past two decades that admissibility ideas have been unfruitful, except for the reduction of data to a sufficient statistic. There is a great communication gap, because a huge number of experiments are analysed by the techniques of the "methods" books, and neither the authors of these books nor the users seem to be bothered by the theoretical fact that they are using "inadmissible" procedures. Let me hasten to add that while I do not understand the deeply mathematical work that has been done, I know enough mathematics to form the opinion that it is very difficult and much of it is beautiful *qua* mathematics. I do not wish to denigrate the work or the workers. I merely question its relevance to problems of interpretation of a given set of data. I also would like to register the plaint that "inadmissible with respect to a particular loss function" becomes through journal space exigencies merely "inadmissible", and then it is quite an easy step to replace this word by "unsatisfactory". And this is just what the authors have done.

In connection with the Stein result, it is surely well realized by workers in the admissibility area that the result, if it is to be taken as having real logical force, involves the experimenter and the statistician in an embarrassing dilemma. Let us suppose that the experimenter has done an experiment comparing two treatments A and B and wishes to record numbers $\hat{\mu}_A, \hat{\mu}_B$. A little later he makes a test of treatment C with independent errors. Then he is told by the admissibility workers that his answers for μ_A and μ_B are no longer good. Surely it is offensive to the experimenter to be told that his opinion (in whatever way he forms this) about μ_A and μ_B will be altered if it turns out later that he obtains data also on μ_C. I believe this sort of idea will not "sell". I believe the idea will "butcher" the accumulation and condensation of investigational information. I doubt strongly that the authors have talked to workers in *experimental* science and tried to sell them the idea. I surmise, furthermore, that if our informed public were aware of the "pulling-in" of test records of their off-spring that they would not be happy. I am glad to see evidence that the public is concerned about the test procedures that are being used. It is high time. Let us try to persuade an "orthodox" Bayesian (e.g. Professor Lindley) that a scientist's opinions about measured physical properties of substance X are *not* and should not be based on what he has observed for substance Y unless errors of measurements are dependent. This is not to say that the scientist will not form opinions such as "substance X is like substance Y with regard to its phase diagram".

Let me also try to persuade the authors that the reporting only of Bayesian estimates, each based on the prior of the person who obtained them, will butcher the processes of science. Here I deliberately again use the word "butcher". This is not to say that workers should not report their own *private subjective* Bayesian estimates. But they should report sufficient statistics so that other workers can construct their own Bayesian estimates. Indeed, let me assert my opinion, with which I believe Professor Lindley must agree, that reported Bayesian estimates, which must be accompanied by reporting of the prior on

which they are based, are useless to any other worker unless they have *not* been marginalized in any way, or unless the other worker has the same prior. Another worker will simply be unable to recover the sufficient statistic and construct his own Bayesian estimate. I believe this view is compelled by the Bayesians themselves, unless they argue (as perhaps some do and as Jeffreys does) that there is one and only one prior that is appropriate (and that one happens in the case of Jeffreys to be improper, so the logical status of any probability conclusions cannot but be obscure). Professor Lindley and the Bayesians should be happy that the conventional "inadmissible" and "unsatisfactory" statistics are reported, because they can then apply whatever prior they wish.

I can well surmise the attitude of scientists whose only data input from other workers consists of other workers' Bayesian estimates. Surely the answer will be, "I do not care what Joe thinks about the parameter: I want to know the observational facts or a good condensation of them". *I believe our present authors are not in touch with the processes of science.* They are not aware of the need for the development of interpersonal validity to opinions. For them life is very lonely. They are concerned about how they alone should make a terminal decision. They are not concerned with the acceptability of their experimental conclusions to other workers. In contrast, I would report the standard statistics with estimates of error and so on. I might also compute my own posterior distribution, but I would be hesitant about burdening the literature with it. If a compelling case could be made for a Jeffreys prior I would use it and report corresponding Bayesian estimates, but the case presented has not been found compelling. In just the same way, if a compelling case could be made for a particular loss function, I would base my thinking on it. But I would not, I hope, merely try to rationalize a loss function that fits with the probability structure for the data that I am using, which the authors do seem to do in connection with equation (3). What loss functions are candidates in the case of the multivariate normal mean? Very little thought suggests $\sum(\hat{\mu}_i-\mu_i)^2$, or $\sum(\hat{\mu}_i-\mu_i)^4$, or $\max_i|\hat{\mu}_i-\mu_i|$ to go to the extreme power. But also, I can readily imagine there being interest in selecting the component of μ that is highest. There are huge possibilities for very interesting and difficult mathematical work.

