Sampling and Bayes' Inference in Scientific Modelling and Robustness
Author(s): George E. P. Box
Source: *Journal of the Royal Statistical Society. Series A (General)*, Vol. 143, No. 4 (1980), pp. 383-430
Published by: Blackwell Publishing for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2982063
Accessed: 24/02/2009 15:16

J. R. Statist. Soc. A, (1980),
143, *Part* 4, *pp*. 383–430

Sampling and Bayes' Inference in Scientific Modelling and Robustness

By George E. P. Box

University of Wisconsin–Madison

[Read before the Royal Statistical Society at a meeting organized by the South Wales Local Group on Thursday, May 15th, 1980, the President Sir Claus Moser in the Chair]

Summary

Scientific learning is an iterative process employing Criticism and Estimation. Correspondingly the formulated model factors into two complementary parts—a predictive part allowing model criticism, and a Bayes posterior part allowing estimation. Implications for significance tests, the theory of precise measurement and for ridge estimates are considered. Predictive checking functions for transformation, serial correlation, bad values, and their relation with Bayesian options are considered. Robustness is seen from a Bayesian viewpoint and examples are given. For the bad value problem a comparison with M estimators is made.

Keywords: ITERATIVE LEARNING; MODEL BUILDING; INFERENCE; BAYES THEOREM; SAMPLING THEORY; PREDICTIVE DISTRIBUTION; DIAGNOSTIC CHECK; TRANSFORMATIONS; SERIAL CORRELATION; BAD VALUES; OUTLIERS; ROBUST ESTIMATION

0. Introduction

No statistical model can safely be assumed adequate. Perspicacious criticism employing diagnostic checks must therefore be applied. But while such checks are always necessary, they may not be sufficient, because some discrepancies may on the one hand be potentially disastrous and on the other be not easily detectable. In addition therefore it is often pertinent to make the developing model robust against contingencies to which it is currently judged sensitive.

The object of this paper is to review the complementary roles in the model building process of the predictive distribution and of the posterior distribution; the former in producing diagnostic checks of parametric as well as residual features of the model, the latter in providing a general basis for robust estimation.

1. Scientific Learning and Statistical Inference

Much of statistics is concerned with extending knowledge by building empirico-mechanistic models that involve probability. A theory about such scientific model building ought to explain what good statisticians and scientists actually do. It seems that scientific knowledge advances by a practice–theory iteration. Known facts (data) suggest a tentative theory or model, implicit or explicit, which in turn suggests a particular examination and analysis of data and/or the need to acquire further data; analysis may then suggest a modified model that may require further practical illumination and so on. I shall suppose that data are acquired from a designed experiment, but the same argument would apply if data acquisition was from a sample survey or even from a visit to the library. New knowledge thus evolves by an interplay between *dual* processes of induction and deduction in which the model is not fixed but is continually developing. The statistician's role is to assist this evolution (see, for example, Box and Youle, 1955; Box, 1976). In doing so he employs two inferential devices: *Criticism*† and *Estimation.*

Suppose that at some stage i of an investigation, model M_i is being entertained.

Criticism can induce model modification. It involves a confrontation of M_i with available data $\mathbf{y}$ (old as well as newly acquired), and asks whether M_i is consonant with $\mathbf{y}$ and, if not, how

† The apt naming of inferential *criticism* is due to Cuthbert Daniel, see also Popper (1959).

not. It employs a process of diagnostic checking (see, for example, Box and Jenkins, 1970), which is often done informally using plots of various kinds of residual quantities, or more formally, with tests of goodness of fit or "tentative overfitting" procedures. When a modification to M_{i+1} has been made, this new model, in addition to confronting the same data, will in some cases be checked against new data generated by a design D_{j+1}. This new design will be chosen to explore those shadowy regions whose illumination is judged currently to be important in view of the nature of the modified model and the needs of independent verification.

Estimation. When the iteration leads to a model worthy to be entertained it may be used to estimate parameters conditional on its truth. In practice such estimation is used not only at the termination of the model building sequence but at many stages throughout it. This is because, to conduct criticism of a model, it is often necessary to estimate provisionally parameters at intermediate stages.

In any such enterprise many subjective choices are made, conscious or unconscious, good or bad. They determine for instance which plots, displays and checks of data and residuals are looked at; and what treatments and variables are included at which levels, over what experimental region, in which transformation, in what design, to illuminate which models. The wisdom of these choices over successive stages of development is the major determinant of how fast the iteration will converge or of whether it converges at all, and distinguishes good scientists and statisticians from bad. It is in this context that theories of inference need be considered. While it is comforting to remember that a good scientific iteration is likely to share the property of a good numerical iteration—that mistakes often are self-correcting, this also implies that the investigator must worry particularly about mistakes which are likely not to be self-correcting.

1.1. *Rival Theories of Inference*

The distinction between model criticism and parameter estimation has not always been made and proponents both of sampling inference and Bayesian inference have long sought for a single comprehensive theory.

I believe that, subject to some overlap discussed later, sampling theory is needed for exploration and ultimate *criticism* of an entertained model in the light of current data, while Bayes' theory is needed for *estimation* of parameters conditional on the adequacy of the entertained model. On this view (see also Box and Tiao, 1973) both processes would have essential roles in the continuing scientific iteration just as the two sexes are required for human reproduction. Attempts to choose between two entities which were not alternative but complementary could certainly be expected to lead to contention, paradox and confusion of the kind we have been experiencing. The view that more than one mode of statistical reasoning can be useful is not, of course, new and was advanced (however with a different emphasis and conclusions) by R. A. Fisher. See also in particular Dempster (1971).

1.2. *The Need for Prior Distributions*

In the past, the need for probabilities expressing prior belief has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief. The interconnection between model assumptions and prior distributions becomes clear when it is remembered that every model can be imagined as embedded in a more complex one. For example, an outright assumption of normality can be modelled by a suitable parametric family of distributions indexed by a parameter β, which has a sharp prior at the normal value. I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters. The model *is* the prior in the wide sense that it is a probability statement of all the assumptions currently to be tentatively entertained *a priori.* On this view, traditional sampling theory was of course not free from assumptions of prior knowledge. Instead it was as if only two states of mind had been allowed—complete certainty or complete uncertainty.

One illustration of how implied prior knowledge which is *implausibly imprecise* can lead to trouble in sampling theory is the famous discovery by Stein (1956) of the inadmissibility of the multivariate sample mean. Consider for example the usual one-way analysis of variance set-up. The prior assumption which justifies the shrinkage estimator (see, for example, Box and Tiao, 1968a; Lindley and Smith, 1972) that the group means μ_j are random samples from some normal super-population having unknown mean and variance might, in appropriate circumstances, be eminently reasonable. It is easy, however, to miss the lesson which is to be learned from such examples. Notice that there are many circumstances in which this "Model II" assumption would not be sensible either. For example, if the μ's were daily batch yields from some production process, it might be much more reasonable to postulate *a priori* that they followed some time series model such as a stationary autoregressive process. The estimators (Tiao and Ali, 1971) then derived from Bayesian means are not Stein's shrinkage estimators, but alternative estimators allowing incorporation of relevant sample information about the *autocorrelation* of the batch means. Thus while for this example, except as a numerical approximation, we ought not to use the sample means as estimates, we ought not to use Stein's shrinkage estimates either. There seems no logical way to avoid trouble except by the explicit prior statement of the model we wish to entertain.

1.3. *Two Complementary Factors from Bayes' Formula*

If the prior probability distribution of parameters is accepted as essential, then a complete statement of the entertained model at any stage of an investigation is provided by the joint density for potential data $\mathbf{y}$ and parameters $\boldsymbol{\theta}$ calculated from

$$p(\mathbf{y},\boldsymbol{\theta}\,|\,A) = p(\mathbf{y}\,|\,\boldsymbol{\theta},A)\,p(\boldsymbol{\theta}\,|\,A). \tag{1.1}$$

In these expressions A is understood to indicate conditionality on all or some of the assumptions in the model specification. This model (1.1) means to me that current belief about the outcome of contemplated data acquisition would be calibrated with adequate approximation by a *physical simulation* involving random sampling from the distributions $p(\mathbf{y}\,|\,\boldsymbol{\theta},A)$ and $p(\boldsymbol{\theta}\,|\,A)$.

The model can also be factored as

$$p(\mathbf{y},\boldsymbol{\theta}\,|\,A) = p(\boldsymbol{\theta}\,|\,\mathbf{y},A)\,p(\mathbf{y}\,|\,A). \tag{1.2}$$

In particular the second factor on the right, which can be computed before any data become available,

$$p(\mathbf{y}\,|\,A) = \int p(\mathbf{y}\,|\,\boldsymbol{\theta},A)\,p(\boldsymbol{\theta}\,|\,A)\,d\boldsymbol{\theta} \tag{1.3}$$

is the *predictive* distribution. It is the distribution of the totality of all possible samples $\mathbf{y}$ that could occur if the assumptions were true.

When an actual data vector $\mathbf{y}_d$ becomes available

$$p(\mathbf{y}_d,\boldsymbol{\theta}\,|\,A) = p(\boldsymbol{\theta}\,|\,\mathbf{y}_d,A)\,p(\mathbf{y}_d\,|\,A) \tag{1.4}$$

and the first factor on the right is Bayes' posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}_d$

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}_d,A) \propto p(\mathbf{y}_d\,|\,\boldsymbol{\theta},A)\,p(\boldsymbol{\theta}\,|\,A). \tag{1.5}$$

But of equal importance is the second factor

$$p(\mathbf{y}_d\,|\,A) = \int p(\mathbf{y}_d\,|\,\boldsymbol{\theta},A)\,p(\boldsymbol{\theta}\,|\,A)\,d\boldsymbol{\theta}, \tag{1.6}$$

the predictive density associated with the particular data $\mathbf{y}_d$ actually obtained. Fig. 1 illustrates for a single parameter θ and a sample $\mathbf{y}_d$ of $n = 2$ observations.

If the model is to be believed, the posterior distribution $p(\boldsymbol{\theta}\,|\,\mathbf{y}_d,A)$ allows all relevant estimation inferences to be made about $\boldsymbol{\theta}$. However, if $\mathbf{y}_d$ were such as would be very unlikely to

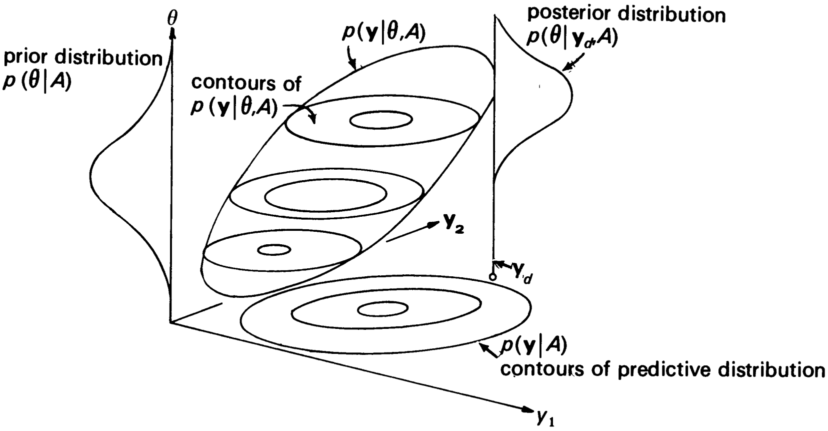

FIG. 1. A representation of the prior distribution, the posterior distribution and the predictive distribution for a single parameter θ and sample $\mathbf{y}_d$ of two observations.

be generated by the model, this could not be shown by any abnormality in this factor, but *could* be assessed by reference of the density $p(\mathbf{y}_d\,|\,A)$ to the predictive reference distribution $p(\mathbf{y}\,|\,A)$ or of the density $p\{g_i(\mathbf{y}_d)\,|\,A\}$ of some relevant checking function $g_i(\mathbf{y}_d)$ to its predictive distribution. The importance of the predictive distribution and the possibility of using it in some way as a model checking device has been discussed by a number of authors. See in particular Roberts (1965), Guttman (1967), Geisser (1971, 1975), Dempster (1971, 1975), Geisser and Eddy (1979) and Kadane *et al.* (1979). Also measures of surprise other than that discussed here have been proposed, for example by Good (1956).

2. Estimation of the Mean of a Normal Distribution

As an example consider a sample of n observations drawn randomly from a normal distribution with unknown mean θ and known variance σ^2 with uncertainty about the mean expressed by supposing that, *a priori*, θ is distributed normally about θ_0 with known variance σ_θ^2. Then *conditional* on the adequacy of the model, θ is estimated by combining data and prior information in the normal posterior distribution

$$p(\theta\,|\,\mathbf{y},A) \propto (I_{\bar{y}}+I_\theta)^{\frac{1}{2}}\exp\{-\tfrac{1}{2}(I_{\bar{y}}+I_\theta)(\theta-\bar{\theta})^2\}, \tag{2.1}$$

where $I_{\bar{y}} = n\sigma^{-2}$, $I_\theta = \sigma_\theta^{-2}$ and $\bar{\theta} = w\bar{y}+(1-w)\theta_0$ is an appropriately weighted average of $\bar{y}$ and θ_0, with $w = I_{\bar{y}}/(I_{\bar{y}}+I_\theta)$ the proportion of information coming from the data.

The predictive distribution allowing criticism of the model by contrasting data and prior information is

$$p(\mathbf{y}\,|\,A) \propto \sigma^{-(n-1)}\left(\frac{\sigma^2}{n}+\sigma_\theta^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\frac{(n-1)s^2}{\sigma^2}+\frac{(\bar{y}-\theta_0)^2}{n^{-1}\sigma^2+\sigma_\theta^2}\right]\right\}. \tag{2.2}$$

An overall predictive check is supplied by calculating

$$\alpha = \Pr\{p(\mathbf{y}\,|\,A)<p(\mathbf{y}_d\,|\,A)\} = \Pr\{\chi_n^2>g(\mathbf{y}_d)\}, \tag{2.3}$$

where

$$g(\mathbf{y}_d) = \frac{(\bar{y}_d-\theta_0)^2}{n^{-1}\sigma^2+\sigma_\theta^2}+\frac{(n-1)s_d^2}{\sigma^2}. \tag{2.4}$$

As an example suppose the sample consists of $n = 4$ analytical tests of yield $\mathbf{y}_d' = (77, 74, 75, 78)$ performed on a single batch from an industrial process for which it is believed that the testing variance $\sigma^2 = 1$, the process mean $\theta_0 = 70$ and the batch to batch variance is $\sigma_\theta^2 = 2$. We wish to estimate the mean θ for *this particular batch.*

In this example $\bar{y}_d = 76$, $\theta_0 = 70$, $s_d^2 = 3{\cdot}33$, $I_{\bar{y}} = 4$, $I_\theta = 0{\cdot}5$, $w = 0{\cdot}89$; so that, given the appropriateness of the model previously discussed, θ is estimated by the normal distribution $N(\bar{\theta}, \bar{\sigma}^2)$ with $\bar{\theta} = (0{\cdot}89 \times 76) + (0{\cdot}11 \times 70) = 75{\cdot}3$, $\bar{\sigma}^2 = (4 + 0{\cdot}5)^{-1} = 0{\cdot}22$.

However, from the predictive check

$$g(\mathbf{y}_d) = \frac{(76-70)^2}{2{\cdot}25} + \frac{3 \times 3{\cdot}3}{1} = 26 \tag{2.5}$$

and

$$\alpha = \Pr\{\chi_4^2 > 26\} < 0{\cdot}001. \tag{2.6}$$

Thus for this example the model, and hence the estimate of θ supplied by the posterior distribution $N(75{\cdot}3, 0{\cdot}22)$, is discredited by the predictive check.

Notice the following: (a) While the posterior distribution *combines* information from data and prior in a manner which is entirely appropriate if the model is to be believed, the predictive distribution *contrasts* these two sources of information and checks their compatibility.

(b) The predictive check formalizes questions that any competent statistician would raise having been presented with the supposed form of the model and the data. The components of $g(\mathbf{y}_d)$, $\{(n-1)s_d^2\}/\sigma^2$ and $(\bar{y}_d - \theta_0)^2/\{n^{-1}\sigma^2 + \sigma_\theta^2\}$ are the standard checking functions for contrasting an estimate of variance with a prior value and contrasting two estimates of the same mean.

(c) In making this predictive check it was not necessary to be specific about an alternative model. This issue is of some importance for it seems a matter of ordinary human experience that an appreciation that a situation is unusual does not necessarily depend on the immediate availability of an alternative.

(d) Whereas in estimating θ assuming the model to be true the posterior distribution makes use only of the single data vector $\mathbf{y}_d$ that has actually occurred, by contrast, an assessment of whether the sample $\mathbf{y}_d$ is likely to have occurred at all is necessarily achieved by relating $\mathbf{y}_d$ to a relevant reference set of *all* data vectors $\mathbf{y}$ which could have occurred with the model true.

Inspection of the global function $g(\mathbf{y}_d)$ alone would rarely ensure adequate checking of the model. In this example, for instance, it would be natural to consider the individual contributions from $\bar{y}_d$ and s_d^2 not only so that they could be separately considered, but also because unusually small values of $(n-1)s_d^2/\sigma^2$ as well as unusually large ones could point to model inadequacy. Also if n were larger, we might wish to consider other functions $g_i(\mathbf{y}_d)$ of the data such as moment coefficients and serial correlation coefficients which could reveal model inadequacies believed important in the current experimental situation. This could be done by referring $p\{g_i(\mathbf{y}_d) \mid A\}$ to the predictive distribution $p\{g_i(\mathbf{y}) \mid A\}$ derived by appropriate integration of $p(\mathbf{y} \mid A)$. Associated with these more specific checks are (possibly vague) model alternatives, the logical consequences of which are discussed in Section 4.6.

In practice, criticism of the model is often conducted by visual inspection of residual displays and other more sophisticated plots. But such a process, although it is informal, seems to me to fall within the logical framework described above. The plots are designed to make manifest certain "features" in the data that would rarely be extreme, if the model were true. If such a feature can be described by a function $g_i(\mathbf{y}_d)$, its unusualness, if formalized, would be measured appropriately by reference to $p\{g_i(\mathbf{y}) \mid A\}$.

For the above example obvious functions for checking individual features of the model are $\bar{y}$, s^2 and suitably chosen functions of standardized residuals $\mathbf{r} = (r_1, \ldots, r_n)'$ with $r_i = (y_i - \bar{y})/s$, $i = 1, \ldots, n$. These would usually include the individual residuals themselves plus other functions which, depending on the context, might include checks for needed transformation, heteroscedasticity, serial correlation, "bad values", skewness and kurtosis. See, for example, Anscombe (1961), Anscombe and Tukey (1963), Andrews (1971a, b).

The standardized residuals can be expressed more conveniently in terms of $n-2$ independently distributed functions obtained by making an orthogonal transformation from $\mathbf{y}$ to $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ with $Y_n = \surd(n)\,\bar{y}$ and then transforming to $\bar{y}$, s^2 and $\mathbf{u}$ where $\mathbf{u}$ is a vector of $n-2$ residual quantities $\mathbf{u} = (u_1, u_2, \ldots, u_{n-2})'$ such that

$$u_j = Y_{j+1} \Big/ \left\{ \sum_{i=1}^{j} Y_i^2 \Big/ j \right\}^{\frac{1}{2}}. \tag{2.7}$$

The Jacobian of the transformation from $\mathbf{y}$ to $\bar{y}$, s^2, $\mathbf{u}$ is proportional to

$$(s^2)^{\frac{1}{2}(n-1)-1} \prod_{j=1}^{n-2} \{1 + u_j^2/j\}^{-\frac{1}{2}(j+1)}.$$

After transformation, the predictive distribution contains n elements distributed independently

$$p(\bar{y}, s^2, \mathbf{u} \mid A) = p(\bar{y} \mid A)\, p(s^2 \mid A)\, p(\mathbf{u} \mid A), \tag{2.8}$$

where

$$p(\bar{y} \mid A) \propto (\sigma_\theta^2 + \sigma^2/n)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\bar{y} - \theta_0)^2/(\sigma_\theta^2 + \sigma^2/n)\}, \tag{2.9}$$

$$p(s^2 \mid A) \propto (\sigma^2)^{-\frac{1}{2}(n-1)} \{s^2\}^{\frac{1}{2}(n-1)-1} \exp\{-\tfrac{1}{2}(n-1)\, s^2/\sigma^2\}, \tag{2.10}$$

$$p(\mathbf{u} \mid A) \propto \prod_{j=1}^{n-2} \left\{1 + \frac{u_j^2}{j}\right\}^{-\frac{1}{2}(j+1)}. \tag{2.11}$$

The unusualness of $g_1 = \bar{y}$, $g_2 = s^2$ and of any residual functions of interest $g_3, g_4, \ldots, g_q$ can then be assessed by computing

$$\Pr\{p(g_j \mid A) < p(g_{jd} \mid A)\}, \quad j = 1, 2, \ldots, q. \tag{2.12}$$

which for unimodal distributions will be tail area probabilities. For this example these would be obtained by referring

(i) $(\bar{y}_d - \theta_0)/(\sigma_\theta^2 + \sigma^2/n)^{\frac{1}{2}}$ to the Normal table;

(ii) $(n-1)\, s_d^2/\sigma^2$ to the χ^2 table;

(iii) $g_{3d}, \ldots, g_{qd}$ to reference distributions obtained by appropriate integration of $p(\mathbf{u} \mid A)$.

These probabilities are of course affected by transformation. Thus the answer will be a little different depending for example on whether we ask a question about s or about s^2. I do not find this particularly disturbing. Slightly different questions can be expected to have slightly different answers. We now illustrate some implications.

2.1. *Significance Tests*

Suppose σ_θ^2 is assumed small compared with σ^2/n, so that w, the relative amount of information supplied by the data, is close to zero. Then, *if this model can be relied upon*, the posterior distribution will be essentially the same as the prior, sharply centered at θ_0. A practical context is one where the statistician is told that the process mean is known to be θ_0 and the batch to batch variance σ_θ^2 is negligible compared with testing variance σ^2. If he believed this model, then any data $\mathbf{y}$ could do very little to change his belief that $\theta \doteq \theta_0$. However, *it could deny the relevance of this model.* In particular $g_1(\mathbf{y}_d)$ now involves essentially the reference of $(\bar{y}_d - \theta_0)/(\sigma/\surd n)$ to normal tables; the failure of this check means that the model is discredited and therefore the Bayes calculation that leads to a sharp posterior distribution at θ_0 may not logically be undertaken.

The above most satisfactorily explains to me the rationale of a significance test.

(a) The tentative model (null hypothesis) implies that θ is close to θ_0.

(b) A check on the compatibility of this model and the data, so far as the mean is concerned, is provided by reference of $(\bar{y}_d - \theta_0)/(\sigma/\surd n)$ to the Normal Table.

(c) If the tail area probability is not small we do not question the model. The *application of Bayes' theorem* then produces a posterior distribution which is sharply centred at θ_0. We have "no reason to question the null hypothesis".

(d) If the tail area probability is small we conclude that the model which postulated that $\theta \doteq \theta_0$ is discredited by the data, i.e., the "null hypothesis is discredited".

(e) Notice too that although the failure of this check would most immediately proscribe the use of Bayes' theorem, the failure of other checks (and of that based on s^2 in particular) would also suggest the need for model modification before proceeding further.

A difficulty that this removes for me is that, as usually formulated, significance tests had seemed to provide *no basis for belief.* On this formulation, however, the significance test provides a means of discrediting a model which *if* accepted would inevitably imply acceptance of the belief that θ lay close to θ_0. It is admitted that this formulation does not cover all possible circumstances in which significance tests have been used (see in particular Cox, 1977), but it is arguable that other applications are best dealt with in other ways.

2.2. *Precise Measurement and Improper Priors*

Suppose now that σ_θ^2 is assumed large compared with σ^2/n, so that $1-w$, which measures the proportion of the information about θ coming from the prior, is close to zero. Then σ_θ^2 dominates the denominator in the predictive checking function $(\sigma_\theta^2+\sigma^2/n)^{-\frac{1}{2}}(\bar{y}-\theta_0)$ implying that the model would not be called into question by sets of data having widely different sample averages. This is the situation where we can invoke what L. J. Savage called the "theory of precise measurement" to justify the very useful numerical approximation of the posterior distribution by $N(\bar{y}, \sigma^2/n)$. Now since the predictive distribution for $\bar{y}$ does not exist at the limit $1-w=0$ when this limiting posterior distribution is obtained, it might be argued that, when precise measurement theory is appropriate, we have a license to apply Bayes' theorem without any restraining checks on the model. Obviously, however, in any imaginable experimental situation there *would* be values of $\bar{y}$ which would rightly be regarded as implausible given the investigator's current beliefs. Thus what is really being verified is that a non-informative prior must, to make practical sense, always be proper, even though the appropriate posterior distribution can, in suitable circumstances, be *numerically approximated* by substituting an improper prior.

