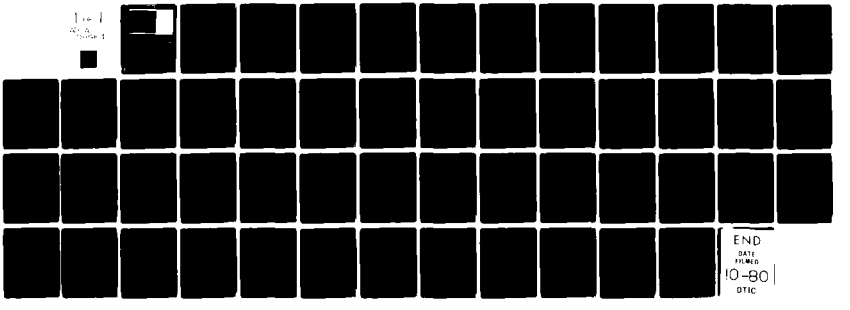
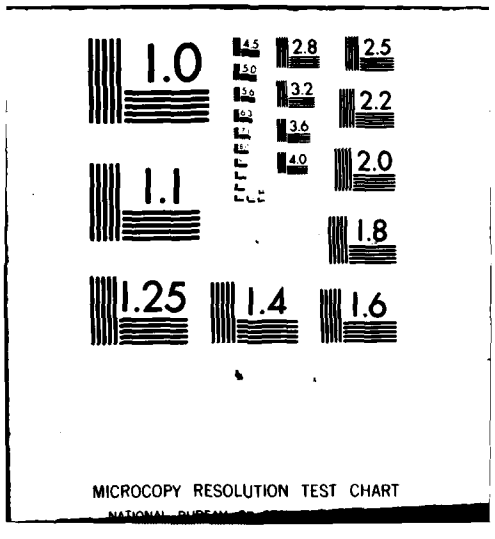
AD-A089 641 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
SOME ASPECTS OF MODEL ESTIMATION AND MODEL CRITICISM.(U)
MAY 80 S P BAILEY, G E BOX DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-2084 NL



END
DATE FILMED
10-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
MICROCOPY RESOLUTION TEST CHART

(9) MRC Technical Summary Report #2084

(6) SOME ASPECTS OF MODEL ESTIMATION AND MODEL CRITICISM.

(10) Steven P. Bailey
George E. P. Box

(12) 54

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

(11) May 80

(Received April 14, 1980)

DAAG29-75-C-0024
DAAG29-80-C-0041

Approved for public release
Distribution unlimited

Sponsored by

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

SOME ASPECTS OF MODEL ESTIMATION AND MODEL CRITICISM

Steven P. Bailey and George E. P. Box

Technical Summary Report # 2084
May 1980

ABSTRACT

The recently advanced philosophy of model building is developed further. It is stressed how Bayesian inferences based on the posterior distribution of the model parameters are appropriate only after sampling theory inferences based on the predictive distribution of the data fail to discredit the model. An example involving the normal distribution is discussed in detail. Diagnostic checking functions are developed which can be applied in an intuitive sequential manner. Careful attention is also given to the nature of the predictive distribution for the extreme situation where information about the parameters is very precise or very vague. For the latter case, it is illustrated how the predictive distribution can simultaneously (i) reflect this vague information in an appropriate manner and (ii) allow for the checking of the adequacy of the basic distributional assumptions such as normality and independence.

A particular problem in the interpretation of predictive distributions arises in situations involving a discrete data-generating distribution with vague prior knowledge about the parameter(s). This problem is explored in depth for the case of the binomial distribution.

AMS(MOS) Subject Classification: 62G35

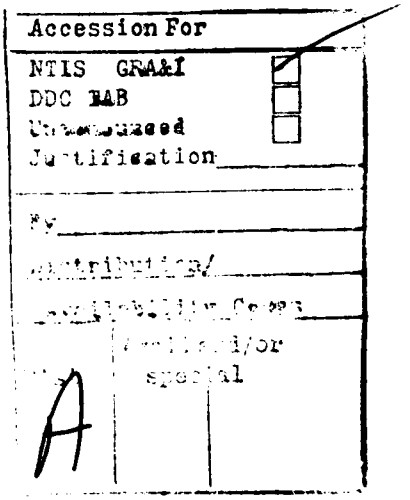
Key Words: Model building, Bayesian inference, sampling theory inference, diagnostic checks, predictive distribution, vague prior knowledge

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract Nos. DAAG29-75-C-0024 and DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

The objective of many scientific studies is to develop a model which will provide a reasonably simple yet sufficiently adequate representation of the phenomenon under consideration. At various stages of a scientific investigation, a confrontation occurs between the model being tentatively entertained at that stage and the data that have been collected up to that stage. Model estimation and model criticism are the two devices which are used by the investigator in performing the dual roles of model sponsor and model critic that are necessary for the advancement of knowledge. This paper explores in further detail a viewpoint of model building, whereby model criticism requires the sampling theory made of statistical inference, while model estimation employs the Bayesian mode of statistical inference. In particular, the implications of having only vague prior information about the model parameters are explored.



SOME ASPECTS OF MODEL ESTIMATION AND MODEL CRITICISM

Steven P. Bailey and George E. P. Box

1. Introduction.

The objective of many scientific studies is to develop a model which will provide a reasonably simple yet sufficiently adequate representation of the phenomenon under consideration. The most useful models of this nature will typically be those which elucidate not only the deterministic relationships among the variables of interest but also the stochastic relationships among the experimental errors associated with these variables. (Here the meaning of the term "deterministic relationship" is not restricted to mechanistic relationships derivable from existing theory; rather suitably developed empirical relationships, such as polynomials, are also considered as being deterministic.) In this paper a theory of model building recently advanced by Box (1979b) will be outlined and studies in further detail. (See also, for example, the following: Box, 1979a; Box, Hunter and Hunter, 1978; Box and Jenkins, 1976; Box and Tiao, 1973; Box and Youle, 1955).

At various stages of a scientific investigation, a confrontation occurs between the model being tentatively entertained at that stage and the data that have been collected up to that stage. Model estimation and model criticism are two inferential devices which aid the

Sponsored by the United States Army under Contract Nos. DAAG29-75-C-0024 and DAAG29-80-C-0041.

investigator in performing the dual role of model sponsor and model critic.

Model criticism techniques focus on the question of whether or not there is approximate concordance between the data currently available and the model in its current form. If some particular aspects of the data seem to be discordant with respect to the model, then either the model will need to be appropriately modified in an attempt to alleviate the model deficiencies, or further data will need to be collected in order to explore the inadequate aspects of the model. The broad spectrum of diagnostic techniques available for model criticism ranges from the informality of examining residual plots to the formality of carrying out goodness of fit tests, but it is argued that all such techniques are justified by the sampling theory mode of inference.

Model estimation is meaningful only when the application of the above model criticism techniques fails to reveal any model inadequacies. If such is the case, then it is appropriate to employ Bayes' Theorem in estimating the unknown model parameters by obtaining their joint posterior distribution. The use of Bayesian inference in model estimation is a logical consequence of the view that a model is, in effect, a joint probability statement of all assumptions, both explicit and implicit, which are to be

tentatively entertained at the current stage of the investigation.

Now, the model building process is iterative by nature, and, as such, there is no single approach that will be appropriate at every stage of this iteration. Thus it is not unreasonable to argue as above that two different kinds of statistical inference are required in order for an investigator to be able to both sponsor and criticize a model.

2. Model specification and subsequent inferences.

Denote by M the model under consideration at a given stage of an investigation. This model can be conveniently expressed as the joint distribution of the potentially observable data vector $\underset{\sim}{y}$ and the unknown parameter vector $\underset{\sim}{\theta}$ and can thus be written

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|M) = p(\underset{\sim}{y}|\underset{\sim}{\theta},M)p(\underset{\sim}{\theta}|M). \qquad (1)$$

When viewed as a function of $\underset{\sim}{y}$ for $\underset{\sim}{\theta}$ given, $p(\underset{\sim}{y}|\underset{\sim}{\theta},M)$ is referred to as the data-generating distribution or, more simply, the data distribution; when viewed as a function of $\underset{\sim}{\theta}$ for $\underset{\sim}{y}$ fixed, $p(\underset{\sim}{y}|\underset{\sim}{\theta},M)$ is the likelihood function. This factor combines with the prior distribution $p(\underset{\sim}{\theta}|M)$, which as argued above is an essential part of the model, to yield the complete model statement given by (1).