On quite a different point, why should one, *as a Bayesian*, be interested in average mean-square error? The concept of repetition in the sample space is dismissed, it appears, by the Bayesians, but like the present authors, they like to discuss averages in a population of repetitions in the sample space. But, at the same time, they castigate others who use the concept of repetitions, namely those of the "sampling theory" school, for so doing. If the purpose of a Bayesian analysis is solely to produce estimators that have good mean-square error properties, there is no basis for argument. If, loosely speaking, standing on one's head aids one in getting an estimator with good repetitive properties, one should stand on one's head. No one has ever questioned this. It is very curious to me that the appeal to a Bayesian process is made on the basis of hard logic from axioms of reasonable behaviour, but often recourse is made for justification to repeated sampling. Or am I alone in thinking there is a grave discrepancy here? On the same point, what is the status of $P(\text{Data}\,|\,\theta)$? Is this a frequency probability? Where did it come from?

I have to comment on the sentence: "It is the greatest strength of the Bayesian argument that it provides a formal system within which any inference or problem can be described". I would like to turn it around and say: "It is the greatest defect of the Bayesian argument that it provides a formal system according to which you can believe what you wish and, furthermore, without any data". I believe the search for the sort of panacea envisaged is a false one, which is based on a *total* misunderstanding of the nature of language and the nature of knowledge. Here again I believe some homework is desirable.

Why, incidentally, should one be interested in the standard error of the posterior distribution? Does Bayesian inference have any predictive value? The posterior distribution, its mean, or median or mode, or whatever, is merely a statistic. Most of the world likes to have some idea of the reliability of a statistic and I believe correctly so. The Bayesians believe quite the opposite and, I believe, incorrectly so.

The use of the exchangeability assumption is warned against. "The estimates (1) are therefore suggested only when this assumption is practically realistic", it is said. But how does one check the assumption? What does it mean to say that an assumption is "practically realistic"? The authors use the phrase. What do they mean? Furthermore, let us turn to the American College Testing Program. Is it "practically realistic" to use an exchangeable prior? Information is available in the records to show that schools differ widely, students of different social and ethnic backgrounds perform differentially on tests, and so on. Information on students is available to show, I think, that exchangeable priors are "practically unrealistic" whatever that means. Are they being used in high school and college testing? It seems that many of our (U.S.A.) societal processes merit outside examination.

I close with brief remarks. First, the paper presented has very interesting material and material that may well be useful to a "legitimate" Bayesian, e.g. an animal breeder. Second, even though I disagree strongly with the underlying background I value the presentation and the discussion. Third, I believe the authors and most Bayesians (but not all) may rather easily be "hoisted with their own petard", to use a delightful Anglicism. Finally, I regard Professor Lindley as a personal friend and have a Bayesian feeling that this is reciprocated. The type of criticism I have voiced is possible, intrinsically, only on some basis of this sort. The same applies to my discussion of the paper by Cornfield that is cited, though I regret to add that the discussion by others and by me in that paper and in the Waterloo paper did not apparently merit discussion. It is only by hard arguing that obscurities in the minds of all of us can be removed and that is why I must push hard and be treated in the same way.