3. The Normal Linear Model

Suppose

$$\mathbf{y} \sim N(\mathbf{1}\mu+\mathbf{X}\boldsymbol{\theta}, \mathbf{I}_n\sigma^2) \tag{3.1}$$

with $\mathbf{1}$ a vector of unities and $\mathbf{X}$ of full rank k such that $\mathbf{X}'\mathbf{1}=\mathbf{0}$ and suppose that prior densities are locally approximated by

$$\mu \sim N(\mu_0, c^{-1}\sigma^2),\ \boldsymbol{\theta} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Gamma}^{-1}\sigma^2),\ \{\sigma^2/v_0 s_0^2\} \sim \chi^{-2}(v_0) \tag{3.2}$$

with μ and $\boldsymbol{\theta}$ independent but conditional on σ^2, and $\chi^{-2}(v_0)$ the inverted χ^2 distribution.

Given a sample $\mathbf{y}_d$, special interest attaches to $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}^2$ which given the assumptions are estimated by $p(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}_d, A)$ with marginal distributions

$$p(\boldsymbol{\theta} \mid \mathbf{y}_d, A) \propto \left\{1+\frac{(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}_d)'(\mathbf{X}'\mathbf{X}+\boldsymbol{\Gamma})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}_d)}{(n+v_0)\hat{\sigma}_d^2}\right\}^{-\frac{1}{2}(n+v_0+k)} \tag{3.3}$$

$$p(\sigma^2 \mid \mathbf{y}_d, A) \propto \sigma^{-(n+v_0+2)} \exp\{-\tfrac{1}{2}(n+v_0)\hat{\sigma}_d^2/\sigma^2\} \tag{3.4}$$

with

$$\bar{\boldsymbol{\theta}}_d = (\mathbf{X}'\mathbf{X}+\boldsymbol{\Gamma})^{-1}(\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}}_d+\boldsymbol{\Gamma}\boldsymbol{\theta}_0),\quad \hat{\boldsymbol{\theta}}_d = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{d'}\quad v = n-k-1,$$

$$(n+v_0)\hat{\sigma}_d^2 = vs_d^2+v_0 s_0^2+(\hat{\boldsymbol{\theta}}_d-\boldsymbol{\theta}_0)'\{(\mathbf{X}'\mathbf{X})^{-1}+\boldsymbol{\Gamma}^{-1}\}^{-1}(\hat{\boldsymbol{\theta}}_d-\boldsymbol{\theta}_0)+(n^{-1}+c^{-1})^{-1}(\bar{y}-\mu_0)^2.$$

Now let s_p^2 be the pooled estimate

$$(v+v_0)^{-1}(vs^2+v_0 s_0^2). \tag{3.5}$$

Then the predictive distributions for $(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)/s_p$, s^2 and the $v-1$ elements of the residual vector $\mathbf{u}$, defined in an analogous manner to that previously employed in (2.7), are independent and are given by

$$p\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)/s_p\,|\,A\} \propto \left\{1+\frac{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\{(\mathbf{X}'\mathbf{X})^{-1}+\boldsymbol{\Gamma}^{-1}\}^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)}{(\nu+\nu_0)\,s_p^2}\right\}^{-\frac{1}{2}(n+\nu_0-1)}, \tag{3.6}$$

$$p(s^2/s_0^2\,|\,A) \propto F^{\frac{1}{2}\nu-1}\left\{1+\frac{\nu}{\nu_0}F\right\}^{-\frac{1}{2}(\nu+\nu_0)}, \quad F = s^2/s_0^2, \tag{3.7}$$

$$p(\mathbf{u}\,|\,A) \propto \prod_{j=1}^{\nu-1}\{1+(u_j^2/j)\}^{-\frac{1}{2}(j+1)}. \tag{3.8}$$

The predictive check derived from (3.6)

$$\begin{aligned}\Pr\{p((\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)/s_p\,|\,A)&<p((\hat{\boldsymbol{\theta}}_d-\boldsymbol{\theta}_0)/s_{pd}\,|\,A)\}\\ &= \Pr\left\{F_{k,\,\nu+\nu_0}>\frac{(\hat{\boldsymbol{\theta}}_d-\boldsymbol{\theta}_0)'\{(\mathbf{X}'\mathbf{X})^{-1}+\boldsymbol{\Gamma}^{-1}\}^{-1}(\hat{\boldsymbol{\theta}}_d-\boldsymbol{\theta}_0)}{ks_{pd}^2}\right\}\end{aligned} \tag{3.9}$$

is the standard analysis of variance check for compatibility of two estimates $\hat{\boldsymbol{\theta}}_d$ and $\boldsymbol{\theta}_0$. It was earlier proposed as a check for compatibility of prior and sample information by Theil (1963). The predictive check derived from (3.7) $\Pr\{p(s^2\,|\,A)<p(s_d^2\,|\,A)\}$ yields the F test having ν and ν_0 degrees of freedom appropriate to check whether the two estimates s_d^2 and s_0^2 are compatible. Residual checks derived from (3.8) are obtainable as before.

3.1. *Ridge Estimates*

Now suppose the $\mathbf{X}$ matrix to be in correlation form and assume $\boldsymbol{\theta}_0 = \mathbf{0}, \boldsymbol{\Gamma} = \mathbf{I}_k\gamma_0, \nu_0 \to 0$ so that $s_p^2 \to s^2$. Then the estimates $\bar{\boldsymbol{\theta}}_d$ are the ridge estimators of Hoerl and Kennard (1970) which, given the assumptions, appropriately combine information from the prior with information from the data. The predictive check (3.9) now yields

$$\alpha = \Pr\left\{F_{k,\nu}>\frac{\hat{\boldsymbol{\theta}}_d'\{(\mathbf{X}'\mathbf{X})^{-1}+\mathbf{I}\gamma_0^{-1}\}^{-1}\hat{\boldsymbol{\theta}}_d}{ks_d^2}\right\} \tag{3.10}$$

allowing any choice of γ_0 to be criticized.

For example, in their original analysis of the data of Gorman and Toman (1966), Hoerl and Kennard (1970) chose a value $\gamma_0 = 0{\cdot}25$. However, substitution of this value in (3.10) yields $\alpha = \Pr\{F_{10,25} > 3{\cdot}59\} < 0{\cdot}01$ which discredits this choice. More recently it has been pointed out (Lindley and Smith, 1972; Hoerl, Kennard and Baldwin, 1975) that given the model, γ can be estimated from the data. If we do this, much smaller values of γ are obtained which of course are not in conflict with the wider model. The two kinds of analysis further illustrate the overlap between predictive checking and Bayesian estimation later discussed in Section 4.6.

The Bayes approach to ridge estimators has the characteristic advantage that the somewhat arbitrary prior assumptions, which have to be made even for compatible values of γ, are uncovered for criticism (see also Draper and Van Nostrand, 1977). If $\gamma_0^{-1} \to 0$, (3.10) yields the standard ANOVA significance test which has a detailed interpretation parallel to that set out in Section 2.1.

4. Diagnostic Checks

It is useful to distinguish two kinds of checks which may be called respectively Overall or Multidirectional checks and Specific or Unidirectional checks. An example of the first would be a general inspection of residuals and the second a Durbin–Watson test for first-order serial correlation. This distinction is made, for example, by Box and Jenkins (1970) in their discussion of the general philosophy of diagnostic checking. Concerning these two kinds of checks these authors say "... although [overall checks] can point out unsuspected peculiarities ... [they] may not be particularly sensitive. Tests for specific departures... are more sensitive, but may fail to warn of trouble other than that specifically anticipated." The two alternatives ought properly to be regarded as extremes on some scale of dependence of checking procedures on specific alternatives. For example, consider the fitting of a parametric time series model. While residuals themselves should always be inspected there are a number of way-stations between this overall but insensitive check and the device called "overfitting" in which a model is tentatively elaborated in a *specific* direction. Thus inspection of, and application of overall tests to, the autocorrelation function and the periodogram of the residuals while still non-specific is less general than the first device and much less specific than the second.

The model checking problem is comparable to that faced by a nation which fears aerial attack that might come from any direction but with certain rather wide zones more likely than others and certain specific directions believed especially likely. How should limited radio detection devices, which are less sensitive the less they are focused, be deployed? The best solution obviously involves some combination of wide and more specific searches, and theoretically could be achieved knowing prior probabilities and expected losses. Correspondingly, the competent statistician must, in a variety of contexts, be able to make intelligent guesses not only of what discrepancies are particularly likely, but which are potentially disastrous, and to allocate his effort accordingly. In practice this is done informally and is part of what an adequate training in statistics achieves.

4.1. *Checking Parametric Features of the Model*

In the examples considered above where sufficient statistics were available parameter preferences evidenced by proper priors were directly challenged, leading without a direct statement of alternatives to appropriate checks. When a specific set of assumptions A_1 alternative to A_0 are in mind then an appropriate checking function might also be obtained from the predictive ratio

$$p(\mathbf{y}_d \mid A_1)/p(\mathbf{y}_d \mid A_0). \tag{4.1}$$

We shall not explore this possibility further here, except to note that this ratio is a component in the direct assessment of Bayesian odds to which we refer briefly in Section 4.6.

4.2. *Checking Residual Features of the Model*

Residual checking functions are sometimes chosen on an *ad hoc* basis and sometimes using specific models. I think the best course is again to employ an iteration—this time between theory and intuition. An empirical procedure that works well invites the question: What kind of model would be needed for its justification? Such a model can then be considered for use in a wider context. For instance, exponential smoothing and the "three term" controller were both empirically developed techniques found to be practically effective. ARIMA time series models are generalizations of the stochastic processes that could justify these methods (Box and Jenkins, 1970). In a similar way the practical usefulness of such things as the *jackknife* and *cross-validation* implies the existence of corresponding models which are worthy of further analysis.

The distinction between parametric features of the model and residual features is of course arbitrary and a matter of convenience. In practice the needs of parsimony urge us to settle for reasonably simple models and to consider possible deviations from them. Consider now therefore an interesting but by no means unique method for obtaining an appropriate function of the data for informal or formal checking for a particular kind of deviation from a current model parametrized by a discrepancy parameter β.

Suppose the predictive distribution conditional on some specific choice of β is $p(\mathbf{y} \mid \beta)$. Then a scaleless function of the data alone, appropriate to measure discrepancies from the value β_0 taken in the current model, is provided by Fisher's score function

$$g_\beta(\mathbf{y}) = \left.\frac{\partial \ln p(\mathbf{y} \mid \beta)}{d\beta}\right|_{\beta=\beta_0}. \tag{4.2}$$

We illustrate by considering some possible discrepancies from the standard normal linear model. First consider the model when there is no discrepancy so that $\beta = \beta_0$, and using the structure of (3.1) write

$$\boldsymbol{\Theta}' = (\mu \,\vdots\, \boldsymbol{\theta}'), \quad \mathscr{X} = (\mathbf{1} \,\vdots\, \mathbf{X}). \tag{4.3}$$

For simplicity we here suppose that the distributions of $\boldsymbol{\Theta}$ and $\ln \sigma$ are locally flat *a priori* so that $p(\boldsymbol{\Theta}, \sigma) \doteq \text{const}\, \sigma^{-1}$. Then $p(\mathbf{y} \mid \beta_0)$ is locally approximated by the singular distribution

$$p(\mathbf{y} \mid \beta_0) \doteq \text{const}\, S^{-\nu}, \tag{4.4}$$

where $S^2 = \Sigma_{i=1}^n (y_i - \hat{y}_i)^2 = \mathbf{y}' \mathbf{R} \mathbf{y}$ and $\mathbf{R} = \mathbf{I} - \mathbf{M}$ with $\mathbf{M} = \mathscr{X}(\mathscr{X}'\mathscr{X})^{-1} \mathscr{X}'$. If we transform to $\hat{\boldsymbol{\Theta}}$, S and $\mathbf{u}$ then the standardized residuals $\mathbf{u}$ which are functions of $\nu - 1$ angles are distributed as in (3.8) and,

$$p(\hat{\boldsymbol{\Theta}}, S, \mathbf{u} \mid \beta_0) \doteq \text{const}\, S^{-1} p(\mathbf{u} \mid \beta_0). \tag{4.5}$$

To see the reasonableness of this set-up notice that by invocation of the linear model the investigator in effect predicts that the sample point $\mathbf{y}$ will lie somewhere close to a hyperplane $\boldsymbol{h}_{\mathscr{X}}$ spanned by the columns of $\mathscr{X}$. The formulation above interprets "somewhere close to" as follows. Consider a future sample $\mathbf{y}$ in relation to $(\hat{\boldsymbol{\Theta}}, S)$ where $\hat{\boldsymbol{\Theta}}$ are the $k+1$ coordinates of the projection $\mathbf{y}$ of $\mathbf{y}$ on $\boldsymbol{h}_{\mathscr{X}}$, and S is the perpendicular distance of $\mathbf{y}$ from $\boldsymbol{h}_{\mathscr{X}}$. Equation (4.5) says that locally any value of $\hat{\boldsymbol{\Theta}}$ is equally acceptable but that the density for the distance S will fall off inversely with S.

To obtain $g_\beta(\mathbf{y})$ we need to determine how $p(\mathbf{y} \mid \beta)$ depends on the discrepancy parameter β in the neighbourhood of $\beta = \beta_0$.

4.3. *A Check for Needed Power Transformation*

Especially when $y_{\max}/y_{\min}$ is large some transformation of the data, for example $y^{(\lambda)} = (y^\lambda - 1)/\lambda$, might permit closer representation. Following the approximate argument of Box and Cox (1964), with λ the discrepancy parameter and $\dot{y}$ the geometric mean of the y's, and for λ close to 1,

$$p(\mathbf{y} \mid \lambda) \,\dot{\propto}\, \dot{y}^{\nu(\lambda - 1)} Q_\lambda^{-\frac{1}{2}\nu}, \tag{4.6}$$

where the omitted constant does not depend on $\mathbf{y}$ or on λ and where $Q_\lambda = \mathbf{y}^{(\lambda)'} \mathbf{R} \mathbf{y}^{(\lambda)}$,

$$g_\lambda(\mathbf{y}) = \left.\frac{\partial \ln p(\mathbf{y} \mid \lambda)}{\partial \lambda}\right|_{\lambda = 1} \doteq \mathbf{z}'\mathbf{R}\mathbf{y}/s^2 = s^{-1} \sum_{i=1}^n z_i r_i \tag{4.7}$$

where $z_i = y_i\{1 - \ln(y_i/\dot{y})\}$, $s^2 = \mathbf{y}' \mathbf{R} \mathbf{y}/\nu$ and $r_i = (y_i - \hat{y}_i)/s$. The predictive check may thus be performed by regressing the residuals $y - \hat{y}$ on the residuals $z - \hat{z}$ of the constructed variable $z = y\{1 - \ln(y/\dot{y})\}$, which accords with a proposal of Atkinson (1973). The check can be made informally by plotting one set of residuals versus the other. More formally the distribution of $g_\lambda(\mathbf{y})$ is not precisely known although an approximate level can be obtained by computer simulation.

Relation to other proposed checks

Related checks proposed by Tukey (1949) and by Andrews (1971a) correlate the original residuals with those from the constructed variables $(\hat{y} - \bar{y})^2$ and $\hat{y} \ln \hat{y}$ respectively. Both possess the advantage of having exactly known sampling distributions.

For illustration we consider:

(a) the biological data of Box and Cox (1964), for which they recommend a reciprocal transformation;

(b) the trapping data of Snedecor and Cochran (1967), for which they recommend a log transformation.

Figures 2(a) and (b) show plots of residuals $y-\hat{y}$ against residuals from $y\{1-\ln(y/\dot{y})\}$ and $-(\hat{y}-\bar{y})^2$. The correlation coefficient for the latter transforms directly to give Tukey's one degree of freedom for non-additivity. Plots for the constructed variable $\hat{y}\ln\hat{y}$ are not shown since they are essentially identical to the Tukey plot. The relationship between these various procedures can be seen by noting that $z = y\{1-\ln(y/\dot{y})\}$ may be closely approximated by

$$z \doteq \dot{y} - B(y-\dot{y})^2. \tag{4.8}$$

Thus after writing

$$\frac{y_i-\dot{y}}{s} = \frac{y_i-\hat{y}_i}{s} + \frac{\hat{y}_i-\bar{y}}{s} + \frac{\bar{y}-\dot{y}}{s} = r_i + Y_i + d, \tag{4.9}$$

$$g_\lambda(\mathbf{y}) \propto -\Sigma(r_i+Y_i+d)^2 r_i = -\{\Sigma r_i^3 + 2\Sigma r_i^2 Y_i + \Sigma r_i Y_i^2 + 2\Sigma r_i^2 d\} \tag{4.10}$$

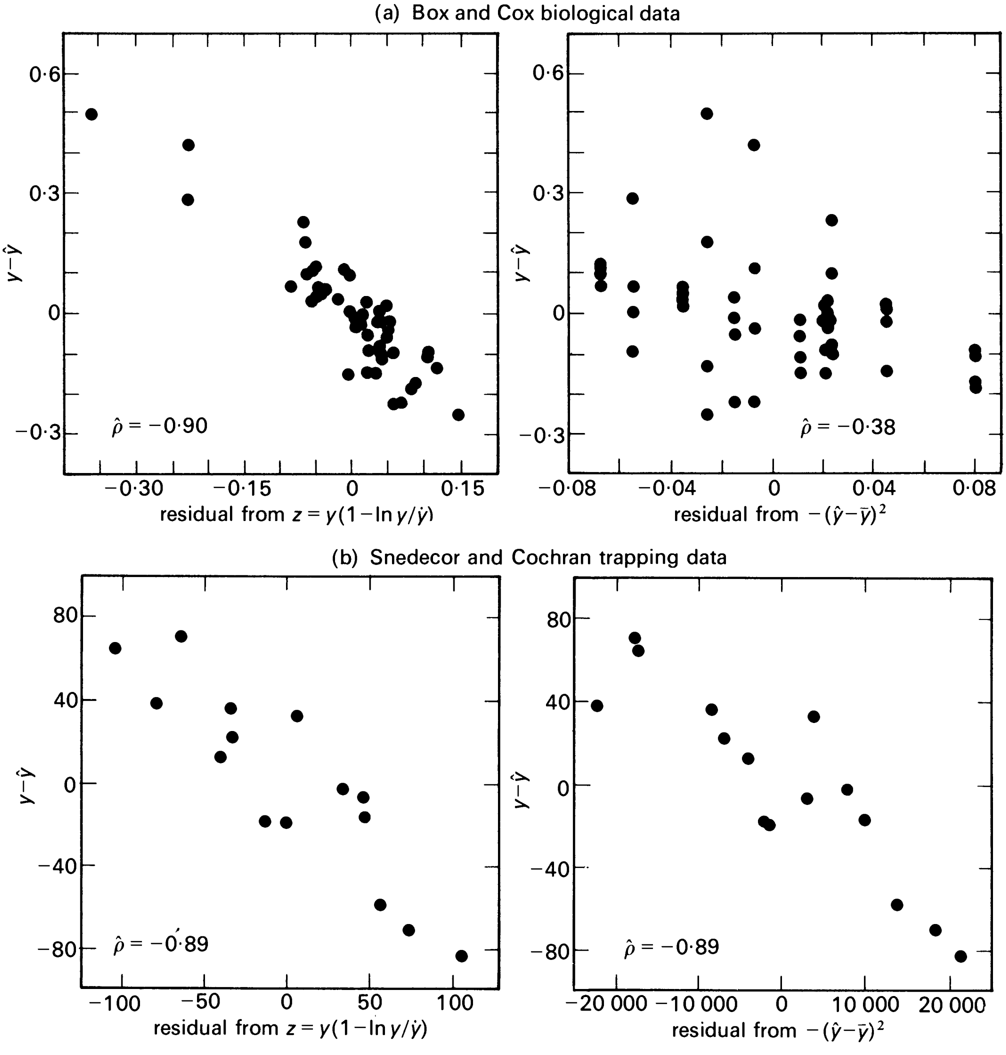

FIG. 2. Plots of residuals $y-\hat{y}$ against residuals from two constructed variables $z = y(1-\ln y/\dot{y})$ and $-(\hat{y}-\bar{y})^2$, with correlation coefficient $\hat{\rho}$ to indicate strength of association.

and

$$g_\lambda(\mathbf{y}) \dot{\propto} -(T_{30}+2T_{21}+T_{12}+2\nu d), \tag{4.11}$$

where the T_{ij} are checking functions proposed by Anscombe (1961) and Anscombe and Tukey (1963). See also Box and Cox (1964). In particular T_{12} is the component associated with Tukey's one degree of freedom for non-additivity. The approximation shows how $g_\lambda(\mathbf{y})$ jointly employs skewness (T_{30}), dependence of variance on level (T_{21}), as well as transformable non-additivity (T_{12}) to indicate the need for transformation.

The Box and Cox data were generated by a 3 × 4 factorial with four-fold replication supplying a good deal of information about the variance as a function of location. It is not surprising therefore (see also Atkinson, 1973) that for this example $g_\lambda(\mathbf{y})$ is considerably more sensitive than T_{12} (or almost equivalently, than Andrew's criterion) as a measure of the need for transformation. By contrast the Snedecor and Cochran data are from an unreplicated 3 × 5 arrangement where most of the information comes from T_{12} measuring non-additivity.

4.4. *A check for Serial Correlation*

For data known or suspected to have been taken in a specific serial order in time or space, a model that permitted the errors to follow a first-order autoregressive process with parameter $|\phi|<1$ might provide an improved approximation. The dispersion matrix for the n-dimensional vector of errors $\mathbf{e}$ would then be $\mathbf{W}_\phi^{-1}\sigma^2$ where $\mathbf{W}_\phi$ is a symmetric continuant with principal diagonal $\{1, 1+\phi^2, 1+\phi^2, \ldots, 1+\phi^2, 1\}$ and with all the elements of super- and sub-diagonals equal to $-\phi$. Thus in particular $\mathbf{W}_0 = \mathbf{I}$. Then

$$p(\mathbf{y}\,|\,\phi) \dot{\propto} (1-\phi^2)^{\frac{1}{2}} Q_\phi^{-\frac{1}{2}\nu}, \tag{4.12}$$

where $Q_\phi = \mathbf{y}'\mathbf{R}_\phi\mathbf{y}$ and $\mathbf{R}_\phi = \mathbf{W}_\phi - \mathbf{M}_\phi$ with $\mathbf{M}_\phi = \mathbf{W}_\phi\mathscr{X}(\mathscr{X}'\mathbf{W}_\phi\mathscr{X})^{-1}\mathscr{X}'\mathbf{W}_\phi$. Then with $\partial\mathbf{W}_\phi/\partial\phi\,|_{\phi=0} = -\mathbf{C}$ where $\mathbf{C}$ is $n \times n$ with unities in super- and sub-diagonals and zeros elsewhere, after some algebraic manipulation,

$$g_\phi(\mathbf{y}) = \left.\frac{\partial \ln p(\mathbf{y}\,|\,\phi)}{\partial\phi}\right|_{\phi=0} \doteq \tfrac{1}{2}(\mathbf{y}'\mathbf{RCRy})/s^2, \tag{4.13}$$

where $R = R_0$. Thus

$$g_\phi(\mathbf{y}) \doteq \sum_{i=1}^{n-1} r_i r_{i+1} \tag{4.14}$$

which is a multiple of the sample first lag autocorrelation of the residuals from the fitted model. This points to the sensitive graphical diagnostic procedure of plotting residuals r_{i+1} against r_i and yields the standard checking function of Durbin and Watson (1950).

4.5. *A Check for Bad Values*

Competent investigators have over the centuries treated data as possibly containing atypical values, see for example Stigler (1973). This imples that they would not really have believed standard textbook models of the kind $y_i = f(\boldsymbol{\theta}, \mathbf{x}_i) + e_i$ $(i = 1, 2, \ldots, n)$ which state that the same structure is appropriate for *every one* of a sample of n observations.