This expression can be subsequently factored as

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|M) = p(\underset{\sim}{\theta}|\underset{\sim}{y},M)p(\underset{\sim}{y}|M) \qquad (2)$$

where

$$p(\underset{\sim}{y}|M) = \int_{\Theta} p(\underset{\sim}{y}|\underset{\sim}{\theta},M)p(\underset{\sim}{\theta}|M)d\underset{\sim}{\theta} \qquad (3)$$

and

$$p(\underset{\sim}{\theta}|\underset{\sim}{y},M) = \frac{p(\underset{\sim}{y}|\underset{\sim}{\theta},M)p(\underset{\sim}{\theta}|M)}{p(\underset{\sim}{y}|M)} \qquad (4)$$

are, respectively, the predictive distribution of $\underset{\sim}{y}$ and the posterior distribution of $\underset{\sim}{\theta}$, with Θ denoting the parameter space of $\underset{\sim}{\theta}$. Upon obtaining the actual data, which will be denoted by $\underset{\sim}{y}_d$, the posterior distribution is

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,M) = \frac{p(\underset{\sim}{y}_d|\underset{\sim}{\theta},M)p(\underset{\sim}{\theta}|M)}{p(\underset{\sim}{y}_d|M)}, \qquad (5)$$

where the normalizing factor in the denominator is seen to be the predictive density (3) evaluated at $\underset{\sim}{y}_d$.

The predictive distribution (3) acts as a reference distribution for the observed data vector $\underset{\sim}{y}_d$, and thus an overall portmanteau check of the "goodness" of the model M is given by the probability

$$\Pr[p(\underset{\sim}{y}|M) \leq p(\underset{\sim}{y}_d|M)|M]. \qquad (6)$$

If this probability is unusually small, then the posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,M)$ provides the means of making inferences about the parameters $\underset{\sim}{\theta}$ of a model M <u>that is of questionable relevance</u>! Stated in this way, the role of the predictive distribution in model criticism is justifiably emphasized.

Of course, due to its portmanteau nature, the check (6) does not comment on the specific ways in which the n-dimensional data vector $\underset{\sim}{y}_d$ may be in discord with the model M . Of more use to the investigator will be individual checks which are sensitive to particular aspects of possible model inadequacy. Thus, if a particular function of the data, say $g(\underset{\sim}{y})$, is appropriate for assessing the validity of a particular model assumption, then the referral of the observed value $g(\underset{\sim}{y}_d)$ to its reference distribution $p(g(\underset{\sim}{y})|M)$, obtainable by integration of the predictive distribution, provides a formal check relative to the as assumption in question. Moreover, informal model criticism techniques such as the examination of residuals can be viewed as a search for any unusual feature of $\underset{\sim}{y}_d$ which suggests model inadequacy; so that, if desired, a formal assessment of the "unusualness" of such a feature can be effected by referring an appropriate summary measure of the feature to its reference distribution derivable from $p(\underset{\sim}{y}|M)$.

To illustrate the above concepts, an example is now considered.

3. The normal distribution.

Suppose that $\underset{\sim}{y} = (y_1,\ldots,y_n)'$ is a vector of n independent observations that are normally distributed with common mean θ and common variance σ^2, where θ and σ^2 are unknown but have a joint prior distribution such that conditional on a given σ^2, θ is normally distributed with mean θ_0 and variance σ^2/n_0 and that marginally $\sigma^2/\nu_0\sigma_0^2$ has an inverted χ^2 distribution with ν_0 degrees of freedom. We will not necessarily assume that n_0 is an integer or that $\nu_0 = n_0 - 1$; however, if this were the case, then the above joint prior distribution would be equivalent to assuming that a relevant set of past data $\underset{\sim}{y}_0 = (y_{01},\ldots,y_{0n_0})'$ had been combined with a non-informative prior for θ and σ^2 so that $\theta_0 = \bar{y}_0 = \frac{1}{n_0}\sum_{u=1}^{n_0} y_{0u}$ and $\sigma_0^2 = s_0^2 = \frac{1}{\nu_0}\sum_{u=1}^{n_0}(y_{0u} - \bar{y}_0)^2$. (See, for example, Box and Tiao, 1973.)

Thus

$$p(\underset{\sim}{y}|\theta,\sigma^2,M) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}[\nu s^2 + n(\bar{y} - \theta)^2]\} \qquad (7)$$

with $\bar{y} = \frac{1}{n}\sum_{u=1}^{n} y_u$, $\nu = n-1$, and $s^2 = \frac{1}{\nu}\sum_{u=1}^{n}(y_u - \bar{y})^2$; and

$$p(\theta,\sigma^2|M) = p(\theta|\sigma^2,M)p(\sigma^2|M) , \tag{8}$$

$$p(\theta|\sigma^2,M) = \left(\frac{2\pi\sigma^2}{n_0}\right)^{-\frac{1}{2}} \exp\left\{-\frac{n_0}{2\sigma^2}(\theta - \theta_0)^2\right\} , \tag{9}$$

$$p(\sigma^2|M) = \left\{\Gamma\left(\frac{\nu_0}{2}\right)2^{\frac{\nu_0}{2}}\right\}^{-1}(\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)}(\nu_0\sigma_0^2)^{\frac{\nu_0}{2}} \exp\left\{\frac{-\nu_0\sigma_0^2}{2\sigma^2}\right\} ; \tag{10}$$

these combine to give the complete model statement

$$p(\underset{\sim}{y},\theta,\sigma^2|M) = p(\underset{\sim}{y}|\theta,\sigma^2,M)p(\theta,\sigma^2|M)$$

$$= p(\underset{\sim}{y}|\theta,\sigma^2,M)p(\theta|\sigma^2,M)p(\sigma^2|M) . \tag{11}$$

Now it is well known that the joint posterior distribution of θ and σ^2 is such that conditional on a given σ^2, θ is normally distributed with mean $\tilde{\theta} = \frac{n_0\theta_0 + n\bar{y}}{n_0 + n}$ and variance $\frac{\sigma^2}{n_0 + n}$ and that, marginally, $\sigma^2/(\nu_0 + n)\tilde{\sigma}^2$ has an inverted χ^2 distribution with $\nu_0 + n$ degrees of freedom, where $(\nu_0 + n)\tilde{\sigma}^2 = \nu_0\sigma_0^2 + \nu s^2 + \frac{n_0 n(\bar{y} - \theta_0)^2}{n_0 + n}$. Thus

$$p(\theta,\sigma^2|\underset{\sim}{y},M) = p(\theta|\underset{\sim}{y},\sigma^2,M)p(\sigma^2|\underset{\sim}{y},M) , \tag{12}$$

$$p(\theta|\underset{\sim}{y},\sigma^2,M) = \left(\frac{2\pi\sigma^2}{n_0 + n}\right)^{-\frac{1}{2}} \exp\left\{-\frac{n_0 + n}{2\sigma^2}(\theta - \tilde{\theta})^2\right\} , \tag{13}$$

and

$$\begin{aligned} p(\sigma^2|\underset{\sim}{y},M) &= \left\{\Gamma\left(\frac{\nu_0 + n}{2}\right) 2^{\frac{\nu_0 + n}{2}}\right\}^{-1} (\sigma^2)^{-\left(\frac{\nu_0 + n}{2} + 1\right)} \\ &\quad \times \{(\nu_0 + n)\tilde{\sigma}^2\}^{\frac{\nu_0 + n}{2}} \exp\left\{\frac{-(\nu_0 + n)\tilde{\sigma}^2}{2\sigma^2}\right\} . \end{aligned} \tag{14}$$

If (11) is factored as

$$\begin{aligned} p(\underset{\sim}{y},\theta,\sigma^2|M) &= p(\theta,\sigma^2|\underset{\sim}{y},M)p(\underset{\sim}{y}|M) \\ &= p(\theta|\underset{\sim}{y},\sigma^2,M)p(\sigma^2|\underset{\sim}{y},M)p(\underset{\sim}{y}|M) , \end{aligned} \tag{15}$$

it follows that

$$\left.\begin{aligned} p(\underset{\sim}{y}|M) &= C_1[(\nu_0 + n)\tilde{\sigma}^2]^{-\left(\frac{\nu_0 + n}{2}\right)} , \\ C_1 &= \pi^{-\frac{n}{2}} \left(\frac{n_0}{n_0 + n}\right)^{\frac{1}{2}} \frac{\Gamma[(\nu_0 + n)/2]}{\Gamma(\nu_0/2)} (\nu_0\sigma_0^2)^{\frac{\nu_0}{2}} . \end{aligned}\right\} \tag{16}$$