Professor LINDLEY and Dr SMITH replied briefly at the meeting and subsequently in writing as follows:

So many interesting points have been raised in the discussion that we cannot, within reasonable space limitations, give them all the attention they deserve. We therefore hope that contributors will not take a one-sentence response to an idea to imply lack of interest on our part: quite the contrary.

Dr Nelder is not correct when he says that our estimates are the same as those derived by Yates. Using an exchangeable prior for the blocks and vague prior for the treatments, the estimates are the same when all variances are *known*. Even with this last condition, an exchangeable prior for the treatments produces different estimates. In both cases when we turn to the *unknown* variance situation, the techniques described in Section 5 lead to estimates different from those usually employed for balanced incomplete blocks. We do not understand methods based on likelihoods (whether prior or otherwise) since they seem to fail in one of the most basic of all statistical problems, namely sampling from a finite population (Godambe, 1966). The remark of Marquardt that was quoted is not strictly correct since the ridge estimates are not *confined* to a sphere. They are a *compromise* between least-squares estimates and others so confined. The Bayesian method does not invoke such restrictions whether to a sphere or a linear subspace: in this case it allows a compromise, the magnitude of which can be estimated from the data.

Dr Barnett and Professor Hill both have doubts concerning our assumption of normality at the second and subsequent stages. We share them. As a partial justification we would say that a prior distribution can sometimes be reasonably taken to be normal, and it may be better to have an approximate answer than none at all. In answer to Dr Barnett's query regarding the relative importance of exchangeability, *per se*, or normality, *per se*, as the basis for Bayesian estimation, we note the following. The representation of exchangeability by the mixture form is by no means confined to the normal hierarchy: we cite as an example the discussion of heteroscedasticity in Lindley (1971a), where it is applied to gamma distributions. On the other hand, the normal hierarchy can be used to incorporate meaningful, non-exchangeable priors; for example, if the means in a one-way

model correspond to different levels of a single factor, we might expect them to lie on a linear or quadratic surface, and this can be represented within the structure (11.1)–(11.3). This will be reported on elsewhere by Dr Smith.

Professor C. A. B. Smith puts forward an interesting argument that suggests the assumption of normality is robust in respect of a shift effect. We note that in the case of a normal likelihood the form of the prior can have a substantial influence on the magnitude of the shift. Let data $x \sim N(\theta, 1)$ and let θ have a Cauchy prior with density proportional to $(1+\theta^2)^{-1}$. The posterior distribution has turning points at the roots of the cubic $(x-\theta) = 2\theta(1+\theta^2)^{-1}$. If x is large an approximate root is x, a better value is $x-(2/x)$. In particular, the shift of the natural estimate, x, towards the prior value, zero, tends to zero as $x \to \infty$. This is quite different from the behaviour of our estimates. We have tried using t-like priors and hope to report on this elsewhere. Our guess is that the methods of the paper are not robust to outliers. Dr Hinkley's suggestions are valuable here but we do not like the abrupt change of form of the estimates that he describes.

Dr Barnett asks about Bayesian devils. Is not the whole point of the Ramsey–Savage–de Finetti approach, leading to the Bayesian position, that it bars them from the beginning? Axiom one is essentially "there shall be no devils". His query about the two posterior distributions does not revolve around which is right and which is wrong, but rather on their appropriateness to the practical situation. A subjectivist Bayesian assesses each problem on its merits. The formal framework makes it clear what has to be specified and what calculations will lead to the required answer. The choice of whether μ is known or not has to be dictated by the practical situation.

Mr Leonard's remarks are valuable. The method shows promise of being capable of extension to any member of the exponential family with its associated conjugate family.