When it is unknown which observations are dubious a more credible "contaminated" model proposed by Jeffreys (1932), Dixon (1953) and by Tukey (1960) supposes that there is a probability α that any given observation is "bad" (cannot be represented by the ideal model). Given α, let $p(\mathbf{y}\,|\,\alpha)$ be the predictive distribution and let $p(b\,|\,\alpha)$ denote the probability of getting b bad values, then (Box and Tiao, 1968b; Bailey and Box, 1980a)

$$p(\mathbf{y}\,|\,\alpha) = \sum_{b=0}^{n} \binom{n}{b} \alpha^b (1-\alpha)^{n-b} p(\mathbf{y}\,|\,b) \tag{4.15}$$

and

$$g_\alpha(\mathbf{y}) = \frac{\partial \ln p(\mathbf{y} \mid \alpha)}{\partial \alpha}\bigg|_{\alpha=0} = n\left\{\frac{p(\mathbf{y} \mid b=1)}{p(\mathbf{y} \mid b=0)} - 1\right\}. \tag{4.16}$$

Now let z_i indicate that the ith observation is bad, then

$$p(\mathbf{y} \mid b=1) = n^{-1} \sum_{i=1}^{n} p(\mathbf{y} \mid z_i) \tag{4.17}$$

so

$$g_\alpha(\mathbf{y}) = \left(\sum_{i=1}^{n} p(\mathbf{y} \mid z_i) \Big/ p(\mathbf{y} \mid b=0)\right) - n.$$

Depending on experimental circumstances, there are a variety of ways in which bad values might be modelled. In particular, contamination could come from increased error variance, unknown bias, and mistaken sign. The last possibility was suggested by Barnard (1978) to account for two suspiciously large outliers in Darwin's data on cross- and self-fertilized plants, quoted by Fisher (1935).

For illustration consider the first possibility. With one bad value, suppose the error covariance matrix $W_i^{-1}\sigma^2$ has all elements equal to σ^2, except for the ith element which is equal to $\kappa^2\sigma^2 (\kappa > 1)$. Then

$$g_\alpha(\mathbf{y}) = \kappa^{-1}\left(\frac{n}{n-q}\right)^{\frac{1}{2}} \sum_{i=1}^{n} \left(\frac{s_i^2}{s^2}\right)^{-\frac{1}{2}\nu} - n, \tag{4.18}$$

where

$$\nu s^2 = \mathbf{y}'\mathbf{R}\mathbf{y}, \quad \nu s_i^2 = \mathbf{y}'\mathbf{R}_i\,\mathbf{y} = \mathbf{y}'\{\mathbf{W}_i - \mathbf{W}_i\mathscr{X}(\mathscr{X}'\mathbf{W}_i\mathscr{X})^{-1}\mathscr{X}'\mathbf{W}_i\}\,\mathbf{y}, \tag{4.19}$$

where $\mathbf{W}_i = \mathbf{I} - q\mathbf{G}_i$, $q = 1-\kappa^{-2}$ and $\mathbf{G}_i$ is an $n \times n$ matrix with a single unity for the ith diagonal element and all other elements zero. Now

$$\nu s_i^2 = \nu s^2 - \frac{q}{1-qv_i}(y_i - \hat{y}_i)^2,$$

where $v_i = \text{var}(\hat{y}_i)/\sigma^2 = x_i'(\mathscr{X}'\mathscr{X})^{-1}x_i$ and $y_i - \hat{y}_i$ is the ith residual from the ideal model fit, $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{R}\mathbf{y}$.

Thus finally $g_\alpha(\mathbf{y}) = \kappa^{-1}\{n/(n-q)\}^{\frac{1}{2}} D - n$ where

$$D = \Sigma\left\{1 - \frac{q}{\nu(1-qv_i)} r_i^2\right\}^{-\frac{1}{2}\nu}. \tag{4.20}$$

This is the simplest form for computation. The nature of this checking function D can however be more clearly seen by writing it in terms of the weighted residuals $\tilde{r}_i = (y_i - \tilde{y}_i)/s_i$ where $\mathbf{y} - \tilde{\mathbf{y}} = \mathbf{R}_i\,\mathbf{y}$. Thus

$$D = \Sigma\left\{1 + \frac{q(1-qv_i)}{\nu}\tilde{r}_i^2\right\}^{\frac{1}{2}\nu}.$$

Thus D is proportional to the sum of the reciprocals of the n residual t ordinates obtained by downweighting (omitting as $q \to 1$) each observation in turn and recomputing the fitted value $\tilde{y}_i$ and the standard deviation s_i.

The situation of most interest is when κ is large (say $\kappa \geqslant 5$). Then q approaches unity and the check may be carried out by calculating

$$D = \Sigma\left(1 - \frac{r_i^2}{\nu(1-v_i)}\right)^{-\frac{1}{2}\nu} = \Sigma\left\{1 + \frac{(1-v_i)}{\nu}\tilde{r}_i^2\right\}^{\frac{1}{2}\nu}. \tag{4.21}$$

Equation (4.16) brings out a feature of the checking function (4.2) which can be a disadvantage. Differentiation at $\alpha = 0$ on the boundary of the parameter space ensures that only the possibility of one bad value is taken account of. Thus as is clear from (4.21), D in its present form would not be expected to be sensitive to the occurrence of two or more bad values. Thus with $\kappa = 5$, we obtain the value $D = 59{\cdot}05$ for Darwin's data. A Monte Carlo study with 5000 samples of 15 observations shows that this value would be exceeded by chance about 14 per cent of the time, which hints only mildly at inadequacy in the standard model, confirming D's insensitivity for this example.

4.6. *Bayesian Options for Specific Alternatives*

When concrete alternatives are in mind, Bayesian options are available. In particular, the predictive ratio $p(\mathbf{y} \mid A_1)/p(\mathbf{y} \mid A_0)$, mentioned earlier, is a component in the posterior odds ratio which, with suitable priors, might be used to assess directly the relative evidence for one model versus another. Also $g_\beta(\mathbf{y})$ of (4.2) has a Bayesian interpretation for, if corresponding to some discrepancy parameter β, the prior distribution $p(\beta)$ was locally flat then the posterior distribution $p(\beta \mid \mathbf{y}_d)$ would be proportional to the predictive density $p(\mathbf{y}_d \mid \beta)$ regarded as a function of β. Furthermore, if that posterior distribution was approximated by a normal distribution, then

$$g_\beta(\mathbf{y}) \doteq \{-(\beta_0 - \hat{\beta})/\sigma_\beta^2\} \tag{4.22}$$

and a second differentiation would produce a standardized variate.

The relation shows how any specific predictive check $g_\beta(\mathbf{y})$ is linked to a posterior distribution. In particular, considering the illustrative examples of the last section, the marginal posterior distribution for λ was given by Box and Cox (1964), for ϕ by Zellner and Tiao (1964), for the ridge regression parameters by Lindley and Smith (1972) and that for α may be obtained using the results of Box and Tiao (1968b) and Bailey and Box (1980a).

Before leaving the topic of diagnostic checks two final points need to be made:

(i) The above discussion illustrates the "overlap" previously mentioned when specific alternatives are in mind. It does not, however, establish the omnipotence of purely Bayesian inference. However far the process of model elaboration is taken by Bayesian methods the final model involving say the mth set of assumptions A_m can still be factored

$$p(\mathbf{y}, \boldsymbol{\theta} \mid A_m) = p(\boldsymbol{\theta} \mid \mathbf{y}, A_m)\, p(\mathbf{y} \mid A_m) \tag{4.23}$$

thus there always remains an unexplored n-dimensional predictive distribution $p(\mathbf{y} \mid A_m)$ in relation to which a small relative value for $p(\mathbf{y}_d \mid A_m)$ could, on a sampling theory argument, discredit the assumptions on which the Bayes analysis was conditional. The same is true of the more plausible of two models chosen using a posterior odds ratio.

(ii) In addition to possible discrepancies to which we have been alerted by experience, other features may appear pointing to inadequacies of a kind not previously suspected. This possibility has sometimes proved perplexing, for while on the one hand the truly unexpected could point the way to precious new knowledge, on the other, associated probabilities would be indeterminate because of the uncountable character of other features that might also have been regarded as surprising. I think the calculation which ignores this difficulty of indeterminate selection is still worth making, for at least it helps to correct a misjudgement of something that appears unusual but really is not. For example, Feller (1968) shows that for a random group of 30 people, the probability that at least two have coincident birthdays is over 70 per cent; this tells us we need look no further for an explanation when we are surprised to find two such people at a party. While the proposed policy will lead to the too frequent pursuit of non-existent assignable causes, the iterative process will quickly terminate this chase.

5. Robust Estimation

Efficient model building requires both *diagnostic checking* and *model robustification*, where by robustification I mean judicious and grudging elaboration of the model to ensure against particular hazards (see also Box, 1979). Robustification becomes necessary when it is known that likely, but not easily detectable, model discrepancies can yield badly misleading analyses. It is well known, for example, that least squares analysis can be dramatically affected by moderate serial correlation of errors.

Recently the serious consequences of bad values on standard least squares analysis has been especially emphasized and numerous authors have proposed methods which rely on abandonment or modification of classical estimation methods. In discussing the rationale for this approach Huber (1977) says "The traditional approach to theoretical statistics was and is to optimize an idealized model and then to rely on a continuity principal: what is optimal at the model should be optimal nearby. Unfortunately, this reliance on continuity is unfounded: the classical optimal procedures tend to be discontinuous in the statistically meaningful topologies."

He then quotes a motivating example given by Tukey (1960), who pointed out for example that if a normal distribution were very mildly contaminated with another which is centrally located but of larger variance, then the sample standard deviation could be a very poor estimate of scale. Tukey's contribution was remarkable because it had previously gone unnoticed that the assumption that the same structure must apply to *every* observation $y_i\,(i = 1, 2, \ldots, n)$ *with absolute certainty* $(1-\alpha = 1)$, not only was unrealistic (since no responsible investigator would make the claim that inadvertent bad values were impossible), but also could have serious consequences. While Huber goes on to say that typical "good data" samples in the physical sciences appear to be well modelled by this contaminated normal model, he does not develop methods based on this more realistic set up. This is presumably because his objection would apply equally to the new as well as to the old model.

I do not agree that the example would support a thesis of the need to abandon model-based procedures. A model that omits the parameter α is, of course, the same as one that includes it but sets its value exactly to zero. A value of $\alpha = 0$, which allows no possibility whatever for bad values, and a value of $\alpha = 0{\cdot}001$ are, I think, *not close* in any statistically meaningful topology. Although 0·001 may look close to zero, an odds ratio of $0{\cdot}999/0{\cdot}001 = 999$ for a "good" to a "bad" value is obviously very different from one of infinity. Such differences in probability distinguish, for example, a lifeless world in which no evolution could possibly occur from the one we live in.

The proper conclusion to draw from Tukey's example is, I think, that for many practical situations in which occasional bad values are to be expected the standard linear model provides an inadequate approximation that is potentially misleading and therefore the model should be appropriately changed to approximate what is believed rather than what is not. The situation is logically the same for a model that implicitly insists there can be no serial correlation, when data have in fact been collected serially, or that no transformation of y could be needed when $y_{\max}/y_{\min}$ is large. As in the classical Stein problem if we know something *a priori* it may be disastrous to omit it. On this view for robust estimation of the parameters of interest we should modify the model which is at fault, rather than the method of estimation which is not.

5.1. *Bayesian Robust Estimation*

As was argued for example by Box and Tiao (1964), all relevant aspects of the problem are brought out in an appropriate Bayes analysis. Supposing that $\boldsymbol{\theta}$ has the same physical interpretation for all β then estimation of $\boldsymbol{\theta}$ which is robust relative to the discrepancy parameter β is supplied by the posterior distribution

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \int p(\boldsymbol{\theta} \mid \beta, \mathbf{y})\, p_u(\beta \mid \mathbf{y})\, p(\beta)\, d\beta. \tag{5.1}$$

This expression contains three key elements that repay individual study:

(a) the sensitivity of inferences about $\boldsymbol{\theta}$ to changes in β is reflected by $p(\boldsymbol{\theta} \mid \beta, \mathbf{y})$ considered as a function of β;

(b) the information about β coming from the data themselves is reflected in the pseudo-likelihood

$$p_u(\beta \mid \mathbf{y}) = p(\beta \mid \mathbf{y})/p(\beta) \propto p(\mathbf{y} \mid \beta); \tag{5.2}$$

(c) the probability of occurrence of different values of β in the real world is represented by $p(\beta)$ which can be chosen to approximate what is believed or feared.

This route was used to explore deviations from the standard normal model for a particular class of heavy-tailed distributions by Box and Tiao (1962, 1964); for the contaminated model of Section 4.5 by Jeffreys (1932) and Box and Tiao (1968b); for a serial correlation model by Zellner and Tiao (1964); for a transformation problem by Box and Cox (1964). Notice that using this approach the parameters $\boldsymbol{\theta}$ of interest are completely estimated in the sense that their distribution rather than merely a point estimate is available. Also the various elements of $p(\boldsymbol{\theta} \mid \mathbf{y})$ which can be studied individually can provide a deep understanding of each robustness problem. A particularly informative display shows contours of the joint distribution $p_u(\theta, \beta \mid \mathbf{y})$ for some parameter θ of interest and the discrepancy parameter β together with the marginal distribution $p_u(\beta \mid \mathbf{y})$. When a less prodigal display is necessary the mean and standard deviation of $p(\theta \mid \beta, \mathbf{y})$ may be shown with $p_u(\beta \mid \mathbf{y})$. For illustration we consider some serial data analysed by Coen, Gomme and Kendall (1969). They regressed quarterly values of the Financial Times Share Index y on detrended lagged values of UK car production X_1, and of the Financial Times Commodity Index X_2 using a model† which could be written (Box and Newbold, 1971) as

$$y_t = \beta_0 + \beta_1 t + \theta_1 X_{1,t-6} + \theta_2 X_{2,t-7} + e_t \quad \text{with } e_t = \phi e_{t-1} + a_t \tag{5.3}$$

with a_t white noise, and ϕ constrained to be equal to zero. Fig. 3 illustrates an analysis made by Pallesen (1977) in which ϕ is unconstrained. It shows the joint posterior distribution for θ_1 and ϕ and the marginal distribution for ϕ assuming locally flat priors for $\boldsymbol{\theta}$, $\ln \sigma$ and ϕ. Although for this example serial correlation could have been easily detected by diagnostic checks, notice the enormous shift (about five standard deviations) of the conditional distribution $p(\theta_1 \mid \phi, \mathbf{y})$ which occurs as ϕ changes from zero to more plausible values. This illustrates the point that smaller serial correlation, of a magnitude difficult to detect with diagnostic checks, could disastrously invalidate estimates of $\boldsymbol{\theta}$.

A second example discussed more fully in Bailey and Box (1980b) further illustrates this approach for the "bad value" problem using the contaminated normal model of Section 4.5. The data were used originally by Box and Behnken (1960) to illustrate the analysis of a balanced incomplete four factor three-level design with $n = 27$ observations arranged in three blocks of nine. A residual plot suggests the possibility of two bad values (y_{10} and y_{13}). However, the small number of residual degrees of freedom and the nature of this particular design would induce large correlations yielding potentially misleading residual patterns.

Table 1 gives Bayesian means and standard deviations for coefficients in the fitted model

$$y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \beta_{ii} x_i^2 + \sum_{i=1}^{4} \sum_{j>1}^{4} \beta_{ij} x_i x_j + e. \tag{5.4}$$

In this analysis κ was set equal to 5 and the values of α varied over the range 0 to 0·091. It has been shown by Chen and Box (1979) that for $\kappa \geqslant 5$ the posterior distribution is mainly a function of $\varepsilon = \alpha/(1-\alpha)\kappa$ so the results are also labelled in terms of this dominant discrepancy parameter ε. It will be noticed:

(a) The large change in each estimated effect and standard deviation occurs when no possibility whatever of bad values ($\varepsilon = 0$) is replaced by a small possibility ($\varepsilon = 0·001$). For good

† For the present purpose we retain the model structure of Coen, Gomme and Kendall. However, its relevance seems dubious, for example, a multivariate time series analysis by Tiao and Box (1980) for these three series shows the stock prices y acting as a weak *leading* indicator for the commodity index X_2.

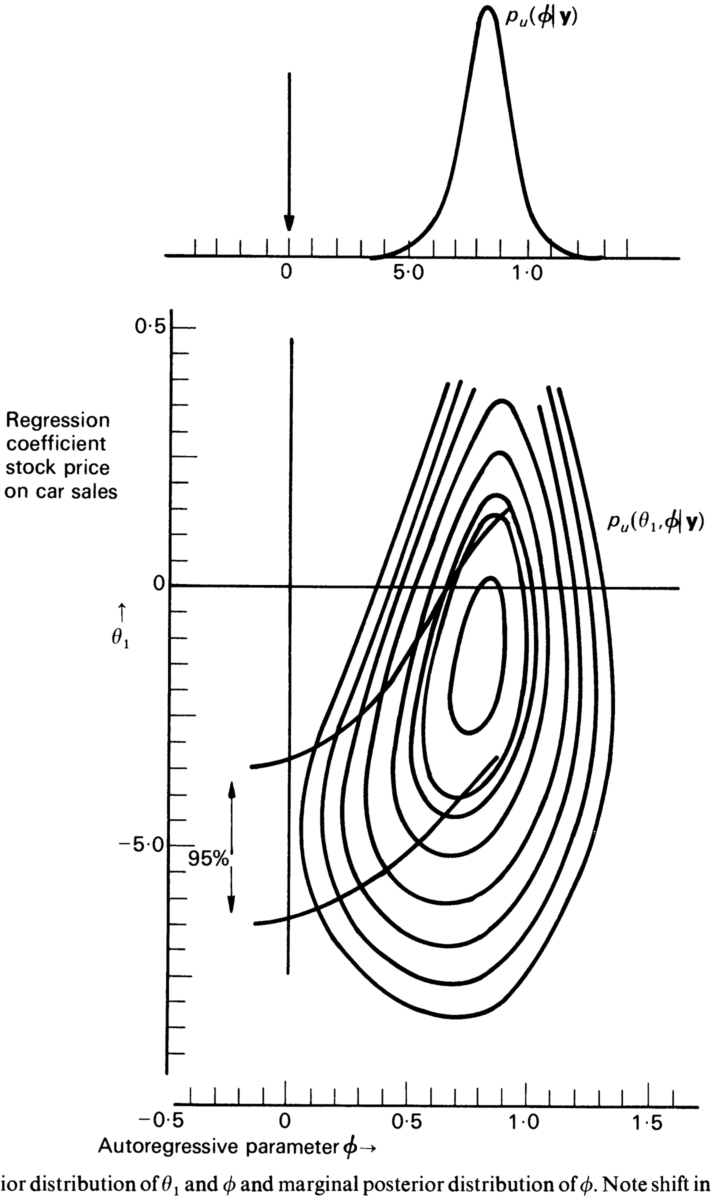

FIG. 3. Joint posterior distribution of θ_1 and ϕ and marginal posterior distribution of ϕ. Note shift in approximate 95 per cent interval as ϕ is changed.

data the typical behaviour of a table of this kind is that only very minor changes in mean and standard deviation occur as ε is changed over the plausible range.

(b) For all the estimates except β_{14} the standard deviations of effects are about halved. Thus for these effects the use of the more appropriate model is equivalent to a four-fold increase in the size/sensitivity of the experiment. This may be compared for example with a parallel analysis by Box and Cox of their biological data where a three-fold increase in sensitivity resulted from the use of an appropriate transformation.

(c) The analysis can be further illuminated by considering other available quantities. In particular a plot of the probability that the ith value is bad, given that one value is bad (see, for example, Abraham and Box, 1978), results in a plot with 94 per cent of the probability associated with the tenth observation and the remainder spread among the remaining 26 observations. It is likely therefore that y_{10} alone is a bad value. It is a deficiency of the design being used here that least squares estimates of interactions employ only four of the 27 observations and so lack robustness to bad observations (see, for example, Box and Draper, 1975). In particular $\hat{\beta}_{14} = 0{\cdot}25\,(y_{10}-y_{11}-y_{12}+y_{13})$ so that the Bayesian down-weighting of y_{10} accounts for the large change in this estimate and the *increase* in the standard deviation.

(d) We saw in the case of ridge regression how failure to take account of observational information could lead to an unrealistic choice of the discrepancy parameter γ. To complete the picture, therefore, a plot of the marginal distribution of the discrepancy parameter ε should be made in conjunction with Table 1 (compare also with the serial correlation example in Fig. 3). For these data the distribution $p_u(\varepsilon\,|\,\mathbf{y})$ has its mode close to $\varepsilon = 0{\cdot}010$.

Table 1

Bayesian means and (standard deviations) for polynomial coefficients using the contaminated model of Section 4.5 with $\kappa = 5$ $(\varepsilon = \alpha/(1-\alpha)\kappa)$

ε	0	0·001	0·005	0·010	0·015	0·020
α	0	0·005	0·024	0·048	0·070	0·091
β_0	90·60 (0·94)	90·60 (0·45)	90·60 (0·41)	90·60 (0·41)	90·60 (0·41)	90·60 (0·41)
β_1	1·93 (0·47)	2·46 (0·28)	2·49 (0·23)	2·49 (0·22)	2·49 (0·22)	2·49 (0·22)
β_2	−1·96 (0·47)	−1·96 (0·22)	−1·96 (0·20)	−1·96 (0·20)	−1·96 (0·20)	−1·96 (0·20)
β_3	1·13 (0·47)	1·13 (0·22)	1·13 (0·20)	1·13 (0·20)	1·13 (0·20)	1·13 (0·20)
β_4	−3·68 (0·47)	−3·15 (0·28)	−3·12 (0·23)	−3·12 (0·22)	−3·12 (0·22)	−3·12 (0·22)
β_{11}	−1·42 (0·70)	−1·88 (0·44)	−1·90 (0·41)	−1·90 (0·41)	−1·89 (0·42)	−1·89 (0·42)
β_{22}	−4·33 (0·70)	−4·10 (0·36)	−4·09 (0·34)	−4·09 (0·34)	−4·09 (0·34)	−4·09 (0·34)
β_{33}	−2·24 (0·70)	−2·01 (0·38)	−2·00 (0·34)	−2·00 (0·34)	−2·00 (0·34)	−2·00 (0·34)
β_{44}	−2·58 (0·70)	−3·05 (0·44)	−3·06 (0·41)	−3·06 (0·41)	−3·06 (0·42)	−3·05 (0·42)
β_{12}	−1·67 (0·81)	−1·67 (0·39)	−1·67 (0·35)	−1·67 (0·35)	−1·67 (0·34)	−1·67 (0·34)
β_{13}	−3·83 (0·81)	−3·82 (0·39)	−3·82 (0·35)	−3·82 (0·35)	−3·82 (0·34)	−3·82 (0·34)
$\rightarrow \beta_{14}$	0·95 (0·81)	−0·45 (0·95)	−0·51 (0·92)	−0·50 (0·93)	−0·49 (0·95)	−0·48 (0·95)
β_{23}	−1 67 (0·81)	−1·67 (0·39)	−1·67 (0·35)	−1·67 (0·35)	−1·67 (0·35)	−1·67 (0·35)
β_{24}	−2·62 (0·81)	−2·62 (0·39)	−2·62 (0·35)	−2·62 (0·35)	−2·62 (0·35)	−2·62 (0·35)
β_{34}	−4·25 (0·81)	−4·25 (0·39)	−4·25 (0·35)	−4·25 (0·34)	−4·25 (0·34)	−4·25 (0·34)

The Bayes approach to robust estimation has the advantage of generality; furthermore it clearly reveals at any given stage, on precisely what assumptions the analysis is conditional. With the increased speed of computers and availability of visual display equipment a general Bayesian computer program, that can analyse any model we wish to entertain, seems a much more attractive prospect than the fresh devising of semi *ad hoc* procedures for each new possibility.

Some parallels in the two approaches are briefly considered below for the "bad value" problem.

5.2. *Robust Estimation for the "bad value" Problem*

For the "bad value". problem a wide variety of semi-empirical estimators have been proposed. Among these are the M, L and R, and various kinds of adaptive estimators. In turn among the M estimators a number of different "ψ" functions have been suggested leading to different ways of downweighting extreme observations.