Model criticism will involve using the above predictive distribution to assess whether or not relevant aspects of the observed data vector $\underset{\sim}{y}_d$ are in concordance with M. For this example, the most useful diagnostic sample quantities can be expected to directly involve $\bar{y}$, s^2, and the standardized residuals $\underset{\sim}{r} = (\underset{\sim}{y} - \bar{y}\underset{\sim}{1})/s$; thus it will be convenient to make the following transformation. First, make an orthogonal transformation from $\underset{\sim}{y}$ to $\underset{\sim}{v} = (v_1,\ldots,v_n)'$ with $v_n = \sqrt{n}\,\bar{y}$, and then transform from $\underset{\sim}{v}$ to $\bar{y}$, s^2, and $\underset{\sim}{u} = (u_1,\ldots,u_{n-2})'$ with

$$u_j = \frac{v_{j+1}}{\left[\frac{1}{j}\sum_{i=1}^{j} v_i^2\right]^{\frac{1}{2}}} \,. \tag{17}$$

The first transformation has unit Jacobian, while the second transformation has Jacobian

$$\left.\begin{aligned} J &= C_2 (s^2)^{\frac{\nu}{2} - 1} \prod_{j=1}^{n-2} \left(1 + \frac{u_j^2}{j}\right)^{-\left(\frac{j+1}{2}\right)} \\ C_2 &= n^{\frac{1}{2}} (n-1)^{\frac{n-1}{2}} \left[\Gamma(n-1)\right]^{-\frac{1}{2}} . \end{aligned}\right\} \tag{18}$$

Hence,

$$p(\bar{y}, s^2, \underset{\sim}{u}|M) = p(\underset{\sim}{y}|M)|J|$$

$$= C_1 C_2 (s^2)^{\frac{\nu}{2} - 1} [(\nu_0 + n)\tilde{\sigma}^2]^{-(\frac{\nu_0 + n}{2})}$$

$$\times \prod_{j=1}^{n-2} (1 + \frac{u_j^2}{j})^{-(\frac{j+1}{2})}. \qquad (19)$$

In the above, the factor involving a given u_j is proportional to the standardized t density with j degrees of freedom, so that

$$\left.\begin{aligned} p(u_j|M) &= a_j (1 + \frac{u_j^2}{j})^{-(\frac{j+1}{2})}, \\ a_j &= \frac{\Gamma[(j+1)/2]}{(j\pi)^{\frac{1}{2}}\Gamma(j/2)}. \end{aligned}\right\} \qquad (20)$$

The u_j's are mutually independent and independent of $\bar{y}$ and s^2, so that

$$\left.\begin{aligned} p(\underset{\sim}{u}|M) &= \prod_{j=1}^{n-2} p(u_j|M) = C_3 \prod_{j=1}^{n-2} (1 + \frac{u_j^2}{j})^{-(\frac{j+1}{2})}, \\ C_3 &= \prod_{j=1}^{n-2} a_j = \pi^{-(\frac{n-1}{2})} \Gamma(\tfrac{n-1}{2}) [\Gamma(n-1)]^{-\frac{1}{2}}, \end{aligned}\right\} \qquad (21)$$

$$p(\bar{y}, s^2|M) = \frac{C_1 C_2}{C_3} (s^2)^{\frac{\nu}{2}-1} [(\nu_0 + n)\tilde{\sigma}^2]^{-(\frac{\nu_0 + n}{2})} . \quad (22)$$

Writing $\nu_p s_p^2 = \nu_0 \sigma_0^2 + \nu s^2$, where $\nu_p = \nu_0 + \nu = \nu_0 + n - 1$, so that $(\nu_0 + n)\tilde{\sigma}^2 = \nu_p s_p^2 + \frac{n_0 n(\bar{y} - \theta_0)^2}{n_0 + n}$, (22) then becomes

$$p(\bar{y}, s^2|M) = (\frac{C_1 C_2}{C_3})(s^2)^{\frac{\nu}{2}-1} (\nu_p s_p^2)^{-(\frac{\nu_0 + n}{2})} (1 + \frac{t^2}{\nu_p})^{-(\frac{\nu_p + 1}{2})} , \quad (23)$$

where

$$t = \frac{(\bar{y} - \theta_0)}{(\frac{1}{n_0} + \frac{1}{n})^{\frac{1}{2}} s_p} . \quad (24)$$

The above establishes that transformation from $\bar{y}$ to t will achieve independence. The Jacobian of this transformation is

$$\left.\begin{aligned} \frac{\partial \bar{y}}{\partial t} &= (\frac{1}{n_0} + \frac{1}{n})^{\frac{1}{2}} s_p = C_4 (\nu_p s_p)^{\frac{1}{2}} , \\ C_4 &= (\frac{1}{n_0} + \frac{1}{n})^{\frac{1}{2}} \nu_p^{-\frac{1}{2}} . \end{aligned}\right\} \quad (25)$$

Hence,

$$p(t,s^2|M) = p(\bar{y},s^2|M)\left|\frac{\partial \bar{y}}{\partial t}\right|$$

$$= \left(\frac{C_1C_2C_4}{C_3}\right)(s^2)^{\frac{\nu}{2}-1}(\nu_p s_p^2)^{-\left(\frac{\nu_0+\nu}{2}\right)}\left(1+\frac{t^2}{\nu_p}\right)^{-\left(\frac{\nu_p+1}{2}\right)}$$

$$= \left(\frac{C_1C_2C_4}{C_3C_5}\right)\left[\frac{(s^2)^{\frac{\nu}{2}-1}}{(\nu_p s_p^2)^{\frac{\nu_0+\nu}{2}}}\right]\left[C_5\left(1+\frac{t^2}{\nu_p}\right)^{-\left(\frac{\nu_p+1}{2}\right)}\right]$$

$$= p(s^2|M)p(t|M)\ , \qquad (26)$$

$$\left.\begin{aligned} p(t|M) &= C_5\left(1+\frac{t^2}{\nu_p}\right)^{-\left(\frac{\nu_p+1}{2}\right)}, \\ C_5 &= \frac{\Gamma[(\nu_p+1)/2]}{(\nu_p\pi)^{\frac{1}{2}}\Gamma(\nu_p/2)}\ , \end{aligned}\right\} \qquad (27)$$

$$p(s^2|M) = \left(\frac{C_1C_2C_4}{C_3C_5}\right)\left[\frac{(s^2)^{\frac{\nu}{2}-1}}{(\nu_p s_p^2)^{\frac{\nu_0+\nu}{2}}}\right]$$

$$= \left(\frac{C_1C_2C_4}{C_3C_5}\right)(\nu_0\sigma_0^2)^{-\left(\frac{\nu_0+\nu}{2}\right)}\left\{\frac{(s^2)^{\frac{\nu}{2}-1}}{\left[1+\frac{\nu_p s_p^2}{\nu_0\sigma_0^2}\right]^{\frac{\nu_0+\nu}{2}}}\right\} \qquad (28)$$

$$= \left(\frac{C_1C_2C_4C_6}{C_3C_5}\right)\left\{\frac{F^{\frac{\nu}{2}-1}}{\left[1+\frac{\nu_0}{\nu}F\right]^{\frac{\nu_0+\nu}{2}}}\right\} ,$$

$$C_6 = \nu_0^{-\left(\frac{\nu_0+\nu}{2}\right)}(\sigma_0^2)^{-\left(\frac{\nu_0}{2}+1\right)} ,$$

with

$$F = \frac{s^2}{\sigma_0^2} , \qquad (29)$$

so that

$$p(F|M) = C_7 \left\{ \frac{F^{\frac{\nu}{2}-1}}{\left[1 + \frac{\nu}{\nu_0} F\right]^{\frac{\nu_0+\nu}{2}}} \right\} , \qquad C_7 = \frac{\Gamma(\frac{\nu_0+\nu}{2})}{\Gamma(\frac{\nu_0}{2})\Gamma(\frac{\nu}{2})}(\frac{\nu}{\nu_0})^{\frac{\nu}{2}} = (\frac{C_1C_2C_4C_6}{C_3C_5})\sigma_0^2 . \qquad (30)$$

In summary, then, by transforming from $\underset{\sim}{y}$ to $t = (\bar{y} - \theta_0)/(1/n_0 + 1/n)^{\frac{1}{2}}s_p^2$, $F = s^2/\sigma_0^2$, and the quantities $\underset{\sim}{u} = (u_1,\ldots,u_{n-2})'$ defined by (17), the predictive distribution of $\underset{\sim}{y}$ in (16) can be alternately expressed in the form

$$p(t,F,\underset{\sim}{u}|M) = p(t|M)p(F|M)\prod_{j=1}^{n-2} p(u_j|M) \qquad (31)$$

which is more convenient for the purposes of model criticism.

Nature of the diagnostic checks.