Professors Novick and Kempthorne take us to task for omitting important references. Their strictures are justified and we apologize to all the authors they cite for failing adequately to acknowledge their contributions. (It is particularly remiss of us to have omitted the work of Box and his colleagues since it is so close to ours and appears in the "usual statistical journals"). The fact is that practitioners in psychology and genetics have been ahead of statisticians in appreciating the significance of related information. Perhaps the only novelty in the paper is the suggestion of using these estimates in wider fields. Both Novick and Fienberg alert us to the difficulties concerned with the estimation of the variances (and covariances). Our guess is that faulty estimation of these does not matter too much when the primary object is to determine the means (just as the correct weights are not too important in standard least squares), but they are surely correct when the variances themselves are of interest. Dickey's work is important here and again we humbly apologize for omitting to refer to it.

Our reason for not using Model II terminology, a point raised by Novick and Hill, is that Model II typically has a decision space involving the variance components whereas ours incorporates the means. Indeed, by looking at the latter we obtain estimates for the former. We agree with the remark made somewhere by Yates that the distinctions between the models are largely irrelevant.

Professor Cox and Dr Copas discuss an empirical Bayes approach. Cox's estimates are clearly unsatisfactory in some situations; for example, Model II analysis of variance where a maximum-likelihood estimate is rather absurdly zero. However, it is interesting to note that in the general linear model with $\mathbf{C}_3^{-1} = 0$ the posterior mean for $\boldsymbol{\theta}_1$, given effectively by equations (16) and (17), can be written explicitly in the form

$$(\mathbf{A}_1^{\mathrm{T}} \mathbf{C}_1^{-1} \mathbf{A}_1 + \mathbf{C}_2^{-1})^{-1} (\mathbf{A}_1^{\mathrm{T}} \mathbf{C}_1^{-1} \mathbf{A}_1 \hat{\boldsymbol{\theta}}_1 + \mathbf{C}_2^{-1} \mathbf{A}_2 \hat{\boldsymbol{\theta}}_2)$$

where $\hat{\boldsymbol{\theta}}_1$ is the standard least-squares estimator derived from equation (11.1), and $\hat{\boldsymbol{\theta}}_2$ is the least-squares estimator of $\boldsymbol{\theta}_2$ which obtains if (11.1) and (11.2) are combined to give the distribution of $\mathbf{y}$ conditional on $\boldsymbol{\theta}_2$. This has an obvious interpretation in Professor Cox's empirical Bayes formulation.

Our limited experience suggests that scientists do often perform experiments with reasonably homogeneous treatments and would be prepared to take similar bets on each of the treatments being used. Dr Copas says the assumption can, *in principle*, be tested. Our assumptions can *actually* be tested—in terms of gambles for example. The point concerning negative correlations is interesting. We do not know the complete answer to his question but point out that the correlation must exceed $-1/n$ in the symmetric case of n means, and so must be small.

Professor Plackett's contribution is totally delightful. It may be dangerous to express prior knowledge in terms of equivalent numbers of observations, but his is a valuable exercise in doing this and throws considerable light on the approach.

We cannot agree to Professor Sprent's term, coefficient of stupidity. The use of prior knowledge becomes even more important when the data are providing relatively little information. In his example ($\tau = 1$, $\sigma = 3$) it is certainly true that the estimate will be less than the true value with high probability, but on the other hand it is more likely to be within ± 1 of the true value (63 per cent for our estimate, 26 per cent for the least squares).

We are most interested in Dr Hinkley's final remarks about Stein's results. One of us (D. V. L.) has obtained similar estimates for a multivariate normal dispersion matrix using a Wishart prior, but there are several snags in its use. Our only reply to Dr Harding can be that we would not be willing to exchange butterflies for ball-bearings, nor is the loss function he quotes the only one for which the usual estimate is inadmissible—the paradox holds for almost all losses.