Now consider the model of Section 4.5 for the simple location structure $E(y_i) = \mu$. Then (see, for example, Box and Tiao, 1968b) the Bayesian mean may be written

$$\hat{\mu} = \sum_{b=0}^{n} p(b \mid \mathbf{y}, \alpha)\bar{y}^{(b)}, \tag{5.5}$$

where $p(b \mid \mathbf{y}, \alpha)$ is the posterior probability that there are b bad values and $\bar{y}^{(b)}$ is the corresponding conditional posterior mean. Consider in particular $\bar{y}^{(1)} = \Sigma\, w_i\, y_i$. Then Chen and Box (1979) show that for $\kappa \geqslant 5$

$$w_i \doteq (n-1)^{-1}(1 - D_i/D), \tag{5.6}$$

$$D_i = \left\{1 - \frac{n r_i^2}{(n-1)^2}\right\}^{-\frac{1}{2}(n-1)} = \{1 + n^{-1}\tilde{r}_i^2\}^{\frac{1}{2}(n-1)}, \tag{5.7}$$

where r_i and $\tilde{r}_i$ are unweighted and weighted residuals defined in Section 4.5. Fig. 4(a) and (b) show plots of $w = w_1$ against r_1 and $\tilde{r}_1$ for three random normal samples of ten observations from a normal distribution when a multiple $0, 1, 2, \ldots,$ of σ is added to the first observation in each sample. Empirical approximations for these weighting curves are provided by the functions

$$w = 0{\cdot}1 \exp\{-|0{\cdot}49\, r_1|^7\} \quad \text{and} \quad w = 0{\cdot}1 \exp\{-|0{\cdot}3\, \tilde{r}_1|^{3{\cdot}5}\}.$$

Also shown in Fig. 4(b) for comparison is Tukey's biweight function $w = 0{\cdot}1\{1 - (\tilde{r}/c)^2\}^2$ for $c = 5{\cdot}3$ (chosen to roughly match the curve). Although the Bayesian weights are sample dependent they remain remarkably stable as is indicated (a) by the smooth manner in which the remaining weight is evenly spread throughout the non-discrepant observations; (b) by the closeness with which points from different samples follow the same curve.

The estimate $\hat{\mu}$ is sample adaptive in another more striking way however. For illustration consider the case where the $p(b \mid \mathbf{y}, \alpha)$ are negligible for $b \geqslant 2$. Then writing $p = p(1 \mid \mathbf{y}, \alpha)$ (5.5) becomes

$$\hat{\mu} = (1-p)\,\bar{y} + p\bar{y}^{(1)} \tag{5.8}$$

and the Bayesian mean is an interpolation between $\bar{y}$ and the "robustified" $\bar{y}^{(1)}$. In this expression the value of p is determined by the posterior odds ratio for one *vs* no bad values

$$p/(1-p) \doteq \varepsilon\{n/(n-1)\}\, D, \quad \varepsilon = \alpha/(1-\alpha)\kappa \tag{5.9}$$

and D is the checking function encountered earlier.

Sample adaptivity is evidenced as follows. For a sample with no outliers $\bar{y}$ and $\bar{y}^{(1)}$ are not very different so that $\hat{\mu}$ is close to $\bar{y}$. But in the presence of an outlier of larger and larger size two

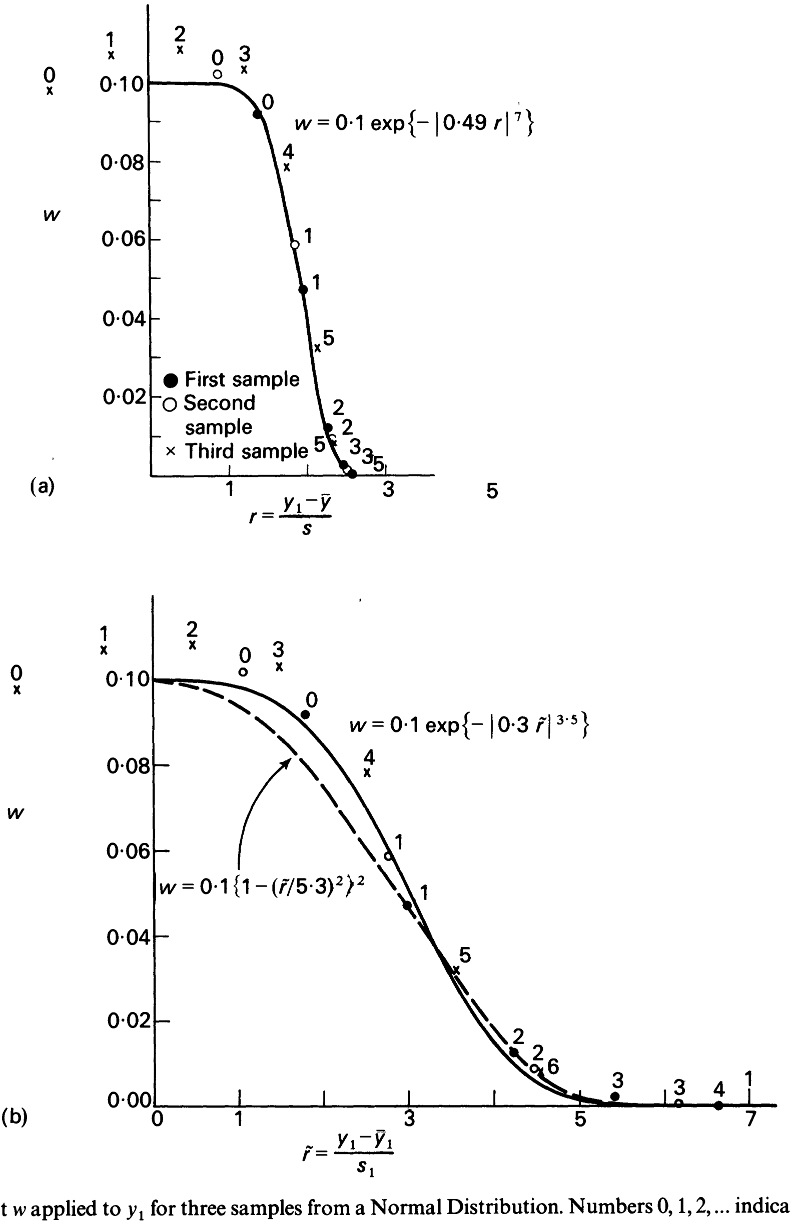

FIG. 4. Weight w applied to y_1 for three samples from a Normal Distribution. Numbers 0, 1, 2, ... indicate that $0, \sigma, 2\sigma, \ldots$ has been added to y_1.

things happen: the outlier is downweighted in $\tilde{y}^{(1)}$ which becomes more and more different from $\bar{y}$ and also p becomes larger placing more and more emphasis on $\tilde{y}^{(1)}$.

The purpose of this discussion is to show that sensible solutions which appropriately downweight suspected bad values may be obtained directly from an appropriate model. From the viewpoint of the traditional M estimator, the weight function (W say) for $\hat{\mu}$ itself is an interpolation $W = (1-p)n^{-1} + pw$ between $1/n$ and w. Thus W will descend to the value $(1-p)/n$ for large $\tilde{r}$. For a sample containing a large outlier, $1-p$ will be negligible and W will approach w plotted in Fig. 4(b) and will descend like Tukey's biweight. However for a more normal-looking sample W will flatten out to some moderate non-zero value and will more closely resemble the weighting originally proposed by Huber.

In choosing robust estimators there is room for empiricism but I think that some of its inspiration should be applied to the choice, study, and consequences of appropriate parsimonious models. The structure of the resulting Bayesian analysis should in each case be carefully analysed, for the great strength of such a model-based approach is that the exact consequences of whatever goes into the model must come out. These consequences will either agree with "common sense" or they will not. If they do not then we know either that what went in was inappropriate in a way we had failed to forsee, or else, as happens quite frequently, that our common sense was too shortsighted. In either case we learn something.

Acknowledgements

This research was sponsored by the United States Army under Contract No. DAAG29–75–C–0024. My thanks are due to Steve Bailey, Gina Chen, Conrad Fung, Kevin Little and Ruey-Shiong Tsay for their help in preparing this manuscript, and to Tom Leonard, a referee, and many other friends for helpful suggestions and comments.

References

Abraham, B. and Box, G. E. P. (1978). Linear models and spurious observations. *Appl. Statist.*, **27**, 131–138.

Andrews, D. F. (1971a). A note on the selection of data transformations. *Biometrika*, **58**, 249–254.

—— (1971b). Significance tests based on residuals. *Biometrika*, **58**, 139–148.

Anscombe, F. J. (1961). Examination of residuals. In *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, **1**, 1–36. Berkeley and Los Angeles : University of California Press.

Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics*, **5**, 141–160.

Atkinson, A. C. (1973). Testing transformations to normality. *J. R. Statist. Soc.* B, **35**, 473–479.

Bailey, S. P. and Box, G. E. P. (1980a). Modeling the nature and frequency of outliers. Technical Report 2085 Math. Research Center, University of Wisconsin at Madison.

—— (1980b). The duality of diagnostic checking and robustification in model building: Some considerations and examples. Technical Report 2086, Math. Research Center, University of Wisconsin at Madison.

Barnard, G. (1978). Personal communication.

Box, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Ass.*, **71**, 791–799.

—— (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pp. 201–236. New York: Academic Press.

Box, G. E. P. and Behnken, D. W. (1960). Some new three-level designs for the study of quantitative variables. *Technometrics*, **2**, 455–475.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc.* B, **26**, 211–243.

Box, G. E. P. and Draper, N. R. (1975). Robust designs. *Biometrika*, **62**, 347–352.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Box, G. E. P. and Newbold, P. (1971). Some comments on a paper by Coen, Gomme and Kendall. *J. R. Statist. Soc.* A, **134**, 229–240.

Box, G. E. P. and Tiao, G. C. (1962). A further look at robustness via Bayes' theorem. *Biometrika*, **49**, 419–432.

—— (1964). A Bayesian approach to the importance of assumptions applied to the comparison of variances. *Biometrika*, **51**, 153–167.

—— (1968a). Bayesian analysis of means for the random effect model. *J. Amer. Statist. Ass.*, **63**, 174–181.

—— (1968b). A Bayesian approach to some outlier problems. *Biometrika*, **55**, 119–129.

—— (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass.: Addison-Wesley.

Box, G. E. P. and Youle, P. V. (1955). The exploration and exploitation of response surfaces: an example of the link between the fitted surface and the basic mechanism of the system. *Biometrics*, **11**, 287–323.

Chen, G. G. and Box, G. E. P. (1979). Further study of robustification via a Bayesian approach. Technical Report 1998, Math. Research Center, University of Wisconsin at Madison.

Coen, P. J., Gomme, E. E. and Kendall, M. G. (1969). Lagged relationships in economic forecasting. *J. R. Statist. Soc.* A, **132**, 133–152.

Cox, D. R. (1977). The role of significance tests. *Scand. J. Statist.*, **4**, 49–70.

Dempster, A. P. (1971). Model searching and estimation in the logic of inference (with Discussion). In *Foundations of Statistical Inference*, pp. 56–81. Toronto: Holt, Rinehart and Winston.

—— (1975). A subjectivist look at robustness. *I.S.I. Bulletin*, **46**, 349–374.

Dixon, W. J. (1953). Processing data for outliers. *Biometrics*, **9**, 74–89.

Draper, N. R. and Van Nostrand, R. C. (1977). Ridge regression: is it worthwhile? Technical Report 501, Department of Statistics, University of Wisconsin at Madison.

DURBIN, J. and WATSON, G. S. (1950). Testing for serial correlation in least square regression I. *Biometrika*, **37**, 409–428.

FELLER, W. (1968). *An Introduction to Probability Theory and its Applications*, Vol. 1. New York: Wiley.

FISHER, R. A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

GEISSER, S. (1971). The inferential use of predictive distributions (with Discussion). In *Foundations of Statistical Inference*, pp. 458–469. Toronto: Holt, Rinehart and Winston.

—— (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Ass.*, **70**, 320–328.

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Ass.*, **74**, 153–160.

GOOD, I. J. (1956). The surprise index for the multivariate normal distribution. *Ann. Math. Statist.*, **27**, 1130–1135.

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

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. R. Statist. Soc.* B, **29**, 83–100.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression : applications to non-orthogonal problems. *Technometrics*, **12**, 69–82.

HOERL, A. E., KENNARD, R. W. and BALDWIN, K. F. (1975). Ridge regression : some simulations. *Comm. in Statist.*, **4**, 105–124.

HUBER, P. J. (1977). Robust statistical procedures. Society for Industrial and Applied Mathematics, **27**, Philadelphia.

JEFFREYS, H. (1932). An alternative to the rejection of observations. *Proc. Roy. Soc.* A, **CXXXVII**, 78–87.

KADANE, J. B., DICKEY, J. M., WINKLER, R. L., SMITH, W. S. and PETERS, S. C. (1979). Interactive elicitation of opinion for a normal linear model. Technical Report 150, Department of Statistics, Carnegie Mellon University.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes' estimates for the linear model (with Discussion). *J. R. Statist. Soc.* B, **34**, 1–41.

PALLESEN, L. C. (1977). Studies in the analysis of serially dependent data. Ph.D. thesis, Department of Statistics, University of Wisconsin at Madison.

POPPER, K. R. (1959). *The Logic of Scientific Discovery.* New York: Harper and Row.

ROBERTS, H. V. (1965). Probabilistic prediction. *J. Amer. Statist. Ass.*, **60**, 50–62.

SNEDECOR, G. W. and COCHRAN, W. G. (1967). *Statistical Methods.* Ames : Iowa State University Press.

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

STIGLER, S. M. (1973). Simon Newcomb, Percy Daniel and the history of robust estimation 1885–1920. *J. Amer. Statist. Ass.*, **68**, 872–879.

THEIL, H. (1963). On the use of incomplete prior information in regression analysis. *J. Amer. Statist. Ass.*, **58**, 401–414.

TIAO, G. C. and ALI, M. M. (1971). Analysis of correlated random effects: linear model with two random components, *Biometrika*, **58**, 37–51.

TIAO, G. C. and BOX, G. E. P. (1980). An introduction to applied multiple time series analysis. Technical Report 582, Department of Statistics, University of Wisconsin at Madison.

TUKEY, J. W. (1949). One degree of freedom for non-additivity, *Biometrics*, **5**, 232–242.

——(1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pp. 448–485. Stanford, Stanford University Press.

ZELLNER, A. and TIAO, G. C. (1964). Bayesian analysis of the regression model with autocorrelated errors. *J. Amer. Statist. Ass.*, **59**, 763–778.

DISCUSSION OF PROFESSOR BOX'S PAPER

Professor G. A. BARNARD (University of Waterloo): I very much welcome this important paper as a further indication that views on the foundations of statistical inference are converging and the worrying prospect that seemed to be opening itself up four or five years ago of hopelessly irreconcilable attitudes being adopted amongst experts in the field can now be thought less probable.

Although I shall follow tradition and emphasize my difference from the author, I must first of all say how strongly I agree with what he has said concerning "robust" estimation. The conditional approach which he adopts—and which follows naturally from a Bayesian approach—is surely the only sound one. We ought not to look for robust *procedures*, but rather for robust *samples.* For a robust sample varying assumptions about the shape of the distribution will have little effect on the inference to be drawn. For a non-robust sample these assumptions have an effect; and in that case the *existence of this effect must form part of the inference.* We must tell our clients that unless they can find out something concerning the shape of the distribution, the message of their sample is ambiguous. Alternatively, by taking further observations, they may convert what was a non-robust sample into a robust sample and so avoid the necessity for investigations about the population shape. Insofar as the Bayesian approach bases itself essentially on using the likelihood function, of course, we would expect a likelihood approach and a Bayesian approach to agree in this respect. It has sometimes been said that the trouble with likelihood is that we need to know the form of the distribution with more precision than is commonly available. Such a view overlooks the fact that if we have doubts about the form of the distribution we can always make a range of suppositions concerning this, and derive a corresponding range of likelihood functions. It will very often be the case that the range of likelihood functions so obtained will be quite narrow, showing that our sample is a robust one and our ignorance of the precise shape of the distribution will do no harm. In the contrary case, where the shape of the distribution does matter it is actually misleading to fail to point this out. It will be no comfort to our current client if we should say that cases such as his rarely occur—he is concerned with what has happened in his particular case, and if the inference drawn proves to be wholly incorrect, he will be entitled to say that he ought to have been warned.

When the conditions for the simple-minded application of the method of maximum likelihood apply—that is when the log likelihood function is nearly parabolic—the sensitivity or otherwise of a particular sample to changes in distribution form can be indicated by the change in the position of the maximum likelihood estimate, primarily, and secondarily by changes in the second derivative of the score function, the information. If the maximum likelihood estimate does not change very much then for "point estimate" purposes our sample is robust, and very often a modest allowance for variations in the information will give the protection that is needed.

Now to return to the main topic. Professor Box had adopted a position very close to that of Emile Borel. Borel's view of the nature of probability was very close to that of de Finetti—the strongest advocate of the personalistic view—but Borel was criticized, I think wrongly, by de Finetti for adopting "Cournot's principle" according to which we behave as if events of small probability do not occur. Another version of Cournot's principle is the so-called empirical law of large numbers, according to which probabilities must agree with long run frequencies.

The fact is that if an individual went through life continually being amazed to find that the events for which his subjective probability had been extremely small regularly turned out to be those that occurred, the individual in question ought to ask himself whether his modes of assessment of probabilities were reliable. The fallacy in the personalistic argument appears to me to arise from the fact that a personalistic view of probability makes no allowance for the limitations of human imagination. In Box's sequence of model criticism, revised model and so on, the pure personalistic view has to treat the propositions "model M_i is true" and "model M_i is false" as propositions on the same footing. Yet this is obviously absurd. If we take as given that model M_i is true we have a very clear picture of the possible events to be observed and the parameter values to be associated with them. But to say simply that model M_i is false give us absolutely no idea of what sort of model might be true—we might or might not be required to introduce further parameters, we might or might not think of a great multiplicity of shapes of distribution that might arise and so on. Observations of incompatible events could well lead us to feel that M_i could not be true, and yet the intellectual effort involved in constructing a model that we think would apply to the situation in hand could form as large a part of the whole intellectual exercise as any other. To take one example, Michelson and Morley gained well-deserved fame for demonstrating the falsity of the simple minded notion of an ether. But their credit is small compared with that of Einstein whose tremendous imagination was needed to set up an alternative model which accounts for their data. To put the point formally, the pure theory of personal probability treats H and not-H as propositions of the same kind. But the problem of calculating the probability of E given H may be quite trivial, whereas the problem of calculating the probability of E given not-E could be quite insoluble.

It is perhaps worthwhile to indicate how the de Finetti theory could be modified to take account of the points that Box is making. To simplify a little, we can regard the model M_i as being our (tentative) view of what constitutes the totality of possibilities before we perform the experiment. If we then observe a result y for which the probability density on the predictive distribution is low, we need to reconsider M_i and ask whether it really does encompass all the possibilities. When y presents a feature which, in the words of Borel., is "en quelque sort remarquable" the fact that it is remarkable points us in a certain direction to try to use our imagination to conjure up possibilities which previously we had not recognized. M_{i+1} will then embody these possibilities and the iteration process begins again. Thus as I see it a strict adherent to de Finetti's views need only admit the obvious fact that we can never think of everything to accept the force of Box's argument.

The modified personalistic view of probability put forward by Box and Borel is, I think, inadequate for statisticians basically because statisticians work for clients. They are therefore not concerned with personal probabilities but with what might be called "agreed probabilities". And whereas personal probabilities may be said always to exist in principle, *agreed* probabilities need not exist. We may have a parameter involved in the statistical problem for which an agreed prior probability distribution does not exist and in such a case I think it will often be wise for a statistician to regard the parameter as simply unknown, capable of taking on a number of values, and for the statistician to see his job as pointing to the probabilistic consequences, in the light of the data, of assuming that the parameter takes various values in its range. This is *not* equivalent to a prior concentrated on a single value. This is my major disagreement with Professor Box.

As an extension of the general theory of likelihood, I have in recent years been developing a theory of what I call "pivotal inference" concerning which I hope to present a paper to the Society shortly. In this theory we start from quantities, called pivotals, which do have agreed probability distributions. Such pivotals will usually be functions both of observations and of parameters. But some of them may also be, or be transformed into, quantities which are functions of the observations only—these then become ancillaries, upon which we should condition when we have the observations—or they may be functions of the parameters only, in which case they represent prior distributions for the functions of the parameters involved. By applying the standard rules of probability, concerned with marginalization and with conditioning in the light of the values of the observations when these become known, we can infer distributions for functions of the basic pivotals which will sometimes give posterior distributions, if the necessary information for a full Bayesian inference is available, but more often will enable us only to make statements to the effect that an assumption that a parameter has a value in some specified range will entail that an event has occurred whose probability is small. We shall be disinclined to swallow such improbabilities.

The pivotal approach does not require us to take a general position on the question, whether or not unknown parameters are required to be endowed with probability distributions. In each specific case, we can exercise judgement as to whether to ascribe such distributions to some or all of the parameters and we can explore the final effects—often small—of adding or removing such assumptions. In any case the basic inferential procedure is always the same—to condition on known, or approximately known quantities, whose distribution is known, to arrive, if we can, at invertible pivotals for the quantities of interest, which enable us to say that accepting the notion that a parameter value lies in a certain range entails accepting that an event of a specifically low probability has occurred. This is all we can derive from a Bayesian approach, unless we have—as is very rarely the case—a well-specified loss function which we aim to minimize.

The neo-Bayesian movement has purged statistical inference of a great many stupidities which arise from neglect of proper conditioning. I hope this paper will come to represent a major step towards a situation where such absurdities no longer plague us, and we are much closer to general agreement on foundations than we have been in the recent past.

It gives me great pleasure to propose the vote of thanks to Professor Box.

Professor A. P. Dawid (The City University): This Society has traditionally recognised the importance of both *Estimation* and *Criticism* as principal features of a vote of thanks. It is my very agreeable task to bring these twin criteria to bear on tonight's paper.

First, then, to register my esteem. Professor Box has given us a compelling account of his search for the guiding principles of scientific learning, and illuminated it with practical examples which are both interesting and important. We are all the richer for having him share his experience and insights with us. Whilst, as he admits, his views are not new, his paper is a valuable and forceful reminder of an important lesson: that there are at least two distinct fundamental functions of statistical reasoning, Criticism and Estimation, and "never the twain shall meet". We should constantly bear the distinction in mind when we construct, select, or teach students about statistical techniques. A hypothesis test, for example, can be used for model testing (Criticism) or model simplification (Estimation). The purpose, interpretation and relevance are quite different for the two different cases.

Whether or not one agrees with Professor Box that Bayesian conditioning is the way to go about Estimation, one must agree that something else is needed for Criticism, where no fully specified alternative model is given. Professor Box makes a valuable contribution in proposing the predictive distribution as the basis of Criticism, but the question of *how* to use it is not clarified, and we are left with familiar *ad hoc* devices such as tail-area tests and (in a framework that makes some concession to Estimation) score statistics. This is not to belittle the usefulness of *ad hoc* solutions in the absence of underlying principles, merely to point to a gaping hole in all our current theoretical formulations.

At a very general level, Box's dualistic view of statistical reasoning is crudely analogous to Thomas Kuhn's account of the progress of scientific theories. In periods of "normal science", a particular paradigm, or model, is taken for granted, and scientists work at refining it and apply it ("Estimation"). But the predictions of the current paradigm are always open to confrontation with the real world ("Criticism"), and if and when discrepancies become unacceptable a "scientific revolution" topples the old paradigm and puts a new one in its place. However, the old paradigm can be discredited even when no workable alternative is in sight.

Kuhn's ideas have made an enormous impact on the philosophy of Science, and it is valuable to be reminded that similar reasoning can, and should, be applied to statistical investigations, however mundane.