The above quantities can be utilized in the following sequential manner:

(1) The vector $\underset{\sim}{r}$ of standardized residuals has a sampling distribution which is uniform over the surface of an n-1 dimensional hypersphere of radius $\sqrt{n-1}$ which lies in the subspace that is orthogonal to the vector $\underset{\sim}{1}$

in n space (Andrews, 1971). Since this sampling distribution is independent of θ and σ^2, it will also be the predictive distribution of the standardized residuals. Now, the purpose of examining residuals, either informally through residual plotting or formally through the explicit consideration of any suitably chosen function $g(\underset{\sim}{r})$, is to assess whether or not $\underset{\sim}{r}$ lies in a "suspicious" direction (Box, 1960) due to some inadequacy in the assumed data-generating distribution (for example, caused by a lack of independence or of normality). Since any function of $\underset{\sim}{r}$ can be equivalently expressed as a function of $\underset{\sim}{u}$, the suspiciousness of the observed value $g(\underset{\sim}{r}_d)(=h(\underset{\sim}{u}_d))$ of any function $g(\underset{\sim}{r})(=h(\underset{\sim}{u}))$ of interest can be assessed by using the reference distribution $p(h(\underset{\sim}{u})|M)$ obtainable from $p(\underset{\sim}{u}|M)$.

For the purpose of illustration, suppose it was suspected that the responses $\underset{\sim}{y}$ might be approximately linearly affected by a variable ξ (say, lab temperature) which takes values $\underset{\sim}{\xi}_d = (\xi_{1d},\ldots\xi_{nd})'$ for the n observations. If, in terms of the orthogonal transformation from $\underset{\sim}{y}$ to $\underset{\sim}{v}$ used earlier, we define $v_{n-1} = \underset{\sim}{c}'\underset{\sim}{y}$ with $\underset{\sim}{c} = (\underset{\sim}{\xi}_d - \bar{\xi}_d\underset{\sim}{1})/S_{\xi d}$, where $\bar{\xi}_d = \frac{1}{n}\sum_{j=1}^{n}\xi_{dj}$ and $S_{\xi d} = \sum_{j=1}^{n}(\xi_{jd} - \bar{\xi}_d)^2$, then a checking function appropriate for this situation would be u_{n-2}, as defined by (17). Referral of the observed quantity $u_{(n-2)d}$ to the t distribution with $n-2$ degrees of freedom provides a check

on this possible departure from the adequacy of $p(\underset{\sim}{y}|\theta,\sigma^2,M)$. Note how this check links up with the techniques of residual plotting (of the r_{jd}'s vs. the ξ_{jd}'s) and analysis of variance (where it turns out that $u^2_{(n-2)d}$ is the mean square ratio $MS_{linear}/MS_{residual}$).

(2) If the above checks based on the u_j's and the quantities derivable from them do not invalidate the part of the model given by the data-generating distribution, then attention can be shifted to the appropriateness of the prior distribution. In particular, referral of the observed quantity $F_d = s_d^2/\sigma_0^2$ to the F distribution with ν and ν_0 degrees of freedom provides a check on the concordance of the sample estimate of σ^2, s_d^2, with the prior mean for σ^2, σ_0^2.

(3) A check on the concordance of $\bar{y}_d$, the sample estimate of θ, with θ_0, the prior mean of θ, is provided by referring the quantity $t_d = (\bar{y}_d - \theta_0)/(\frac{1}{n_0} + \frac{1}{n})s_{pd}$ to the t distribution with $\nu_p = \nu_0 + \nu$ degrees of freedom. Note that the denominator of t_d utilizes the estimate s_{pd}^2 of σ^2 which results from pooling s_d^2 and σ_0^2.

(4) If all of the above checks do not indicate model inadequacy, then the investigator can proceed to make inferences about θ and σ^2 from $p(\theta,\sigma^2|\underset{\sim}{y}_d,M)$.

In particular,

(i) $\tilde{\theta}$, the posterior mean of θ , is obtained as a weighted average of $\bar{y}_d$ and θ_0 , and

(ii) $\tilde{\sigma}^2$, the posterior mean of σ^2 , is obtained by pooling s^2_{pd} with the single degree of freedom estimate $n_0 n(\bar{y}_d - \theta_0)^2/(n_0 + n)$ of σ^2 .

A numerical example.

DeGroot (1970, p. 171) gives an example involving the present model, where the joint prior distribution of θ and σ^2 is chosen to satisfy $E(\theta|M) = 2$, $Var(\theta|M) = 5$, $E(\sigma^{-2}|M) = 3$, and $Var(\sigma^{-2}|M) = 3$. In our notation, this corresponds to

$$\left.\begin{aligned} \theta_0 &= E(\theta|M) = 2 , \\ \sigma_0^2 &= [E(\sigma^{-2}|M)]^{-1} = \tfrac{1}{3} , \\ \nu_0 &= 2[\sigma_0^4 \, Var(\sigma^{-2}|M)]^{-1} = 2[(\tfrac{1}{3})^2 3]^{-1} = 6 , \\ n_0 &= \left(\frac{\nu_0}{\nu_0 - 2}\right)\sigma_0^2[Var(\theta|M)]^{-1} = (\tfrac{6}{4})(\tfrac{1}{3})(\tfrac{1}{5}) = \tfrac{1}{10} . \end{aligned}\right\} \quad (32)$$

DeGroot then uses the summary statistics $\bar{y}_d = 4.20$ and $\sum_{u=1}^{n}(y_{ud} - \bar{y}_d)^2 = \nu s_d^2 = 5.40$ from a vector $\underset{\sim}{y}_d$ of n=10 observations to obtain the joint posterior distribution of θ and σ^2, which, in our notation, is such that conditional on a given σ^2, θ is normally distributed with mean

$$\tilde{\theta}_d = \frac{n_0\theta_0 + n\bar{y}_d}{n_0 + n} = \frac{(.1)(2) + (10)(4.20)}{.1 + 10} = \frac{42.2}{10.1} = 4.18$$

and variance $\sigma^2/(n_0+n) = \sigma^2/10.1$, and that marginally $\sigma^2/(\nu_0 + n)\tilde{\sigma}_d^2$ has an inverted χ^2 distribution with $\nu_0 + n = 16$ degrees of freedom, where

$$(\nu_0 + n)\tilde{\sigma}_d^2 = \nu_0\sigma_0^2 + \nu s_d^2 + \frac{n_0 n(\bar{y}_d - \theta_0)^2}{n_0 + n}$$

$$= 6\left(\tfrac{1}{3}\right) + 5.40 + \frac{(.1)(10)(4.20 - 2)^2}{.1 + 10} = 7.88 .$$

However, before making inferences based on $p(\theta,\sigma^2|\underset{\sim}{y}_d,M)$, the diagnostic checks discussed above should be employed to see whether or not $\underset{\sim}{y}_d$ is concordant with M. In particular, by referring

$$F_d = \frac{s_d^2}{\sigma_0^2} = \frac{.60}{\left(\frac{1}{3}\right)} = 1.80$$

to the F distribution with $\nu = 9$ and $\nu_0 = 6$ degrees of freedom and referring

$$t_d = \frac{(\bar{y}_d - \theta_0)}{(\frac{1}{n_0} + \frac{1}{n})^{\frac{1}{2}} s_{pd}} = \frac{4.20 - 2}{(\frac{1}{.1} + \frac{1}{10})^{\frac{1}{2}} \left[\frac{6(\frac{1}{3}) + 9(.60)}{6 + 9}\right]^{\frac{1}{2}}}$$

$$= \frac{2.20}{\left[\frac{(10.1)(7.40)}{15}\right]^{\frac{1}{2}}} = .986$$

to the t distribution with $\nu_0 + \nu = 15$ degrees of freedom, no evidence is seen of model inadequacy with respect to the prior distribution. Of course, a thorough criticism of the model M would require that checks based on the standardized residuals $\underset{\sim}{r}_d$ be carried out to assess the appropriateness of $p(\underset{\sim}{y}|\theta,\sigma^2,M)$; however only the minimal sufficient statistics $\bar{y}_d$ and s_d^2 are given by DeGroot.

Extreme cases of precise or vague prior knowledge.

The diagnostic checks developed for the present model are summarized in the center box of Table 1 . The rest of this table indicates the behavior of these checks when the prior information about θ and/or σ^2 is of an extreme nature (either very precise or very vague).