Why did not Professor Fienberg's memorandum get published? Since writing this paper D. V. L. has obtained estimates closely similar to those he mentions, by extending the arguments of Sections 3.1 and 5.1 to include an interaction component. The method is particularly attractive because it provides estimates of the cell means, $\mu+\alpha_i+\beta_j+\gamma_{ij}$, which depend on estimates of the relative sizes of the main effects and interactions, thereby quite avoiding the usual significance tests. The estimates of variances that he proposes do not seem so sensible because they use the *data* y_{ij}. etc. rather than the *estimates* of the cell means. Stein's famous estimate $x_i[1+(n-2)/\sum x_i^2)$ can probably be improved by using $Z_i = x_i[1+(n-2)/\sum Z_i^2]$ in accordance with the iteration procedure suggested in the paper. In answer to his query concerning incomplete observation vectors, some work has been done on this and although, as always, the Bayesian procedure is clear, its implementation presents technical problems that we have not been able to surmount. As a partial justification of the modal approximation we cite the work of Tiao and Zellner (1964) which shows that posterior distributions which are proportional to products of t-kernels are well approximated by normal distributions, when reasonable amounts of data are available. We note that Dickey's reduction in dimensionality of the integral becomes rather intractable when more than two such products are involved.

We do not understand Dr Thompson's numerical solution to equation (32) since he does not appear to have calculated the values of α_i^* and β_j^*. Nor do we understand his point concerning singular matrices for the within-regression example. The estimates are found directly from equations (16) and (17).

Mr Dawid makes an important point in relating our work to that of Ericson. In a personal communication he has shown how, using a result of Ericson, it is possible in some cases to write down the posterior mean quite simply, thereby avoiding the tedious matrix manipulations of this paper. To answer his second point about the status of $\boldsymbol{\theta}_1$, consider the case where n varieties of wheat are being tested in a trial. The components of $\boldsymbol{\theta}_1$ are essentially measures of the qualities of the varieties. In a sense they are never needed since all we want to know is how a variety will perform in a specified situation. Nevertheless it is practically convenient to have such a concept as an average yield. We do not know the answer to his problem of two-way exchangeability, nor to that of error-rates.

It was good to have Professor de Finetti's kind remarks since all Bayesian work owes so much to his important and original views on probability.

The Royal Statistical Society is well known for the discussions that take place at its meetings and for the invective that often accompanies them. We are particularly grateful for the kindness that contributors have shown towards our ideas. Most of them have been prepared to look at the estimates and judge them as estimates, whether from a likelihood, least squares or other approach, and not get involved in continual arguments on philosophy. This is a healthy approach. Let us see what the results amount to and temper our theory with a little pragmatism.

The two exceptions to this are Dr Chatfield and Professor Kempthorne. If a reduction by 75 per cent is not of practical importance then we would like to know what does measure up to Dr Chatfield's peculiar criteria. Professor Kempthorne twice complains that his previous discussions of our work and Cornfield's have been ignored. The reason is very simple. His fellow Iowan, Aydelotte (1966), has said, "an imprecise or slipshod formulation is impregnable: a statement that has no exact meaning cannot be disproved". Professor Kempthorne is impregnable because he will not say precisely what he would do. He knocks everyone about, Bayesians, sampling theorists, the lot, because we make our arguments precise. He does none of this and then has the effrontery to tell us we are not scientific. Precision is an essential ingredient of science.

On the one hand, he tells us that we are doing what some scientists, namely geneticists, have been doing for a long time; and later he says we are out of touch with scientific thinking. He cannot have it both ways. He casts doubts on Stein's results on the grounds that the range of means that might be involved when considering one of them is uncertain. We quote a little example that came our way recently from a physicist. He had four sample means, 79, 82, 89 and 126. He was puzzled by the last and thought it overestimated. He then obtained other means: 83, 91, 104, 111 and 112. He then revised his opinion about the 126: a high value certainly, but not excessive. Is not this just what a Stein estimate does in precise terms.

It would impose unreasonably upon our Society's space to answer every one of his points in detail. We conclude by remarking that his penultimate paragraph reveals that he has not, despite de Finetti's expectation, understood the idea of exchangeability (it does not mean the units are the same), and by expressing the hope that the American College Testing Program can deal with his accusations.