Let me now switch to Criticism mode and consider some details of tonight's paper. Box distinguishes between checks on parametric and residual features of the model. I believe that the former, which is basically a check on the prior distribution, as in (3.9), will be the most useful practical contribution of this paper, but it may not be much appreciated by non-Bayesians who do not accept, with Box, that a model cannot exist in isolation from a prior distribution. Moreover, I do not find the attempted Bayesian justification of significance tests (Section 2.1) any advance on the non-Bayesian interpretation. As for residual checks, these have long been available in a non-Bayesian setting. Indeed, with the exception of Section 4.3 (to which I shall return), Box's analysis has merely reproduced classical tests. These examples all involve models of the form

$$p(\hat{\boldsymbol{\Theta}}, S, \mathbf{u} \mid \boldsymbol{\Theta}, \sigma, \beta) = S^{-2} \,.\, p((\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})/\sigma, S/\sigma \mid \mathbf{u}; \beta) \,.\, p(\mathbf{u} \mid \beta).$$

Box takes his prior for $(\boldsymbol{\Theta}, \sigma)$ given β to be "locally uniform":

$$p(\boldsymbol{\Theta}, \sigma \mid \beta) \doteqdot c(\beta) \,.\, \sigma^{-1}.$$

(We must of course allow the general level, determined by $c(\beta)$, to depend on β). Integrating out $(\boldsymbol{\Theta}, \sigma)$ gives

$$p(\hat{\boldsymbol{\Theta}}, S, \mathbf{u} \mid \beta) \doteqdot S^{-1} \,.\, c(\beta)\, p(\mathbf{u} \mid \beta),$$

and so the score-statistic $g_\beta(\mathbf{y})$, from (4.2), is just $\{c'(\beta_0)/c(\beta_0)\} + h_\beta(\mathbf{u})$, a simple transform of $h_\beta(\mathbf{u})$, the score statistic derived from the purely classical approach of considering the standardized residuals only, and having a known null distribution. A generalization of this remark holds for arbitrary group-invariant models where the underlying pivot has a distribution governed by the discrepancy parameter β.

At least the above theory leads us to hope that the Box approach to checking residual features, when applied (as it should be) with a genuinely informative prior distribution, may lead to a test statistic approximating $h_\beta(\mathbf{u})$, with null distribution not over-dependent on the prior. The situation is not so clear for Section 4.3, however. If we again take a "locally uniform" prior of the form $p(\boldsymbol{\Theta}, \sigma \mid \lambda) \doteqdot c(\lambda)\, \sigma^{-1}$, we find

$$p(\mathbf{y} \mid \lambda) \dot{\propto} c(\lambda)\, \dot{y}^{\cdot n(\lambda-1)} Q_\lambda^{-\frac{1}{2}\nu}.$$

(The slightly different expression (4.6) results from an interesting attempt by Box and Cox to specify $c(\lambda)$ reasonably; for present purposes, this can be avoided.) Then we find

$$g_\lambda(\mathbf{y}) \doteqdot \{c'(1)/c(1)\} + (k+1) \log \dot{y} + s^{-1} \sum r_i z_i.$$

The constant term is irrelevant to the test statistic, leaving the extra term $(k+1)\log \dot{y}$ as a correction to (4.7).

Even if we stick to (4.7), what are we to do with it? Box suggests an informal graphical analysis, but we have to know what features of the diagram to look for. It seems implicit that a substantial sample correlation between $z - \hat{z}$ and $y - \hat{y}$ is to be regarded as evidence against $\lambda = 1$. In fact, (4.7) is ν × the slope of the regression of $z - \hat{z}$ on $y - \hat{y}$, and is not a function of correlation. And is zero the appropriate "null value" for (4.7)? What departures are "significant"? These problems do not disappear on "actual reference of $p(g_\lambda(\mathbf{y}_d))$ to its sampling distribution", since the null sampling distribution depends on the unknown $(\boldsymbol{\Theta}, \sigma)$. We could marginalize it with respect to the prior distribution (for $(\boldsymbol{\Theta}, \sigma)$ given $\lambda = 1$), but the answer would be critically dependent on the (proper, locally uniform) prior chosen, and cannot be approximated by using an improper prior. Nevertheless, this marginalization would be the right course for the Bayesian. The moral is that the need for transformation can only be assessed in the light of prior knowledge.

Box's representation (4.11) is in error, a factor $(s/\dot{y})$ being omitted ($B = 1/\dot{y}$ in (4.8)). For the simplest case $k = 0$, we get $g_\lambda(\mathbf{y}) \doteqdot -\frac{1}{2}(s/\dot{y})\, T_{30}$, and the variance of this is approximately proportional to σ^2/μ^2, making assessment of significance impossible without prior information about (μ, σ). If we adjust by omitting the $-\frac{1}{2}(s/\dot{y})$, we get T_{30}, a function of the normalized residuals with a known null distribution; but after all these *ad-hockeries*, have we gained anything from a pseudo-Bayesian approach?

Finally, let me remove my critic's hat and welcome Box's whole-hearted Bayesian account of robust estimation (Section 5). This captures all the common-sense features one would want: for example, there is no point in guarding against model departures of a size which the data themselves suggest is implausible. Even non-Bayesians could learn useful lessons from this analysis.

Altogether I regard tonight's paper as of the greatest importance, and second the vote of thanks to its author, George Box, most warmly.

The vote of thanks was passed by acclamation.

Dr A. O'Hagan (University of Warwick): Like all Professor Box's work, this paper is suffused with common-sense and insight. I applaud all that he has done in identifying and illuminating very clearly a difficult problem. I have two complaints, one small and one I think rather larger. Both concern his curious belief that Bayesian methods are appropriate to the selection of sensible parameter values but not to the selection of models. The obvious Bayesian solution to the latter is to compute the posterior probability of model M_i via

$$P(M_i \mid y_d) = \frac{p(y_d \mid M_i)\,P(M_i)}{\sum_j p(y_d \mid M_j)\,P(M_j)} \qquad (*)$$

Of course this is conditional on one of the stated M_j being the true model, but for the moment let us suppose that this is so. Professor Box argues that if $p(y_d \mid M_i)$ is sufficiently small, that is if the model M_i completely fails to fit the data y_d, then M_i is discredited. But the appearance of the prior probabilities $P(M_j)$ in (*) shows that these are also relevant. We cannot say that $P(M_i \mid y_d)$ is small unless we can find a model M_j which fits the data well *and* is credible *a priori*. Otherwise the denominator in (*) will also be small and $P(M_i \mid y_d)$ may even be near to one. Professor Box on his first page tells us that "A theory about scientific model building ought to explain what good statisticians and scientists actually do". Practising statisticians when criticizing a model look at things like residual plots and sample autocorrelations, which Professor Box has shown us carry implicit ideas of alternative models. Every way the statistician chooses to look at the data corresponds to a feeling, which perhaps never becomes conscious and explicit, for a specific kind of alternative. Furthermore, his implied alternative is *a priori credible*—if he couldn't conceive of autocorrelation then he wouldn't bother to look for it.

When good statisticians *compare* models they should and effectively do employ (*). The new idea in Professor Box's approach is that we can criticize a model without having any alternatives in mind. This is very tempting because of course we can never think of more than a few of the countless possible models. Certainly if $p(y_d \mid M_i)$ is small it sounds a warning to me and I start to look for sensible alternatives, but I do not feel that I can gauge that smallness or reject M_i until I've found a better alternative. Professor Box judges the size of $p(y_d \mid M_i)$ by means of a "tail-area", but this is wrong in this context for the same reasons that it is wrong in other stages of the analysis—the role played by $p(y_d \mid M_i)$ in (*) is that of the likelihood. Studying $p(y \mid M)$ as y varies is useless because we are interested in it only as M varies. Of course, in other contexts sampling-theory methods based on tail-areas have often been found to approximate closely to Bayesian procedures. But tail-areas are also known to be unreliable indicators, and I feel sure that uncritical use of them to compare models will lead to familiar pitfalls. Just as in parameter estimation, there is no substitute for a proper Bayesian analysis. We must consider all the alternatives which occur to us, and if our original model M_i is not bettered then we cannot in any sense reject it yet, however "small" $p(y_d \mid M_i)$ may appear to be. Nor should we forget that if we are satisfied with our current model we still cannot rule out the possibility of something better being proposed.

In conclusion, I am sure that Professor Box is a far better practical statistician than I, and if he tells me that he can persuade a thing for taking stones out of horses' hooves to turn screws I believe him. *I* will stick to my screwdriver.

Dr A. C. Atkinson (Department of Mathematics, Imperial College, London): I can still recall the intellectual excitement of first reading Chapter 11 of "Big Davies" (Davies, 1956). More than any other, this was the experience which inspired me to become a statistician. It is a pleasure to be able to thank Professor Box in person for the initial and continuing stimulus of his books and papers.

My comments on tonight's paper concern the diagnostic checking of regression models. If outlying values of the carriers x are suspected, residual plots should be augmented by functions which respond to the influence of the individual observations. Fig. D1 shows a half normal plot, for the Box and Behnken data, of a modification of a statistic due to Cook (1977) which I call T_i. For this designed experiment the plot is similar to a residual plot, with identity for a D optimum design. There is clearly something strange about one of the observations.

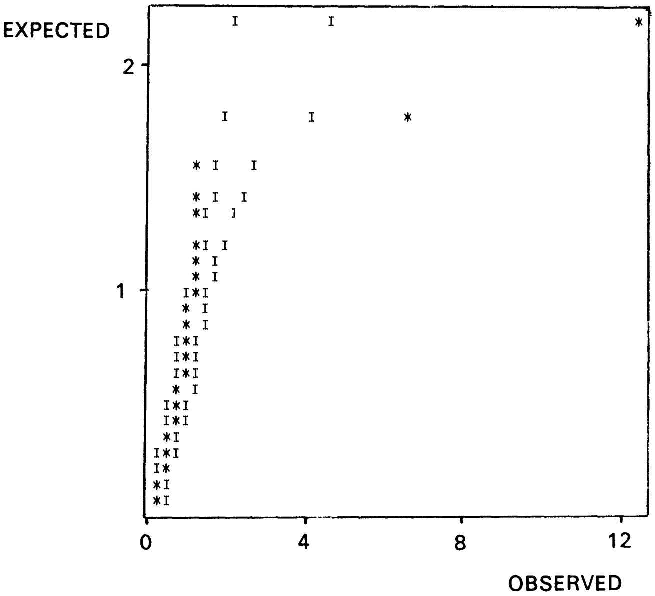

FIG. D1. Half normal plot of modified Cook statistic T_i for 27 observations from Box and Behnken. * Observed. I, Envelope from 19 simulations.

In several standard examples (Atkinson, 1980, 1981) transformation of the data provides an alternative to rejection of outliers. In this case the asymptotically standard normal score statistic for transformations, T_D, has the value $-0{\cdot}360$, so no transformation is needed. Incidentally, I do not understand the remark after (4.7) about the distribution of this statistic. Perhaps Professor Box was thinking of the apparently anomalous results of Schlesselman (1973) which are due to a programming error (Fuchs, 1979). Other transformations of the Box and Behnken data, such as considering $(100-y)^{\lambda}$ and replacement of

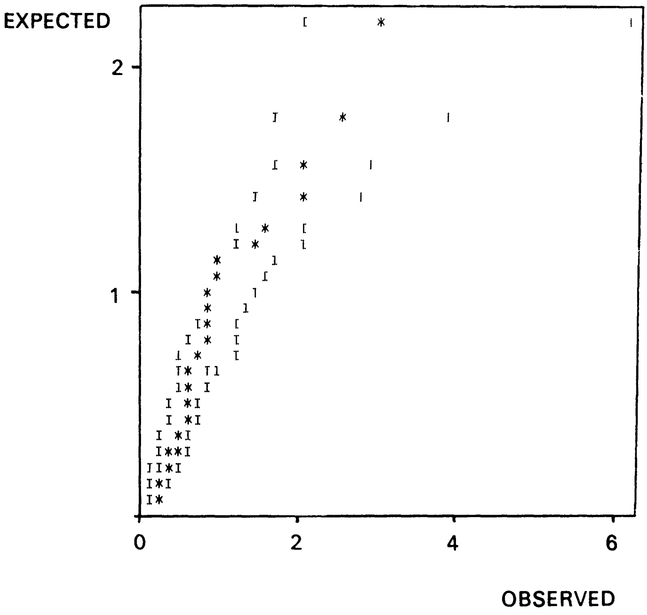

FIG. D2. Plot of T_i with y_{10} deleted. * Observed. I, Envelope from 19 simulations.

$y_{10} = 89{\cdot}4$ by 84·9, also fail to remove the outlier. If the observation is rejected the plot of T_i for the remaining 26 observations, Fig. D2, shows no further features of interest.

Table D1 gives the least squares estimates of the coefficients and the associated standard errors for the full second-order model fitted to the 26 observations. If only two factor interactions are considered, with the exception of β_{14}, the estimated coefficients agree to 3 significant figures with those in the last column of Table 1. Complete agreement can be achieved by interchanging the two rows for β_{14} in Table 1, which also brings the standard errors more nearly into line. I wonder whether the observed behaviour for β_{14} may not be due to a transcription error for at least some of the columns of these two rows.

In conclusion I would like to echo Professor Box's remarks about the importance of computer graphics in model checking. It is ironical that at a time when we are technically able to produce interesting plots with great ease, economic pressures are such as to discourage the publication of figures and graphs.

Table D1

Parameter estimates and associated standard deviations for Box and Behnken's data with y_{10} deleted

Parameter	*Least squares estimate*	*Standard deviation*
β_1	2·568	0·124
β_2	−1·958	0·111
β_3	1·133	0·111
β_4	−3·041	0·124
β_{11}	−2·051	0·175
β_{22}	−4·012	0·168
β_{33}	−1·924	0·168
β_{44}	−3·241	0·175
β_{12}	−1·675	0·192
β_{13}	−3·825	0·192
β_{14}	−0·953	0·252
β_{23}	−1·675	0·192
β_{24}	−2·625	0·192
β_{34}	−4·250	0·192

Professor D. R. Cox (Department of Mathematics, Imperial College, London): I admire the paper for its combination of important general discussion with interesting examples.

Equation (2.12) derives a tail area by integration over all sample points with a density equal to or smaller than that of the observed point. While, as stated just below the equation, relatively minor changes such as from s to s^2 (or $s^{2/3}$ or $\log s$) will make relatively minor changes to the answer, in fact adjusting the relation between two tails, it is not clear why they should make any difference at all: or to put the point differently, how do we decide which is appropriate in a given instance? There is, of course, also the theoretical possibility that radically non-linear transformations, or the use of quantities with very spiky distributions, would lead to strange regions of integration. Would it not be better to define the test quantity so as to order the sample points in order of increasing discrepancy in some respect and to integrate over large values of the test quantity? I appreciate that this involves an implied qualitative specification of alternatives which it might be desirable to avoid, but is some such specification really avoidable?

The section on robust estimation is very appealing. The idea of modelling suspected complications and exploring the consequences by general theory is excellent for broad guidance, but if undue complication is to be avoided, this idea would presumably have to be severely restrained in applications. We all recall what Lord Kelvin said.

Professor M. Stone (University College London): The emphasis, in this authoritative and readable paper, on the concept of compatibility between data and their marginal distribution is very welcome. I hope that Professor Box's message will be studied by those who have up to now ignored the value of the compatibility concept in constructive criticism of a Bayesian model and elucidation of any paradoxical aspects it may have when improper priors are used.

In Section 2.2, Professor Box rightly advises us that impropriety of the marginal data distribution does not dispense with the need for criticism but he does not give much guidance to the potential critic. One analytical method available to deal with this troublesome corner is roughly as follows. Find, if you can, a sequence of proper priors such that, for every $\varepsilon > 0$, the (proper) marginal probability that the (proper) posterior differs by "more than ε" from the improper one of the model tends to unity down the sequence. You can then say that the improper posteriors are justifiable in marginal probability but you cannot conclude that the improper posterior for fixed data (i.e. what you wish to use) is also justified. Whether or not it is acceptable may be determined by (a) the way in which the posterior is used (b) whether the (fixed) data are in some relevant way asymptotically incompatible with the proper marginals. In some cases, the test of asymptotic incompatibility is dramatically simple: any fixed data are unacceptable!

In a paper so rich in suggestions, I should not be surprised that the author has touched on another problem that has worried me for some time. That is the question of the possible link between cross-validatory procedures and the Bayesian models to whose output they bear a striking resemblance. The similarity is most striking for Bayesian models that incorporate flexible priors of the multistage variety and that are therefore rendered data-adaptive. By construction, cross-validatory procedures are data-adaptive within the ambit prescribed for choice. However, I have been unable to obtain any technical insight into problems of real significance. It is tempting to try to relate the cross-validatory weights in Modelmix to posterior probabilities of the component "models". A simple example with discontinuous weights, (1, 0) or (0, 1), is that of estimating the true mean μ from a random sample of size n when it is known that μ is either μ_1 or μ_2 with $\mu_1 < \mu_2$. If the cross-validatory prescription is to say $\hat{\mu} = \mu_1$ if $\bar{y} < \alpha$ and $\hat{\mu} = \mu_2$ if $\bar{y} \geqslant \alpha$ and if the "loss function" is quadratic between $\hat{\mu}$ and an observation, a cross-validatory choice is to take $\alpha^{\dagger} = \infty$ or $-\infty$ according to whether $\bar{y} < \frac{1}{2}(\mu_1 + \mu_2)$ or $\bar{y} \geqslant \frac{1}{2}(\mu_1 + \mu_2)$, that is $\hat{\mu} = \mu_1$ if and only if $\bar{y} < \frac{1}{2}(\mu_1 + \mu_2)$. A model for which this would be the posterior modal estimate of μ would be that of normality and a prior with $P(\mu = \mu_1 \mid \sigma^2) \equiv P(\mu = \mu_2 \mid \sigma^2)$. Is the normality implied by the use of $\bar{y}$ in the prescription and are the equiprior probabilities mildly expressive of maximal data-adaptivity? In some sense, I believe they are and that it ought to be possible to develop better illustrations of the relationship

Professor A. F. M. SMITH (University of Nottingham): Professor Box is a wise, practical statistician and tonight's distillation of his wisdom merits careful consideration by everyone genuinely concerned with statistical methodology. In particular, it should be required reading for all those statistical Yahoos who stridently proclaim the Bayesian approach to be misguided or irrelevant without bothering to study in detail what it has to offer. Perhaps there *are* ad hoc "non-Bayesian" analogues of everything Box mentions (and no doubt many of these *ad hockeries* "came first"), but Box's *unified perspective* is surely more satisfying and suggestive as a framework for overall understanding and further advance—even if, as various other, basically sympathetic, discussants have indicated, there are still many issues to be clarified *within* the framework.

So far as the details of the paper are concerned, I shall confine myself to a brief comment on the Bayesian approach to Robust Estimation (Section 5). The author's discussion concentrates on robustness against bad values, where his approach is shown to have close links with non-Bayesian procedures. In fact, an even more direct association with ideas like "M-estimates" and "influence functions" can be established using the fact that if $y = \theta + \varepsilon$, $p_\varepsilon(\cdot)$ arbitrary, with $\theta \sim N(m, c)$, then, using Masreliez (1975),

$$E(\theta \mid y) = m + c\left[-\frac{\partial}{\partial y}\log p(y)\right]$$

$$\approx m + c\left[-\frac{\partial}{\partial y}\log p_\varepsilon(y - m)\right],$$

where $p(y) = \int p_\varepsilon(y - \theta)\, p(\theta)\, d\theta$. Using the approximate form, we see that the way in which the "innovation" is used to update the prior mean depends on the choice of $p_\varepsilon(\cdot)$ through the score (or "influence") function. Interesting families of error distributions—such as the t-family, exponential power family, normal centre/Laplace tails family—can be investigated for use as (model-based) robust estimation procedures. The approach can be extended to provide robust Bayesian sequential learning within the Kalman filter framework, with both location and covariance structure unknown. An account of these ideas is being written as a Ph.D. thesis by Michael West at Nottingham University. Finally, it is worth remarking that these and other robustifying and checking devices tend to lead to some tricky *numerical* problems if integration over several (perhaps highly correlated) parameters is required in order to isolate a marginal feature. An account of some recent progress in this area will be given in a paper currently being written by myself and John Naylor, based on material which will form part of a Ph.D. thesis at Nottingham.

Professor Jose M. Bernardo (Department of Biostatistics, University of Valencia, Spain): Although I certainly agree with Professor Box in the cyclical nature of scientific experimentation through model criticism and parameter estimation, I do not understand his claim that "sampling theory is needed for ultimate criticism of a model". Indeed, when for this purpose he computes (2.3), that is $\alpha = \Pr\{p(y\,|\,A) < p(y_d\,|\,A)\}$, he is making use of a *predictive* distribution, which is not defined unless one has a prior. The (interesting) mathematical accident by which the use of a reference (non-informative) prior often leads to the same numerical manipulations as classical tests should not obscure the fact that one is arguing within a Bayesian framework. Incidentally, reference should be made to the use of the *surprise index* (2.3) by Barnard (1967, p. 28) and by Aitchison and Dunsmore (1975, p. 224).

At a more concrete level, I find unsatisfactory the use of Fisher's score function (4.2) to measure discrepancies from β_0 from a current model $p(y\,|\,\beta)$. Indeed, as its Bayesian interpretation (4.22) openly shows, (i) $g_\beta(y)$ is *not* invariant under scale changes in β, as one might wish, and (ii) it assumes implicitly a uniform prior for β, that would only be appropriate if β were a location parameter. An alternative definition could be

$$g'_\beta(y) = \frac{E(\beta\,|\,y)-\beta_0}{D(\beta\,|\,y)}.$$

Here, $E(\beta\,|\,y)$ and $D(\beta\,|\,y)$ are the mean and standard deviation of the *reference posterior* distribution (Bernardo, 1979) $\pi(\beta\,|\,y)$ defined as $\pi(\beta\,|\,y) \propto p(y\,|\,\beta)\,\pi(\beta)$ where $\pi(\beta)$ is a reference (non-informative) prior. If β is a continuous one-dimensional variable and $p(y\,|\,\beta)$ is well behaved, then the appropriate choice for $\pi(\beta)$ is Jeffreys' prior

$$\pi(\beta) \propto \left\{-\int p(y\,|\,\beta)\frac{\partial^2}{\partial\beta^2}\log p(y\,|\,\beta)\,dy\right\}^{\frac{1}{2}}.$$

If only distance from β to β_0 matters, I would probably use $\{g'_\beta(y)\}^2$ rather than $g'_\beta(y)$ as a measure of discrepancy.

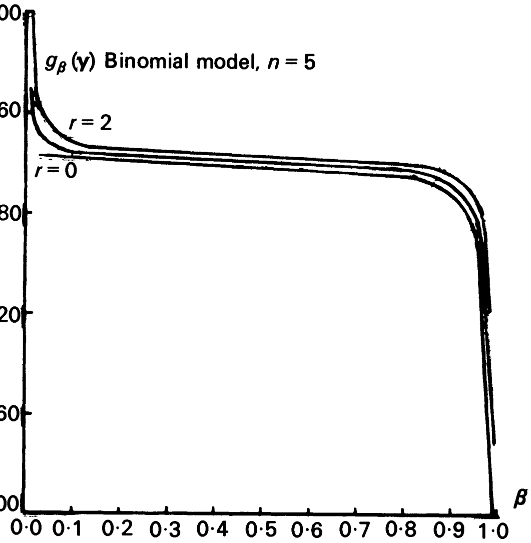

Fig. D3. Box's discrepancy.

In Figs D3 and D4 the behaviour of Box's $g_\beta(y)$ and the proposed measure $\{g'_\beta(y)\}^2$ is shown for the simple situation in which the value β_0 as the proportion of elements in a population which possess a given feature, is to be tested using a binomial model $p(y\,|\,\beta) = p(r\,|\,\beta, n) \propto \beta^r(1-\beta)^{n-r}$. Here, one has

$$g_\beta(y) = \frac{r}{\beta} - \frac{n-r}{1-\beta}$$

$$\pi(\beta) \propto \beta^{-\frac{1}{2}}(1-\beta)^{-\frac{1}{2}}, \quad \pi(\beta\,|\,r) \propto \beta^{r-\frac{1}{2}}(1-\beta)^{n-r-\frac{1}{2}},$$

$$E(\beta\,|\,r) = (r+1/2)/(n+1), \quad D^2(\beta\,|\,r) = \{(r+1/2)(n-r+1/2)\}/\{(n+1)^2(n+2)\},$$

$$\{g'_\beta(\mathbf{y})\}^2 = \frac{(n+1)(n+2)}{(r+1/2)(n-r+1/2)}\left(\frac{r+1/2}{n+1}-\beta\right)^2.$$

I believe most people would prefer the behaviour of $g'_\beta(\mathbf{y})$ to that of $g_\beta(\mathbf{y})$.