TABLE 1 THE NATURE OF SOME DIAGNOSTIC CHECKS FOR THE NORMAL DISTRIBUTION MODEL WITH RESPECT TO THE TYPE OF PRIOR INFORMATION

info on σ^2: / info on θ:	precise ($\nu_0 \to \infty$)	informative	vague ($\nu_0 \to 0$)
precise ($n_0 \to \infty$)	$t = \frac{\sqrt{n}(\bar{y}-\theta_0)}{\sigma_0} \sim N(0,1)$ $F = \frac{s^2}{\sigma_0^2} \sim \nu^{-1}\chi^2_\nu$ $u_j \sim t_j$; $j = 1,\ldots,n-2$	$t = \frac{\sqrt{n}(\bar{y}-\theta_0)}{s_p} \sim t_{\nu_p}$ $F = \frac{s^2}{\sigma_0^2} \sim F_{\nu,\nu_0}$ $u_j \sim t_j$; $j = 1,\ldots,n-2$	$t = \frac{\sqrt{n}(\bar{y}-\theta_0)}{s} \sim t_\nu$ "informal F" $u_j \sim t_j$; $j = 1,\ldots,n-2$
informative	$t = (\frac{1}{n_0} + \frac{1}{n})^{-\frac{1}{2}} \frac{(\bar{y}-\theta_0)}{\sigma_0} \sim N(0,1)$ $F = \frac{s^2}{\sigma_0^2} \sim \nu^{-1}\chi^2_\nu$ $u_j \sim t_j$; $j = 1,\ldots,n-2$	$t = (\frac{1}{n_0} + \frac{1}{n})^{-\frac{1}{2}} \frac{(\bar{y}-\theta_0)}{s_p} \sim t_{\nu_p}$ $F = \frac{s^2}{\sigma_0^2} \sim F_{\nu,\nu_0}$ $u_j \sim t_j$; $j = 1,\ldots,n-2$	$t = (\frac{1}{n_0} + \frac{1}{n})^{-\frac{1}{2}} \frac{(\bar{y}-\theta_0)}{s} \sim t_\nu$ " informal F" $u_j \sim t_j$; $j = 1,\ldots,n-2$
vague ($n_0 \to 0$)	"informal t" $F = \frac{s^2}{\sigma_0^2} \sim \nu^{-1}\chi^2_\nu$ $u_j \sim t_j$; $j = 1,\ldots,n-2$	"informal t" $F = \frac{s^2}{\sigma_0^2} \sim F_{\nu,\nu_0}$ $u_j \sim t_j$; $j = 1,\ldots,n-2$	"informal t" "informal F" $u_j \sim t_j$; $j = 1,\ldots,n-2$

The special cases where the prior knowledge about σ^2 is so precise that, in effect, it is assumed that σ^2 is known to equal σ_0^2, have been considered by Box (1979b, pages 13-19, where σ_θ^2 in his notation corresponds to σ_0^2/n_0 in the present notation). Also, the special cases where the prior knowledge about θ is so precise that, in effect, it is assumed that θ is known, have been discussed by Box (1979b, pages 24-26, where the assumed value of θ is denoted by θ_0 in the present notation).

In particular, consider the case where there is very precise information about both θ and σ^2. This situation can be approximated using a limiting argument where $n_0 \to \infty$ and $\nu_0 \to \infty$. On this argument, the portmanteau predictive check given by (6) will correspond to a test of the goodness of fit of the data to the $N(\theta_0, \sigma_0^2)$ distribution. Specifically, this test can be carried out by referring the observed value of the statistic

$$Q = \sum_{u=1}^{n} \left(\frac{y_u - \theta_0}{\sigma_0}\right)^2 = t^2 + \nu F \qquad (33)$$

to the χ^2 distribution with n degrees of freedom, where the limiting forms of t and F are as given in Table 1. For a discussion of the relationship between precise prior knowledge and significance tests, see Box (1979b).

Alternatively, consider the situation where there is relatively little prior information about either θ or σ^2. This could be reflected by values of n_0 and ν_0 which are very small relative to n. However, in lieu of an explicit specification of n_0 and ν_0, the posterior distribution of θ and σ^2 can be numerically approximated in a suitable manner by using a limiting argument such as that which will be developed in the next section. The consequence of this, though, would be that the predictive checks involving t and F cannot be <u>formally</u> made. This should not deter the investigator from rejecting the model M based on observed values of $\bar{y}_d$ and/or s_d^2 that he considers to be extremely unlikely, since such an action can be viewed as resulting from an informal check which, if formalized through the explicit consideration of n_0 and ν_0, would result in unusually small probabilities $\Pr[p(t|M) \leq p(t_d|M)|M]$ and/or $\Pr[p(F|M) \leq p(F_d|M)|M]$. Further discussion of this point is given by Box (1979b).

Finally, it is noted in Table 1 that, regardless of the nature of the prior distribution $p(\theta,\sigma^2|M)$, the model checks involving the standardized residuals $\underset{\sim}{r}$ (or, equivalently, the residual functions $u_1,\ldots,u_{n-2}$) are always available for the investigator to use in assessing the concordance between the observed data $\underset{\sim}{y}_d$ and the distributional form $p(\underset{\sim}{y}|\theta,\sigma^2,M)$.

4. Discussion.

Further insight into the impact that vague prior knowledge has on the predictive distribution in the example of the previous section can be gained by considering the following argument.

Suppose that the investigator wishes to characterize the relatively little prior information about θ and σ^2 by utilizing a prior in which θ and $\phi_q(\sigma^2)$ are locally uniform and independent, where

$$\phi_q(\sigma^2) = \begin{cases} \sigma^q & , \quad q \neq 0 \\ \ell n\ \sigma & , \quad q = 0 \end{cases} \tag{34}$$

so that

$$p(\theta,\sigma^2|M) \propto (\sigma^2)^{\frac{q}{2} - 1} . \tag{35}$$

The relevant posterior distributions based on the above prior are given by Box and Tiao (1973, Section 2.4.6). In the context of the previous section, $p(\theta,\sigma^2|\underset{\sim}{y},M)$ for the above situation can be obtained as the limiting case of (12) - (14)*, where $n_0 \to 0$, $\nu_0 \to -(q+1)$ and

*Note in passing that the limiting case where $n_0 \to 0$ and $\nu_0 \to 0$ will correspond to using a prior obtained from Jeffreys' Rule in the case there θ and σ^2 are not assumed to be independent a priori. (See, for example, Box and Tiao, Section 1.3.6.) However, somewhat paradoxically, the resulting prior $p(\theta,\sigma^2|M) \propto (\sigma^2)^{3/2}$ would have, from (34) and (35), the interpretation that θ and σ^{-1} are locally uniform and independent a priori.

$\sigma_0^2 \to 0$, since $p(\theta,\sigma^2|M)$ as given by (8) - (10) corresponds in the limit to the prior (35) above when terms not involving θ and σ^2 are ignored.

Using (22) and ignoring terms that do not involve either $\bar{y}$ or s^2 , it follows that

$$p(\bar{y},s^2|M) \propto (s^2)^{\frac{q}{2}-1} \tag{36}$$

in the above limiting case, so that the predictive distribution of the maximum likelihood estimates $\hat{\theta} = \bar{y}$ and $\hat{\sigma}^2 = (n-1)s^2/n$ behaves like

$$p(\hat{\theta},\hat{\sigma}^2|M) \propto (\hat{\sigma}^2)^{\frac{q}{2}-1} \tag{37}$$

By comparing (37) with (34), the following intuitively reasonable observation can be made: If the local uniformity and independence of θ and ϕ_q is assumed a priori, the behavior of the predictive distribution is such that the maximum likelihood estimates $\hat{\theta}$ and $\hat{\phi}_q = \phi_q(\hat{\sigma}^2)$ are similarly uniform and independent. Stated another way, if the investigator believes a priori that a wide range of values for (θ,ϕ_q) are equally plausible, then the predictive distribution for $(\hat{\theta},\hat{\phi}_q)$ will appropriately reflect this state of indifference.

It is of interest to consider whether similar results hold in other modeling situations. In particular, an example involving a discrete data-generating distribution is now investigated.

5. The binomial distribution.

Consider a situation in which

(i) N Bernoulli trials are performed, in each of which the probability of success is θ,

(ii) uncertainty about θ is expressed as a beta prior with parameters b_1 and b_2, and

(iii) the investigator can observe the number of successes, say Y, that occur in the N trials.