References in the Discussion

Aydelotte, W. O. (1966). Quantification in history. *Amer. Hist. Rev.*, **71**, 814–833.

Box, G. E. P. and Tiao, G. C. (1968). Bayesian estimation of means for the random effects model. *J. Amer. Statist. Ass.*, **63**, 174–181.

Copas, J. B. (1969). Compound decisions and empirical Bayes. (With Discussion). *J. R. Statist Soc.* B, **31**, 397–425.

Dempster, A. P. (1971). Model searching and estimations in the logic of inference. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds) pp. 56–77. Toronto: Holt, Rinehart and Winston.

Dickey, J. M. (1968). Three multidimensional-integral identities with Bayesian applications. *Ann. Math. Statist.*, **39**, 1615–1627.

Edwards, A. W. F. (1969). Statistical methods in scientific inference. *Nature, Lond.* **222**, 1233–1237.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations. *Biometrika*, **59** (in press).

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations (with Discussion). *J. R. Statist. Soc.*, B, **31**, 195–233.

Fienberg, S. F. (1967). Cell estimates for one-way and two-way analysis of variance tables. Memorandum NS-69, Department of Statistics, Harvard University.

—— (1971). Discussion of a paper by H. O. Hartley and R. R. Hocking on incomplete data analysis. *Biometrics*, **27**, 813–817.

Godambe, V. P. (1966). A new approach to sampling from finite populations. I. Sufficiency and linear estimation. *J. R. Statist. Soc.*, B, **28**, 310–319.

HENDERSON, C. R., KEMPTHORNE, O., SEARLE, S. R. and VON KROSIGK, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

HILL, B. M. (1969). Foundations for the theory of least squares. *J. R. Statist. Soc.* B, **31**, 89–97.

JACKSON, P. H., NOVICK, M. R. and THAYER, DOROTHY T. (1971). Estimating regressions in *m*-groups. *Brit. J. Math. Statist. Psychol.*, **24**, 129–153.

KELLEY, T. L. (1927). *The interpretation of Educational Measurements*. New York: World Books.

KEMPTHORNE, O. (1957). *An Introduction to Genetic Statistics*. New York: Wiley. Reprinted in 1968 by the Iowa State University Press.

LUSH, J. L. (1937). *Animal Breeding Plans*. Ames, Iowa: Iowa State University Press.

MARQUARDT, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, **12**, 591–612.

MOSTELLER, F. and WALLACE, D. L. (1964). *Inference and Disputed Authorship: the Federalist Papers*, Chapter 4. Reading, Mass.: Addison-Wesley.

NOVICK, M. R. and JACKSON, P. H. (1970). Bayesian guidance technology. *Rev. Educ. Res.*, **40**, 459–494.

NOVICK, M. R., JACKSON, P. H. and THAYER, DOROTHY T. (1971). Bayesian inference and the classical test theory model: reliability and true scores. *Psychometrika*, **36**, 207–328.

NOVICK, M. R., JACKSON, P. H., THAYER, DOROTHY T. and COLE, NANCY S. (1972). Estimating multiple regressions in *m*-groups; a cross-validation study. *Brit. J. Math. Statist. Psychol.*, **25**, (in press).

PLACKETT, R. L. (1950). Some theorems in least squares. *Biometrika*, **37**, 149–157.

ROBERTSON, A. (1955). Prediction equations in quantitative genetics. *Biometrics*, **11**, 95–98.

SMITH, C. A. B. (1970). Discussion of a paper by A. W. F. Edwards. *J. R. Statist. Soc.*, B, **32**, 165–166.

STEIN, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In *Festschrift for J. Neyman: Research Papers in Statistics* (F. N. David, ed.), pp. 351–366. New York: Wiley.

TIAO, G. C. and ZELLNER, A. (1964). Bayes's theorem and the use of prior knowledge in regression analysis. *Biometrika*, **51**, 219–230.