An even more attractive possibility which does **not** invoke approximate normality would be to test the data against an assumed *distribution* $p_0(\beta)$ (maybe centred in β_0). A good measure of discrepancy would then be the (positive, invariant) directed divergence

$$\int \pi(\beta\,|\,\mathbf{y}) \log \frac{\pi(\beta\,|\,\mathbf{y})}{p_0(\beta)}\, d\beta$$

where $\pi(\beta\,|\,\mathbf{y})$ is the reference posterior mentioned before.

I would like to finish by congratulating Professor Box by this very interesting, thought-provoking paper.

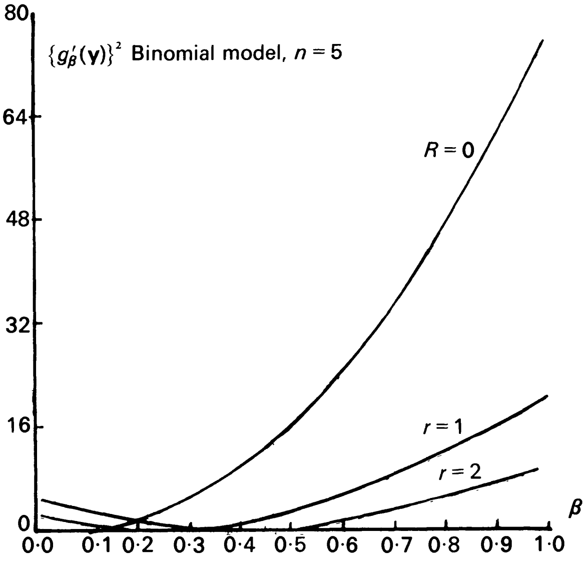

Fig. D4. Alternative discrepancy.

Professor Bovas Abraham (University of Waterloo, Canada): I wish to congratulate Professor Box for this excellent paper in which he discusses how Bayesian and Sampling Theory methods can complement each other in the scientific modelling process. It seems implied in this paper as well as in many others in the literature that the Bayesians do not perform model checking. I submit that any practically minded, and realistic Bayesian would perform diagnostic checks (residual plots, etc.) before using the appropriate posterior distributions obtained. After all, these distributions are derived under certain model assumptions. It is, therefore, comforting to see that this paper, using the concept of predictive distributions and checks, has formally justified the use of significance tests for model checking.

I agree fully with the author in his preference to parametric models over the *ad hoc* procedures suggested in the literature for "robustification". The parametric analysis has the flexibility of presenting the analyst with various alternatives. For instance, a parametric model containing a contamination parameter gives the analyst the opportunity to examine the analysis corresponding to various degrees (including zero) of contamination. Very often there is a dramatic difference in the analyses between models with extremely small and zero amounts of contamination. Of course this difference will not be revealed in an analysis of a model which excludes this possibility. This latter model is the same as the one which contains this parameter but with the value zero. It should be noted that this type of model including the contamination parameter also lends itself for further predictive checking.

I am sure that this paper will contribute significantly towards the progress of statistical science.

Professor Peter J. Bickel (University of California): I very much agree with the formulation in this paper of the process of scientific inference and even with some of the critique of traditional sampling theory—the lack of recognition that prior assumptions are prior assumptions and specifying the class of distributions which the data are presumed to come from (the parametric model) may be a giant step compared to the further specification of prior distributions on the parameters.

Where I differ from Professor Box is in his next step of requiring that prior beliefs on the parameters are made precise in terms of prior distributions and that subsequent inference be expressed in terms of corrections to such prior beliefs. It seems to me there is middle ground between what he calls implausibly imprecise prior knowledge i.e., Haar priors or global minimax criteria and proper Bayesian priors and posterior inference. These types of compromises have been explored by Hodges and Lehmann (1952) and Efron and Morris (1971) among others. I am also exploring such compromises.

I find his use of tests based on the predictive distribution very stimulating though perhaps a little inconsistent with what I think of as a purist Bayesian point of view which should try to find a prior on the space of all possible models. While the predictive distribution approach is very useful from a Bayesian point of view it is not clear to me what advantages it possesses over the classical tests of model adequacy. It can fail even as they do in detecting departures which can lead to overconservatism in inferences about parameters of interest.

Suppose, for example, in the model of Section 2 that the assumptions are off in two respects.

1. The observations are not quite normal say they are better approximated by a long-tailed symmetric density f.

2. The true variance σ_0^2 (say) is less than σ^2, the assumed variance.

Then the significance test based on (2.6) leads to large (non-significant) values of α. To see this note

$$(n-1)\frac{s_d^2}{\sigma^2} \simeq \sum_{i=1}^{n}\left\{\frac{y_i-\theta}{\sigma}\right\}^2$$

$$= n\frac{\sigma_0^2}{\sigma^2}+O_p(\sqrt{n}).$$

Since $\chi_{n-1}^2 = (n-1)+O_p(\sqrt{n})$, $\sigma_0^2<\sigma^2$ implies that α will tend to be larger than $\frac{1}{2}$ for n large.

On the other hand, the true posterior distribution of θ will be approximately normal $(\hat{\theta}, I(f)/n)$ where $\hat{\theta}$ is the MLE for f and $I(f)$ is the Fisher information.

A posterior probability such as

$$P\left[\bar{Y}+\frac{\sigma}{\sqrt{n}}\leqslant\theta\leqslant\bar{Y}+\frac{\sigma}{\sqrt{n}}\,\middle|\,Y_1,\ldots,Y_n\right]$$

will tend to a limit law, that of,

$$V=\Phi((A+\sigma)\sqrt{(I(f))})-\Phi((A-\sigma)\sqrt{(I(f))}),$$

where $A\sim$ normal $(0,\sigma_0^2-I^{-1}(f))$ (in this case).

Clearly inference is still conservative on the average (Box's robustness of validity more or less)

$$E(U)=2\Phi(\sigma/\sigma_0)-1$$

but intervals centred on a robust estimate with length proportional to its estimated standard deviation could be much shorter and also accurate.

I realize that the procedure discussed in 5.2 in which the prior takes other models into account should do better (though I am a bit worried about masking of bad values by one another). However, the point I am making is that these methods have the same shortcomings as purely frequentist methods do. In fact they can do worse. In the example I just discussed a standard goodness of fit test of normality with variance σ^2 would have (for very large sample sizes!) detected the lack of fit of the model. The approach in Section 5.2 is a help with the "bad value—long tails" problem but so is Huber's and I guess I do not see why one "fix" is preferable to the other.

Professor Bruno de Finetti (University of Rome, Mathematical Institute): I have always felt myself to be on the whole in agreement with the point of view of G. E. P. Box, and this paper seems to confirm strongly such a feeling.

Its illustration of "Bayesianism"—interpreted not simply as a systematic (and maybe sometimes careless) application of a formula—should make clear that what is essential is an exploratory way of thinking and of weighing the evidence. This involves considering when to use or not to use theorems and formulae rather than ideas expressible in words or simply at the level of instinctive propensities.

The Bayesian way of thinking must be followed everywhere, but this happens also under the guidance of unconscious mental processes (often better than conscious ones). Mathematical machinery may be a proper tool in some cases (those, for example, with a high degree of complexity, but which permit exploration), but not in others: in conditions of lack of knowledge, of haste, or of pain, its use might be disastrous.

Professor A. P. DEMPSTER (Harvard University): Professor Box displays an important trail of evolution over the 25 years since R. A. Fisher's *Statistical Methods and Scientific Inference.* The major change of principle is that scattered varieties of likelihood and fiducial methods are subordinated to Bayesian inference as the central principle of good estimation. The major practical change is a greatly heightened concern for potentially damaging effects of model-dependence, while holding firm on the necessity of models. I am in basic agreement with Box.

In the next 25 years we need much more emphasis on what a range of models tells us about the real world as partially reflected in a data set, and less on parameter estimation. The computing revolution will gradually but substantially expand the tool kit of models for which *a posteriori* analysis is available. I am pessimistic about the prospects for robustification, since an extended range of available models may show that nonrobustness is too often built into the limitiations of data and design. Checking sensitivity of results to additional parameters is often inadequate, for there may be sensitivity to simple models rather differently parametrized. For example, the extended discussion of robustness of location estimates for symmetric populations has distracted attention from the less tractable but more important task of estimating the population mean from asymmetric long-tailed populations. It hardly behoves us to be "grudging" in the pursuit of such problems.

If statisticians are to address real questions they will need formal structures to assess and incorporate sources of knowledge outside the current data set, whether from related data sets or rational arguments based on tentatively accepted science. I look to extended development of such formal structures over the next quarter century.

In summary, I applaud Professor Box's report of progress, but am impatient for more revolutionary changes ahead.

Mr R. GATHERCOLE (University College London): In general, the use of predictive checks is to be commended. Certainly, any statistical analysis should contain a review of the prior assumptions, with the aim of better understanding of the process in hand. It is useful to see the use of the Predictive Distribution in this paper, and I would like to add some remarks, together with some further ideas.

With reference to the example of Section 2, the batch mean, the preoccupation with the batch mean implies that the decision-maker has a quadratic loss function. Thus, the actual loss incurred in using the prior mean is of magnitude 36, and the expected loss is of magnitude 2·25. This as a ratio appears in the first half of the test statistic $g(y_d)$. With the same air of vagueness that surrounds the test statistic, we can use the discrepancy in the two losses to show the weakness of the prior as a predictive model. Now consider the case where the data are a random permutation of the sequence 70, 70, 70, 74. The posterior distribution is $N(70{\cdot}8, 0{\cdot}22)$, which is (informally) quite reasonable in its proximity to the prior. The test $g(y_d)$ yields a value of 12·44. Considering that $\Pr(X^2(4)\, 11{\cdot}07) = 0{\cdot}05$, what inference do we draw now about the prior?

For those of us who want to test our assumptions against completely specified alternatives it may be of interest to consider using a loss function of the sort;

$$\begin{aligned} L(d,\theta) &= 1, \quad \text{if } |d-\theta| \leqslant b, \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

This is a step loss function, gauge b, and it is only a convenience that the upper bound of the loss is one. This is not dissimilar to establishing a confidence region. In particular, a "$1-z$" confidence region may be interpreted as a decision rule with constant risk, z. This loss function has the bonus that it may be used as an indicator of how well a model performs. If we have two competing prior formulations, we may want to choose a value of the gauge that will maximize the difference in their respective expected losses. As an illustration, take two models which are univariate normal, with respective variances, c_1 and c_2. Under the stated criterion, the optimal choice of gauge is

$$b^* = [\{c_1\, c_2 \,.\, \ln(c_1/c_2)\}/(c_1 - c_2)]^{\frac{1}{2}}.$$

If we have n models with ranked variances $c_1, \ldots, c_n$, then the same formula applies, with c_n substituted for c_2.

Although this may be of little worth in "one off" cases, such as the batch mean, in sequential sampling, there will be established a string of losses (binary) for each model, which can be used in conjunction with the familiar, conventional odds ratio etc. to make an informed decision about the appropriateness of a model. No such analogy exists with validation under unbounded loss.

Professor SEYMOUR GEISSER (University of Minnesota): Professor Box considers a variety of model checking functions on the joint predictive (marginal) distribution of the entire data set. One can carry this further and calculate

$$p(\mathbf{y} \mid A) = p(\mathbf{y}_2 \mid \mathbf{y}_1, A)\, p(\mathbf{y}_1 \mid A), \tag{1}$$

where $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ for subsets $\mathbf{y}_1$ and $\mathbf{y}_2$. Suppose it is supected that the discrediting of the model is due to some outliers or "bad" observations say, the set $\mathbf{y}_2$. One can check first whether the set $\mathbf{y}_1$ adheres to the model in Boxist fashion and assuming it does, then conditional on $\mathbf{y}_1$, we can now test whether $\mathbf{y}_2$ would have been an appropriate "prediction" from the model after observing $\mathbf{y}_1$, without necessarily a discernible alternative in mind. For example, if there is one observation at issue, e.g. $\mathbf{y}_2 = y_j$ and the model is as given in Section 2, then Y_j conditional on $\mathbf{Y}_{(j)} = \mathbf{y}_{(j)}$ which represents the set $\mathbf{y}$ with y_j deleted, is normally distributed with

$$\mathrm{E}(\mathrm{Y}_j \mid \mathbf{y}_{(j)}) = \frac{(n-1)\sigma_\theta^2\, \bar{y}_{(j)} + \sigma^2 \theta_0}{(n-1)\sigma_\theta^2 + \sigma^2} = \mu_j, \tag{2}$$

$$\mathrm{var}(Y_j \mid \mathbf{y}_{(j)}) = \frac{\sigma^2(\sigma^2 + n\sigma_\theta^2)}{\sigma^2 + (n-1)\sigma_\theta^2} = \sigma_j^2, \tag{3}$$

where $\bar{y}_{(j)}$ is the mean with y_j deleted. Note that in this distribution, as σ_θ^2 grows, μ_j tends to $\bar{y}_{(j)}$ and σ_j^2 to $\sigma^2(1 + 1/(n-1))$. So analysis based on this distribution may still be useful in the sò-called "precise measurement" situation.

Now suppose the four observations were given as

$$y_1 = 71, \quad y_2 = 68, \quad y_3 = 69, \quad y_4 = 76 \quad \text{with} \quad \sigma^2 = 1, \quad \theta_0 = 70, \quad \sigma_\theta^2 = 1.$$

The Boxist computation

$$\bar{y}_4 = 71, \quad 3s_4^2 = 38, \quad g(\mathbf{y}) = \frac{(71-70)^2}{1{\cdot}25} + 38 = 38{\cdot}8, \tag{4}$$

presumably discredits the model. Notice that all of the so-called discreditation accrues to the second part of (4)—the component with three degrees of freedom. On the first glance one might be tempted to patch up the model by increasing σ^2. But an increase in σ^2 while reducing the second component also reduces the first component to such a degree as to make it unbelievably small unless one displaced θ_0. One could, in many instances, vary θ_0, σ_θ^2 and σ^2 simultaneously so that a compatible model resulted. However, the danger of deluding oneself by such Procrustean activities is obvious. In looking at the data however, one is immediately struck by y_4 which appears to be somewhat discrepant from the other three observations.

Checking to see if the first three observations discredit the model, we obtain

$$g(y_1, y_2, y_3) \doteq \frac{(69{\cdot}33 - 70)^2}{1{\cdot}33} + 2{\cdot}43 = 2{\cdot}8,$$

which is certainly in line with a χ_3^2 random variable. Now notice that Y_4 conditional on y_1, y_2, y_3 has

$$E(Y_4 \mid y_1, y_2, y_3) = 69{\cdot}5, \quad V(Y_4 \mid y_1, y_2, y_3) = 1{\cdot}25,$$

so that the observation $y_4 = 76$ is more than 5 standard deviations from the centre of its conditional predictive distribution. So perhaps the model may not be wrong but y_4 may be aberrant. One then might look for reasons why y_4 appears aberrant if one had confidence in the original model. This type of analysis could be of some value in the "precise measurement" situation where the model was assumed true but one or more observations were made under suspicious circumstances and then aberrancy needed checking. Since Box is not averse to probabilities affected by transformations, he could also entertain a diagnostic

(suppressing A)

$$d_j = \frac{p(\mathbf{y})}{p(\mathbf{y}_{(j)})} = p(y_j \mid \mathbf{y}_{(j)}), \tag{5}$$

to hunt for an observation which may be aberrant from the assumed model. In the "precise measurement" case, neither numerator or denominator need exist whilst the ratio often does.

Another diagnostic put forth by Johnson and Geisser (1979, 1980) compares the predictive distribution $p(z \mid \mathbf{y})$ of a future observation z based on $\mathbf{y}$ with the predictive distribution $p(z \mid \mathbf{y}_{(j)})$ based on $\mathbf{y}_{(j)}$. One natural measure for how these distributions differ is the Kullback–Leibler–Good–Turing measure of divergence:

$$I(p, p_j) = \int p(z \mid \mathbf{y}) \log \frac{p(z \mid \mathbf{y})}{p(z \mid \mathbf{y}_{(j)})} dz, \quad j = 1, \ldots, n. \tag{6}$$

A single "bad" observation should yield an $I(p, p_j)$ which stands out from the rest. The first diagnostic d_j stresses how poorly will the "bad" observation be predicted given the remaining data; the second $I(p, p_j)$ can be considered an overall measure of the influence of y_j for predicting future observations from the data. It reflects how the predictive distributions differ with and without the "bad" observation—which is usually the more critical issue in regard to the ultimate use of the data. Sometimes even if $I(p, p_j)$ is an order of magnitude above the others it still may be so small that the discrepancy in the probability that a future observation lies in a particular region of interest with and without the offending observation is negligible enough so that its inclusion is of little or no consequence. How it affects what you are going to conclude, infer or decide is the final arbiter of the observation's influence.

Box notes that for the Bayesian, the "bad" value problem is in a sense manageable by modelling. For one who is predictively oriented the "bad" value problem depends on what the predictive intent is. For example, bad values may come about because of circumstances that may be considered *sui generis* in a particular experiment so that no modelling apparatus is appropriate and the intent is to make a prediction that does not permit the recurrence of such circumstances. Such observations then are to be excised from influencing the predictive distribution of a future observation. Contaminated models are of a different nature and two views are possible towards them. For example, if you assume the sampling in the experiment was from

$$f(y \mid \alpha, \beta, \theta) = (1-\alpha) f_1(y \mid \beta) + \alpha f_2(y \mid \theta),$$

and there is no reason to suppose that your future observations will not be from this model, then you calculate

$$P(y_{n+1} \mid y_1, \ldots, y_n) \alpha \int \prod_{i=1}^{n+1} f(y_i \mid \alpha, \beta, \theta) g(\alpha, \beta, \theta) \, d\alpha \, d\beta \, d\theta,$$

where $g(\alpha, \beta, \theta)$ is the prior density, and that is an end to the matter.

On the other hand, as seems to be implicit in many situations if it is your intention to predict a value from $f_1(y \mid \beta)$ then the appropriate density is

$$P_1(y_{n+1} \mid y_1, \ldots, y_n) = \int f_1(y_{n+1} \mid \beta) p(\beta \mid y_1, \ldots, y_n) \, d\beta$$

where $p(\beta \mid y_1, \ldots, y_n)$ is the posterior marginal density of β. As other models arise they can be dealt with as long as the predictive intent is clear.

With regard to robustness, I believe that Table 1 is incomplete without a demonstration of the effect α or ε has on the predictive distribution of future observables at appropriate values of the independent variables.

Formula (4.1) stressed by Geisser (1969, 1971) for comparing alternatives was modified by Geisser and Eddy (1979) to

$$\prod_{j=1}^{n} p(y_j \mid \mathbf{y}_{(j)}, A_1) \Big/ \prod_{j=1}^{n} p(\mathbf{y}_j \mid y_{(j)}, A_0)$$

for "precise measurement" cases and is also viewed as a sample reuse procedure.

Professor Box has allowed us a glimpse at the contents of his artful and eclectic box of techniques and though purists of various persuasions may perceive a similarity to the legendary Box, those less apprehensive will enthusiastically welcome further revelation.

Professor V. P. Godambe and Mr P. Ferreira (University of Waterloo, Canada): We are thankful to Professor Box for explaining in so many details how the predictive distribution could be used for *model criticism.* His "model" includes a *prior distribution* (Section 1.2) and "criticism" includes *model modification* (Section 1). Thus criticism must also include estimation of the prior distribution, partly or fully. This points to the following problematic area concerning Bayesian logic and practice.

With a notation slightly extended from the author's, let a model $M \equiv (X, \Omega, \mathscr{P}, \xi)$ where $X = \{x\}$ is the sample space, $\Omega = \{\theta\}$ the parameter space, $\mathscr{P} = \{p_\theta\colon \theta \in \Omega\}$ the class of distributions and ξ is a (prior) distribution on Ω. Assuming all distributions here to be discrete we can write the predictive distribution $p_\xi(x) = \Sigma_\Omega p_\theta(x)\, \xi(\theta) \ldots (1)$. If $x = x_0$ is the (present) data, one may estimate ξ by maximizing $p_\xi(x_0)$ in (1) for the variations of ξ. Similarly we can estimate θ from $p_\theta(x_0)$. If $\hat{\xi}$ and $\hat{\theta}$ $(\hat{\theta} \in \Omega)$ are the corresponding estimates we have

$$[p_{\hat{\xi}}(x_0) \geqslant p_\xi(x_0), \forall\, \xi \text{ and } p_{\hat{\theta}}(x_0) \geqslant p_\theta(x_0),\ \theta \in \Omega] \Rightarrow [\hat{\xi}(\hat{\theta}) = 1].$$

That is, the maximum likelihood estimate $\hat{\xi}$ of the prior distribution ξ will have all its mass concentrated at the maximum likelihood estimate $\hat{\theta}$ of θ. It then follows that $p_{\xi}(\theta = \hat{\theta} \mid x_0) = 1$. Thus one may as well use just the maximum likelihood estimate of θ ignoring the Bayesian prior distribution and the methodology completely.

The above extreme example illustrates the general problem. There is nothing in Bayesian logic which can tell in any given situation how to distinguish the "prior or past knowledge" on the one hand and the "present data" on the other (Godambe, 1974, 1980). Yet such a distinction surely underlies all the conventional Bayesian applications; the past knowledge is used (informally) to construct a prior distribution and the present data to compute (formally) the posterior distribution. Worse still, it is not unusual to see a model constructed after (what is judged rightly or wrongly to be) the present data, are at hand. The data suggest some interesting investigation leading to a construction of an appropriate model.

Another related question concerns the *Bayesian options* (Section 4.6), for choosing between the alternative models. What should one do if the significance testing based on the predictive distribution (Section 2.1) rejects model M_1 but not M_2 while the posterior odds ratio based on some assumed (Bayesian) prior probabilities for the models M_1 and M_2 prefers M_1 to M_2? Such a situation can certainly arise in practice. Then one could not avoid the question, which comes first, the result of significance testing or the assumed prior probabilities?

Professor Box's suggestion for the use of the predictive distribution is unquestionably very persuasive. But its satisfactory implementation, obviously depends on how one resolves the problematic situation mentioned in the above paragraphs.

Professor Peter J. Huber (Harvard University): The first and main part of Professor Box's paper is a masterful exposition of the learning process in applied statistics, explaining how knowledge is extended by a spiralling move through modelling, design, data acquisition, data analysis, inference, then again model modification, with various minor loops, and so on. It is certainly the best such article I have ever seen.

I cannot extend this praise without reservations to the last section of the paper, on robustness from a Bayesian viewpoint. I have often wondered why Bayesian robustness did not develop as vigorously as its non-Bayesian counterpart; after all, the term "robust" was coined some 27 years ago by Professor Box, and the early childhood of robustness had quite some Bayesian flavour. For me (and also for Frank Hampel, see Hampel (1973), p. 95) some open robustness problems seem to be ideally fit for a Bayesian approach; I can only surmise they were let lie fallow because some of the goals of robustness may clash with some of the Bayesian dogmas—if the latter are interpreted narrowly. The present paper seems to offer some interesting clues on these "theological" difficulties.

The first issue is rather deeply seated in the foundations and concerns a curious misunderstanding about the "frequentist" probability interpretation. Professor Box alludes to it in Section 1.2, but I think he misses the real point. It is not that the frequentist is unaware of the subjective nature of his probability interpretation, he is only more restrictive than the Bayesian and admits a subjective interpretation only for probabilities which are sufficiently close to zero or one. For intermediate probabilities, where he does not have a direct interpretation, he takes recourse to the weak law of large numbers. This is tersely and lucidly enunciated in Kolmogorov's basic "Ergebnisheft" (1933) and the passage is worth quoting in full (from the English translation (1956), p. 4):

> "Under certain conditions, which we shall not discuss here, we may assume that to an event A which may or may not occur under conditions C, is assigned a real number $P(A)$ which has the following characteristics:
> (a) One can be practically certain that if the complex of conditions C is repeated a large number of times, n, then if m be the number of occurrences of event A, the ratio m/n will differ very slightly from $P(A)$.
> (b) If $P(A)$ is very small, one can be practically certain that when conditions C are realized only once, the event A would not occur at all."

Thus, in a certain sense, the frequentists constitute a most austere fraction of Bayesians!