The model M corresponding to the above is thus

$$p(Y,\theta|M) = p(Y|\theta,M)p(\theta|M) , \tag{38}$$

where

$$p(Y|\theta,M) = \binom{N}{Y}\theta^{Y}(1 - \theta)^{N-Y} \quad (Y = 0,\ldots,N) \tag{39}$$

and

$$p(\theta|M) = \frac{\Gamma(b_1 + b_2)}{\Gamma(b_1)\Gamma(b_2)}\,\theta^{b_1-1}(1 - \theta)^{b_2-1} . \tag{40}$$

Subsequently,

$$p(\theta|Y,M) = \frac{\Gamma(N+b_1+b_2)}{\Gamma(Y+b_1)\Gamma(N-Y+b_2)} \theta^{Y+b_1-1}(1-\theta)^{N-Y+b_2-1} \qquad (41)$$

and

$$p(Y|M) = \binom{N}{Y} \frac{\Gamma(b_1+b_2)}{\Gamma(b_1)\Gamma(b_2)} \frac{\Gamma(Y+b_1)\Gamma(N-Y+b_2)}{\Gamma(N+b_1+b_2)}, \qquad (42)$$

so that once the actual observation Y_d becomes available to the investigator, the posterior distribution $p(\theta|Y_d,M)$ can be obtained from (41) and a check on its relevance can be made by referring Y_d to the predictive distribution (42). This predictive distribution, which is sometimes called the beta-binomial (see, for example, Kendall and Stuart, 1969, page 146) is such that

$$\left.\begin{aligned} E(Y|M) &= N\frac{b_1}{b_1+b_2} = N\theta_0 , \\ Var(Y|M) &= N\frac{b_1 b_2(b_1+b_2+N)}{(b_1+b_2)^2(b_1+b_2+1)} = N\theta_0(1-\theta_0)\frac{N_0+N}{N_0+1} \end{aligned}\right\} \qquad (43)$$

where $\theta_0 = b_1/(b_1+b_2)$ and $N_0 = b_1+b_2$.

For the situation where there is relatively precise prior knowledge about θ in that N_0 is very large in comparison to N, the predictive check involving Y will approximately correspond to the Neyman-Pearson significance

test of the null hypothesis $\theta = \theta_0$. (This correspondence becomes exact as $N_0 \to \infty$.)

Of more practical interest is the situation where there is relatively little prior knowledge about θ in that N_0 is small in comparison to N . Arguing as in the example of the previous section, the investigator may wish to characterize this lack of prior information by employing a prior which is locally uniform in some appropriate metric $\phi = \phi(\theta)$. Three particular choices for this metric are now considered in detail.

Case 1: $\phi = \theta$.

Since the admissible range $0 \leq \theta \leq 1$ is finite, a prior for θ which is globally uniform (rather than just locally uniform) can be used. This prior would correspond to the choices $b_1 = b_2 = 1$ in (40), so that $\theta_0 = .5$ and $N_0 = 2$ in (43).

Figure 1 shows, for $N = 10$, the joint distribution $p(\theta,\hat{\theta}|M)$ for this situation, where $\hat{\theta} = Y/N$ is the maximum likelihood estimate of θ . From (42),

$$p(Y|M) = \frac{1}{N+1} \quad (Y = 0,\ldots,N) \tag{44}$$

so that a uniform prior on θ over the interval $0 \leq \theta \leq 1$

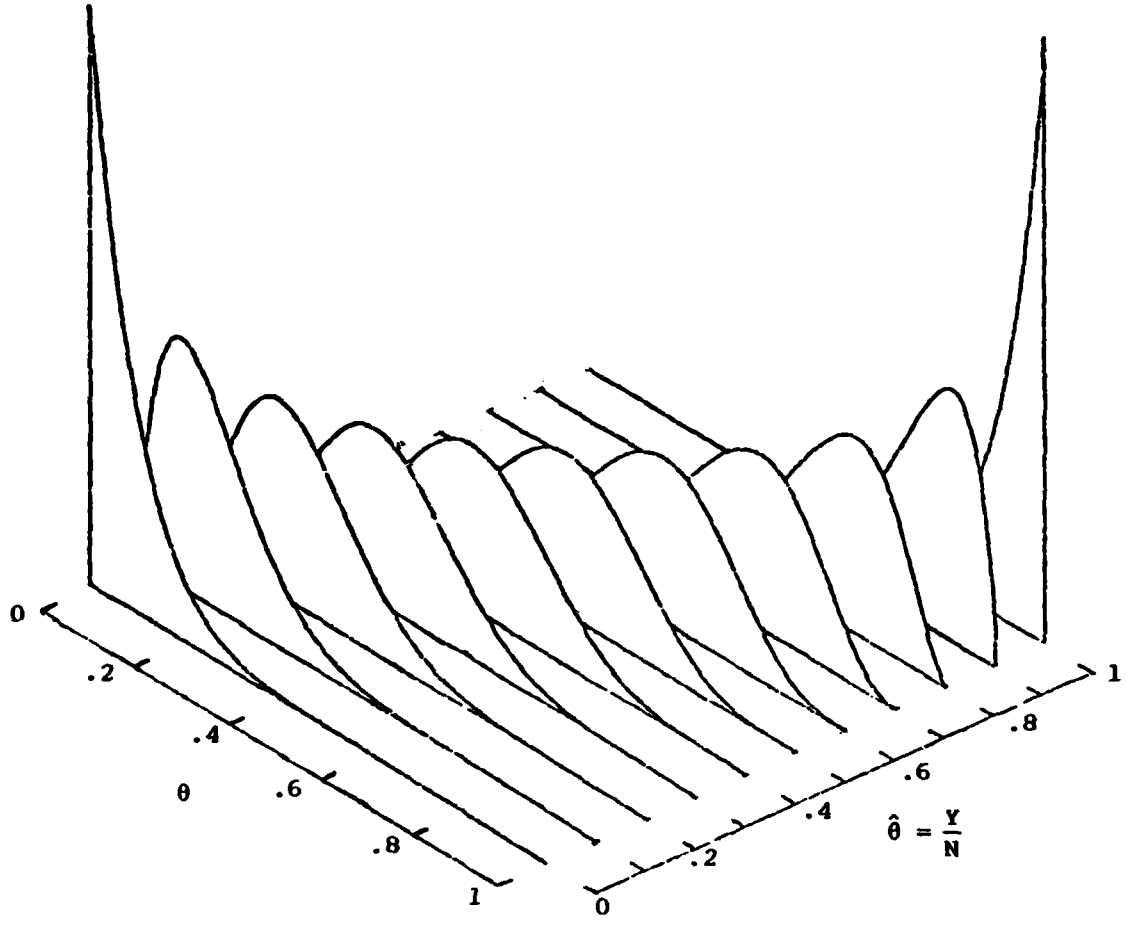


FIGURE 1 JOINT DIST. OF $\hat{\theta}$ AND θ FOR N = 10, BASED ON UNIFORM PRIOR FOR θ.

results in a discrete uniform predictive for $\hat{\theta}$, assigning probability $1/(N+1)$ to each of $\hat{\theta} = 0, \frac{1}{N}, \ldots, 1$. Note how this relationship between the prior for θ and the predictive for $\hat{\theta}$ is analogous to the relationship between the prior for the normal model parameters and the predictive for their maximum likelihood estimates that was discussed in the previous section.

To further develop the prior-predictive correspondence in the present example, consider the predictive cumulative distribution function of $\hat{\theta}$

$$F_{\hat{\theta}}(t) = \begin{cases} 0 & , \ t < 0 \\ \frac{[Nt]+1}{N+1} & , \ 0 \leq t \leq 1 \\ 1 & , \ t > 1 \end{cases} \qquad (45)$$

where $[Nt]$ denotes the integer part of Nt . Since the prior cumulative distribution function of θ is

$$F_{\theta}(t) = \begin{cases} 0 & , \ t < 0 \\ t & , \ 0 \leq t \leq 1 \\ 1 & , \ t > 1 \end{cases} \qquad (46)$$

it is easy to see that $F_{\hat{\theta}}(t)$ converges in distribution to $F_{\theta}(t)$, since $([Nt]+1)/(N+1) \to t$ as $N \to \infty$ for all $0 \leq t \leq 1$.

Case 2: $\phi = \sin^{-1}\sqrt{\theta}$.

This metric is recognized as the familiar asymptotic variance stabilizing transformation for the binomial distribution. It is also the metric in which an approximately noninformative prior distribution, as determined by Jeffreys'

Rule, will be locally uniform. (See, for example, Box and Tiao, 1973, Sections 1.3.4 and 1.3.5.) However, since the admissible range $0 \leq \sin^{-1}\sqrt{\theta} \leq \pi/2$ is finite for this metric, a globally uniform prior can be considered. In terms of the original metric θ, this prior will correspond to the choices $b_1 = b_2 = \frac{1}{2}$ in (40), so that $\theta_0 = .5$ and $N_0 = 1$ in (43).

From (42),

$$p(Y|M) = \frac{1}{\pi} \frac{\Gamma(Y + \frac{1}{2})}{\Gamma(Y + 1)} \frac{\Gamma(N - Y + \frac{1}{2})}{\Gamma(N - Y + 1)} \quad (Y = 0,\ldots,N). \qquad (47)$$

Note that this is a symmetric, "u-shaped" discrete distribution and, as such, the value(s) of Y having smallest probability will be $Y = N/2$ for N even and $Y = (N \pm 1)/2$ for N odd. Hence, all values of the maximum likelihood estimate $\phi = \sin^{-1}\sqrt{\theta} = \sin^{-1}\sqrt{Y/N}$ will not be equiprobable in the predictive sense. The failure of the uniform prior for $\phi = \sin^{-1}\sqrt{\theta}$ to produce a uniform predictive distribtuion for $\hat{\phi}$ in this case is seemingly inconsistent with the logical findings in the examples previously discussed (the normal example in Section 4 and Case 1 above for the binomial example).