As a consequence, modelling by frequentist dogma is restricted to choosing either a single point, or more generally, a point set in a space of probability distributions, in which the unknown true element lies with practical certainty.

At the bottom level, the Bayesian has more freedom: he can choose both a pointset and a prior distribution supported by it. But by Bayesian dogma, he must formalize his ignorance in terms of a prior distribution. This can create problems on the next higher level with robustness, since it is well-nigh impossible to condense some beliefs ("I expect up to about 5 per cent gross errors, which could be anywhere") into a single prior distribution, without making very arbitrary choices; these choices might matter more than what intuition tells us.

Just as in topology, it is not possible to specify smallness in absolute terms. The widespread use of 5 per cent levels in statistical testing shows that for many purposes a probability 0·05 is already close enough to zero for an investigator to accept something as a practical certainty (at least provisionally, as a working hypothesis). Sometimes, 10^{-9} is not small enough. But from the point of view of interpretation, it is absolutely essential that sufficiently small probability values are considered to be topologically close to zero. The statements made by Box in the first part of Section 5 (on $\alpha = 0{\cdot}001$) in essence negate Kolmogorov's interpretation of probability. If we take them literally, they would even knock off the props underneath the practical translation of probabilities into actions (does the probability of a fatal accident have to be exactly zero for you to risk a plane trip?).

The second issue is related and is specific to the foundations of robustness. It takes off from the very opening line of the paper: "No statistical model can safely be assumed adequate." Among other things, this dictum (with which I wholeheartedly agree) implies that the infinite regress of model improvement has to be cut short somewhere (by a judicious application of Occam's razor), and that one has to rely on robustness to take care of the remaining inadequacies.

Here Bayesian robustness fails by too strictly adhering to the dogma that uncertainty has to be formalized by a prior distribution, pulled from one's mind by introspection. This is curious, because otherwise Bayesians are much less strict and often choose priors out of mathematical convenience (e.g. the so-called conjugate priors).

Typically, Bayesians try to achieve robustness with a last model, which contains some contamination or tail-length parameters, together with some prior on these parameters. Let us call this the "super model". The super model is somewhat arbitrary, because typically neither the data nor the prior knowledge offer adequate information at this late modelling stage. Estimation then proceeds in the pious but unwarranted hope that the super model provides robustness. If the statistician has not tired yet, and still is keeping up the good work, he should note a possible lack of robustness in the next criticism stage and he will only waste a few iterations of the learning process.

Such an approach may have been the best available until about 10–12 years ago, but in between the demands and potentialities of robustness have progressed a stage further. Essentially, by now the Bayesian approach should be concerned not with the *ad hoc* construction of super models, but with deriving reliable guide-lines on how to choose the super model (within the inherent arbitrariness) so as to guarantee robustness, and how to do so in a best possible fashion. The paper by Rubin (1977) is a first small step toward such a goal.

All this has to do with modelling within a last speck of probability, just before both the frequentist and the Bayesian would switch to practical certainty, and I believe that in this region some ideas that ordinarily are anathema to Bayesians (like minimax strategies) may be much preferable to picking arbitrary distribution out of one's zone of indifference. There may lie the cause of my disagreement with Professor Box: while the $\alpha = 0{\cdot}001$ in the beginning of Section 5 may be too large to be lumped with zero, it is too small for a reliable, data- or intuition-based modelling. If we cannot change the model to approximate what is believed, because our belief is too fuzzy to stand a 1000-fold magnification, we better change the model into the worst case compatible with our belief (namely, the belief that there might be a fraction of up to about α unspecified bad values). It is curious that the Bayesian version of this (i.e. formalizing the belief about α in terms of a non-degenerate prior) to my knowledge has never been seriously explored.

Professor W. G. HUNTER (University of Wisconsin): I am reminded of something said some time ago by J. Williard Gibbs: "One of the principal objects of theoretical research in any department of knowledge is to find the point of view from which the subject appears in its greatest simplicity." I believe that Professor Box has taken an important step in that direction. He has found a vantage point from which it seems sensible essentially to divide statistical real estate between those who use Bayesian methods and those who use sampling theory. (It will be interesting to see to what extent this proposal actually leads to a reduction in territorial disputes that have enlivened the statistical landscape over the years.) Note that adjoining lands are occupied by experimenters and other research workers. I would like to comment on the value of randomization as perceived by these three different groups.

Leaving aside refinements, the gist of the story is as follows. Suppose the experimenter gives the Bayesian the data necessary for estimating the parameter $\boldsymbol{\theta}$ in a cause-and-effect model A that purportedly connects **x** to **y**. Suppose, as is usually the case, that the variance σ^2 is unknown. The Bayesian's job is to produce the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}, A)$ for the experimenter, and the residuals $\mathbf{y}-\hat{\mathbf{y}}$ for the critic. The critic is provided, by the experimenter, with additional information about how the data $(\mathbf{x}, \mathbf{y})$ were collected, including measurements on variables not contained in model A. The critic's job is to report to the experimenter whether any defects can be detected in A. The Bayesian uses (1.5), the critic (2.3).

Given **x**, **y**, and A, the added knowledge about whether randomization was used would not influence any of the Bayesian's calculations. To the Bayesian, therefore, randomization has no value. By contrast, randomization stands between the critic and unemployment and so, to the critic, it has high value. The critic's job simply cannot be done if the data have not arisen from a randomized experiment. The critic must consider whether the residual vector $\mathbf{y}-\hat{\mathbf{y}}$ resembles random error, in particular, whether it can reasonably be regarded as an approximation to a random sample from the error distribution $p(\boldsymbol{\varepsilon})$. Unlike the Bayesian, the critic does not adhere to the likelihood principle. In calculating a significance level, which is the critic's stock-in-trade, a reference distribution is required. Consequently, the critic must consider information other than **x** and **y**, basically for the same reason that Samuel Johnson said that "among the works of Nature no man can properly call a river deep, or a mountain high, without the knowledge of many mountains, many rivers".

The experimenter wants information about $\boldsymbol{\theta}$ and A, and the statistical task is finished when the Bayesian supplies the posterior distribution for $\boldsymbol{\theta}$ and the critic reports that no defects can be found in A. But the Bayesian and the critic are not infallible; even if they make no mistakes in calculation and conclude that A does not appear to be inadequate, A in fact may be grossly inadequate for future use because of the presence of an undetected lurking variable x_0. Note, in this eventuality, that randomization helps the experimenter in two ways.

(1) It tends to produce an orthogonal "design" in the sense of making the lurking variable vector perpendicular to the space defined by the vectors of the x's in the model. Suppose measurements on x_0 are available but they have thus far been ignored in the analysis of the data. If they are later brought forward, then randomization having been used will improve the experimenter's chances of detecting x_0's importance. In practice, if randomization is not used, the vector $\mathbf{x}_0$ will often tend to be close to the space defined by the vectors of the x's in the model (because of internal feedback or regulatory mechanisms or other linkages within the system being studied); accordingly, to discover the effect of x_0 will be extremely difficult, if not impossible.

(2) Even if the existence of x_0 remains undetected and it thereby biases the estimate of $\boldsymbol{\theta}$, proper randomization will tend to reduce the amount of this bias. The price paid will be to increase the variance of the posterior distribution of $\boldsymbol{\theta}$. This is a desirable trade-off. Note that the Bayesian and the critic are blind to advantages (1) and (2).

Thus, randomization should be used by the experimenter, even though the Bayesian cannot see any sense in it. Although the critic knows it is a good idea, the critic does not realize how good it is. Rather than use either Bayesian (B) or sampling theory (ST) methods, a statistician should use both. A statistician will be even more effective by learning at least some elements of the subject matter field from which the data arise, thereby becoming somewhat of an experimenter (E), too. This combination is clearly best.

Mr P. H. JACKSON (University College of Wales, Aberystwyth): I wonder whether Professor Box could be persuaded to substitute the expression "the model is called into question" for "the model is discredited" throughout Section 2 of the paper?

Practising statisticians of any school of inference will react to surprising data by asking "Have I overlooked something?", and the formal checks for surprising features proposed will be valuable in provoking this question. Usually the next questions to ask oneself are "Have I misunderstood the data?" and "Do the data contain gross recording errors?", followed by "Am I using an inappropriate model?" In considering the last question the Bayesian will recall that his theory requires him to assign prior probabilities to *all* states of nature (models in the present application), whereas his human finiteness has led him to assign zero probability to models which should really have received at least epsilon. He will therefore enquire whether there are models for which the likelihood ratio calculated from the data would have increased even an epsilon's worth of prior probability to a posterior probability greater than that for the model employed. This is especially likely to be the case when, as in many examples given in the paper, parsimony rather than genuine prior belief was the reason for selecting the model.

The objection to saying that "the model is discredited" is that it implies a decision, not a reconsideration. It will be widely interpreted to mean that the model, regarded as a hypothesis, is rejected; Professor Box himself encourages this interpretation in 2.1(d). Suppose that in my first season as captain of a cricket team I lose the toss at all twelve matches of the season. An event has occurred for which almost any relevant tail-area calculation, frequentist or Bayesian, will give a probability less than 0·001. But what is discredited? The toss as a fair way to start a match? My suitability as captain? Certainly there are models which make the data less surprising: telekinesis, "bad vibes", or the ever-available divine intervention. Believers in any of these theories might consider it wise to relieve me of the captaincy. Most of us, however, having satisfied ourselves that there was nothing fishy about the way the tosses were conducted, would still conclude that I had simply had a run of bad luck; that is, we would assign such a small probability to these alternative models *a priori* that even data as extreme as this would not give them a large probability *a posteriori.*

Professor Joseph B. Kadane (Carnegie–Mellon University): I am in sympathy with much of what Professor Box has proposed in this paper, and yet there are parts I cannot accept. His claim for the need for both model estimation and model criticism is well taken. His discussion of examples is illuminating, and his remarks on robustness in the Bayesian context are important and insightful.

My difficulty with this paper lies in his proposal to resurrect significance testing, now to be done with respect to the predictive distribution, as a method of model criticism. My difficulty does not have to do with the use of predictive distribution itself. After all, an equivalent Bayesian analysis can be performed without mentioning parameters at all, using Bayes' Theorem to update the predictive distribution,

$$p(y_{n+1}, y_{n+2}, \dots, | y_1, y_2, y_n) = \frac{p(y_1, y_2, \dots, y_n, y_{n+1}, \dots)}{P(y_1, y_2, \dots, y_n)}.$$

Rather, my difficulty comes in interpreting his equation (2.3). How shall we choose an appropriate level α, below which we decide that the model is discredited? Apparently, from equation (2.6), 0·001 is too small. Why is that? What coherent theory can justify the use of such a critical value α_0, without reference to the size of the likely discrepancies, their impact on the conclusions drawn from the analysis, etc., in other words, without a full decision-theoretic treatment?

A mathematical way of putting the same question is to point out that α, as computed in (2.3) is not invariant to changes in the underlying measure μ with respect to which the density p is a Radon–Nikodym derivative. Thus if $\mu(x)$ is taken to be Lebesgue measure if $x<0$ and k times Lebesgue measure if $x \geqslant 0$, a different interval $p(y|A) < p(y_d|A)$ results for each k, and hence a different α. Which is to be used? Generalizations of this device can lead to α's arbitrarily close to 0 and 1, by changing μ. Does Professor Box wish to argue that only densities with respect to Lebesgue measure are legitimate? By contrast, the predictive ratio (4.1) is invariant to such changes, which suggests to me that it is on a more solid footing.

Dr Ron Kenett (University of Wisconsin): A testimony to the illuminating nature of this work is that while reading it I kept asking myself why it is that this natural compromise between pure Bayesian inference and pure Sampling inference was not already widely recognized.

My comments are on the general nature of model building and specifically on paragraph 1.2 describing the need for prior distributions. I tend to agree with the author that for most problems attacked by statistical means "there seems no logical way to avoid trouble except by the explicit prior statement of the model we wish to entertain" but this might sometimes be impossible to achieve.

The alternative I have in mind stands half-way between Tukey's exploratory data analysis and Box's proposals. In other words, it seems to me that there are situations when no prior elicitation (even of a model) is possible but still there is some relevant prior information that can be used. If the model is the prior in the wide sense we might have a very "diffuse" prior. One might look at an analysis in such circumstances as a structural exploratory data analysis.

To illustrate my point let me describe an analysis in which I was involved (Karlin, Kenett and Bonne'-Tamir, 1979). Professor Bonne'-Tamir of the Human Genetics Department at Tel-Aviv University collected frequencies of various biochemical genetic traits in various Jewish populations living now in Israel and sharing a common origin such as Jews from Iraq, Poland, etc.... An investigative look at these data for similarities and differences between populations and between and within standard demographic–anthropological classifications of these populations can provide clues relevant to the study of genetic diseases and set a basic framework for successful genetic counselling. We had available information on the history of these populations, where they lived, cultural exchanges, migrations, admixture and more. A reasonable analysis of such data should incorporate this information but 2000 years of an eventful history cannot be summarized in a meaningful model. This situation does not give an initial stage for criticism to start iterating on.

My point therefore is that in some situations one will have to use *ad hoc* techniques incorporating features appropriate to the data, such as measures of particular meaning in genetics, as was done in the analysis mentioned above, and making use of prior information, such as historical knowledge, without putting it in an analytical model.

Professor TOM LEONARD (University of Wisconsin): I would like to add my congratulations to Professor Box for this highly creative landmark paper which pioneers the unification of the Bayesian and frequentist elements of parametric statistical methodology. It is by now fairly widely accepted that Bayes is very good when the statistician conditions upon the truth of his assumed sampling model; for example the dualities between Bayes and admissible procedures lead us to some of the best properties available under a frequentist philosophy. However, practical statistics is primarily concerned with aspects of modelling; whilst standard Bayesian approaches, based upon finite mixtures of specified models, are helpful, they do not seem to provide the final answer (e.g. the statistician needs to specify accurately several priors corresponding to several sampling distributions). Two alternative choices involve (a) proceeding pragmatically in the manner recommended by Professor Box, or (b) referring to a non-parametric procedure.

Some Bayesian non-parametric procedures (e.g. Ferguson, 1973; Leonard, 1978) parallel Professor Box's Bayes/non-Bayes compromise; they fit in very naturally with his important philosophy of iterative model-building. A hypothesized model is introduced as a prior estimate; the posterior estimate then indicates both local and overall differences from the hypothesized model and also suggests how the hypothesized model could be revised. One of the further prior parameters measures the degree of belief in the hypothesized model and parallel's Professor Box's significance level; it may itself be estimated from the data by either an empirical or a hierarchical Bayesian procedure.

A beneficial conclusion to be drawn by intermingling the ideas in (a) and (b) is that it might be a bit overambitious for a statistician to try to check out his model against a finite data set unless either (i) he also refers to the scientific background of the data (e.g. by interacting with a client), or (ii) he makes some particular assumptions (e.g. independence and homogeneity of errors) about the true model (or equivalently about available alternative models). In the context of non-parametrics suppose that the true density of the observation vector $\mathbf{y}$ is given by

$$f(\mathbf{y}) = \frac{\exp\{g_0(\mathbf{y})+\varepsilon(\mathbf{y})\}}{\int \exp\{g_0(\mathbf{u})+\varepsilon(\mathbf{u})\}\, d\mathbf{u}}, \tag{1}$$

where g_0 is the logistic transform of the hypothesized density

$$f_0(\mathbf{y}) = \frac{\exp\{g_0(\mathbf{y})\}}{\int \exp\{g_0(\mathbf{u})\}\, du}, \tag{2}$$

and the multi-dimensional function ε measures the departure of f from f_0.

The expression in (1) also provides the likelihood functional of ε and therefore concisely summarizes all information contained in the data about departures of f from f_0. It seems (e.g. by trying to maximize (1) with respect to ε) that this information can only be sensibly utilized on its own by making some specific assumptions about ε which effectively reduce its dimensionality. We could, for example, assume ε to take the form

$$\varepsilon(\mathbf{y}) = \sum_{i=1}^{n} \eta(y_i),$$

and then estimate the one-dimensional function η, e.g. by maximum likelihood or a Bayesian smoothing procedure or an approximation based upon a polynomial or a linear combination of basic splines. This involves an assumption analogous to independence of error terms of the true model. Without a simplifying assumption like this, on the true or alternative models, it seems that we would need to bring in information external to the likelihood functional of ε in order to reach a viable conclusion. This parallels Professor Box's choice of diagnostic statistics, which may be based upon background considerations; also his choice of significance level may be made pragmatically.

In summary, either a Bayes/non-Bayes compromise or a suitably chosen non-parametric procedure is useful in modelling situations. George Box's brilliant approach will prove historically to be a splendid addition to this area.

Professor D. V. LINDLEY (Somerset): Professor Box would have us abandon the likelihood principle when it is a question of testing the adequacy of fit of a statistical model. Even outside the Bayesian paradigm, the arguments in support of the principle by Birnbaum (1962) and Basu (1975) are most convincing and it would be interesting to know why they have been implicitly rejected tonight. *Any* test of a model surely requires some consideration of alternatives as the following example illustrates. A statistician judges a sequence of 12 0's and 1's to be Bernoulli (the model). On observing the sequence he sees 010101010101 and a plausible alternative immediately suggests itself. Suppose, however, he had observed 111010100010, then the alternative that trials of prime (composite) order always give a 1(0) is not seriously entertained because it has low probability. Yet both these sequences have the same probability on the model.

A curious feature of abandoning the likelihood principle is the need to appeal to aspecıs of the data previously thought to be irrelevant, namely the data values that were not obtained originally but might have been. For example, (2.3) implicitly assumes the sample size was fixed. Curiously, were the alternatives themselves to be parameterized, this information would again not be necessary.

The bulk of the paper is not, however concerned with such issues, but with the task of developing workable tests for the adequacy of a model and, master craftsman that he is, Box succeeds admirably. It is interesting to notice that the valuable procedures derived from the $g_\beta(y)$ criterion scarcely depend on the viewpoint adopted. Only casually, as at the end of Section 4.5 when a Monte Carlo study is mentioned, does the sampling attitude surface. And even there, under rather special assumptions, the procedure has a Bayesian interpretation given by (4.22). (Incidentally, in that argument, would it be reasonable to suppose $p(\beta)$ flat since the use of β_0 rather suggests high probability for that value?) The sampling approach is typically quite satisfactory in suggesting a criterion to consider—for, after all, even it admits that the only solutions worth considering are Bayesian solutions—where it fails is in saying what to do with the criterion, or with the plots. Tail-areas are preferred to probability ratios and it is there that the sampling argument is dangerous. Let us use the valuable results derived in tonight's paper but let us judge them by coherent standards.

Dr ROBERT B. MILLER (University of Wisconsin): Our purpose as scientists is to draw conclusions about the world from data. Scientists engage in modelling in order to guide the data analysis and data collection processes, which in turn inform the process of drawing conclusions. Thus a model must be viewed as an expression of a scientist's thinking at a particular moment and not as an objective entity. The fact that a model is widely accepted among scientists makes it neither objective nor true, only widely accepted. The only objective entity is data, and even data are corrupted by measurement error, selection bias, processing error, etc.

Professor Box rightly reminds us that the predictive distribution provides a formal mechanism for checking conclusions against objective evidence. I have reservations about equating this mechanism with sampling theory. Whereas contemplation of hypothetical, identical repetitions of an experiment is very useful in model conception, it does not enter so naturally into model validation. At least I cannot think of a natural sampling theory way to deal with structural shifts in either a model or its parameters or, at a more elementary level, with even the validation of a stationary time series model.

I believe a very important principle is embodied in Professor Box's definition of robustification as "judicious and grudging elaboration of the model to ensure against particular hazards". The principle is that robustness is more in the scientist's frame of mind than in any particular model. While statisticians will inevitably speak of robust models and robust procedures, they will, consciously or unconsciously, be speaking about the one who guides the application of these models and procedures toward conclusions about the world.

Perhaps robustness should be defined as a scientists' ability to ferret simple, lasting structures from data that are subject to myriad sources of variation. In keeping with this point of view it is well to remember that a surprisingly large variety of data patterns is consistent with the assumption of a fixed (and relatively simple) model structure if the parameters are allowed to fluctuate randomly over groups or over time or both. If we make a parameter process a part of our model, then Bayesian analysis becomes hierarchical. Some will say heretical, but I believe this point of view is the best hope for robust data analysis in such volatile fields as business, economics, and environmental modelling. Significantly, random coefficient models already have wide currency in these fields.

Finally, I wish to thank Professor Box for his very useful contribution to both the philosophical and the practical sides of statistics.

Dr D. J. Spiegelhalter (University of Nottingham): Professor Box has suggested the use of Fisher's score function to measure discrepancies from a specific model assumption. One application is within the context of testing for the shape of a univariate distribution, when, for example, normality may be embedded in the exponential power family. For known location and assuming $p(\sigma) \propto \sigma^{-1}$, Fisher's score has a simple form, and for unknown location the statistic may be closely approximated by $\Sigma\, z_i^2 \ln z_i^2$, where $z_i = (x_i - \bar{x})/s$.

Similarly, the exponential shape may be embedded in the gamma or Weibull family and simple "locally", in a specific sense, most powerful invariant tests obtained. The sampling characteristics of these tests are currently under investigation.

Another Bayes/sampling theory combination involves using the posterior probability of a specific model, embedded in a finite family of alternatives, as a test statistic. This has been shown to be successful as a small sample test for normality (Spiegelhalter, 1977, 1980).

Professor S. M. Stigler (University of Chicago): George Box has made a bold, and I think largely successful, attempt to spell out the compatible and complementary roles Bayesian inference and significance tests may have in scientific investigation. There is one point that I think is implicit in his *tour de force* that, I feel, deserves greater emphasis. He notes the importance to model criticism of the predictive distribution,

$$p(\mathbf{y} \mid A) = \int p(\mathbf{y} \mid \boldsymbol{\theta}, A)\, p(\boldsymbol{\theta} \mid A)\, d\boldsymbol{\theta}.$$

But the predictive distribution is, of course, not unique. Let $\mathbf{z} = \psi(\mathbf{y})$. Then the predictive distribution of $\mathbf{z}$ may be very different from that of $\mathbf{y}$, even though if ψ is $1-1$, the data given by $\mathbf{z}$ are equivalent to those given by $\mathbf{y}$. Different choices of ψ will render the significance test sensitive to different departures from the model. In the example that begins Section 2, we have a likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ that is a function of a sufficient statistic $T(\mathbf{y}) = (\bar{y}, s^2)$; $p(\mathbf{y} \mid \theta) = h(\mathbf{y}) \cdot g(T(\mathbf{y}) \mid \boldsymbol{\theta})$, where here $h(\mathbf{y}) \equiv 1$. This yields (2.2), a predictive distribution that is also a function of $T(\mathbf{y})$, and thus a significance test that is only sensitive to model departures that perturb the distribution of $T(\mathbf{y})$. Many other formulations are possible, such as the one Professor Box cleverly exploits later in Section 2: in effect, he works in (2.7)–(2.12) with the predictive distribution of $\mathbf{z} = \psi(\mathbf{y}) = (T(\mathbf{y}), \mathbf{u})$. He is then led to a different set of significance tests, remarking that while the test (and its outcome) is affected by transformation, this is not particularly disturbing. I expect many of his readers will be disturbed, but I am not. I think he is quite correct in expecting different answers to different questions. Still, I think he leaves us with a dilemma, namely what question should we ask?

If the predictive distribution and the resulting significance test are affected in important ways by transformation (and they are), we need guidance in the choice of transformation. Professor Box has shared his experience and considerable intuition in providing some of the needed guidance, but I feel he should give more emphasis to a formal consideration of alternative hypotheses. The tests discussed are introduced without specific (only vague) alternatives in mind, but they are in fact likelihood ratio tests for specific families of alternative hypotheses (as can be seen from De Groot, 1973, for example). I would suggest that we could borrow a clue from Professor Box's last section, where he increases our understanding of Tukey's Biweight by relating it to a Bayesian procedure, and gain some of the needed guidance in choice of transformation by studying the classes of alternatives for which the suggested tests are likelihood ratio tests. Whether or not such an approach will be both feasible and enlightening remains to be seen, but I do think further guidance in choice of transformation (or, equivalently, choice of significance test) is needed. Otherwise, we risk asking either the wrong questions, or too many questions. In many important situations involving social data, non-stationary aspects of the underlying processes will prohibit any practical appeal to the iteration that Professor Box correctly notes (at the end of Section 4) will "quickly terminate" the chase in much industrial or laboratory experimentation.

Two centuries have passed since the first statistician, Laplace, embraced both Bayesian inference and significance tests. It is a pleasure to learn at last that this was not an adulterous relationship, but one that can be defended on principle.

Professor A. ZELLNER (University of Chicago): This stimulating paper prompts the following observations. First, I believe that a scientific model is better represented by $p(\mathbf{y},\boldsymbol{\theta}|A) = p(\mathbf{y}|\boldsymbol{\theta},A)p(\boldsymbol{\theta}|A)$ than by $p(\mathbf{y}|\boldsymbol{\theta},A)$, the likelihood function since the form and content of $p(\boldsymbol{\theta}|A)$, the prior distribution are significant parts of any scientific model. Second, sampling theory considerations appear relevant before we observe the data in establishing sampling properties of Bayesian estimation and significance testing procedures, for example admissibility and average risk properties of Bayesian estimators and operating characteristics of Bayesian significance testing procedures based on posterior odds ratios; on this latter topic see, for example, Dyer (1973), Jeffreys (1967, p. 396) and Zidek (1969). After data are observed, posterior distributions and odds ratios provide a basis for inference about parameter values and hypotheses, sharp or non-sharp. Significance testing based on posterior odds ratios, mentioned briefly by Box in Sections 4.1 and 4.6 is a basic Bayesian procedure for model criticism when specific alternative hypotheses are available, as they usually are. Jeffreys (1967, Ch. V and Ch. VI) provides posterior odds ratios for many problems including hypotheses about a normal mean and standard deviation that are relevant for the normal mean problem that Box analysed in Section 2. Third, for any given sample of data, there will be some function of the observations, including the whole set that will be improbable or unusual given the model, a circumstance that led Jeffreys to write (1967, p. 385), "If mere improbability of the observations, given the hypothesis, was the criterion, any hypothesis whatever would be rejected." Then, without an alternative hypothesis, we are left with no model at all. Also, his critique of the rationale for the use of tail areas or P-values that Box employs warrants attention. Last, the basic idea of significance testing or model criticism is of great importance as Box emphasizes and thus it is fortunate that Bayesian posterior odds ratios are available for this important aspect of inference and that they generally have very good sampling properties.

The AUTHOR replied later, in writing, as follows.

I need hardly say how happy I am at the reception afforded my paper which I was particularly anxious to present here because of the unique vitality of this Society and its well known willingness to entertain and criticize ideas.

To clear up some misunderstandings and to set my reply in context, let me first make clear what I regard as the proper role of a statistican. This is not as the analyst of a single set of data, nor even as the designer and analyser of a single experiment, but rather as a colleague working with an investigator throughout the whole course of iterative deductive–inductive investigation. As a general rule he should, I think, not settle for less. In some examples the statistician is a member of a research team. In others the statistician and the investigator are the same person but it is still of value to separate his dual functions. Also I have tended to set the scene in the physical sciences where designed experiments are possible. I would however argue that the scientific process is the same for, say, an investigation in economics or sociology where the investigator is led along a path, unpredictable *a priori*, but leading to (a) the study of a number of different sets of already existing data and/or (b) the devising of appropriate surveys.

The objective taking precedence over all others is that the scientific iteration converge as surely and as quickly as possible. In this endeavour the statistician has an alternating role as sponsor and critic of the evolving model. Deduction, based on the temporary pretense that the current model is true, is attractive because it involves the statistician in "exact" estimation calculations which he alone controls. By contrast induction resting on the idea that the current model may not be true is messy and the statistician is much less in control. His role is now to present analyses in such a form, both numerical and graphical, as will accurately portray the current situation to the investigator's mind, and appropriately stimulate his colleague's imagination, leading to the next step. Although this inductive jump is the only creative part of the cycle and hence is scientifically the most important, the statistician's role in it may appear inexact and indirect.

If he finds these facts not to his liking, or if his training has left him unfamiliar with them, the statistician can construct an imaginary world consisting of only the clean deductive half of the scientific process. This has undoubted advantages. A model dubbed true remains so, all alternative models are known *a priori*, the likelihood principle and the principle of coherence reign supreme. It is possible to devise rigid "optimal" rules with known operating characteristics which aspire to elevate the statistician from a mere subjective artist to an objective automaton. But there are disadvantages. Deduction alone is sterile—by cutting the iterative process in two you kill it. What is left can have little to do with the never-ending process of model evolution which is the essence of Science.

My object then is to suggest a theory which can fully explain both the inductive and deductive statistical aspects of investigation. I argue that fcr this we need look *no further* than the factorization of the model $p(\mathbf{y}, \boldsymbol{\theta} \mid A_i)$, expressing all aspects of currently held belief at some stage i, into its Bayesian and predictive parts. In particular the predictive distribution $p(\mathbf{y} \mid A_i)$, since it is the distribution of all possible samples generated by the current model, provides, free from nuisance parameters $\boldsymbol{\theta}$, the appropriate reference set for $\mathbf{y}_d$ or for any diagnostic checking function $g(\mathbf{y}_d)$.

Some contributors have found confusing my speaking of a "sampling theory argument". I mean by this, reference of a function of the data to an appropriate reference distribution implied by the model as in a significance test. The problem of what *is* the appropriate reference set is completely resolved in the present context: it is the set defined by the predictive distribution. Notice that on this basis a "sampling theory argument" is no less so because a prior distribution is involved in the generation of the (predictive) reference set.

Although I do not think that Professor Barnard and I are far apart in our basic philosophy, I fear that his final remarks about subjectivity and relations with clients (or as I would say investigators) may be misunderstood. Surely what he says applies only to perhaps the last 5 per cent of the experimental effort when it is to be demonstrated that the final destination reached is where it is claimed to be. The other 95 per cent—the wandering journey that has finally led to that destination—involves, as I have said, many heroic subjective choices (what variables? what levels? which scales? etc., etc.) at every stage. So far as this major effort is concerned then, since we must swallow the subjective camel, why strain at the subjective gnat?

Professor Dawid complains that much of my analysis in Section 4 has reproduced classical tests, and asks what has been gained. The answer is that my proposals come from a theory having general application, in which the Predictive Eve is no longer separated from the Bayesian Adam. While highly specialized structure is not a requirement, when it exists it is appropriately used, and then reproduces sensible classical tests. So far as it goes I suspect my readers will find this more reassuring than if the contrary had been the case. For ordinary well-behaved predictive distributions zero will be the appropriate value for the score functions of Section 4 since this will imply that β_0 maximizes the predictive density.† In particular examination of the plots in Box and Cox (1964) and elsewhere verifies that, for all examples I have seen, this is sensible for $g_\lambda(\mathbf{y})$ of (4.7). The phrase "reference of $p\{g_\lambda(\mathbf{y}_d)\}$ to its sampling distribution" does not appear in the paper presented to this Society. It appeared in an earlier draft which Professor Dawid saw but was changed precisely because the earlier version did not make it sufficiently clear that the sampling involved was from the predictive reference set. It is quite evident in Sections 1, 2 and 3 that a formal check is made by referring the predictive densities $p(\mathbf{y}_d)$ and $p\{g_\beta(\mathbf{y}_d)\}$ to their predictive distributions, where necessary by computer simulation. These distributions cannot of course contain nuisance parameters.

Concerning (4.8), trials with actual samples showed that although local quadratic approximation of z was good, the approximation $B = \dot{y}^{-1}$ was exceptionally poor. It was not used therefore. Of course B is data dependent but the only purpose of the analysis leading to (4.11) is to show that (making only the assumption that z is locally quadratic in y), $g_\lambda(\mathbf{y})$ is a function not only of Tukey's statistic T_{12} but of T_{21} and T_{30} as well; as common sense says it ought to be. This analysis also sets Andrews' criterion in its proper context.

Suppose, following Professors Bernardo and Lindley my proposals are applied to the binomial, and for illustration, using Bernardo's notation, assume a beta function prior

$$p(\beta \mid A) = \beta^{m\beta_0 - 1}(1-\beta_0)^{m(1-\beta_0)-1}/B\{m\beta_0, m(1-\beta_0)\}.$$

† Which, with some advantages, replaces likelihood.

Then the predictive distribution is

$$p(r\,|\,A) = \binom{n}{r} B\{m\beta_0 + r, m(1-\beta_0)+n-r\} \Big/ B\{m\beta_0, m(1-\beta_0)\}.$$

Referral of $p(r_d\,|\,A)$ to this distribution can, without specification of alternatives, discredit the model and subsequent Bayesian analysis. In particular, as $m \to \infty$, $p(\beta\,|\,A)$ becomes concentrated at β_0 and the predictive check consists of the standard binomial significance test in which r_d is referred to

$$p(r\,|\,A) = \binom{n}{r} \beta_0^r (1-\beta_0)^{n-r}.$$

This is despite Lindley's remark that any test of a model requires alternatives. However, as I say, although such an analysis can discredit the model, it cannot adequately check it. In particular suspected deviations, such as that implied by Lindley's example, from binomial sampling would require checking functions which, I grant, would imply specific alternatives. In particular a suitable function of the data $g_\delta(r)$ with δ measuring some discrepancy from binomial sampling could be obtained using the ideas in Section 4 of the paper and would then be referred to its predictive distribution.

The analysis via $p(r\,|\,A)$ discussed above, and not that fathered on me by Professor Bernardo, follows the suggestions made in this paper. If, however, we do look at

$$g_\beta(y) = n\left(\frac{r}{n} - \beta_0\right) \Big/ \beta_0(1-\beta_0)$$

this simply confirms the point (in this case trivial) that the appropriate function of the data to consider is r itself. Reference to the predictive distribution brings us once again to the previous result. Concerning Professor Bernardo's Figs D3 and D4, I can only say that outside the famous classic by Darrell Huff (1954) it is unusual to see comparisons made between unstandardised unsquared quantities and standardised squared ones.

Returning to Professor Lindley's remarks, there is in my mind no question of abandoning the likelihood principle so far as estimation is concerned. When we are asking what can be said of a variable $\boldsymbol{\theta}$ in relation to a single vector $\mathbf{y}_d$, I do not find it surprising that the sampling rule should be irrelevant. But if we aim to judge whether samples resembling in some relevant respect the one we have observed are or are not rare, it seems to me essential to know (as part of the model) the rule by which the samples were generated. In most cases this information is available. But if it is not, then I believe no absolute check on the model is possible. In particular as Professor Hunter points out this allows an appropriate role for randomization.

Concerning tail areas and probability ratios, my prescription is that low predictive density, not location in tail areas *per se*, casts doubt on the model. It is not difficult to produce examples where the predictive distribution of a sensible checking function could have two humps, with extreme *and intermediate* values appropriately suspect. I acknowledge that I prefer to sit on a different horn of this dilemma than the one favoured by Professor Cox. My difficulty with probability ratios (and likelihood ratios and predictive ratios) is of course that while I grant that it may be useful to know that it is a million times more probable that the first man I meet when I walk down the street will be called John Smith rather than Jeremiah Hezekiah Bramblebottom, this in itself tells me little about the chance of meeting a man called John Smith. Invent a few more names and one sees the difficulty with the formula which Dr O'Hagan displays and with whose uses and difficulties I am not unfamiliar (see, for example, Box and Hill, 1967). I do find it astonishing, however, that he regards as trivial the assumption that all possible models $M_1, M_2, \ldots, M_k$ are known *a priori*. If he lacks personal experience of scientific investigation, he would need only to read any moderately accurate account of one such (e.g. Watson's *The Double Helix*, 1968) to know that models *evolve*. Dr O'Hagan is welcome to his screwdriver; I am saying that it is an unsuitable instrument for driving in a nail. Once more let me say that the difficulty with any attempt to use the Bayesian half of the model alone, is that it is eternally conditional. We can move the conditionality around but we cannot lose it. The buck stops when we cease to ignore the other half of the model $p(\mathbf{y}\,|\,A)$.

Dr Atkinson's generous attribution to me of some part in his becoming a statistician is very flattering. In response to his further remarks, Steve Bailey and I have also found that on suitable transformation apparent outliers can vanish away (Bailey and Box, 1980b), but, as he says, this does not seem to be true for the Box and Behnken data. Dr Atkinson has doubts about the accuracy of Table 1; these calculations have now been carefully rechecked and apart from very minor discrepancies they seem not to be in error but remarkably sensible as evidenced by Table D2.

Table D2

	β_1	β_2	β_3	β_4	β_{11}
No omissions	1·93 (0·42)	−1·96 (0·42)	1·13 (0·42)	−3·68 (0·42)	−1·42 (0·63)
$\varepsilon = 0·001$	2·46 (0·28)	−1·96 (0·22)	1·13 (0·22)	−3·15 (0·28)	−1·88 (0·44)
$\varepsilon = 0·020$	2·49 (0·22)	−1·96 (0·20)	1·13 (0·20)	−3·12 (0·22)	−1·89 (0·42)
y_{10} deleted	2·57 (0·12)	−1·96 (0·11)	1·13 (0·11)	−3·04 (0·12)	−2·05 (0·18)
	β_{22}	β_{33}	β_{44}	β_{12}	β_{13}
No omissions	−4·33 (0·63)	−2·24 (0·63)	−2·58 (0·63)	−1·67 (0·73)	−3·82 (0·73)
$\varepsilon = 0·001$	−4·10 (0·36)	−2·01 (0·38)	−3·05 (0·44)	−1·67 (0·39)	−3·82 (0·39)
$\varepsilon = 0·020$	−4·09 (0·34)	−2·00 (0·34)	−3·05 (0·42)	−1·67 (0·34)	−3·82 (0·34)
y_{10} deleted	−4·01 (0·17)	−1·92 (0·17)	−3·21 (0·18)	−1·68 (0·19)	−3·82 (0·19)
	β_{14}	β_{23}	β_{24}	β_{34}	
No omissions	0·95 (0·73)	−1·68 (0·73)	−2·62 (0·73)	−4·25 (0·73)	
$\varepsilon = 0·001$	−0·45 (0·95)	−1·67 (0·39)	−2·62 (0·39)	−4·25 (0·39)	
$\varepsilon = 0·020$	−0·48 (0·95)	−1·67 (0·35)	−2·62 (0·35)	−4·25 (0·34)	
y_{10} deleted	−0·95 (0·25)	−1·68 (0·19)	−2·62 (0·19)	−4·25 (0·19)	

The four rows of the table show (i) least squares estimates with no omissions; the Bailey and Box analysis for (ii) $\varepsilon = 0·001$; (iii) $\varepsilon = 0·020$ taken from Table 1 and (iv) the least squares estimates with y_{10} omitted. Comparison of the estimates shows that over the very wide range $\varepsilon = 0·001$ to $\varepsilon = 0·020$ ($\alpha = 0·005$ to $\alpha = 0·091$) the Bayesian means and standard deviations do not change very much. Also that when the four estimates differ appreciably the Bayesian means occupy a position between, but sharply different from estimates which assume for certain that there are no bad values, and from those which assume for certain that y_{10} is bad. The Bayes analysis thus produces a stable compromise which correctly acknowledges that since we do not know for sure whether y_{10} is bad or good, we should neither accept nor reject it but only downweight it. In fact it is a little more subtle than this because of course it allows appropriately for other possible bad values. In this case this mostly means that y_{13} also is downweighted slightly. The large standard deviation associated with Bayesian means for β_{14} occurs because two of the four observations essentially involved (y_{10} and y_{13}) are the major suspects.

If I understand Dr Gathercole correctly, he believes that the results from the analysis in Section 2 would be anomalous if the data had been 70, 70, 70, 74. I do not see why. In that case

$$g(\mathbf{y}_d) = \frac{(\bar{y}_d - \theta_0)^2}{n^{-1}\sigma^2 + \sigma_\theta^2} + \frac{(n-1)s_d^2}{\sigma^2} = 0·44 + 12·00$$

correctly indicating that the supposition that the test results can be treated as independently and identically distributed random variables with $\sigma^2 = 1$ is called into question, but that the other structure is not. Incidentally, responding to Mr Jackson, I agree that to speak of a model being "called into question" rather than of its being "discredited" would have reduced the chance of my intention being misunderstood. We certainly need (as Fisher clearly intended by his use of the term "discredit") to distance ourselves as far as possible from the terminology "*reject* the model/hypothesis".

Since I am not proposing to estimate the prior distribution in the manner proposed by Professor Godambe, I do not see the relevance of the first part of his contribution to this discussion. Concerning the second part, while, as I have said, I do not "reject" models on the basis he suggests and am suspicious of probability ratios, I acknowledge that it is possible to construct conundrums of the kind he mentions. But when he talks of how these should be "resolved" I assume he does not mean as some kind of mathematical puzzle but in the context of a real scientific investigation. In that case if at first there seemed to be inconsistencies in the quantities I had calculated or the plots I had made, I would try to see why, by reviewing the assumptions behind and the meanings of my various quantities. Then in cooperation with the investigator and taking into account many other vexing indeterminacies which were perhaps more relevant, I would help him to decide what to do next.

My response to Professor Kadane about choice of "significance levels" follows similar lines. In (2.6) I did not mean that 0·001 was to be taken as a critical value but only that the approximate probability should be recorded and in practice would be considered in relation to (among other things) "the size of likely discrepancies, their impact on the conclusions . . ." which he mentions. I believe however that in practice this has to be done informally and not by formal appeal to decision theory. I believe that the question: how small must α be to be small? unrealistically selects one aspect of necessary scientific subjectivity for criticism in the mistaken belief that scientific research can be made wholly objective. It seems to me that the significance test idea is natural and indeed a necessary part of the conduct and management of everyday life, and I find it hard to understand the horror with which it is sometimes greeted nowadays. The process of modification of belief occurs in two stages: (a) the recognition that the data do not fit with the presently entertained model of the world, (b) the later consideration of what are alternative models that might better explain the data. For example, suppose I have an office that looks onto, say, Oxford Street in London, normally thronged with people. One day I look out of the window at 11 o'clock in the morning and notice that there are only two people in the whole street. My initial reaction surely is that on the null (*status quo*) model this is an unusual event possibly worthy of further investigation. Alternative models that might explain the phenomenon come later. These might posit that the street has been blocked off for a ceremonial occasion, that there is a bomb scare, or that it is a Sunday, etc. But notice that the basis of the initial reaction, which requires no alternatives, is surely that I have (or could have) looked out of the window on many previous occasions and rarely have I (would I have) seen as few as two people in the street. The motivation is economy of effort and is employed by all of us hundreds of times in our daily lives—when the null model is plausible I will not worry, but when data make it implausible perhaps I should be concerned. Quality control charts and the principle of "management by exception" also employ this concept, thus ensuring that we are not often distracted except when we should be.

Professor Huber asks why Bayesian robustness has not developed as rapidly as its non-Bayesian counterpart. I think that this is a temporary situation arising from inferior publicity, lack of computer programs and (if we allow "develop" to mean proliferation of theory rather than use) to the fact that non-Bayesian robustness lends itself to mathematistry. Indeed, the problem of arbitrary choice which he lays at the Bayesian door would, I should have thought, have been a greater embarrassment to non-Bayesians. Even for the single problem of robustification against systematic heavy-tailed distributions they have already produced a bewildering plethora of solutions and the end seems nowhere in sight.

I find Professor Huber's objections to my statements in the first part of Section 5 incomprehensible, and my head is unbowed even after the invocation of Kolmogorov and the weak law of large numbers. In relation to the small probability of mishap on a plane trip "the practical translation of probabilities into actions" takes the form before each flight of my fastening my seat belt and attending to the cabin demonstration of the safety equipment, not of my refusing to fly. In the same way, the knowledge that there is a small finite probability that a sample could contain one or more bad values results in my using a robust procedure, not in my refusal to analyse the data.

The idea of using a prior distribution for α is discussed in Bailey and Box (1980a); see also (d) after (5.4) in the present paper. The results are not very different from those presented here. I believe Professor Huber's fear about the possibility of my suggestion's wasting iterations is groundless. My experience has been that to produce a satisfactory model it is not necessary that it be exactly right (no model/procedure is perfect) but rather that it not be grossly wrong in the context in which it is to be used. Examples of models that can be grossly wrong in specific contexts are the standard linear model when data are collected serially (because the model totally discounts the possibility of serial correlation), the same model in the common situation where bad values can occur (because it says they cannot happen), the normal model for the comparison of variances (because it uses, directly or indirectly, the fact that for exact normality $\mu_4 = 3\mu_2^2$). Simple and obvious repairs for these and similar cases can, I believe, be extremely effective. We do not need to throw away our traditional approaches to estimation and statistical models in an orgy of *ad hockery*.

Concerning Dr Kenett's interesting observation, I think it necessary to recognise that the scientific iteration is not necessarily completed at one location (in a wide sense it is never completed). The results from his publication based on vague models may well make possible a course of investigation, perhaps by others, resulting in models that are much more precise.

I thank the remaining contributors for their many kind remarks. Where we differ I find I have already stated my side of the case as well as I can. It therefore only remains for me to thank the Society for tentatively entertaining me and my paper in such a generous spirit.

References in the Discussion

Aitchison, J. and Dunsmore, I. R. (1965). *Statistical Prediction Analysis.* Cambridge University Press.

Atkinson, A. C. (1980). Examples showing the use of two graphical displays for the detection of influential and outlying observations in regression. In *COMPSTAT 1980* (M. M. Baritt and W. Wishart, eds). Vienna: Physica Verlag.

—— (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, in press.

Barnard, G. A. (1967). The use of the likelihood function in statistical practice. *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, **1**, pp. 27–40.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā* A, **37**, 1–71.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with Discussion). *J. R. Statist. Soc.* B, **41**, 113–147.

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

Box, G. E. P. and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, **9**, 57–71.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.

Davies, O. L. (1956). *The Design and Analysis of Industrial Experiments.* London: Oliver and Boyd.

De Groot, M. H. (1973). Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *J. Amer. Statist. Ass.*, **68**, 966–969.

Dyer, A. R. (1973). Discrimination procedures for separate families of hypotheses. *J. Amer. Statist. Ass.*, **68**, 970–974.

Ferguson, T. S. (1973). A Bayesian analysis of some non-parametric problems. *Ann. Statist.*, **1**, 209–230.

Fuchs, C. (1979). Comments on a criterion of transformation proposed by Schlesselman. *J. Amer. Statist. Ass.*, **74**, 238–239.

Geisser, S. (1969). Invited discussion on "The Bayesian outlook and its application" by J. Cornfield. *Biometrics*, **4**, 643–647.

Godambe, V. P. (1974). Review of de Finetti's *Probability Induction and Statistics. J. Amer. Statist. Ass.*, **69**, 578–580.

Hampel, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **27**, 87–104.

Huff, D. (1954). *How to Lie with Statistics.* New York: Atheneum.

Jeffreys, H. (1967). *Theory of Probability*, 3rd revised edition. London: Oxford University Press.

Johnson, W. and Geisser, S. (1979). Assessing the predictive influence of observations, University of Minnesota Technical Report No. 355 (to be published in the C. R. Rao Birthday volume).

—— (1980). A predictive view of the detection and characterization of influential observations in regression analysis, University of Minnesota Technical Report No. 365.

Karlin, S., Kenett, R. S. and Bonne'-Tamir, B. (1979). Analysis of biochemical genetic data on Jewish populations II. Results and interpretations of heterogeneity indices and distance measures with respect to standards. *Amer. J. Hum. Genet.*, **31**, 341–365.

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability.* New York: Chelsea Publications Co.

Leonard, T. (1978). Density estimation, stochastic processes and prior information (with Discussion). *J. R. Statist. Soc.* B, **40**, 113–146.

Masreliez, C. J. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE Trans. Autom. Control*, **AC-20**, 107–110.

Rubin, H. (1977). Robust Bayesian estimation. In *Statistical Decision Theory and Related Topics, II* (S. S. Gupta and D. S. Moore, eds). New York: Academic Press.

Schlesselman, J. J. (1973). Data transformation in two way analysis of variance. *J. Amer. Statist. Ass.*, **68**, 369–378.

Spiegelhalter, D. J. (1977). A test for normality against symmetric alternatives. *Biometrika*, **64**, 415–418.

—— (1980). An omnibus test for normality for small samples. *Biometrika*, **67**, 493–496.

Watson, J. D. (1968). *The Double Helix.* New York: Atheneum.

Zidek, J. V. (1969). A representation of Bayes invariant procedures in terms of Haar measure. *Ann. Inst. Statist. Maths*, **21**, 291–308.