Further comparison of the present case ($\phi = \sin^{-1}\sqrt{\theta}$) with the previous case ($\phi = \theta$) reveals a possible source of the above apparent inconsistency. For the $\phi = \theta$ case, with N fixed, the $N+1$ possible realizations of the estimate $\hat{\theta}$ (i.e., $\hat{\theta} = i/N$ for $i = 0,\ldots,N$) are evenly spread over the interval $0 \leq \hat{\theta} \leq 1$ corresponding to the

admissible range for θ . This, along with the uniform prior for θ , results in predictive probabilities for $\hat{\theta}$ which are uniformly distributed over these N+1 possible values. However, for the $\phi = \sin^{-1}\sqrt{\theta}$ case, the N+1 possible realizations of the estimate $\hat{\phi}$ (i.e., $\hat{\phi} = \sin^{-1}\sqrt{i/N}$ for $i = 0,\ldots,N$) are <u>unequally</u> spaced over the interval $0 \leq \hat{\phi} \leq \frac{\pi}{2}$ corresponding to the admissible range for ϕ (although they are spaced symmetrically about the midpoint of this range, $\pi/4$). Note that this unequal spacing is such that the possible values for $\hat{\phi}$ become more spread out as one moves away from the midpoint and towards either end of the admissible interval. It can thus be heuristically argued that, due to the continuous uniform prior for ϕ , the discrete predictive distribution of $\hat{\phi}$ compensates for the nonuniformity, per se, of the spacing of possible $\hat{\phi}$ values by assigning larger probabilities to those values which are further away from the midpoint $\pi/4$, in accordance with the increasingly spread out nature of these values. The result of this compensation is that the predictive probabilities of <u>different intervals</u> for $\hat{\phi}$ having the same length are more nearly equal than they would be if, say, the predictive probabilities of <u>the N+1 possible values</u> for $\hat{\phi}$ were all equal.

A formal justification of the above heuristic argument can be developed by comparing the prior cumulative distribution function of ϕ

$$F_{\phi}(t) = \begin{cases} 0 & , \quad t < 0 \\ \dfrac{t}{\pi/2} & , \quad 0 \leq t \leq \dfrac{\pi}{2} \\ 1 & , \quad t > \dfrac{\pi}{2} \end{cases} \tag{48}$$

with the predictive cumulative distribution function of $\hat{\phi}$

$$F_{\hat{\phi}}(t) = \begin{cases} 0 & , \quad t < 0 \\ \sum_{i=0}^{[N\sin^2 t]} \dfrac{1}{\pi} \dfrac{\Gamma(i + \frac{1}{2})}{\Gamma(i + 1)} \dfrac{\Gamma(N - i + \frac{1}{2})}{\Gamma(N - i + 1)} & , \quad 0 \leq t \leq \dfrac{\pi}{2} \\ 1 & , \quad t > \dfrac{\pi}{2} . \end{cases} \tag{49}$$

It turns out that $F_{\hat{\phi}}(t)$ converges in distribution to $F_{\phi}(t)$, a result which immediately follows from the fact that the predictive distribution of $\hat{\theta}$ converges to the prior distribution of θ for <u>any</u> choice of $b_1 > 0$ and $b_2 > 0$ in (40) (and, in particular, for the choice $b_1 = b_2 = \frac{1}{2}$ pertaining to the present case). The proof of this fact is given in the Appendix; verification of this fact for the particular choice $b_1 = b_2 = 1$ was given in the discussion of Case 1 above.

Visual insight as to the nature of the above result can be gained by considering Figures 2a, b and c, which show for $N = 20$, 50 and 100, respectively, the predictive cumulative distribution function of $\hat{\phi} = \sin^{-1}\sqrt{Y/N}$ (the solid line in each figure) as compared with the uniform prior cumulative distribution function of $\phi = \sin^{-1}\sqrt{\theta}$ (the broken line in each figure). Notice that, even for $N = 20$, these two distribution functions are in close agreement except at the extreme ends of the interval $0 \leq \hat{\phi} \leq \frac{\pi}{2}$, where the disagreement will necessarily be accentuated due to the discrete nature of $\hat{\phi}$ and the way in which the possible realizations of $\hat{\phi}$ are spread along this interval.

Thus, to summarize this second case, it has been argued that the predictive distribution (47), which at first glance seems to betray the vague prior information about $\phi = \sin^{-1}\sqrt{\theta}$ that is used in generating it, can upon closer look be interpreted in a sensible manner when expressed in terms of $\hat{\phi} = \sin^{-1}\sqrt{Y/N}$, since it is the unequal spacing of the possible $\hat{\phi}$ values that produces this u-shaped distribution,

Case 3: $\phi = \ell n\left(\frac{\theta}{1-\theta}\right)$.

This metric corresponds to the logarithm of the odds in favor of observing a success as the outcome of a single Bernoulli trial. Consideration of the "log odds" as an

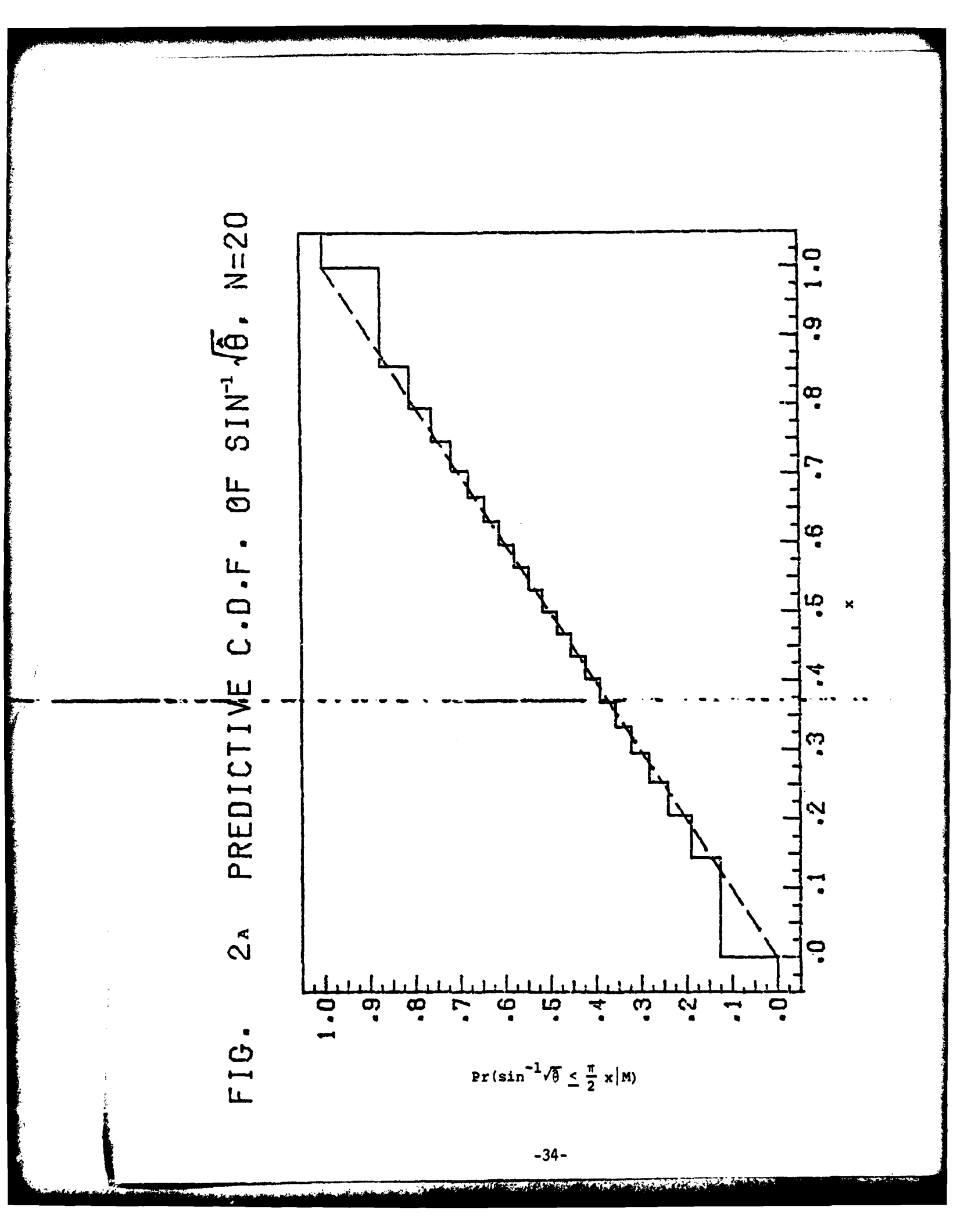


FIG. 2A PREDICTIVE C.D.F. OF $\sin^{-1}\sqrt{\hat{\theta}}$, N=20
$\Pr(\sin^{-1}\sqrt{\hat{\theta}} \leq \frac{\pi}{2} x | M)$
x
.0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0

FIG. 2B PREDICTIVE C.D.F. OF $\sin^{-1}\sqrt{\hat{\theta}}$, N=50

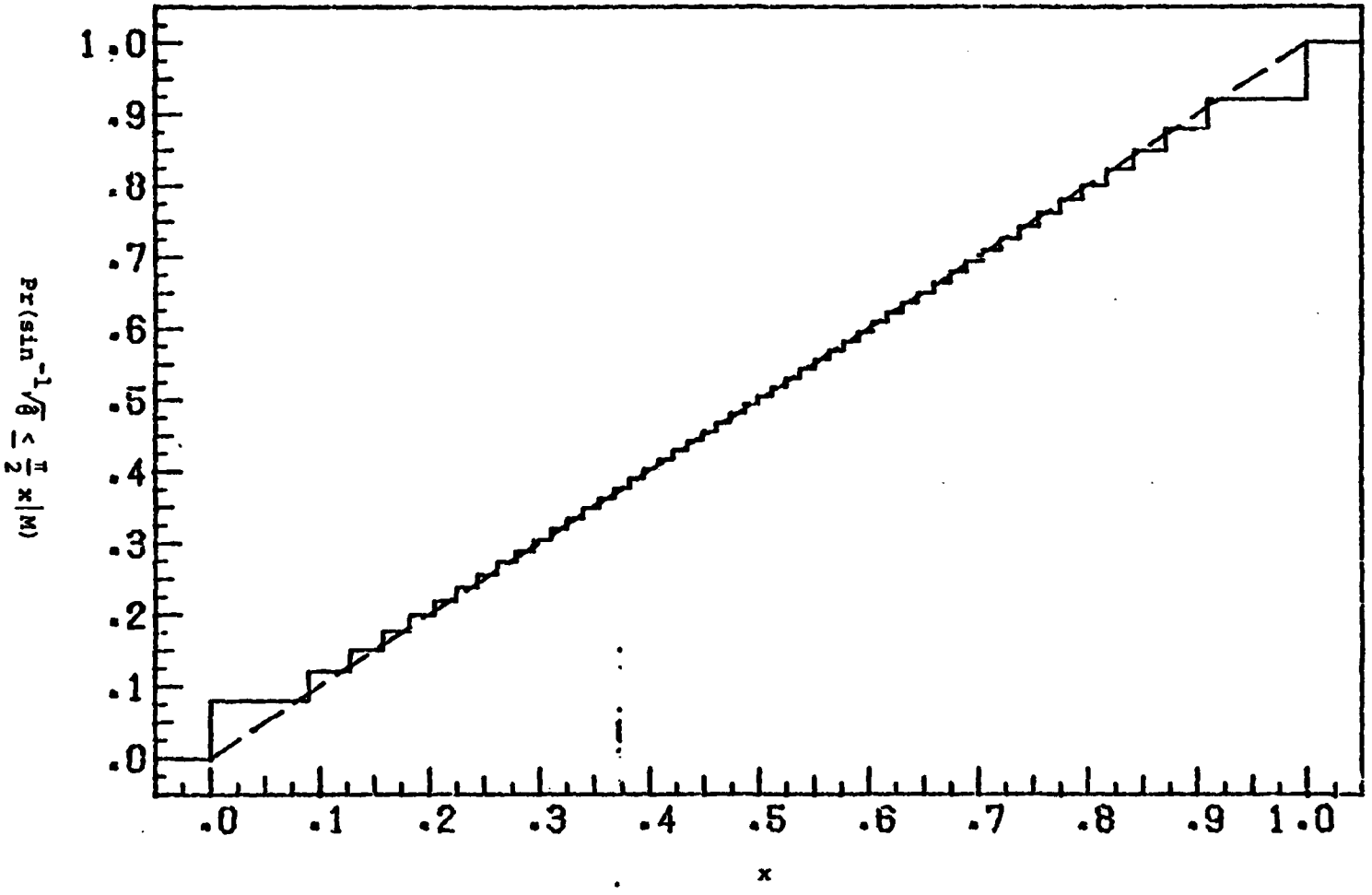
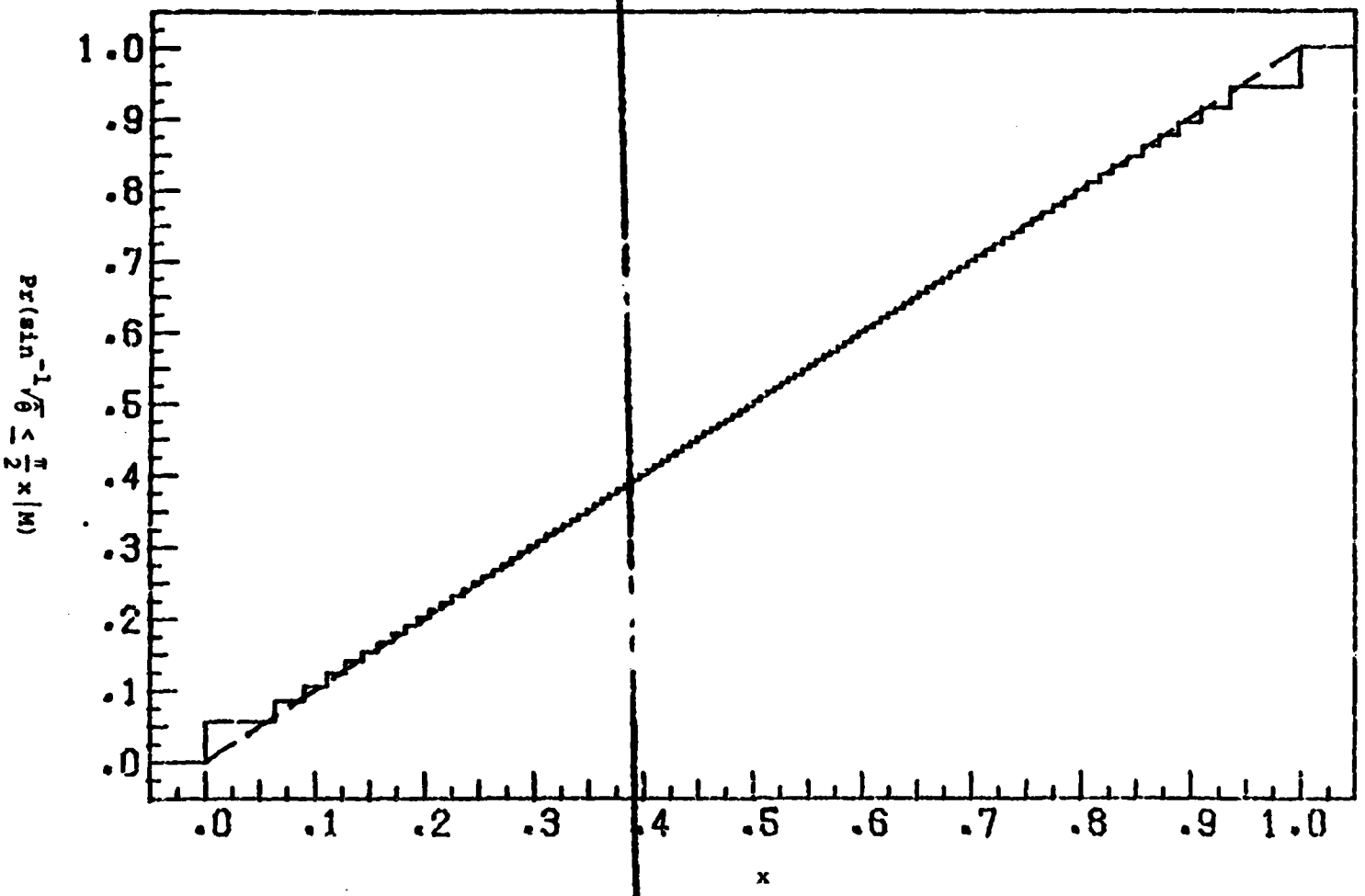


FIG. 2c PREDICTIVE C.D.F. OF $\sin^{-1}\sqrt{\hat{\theta}}$, N=100



appropriate metric in which to express prior ignorance has been advocated by several authors. (See, for example, Lindley, 1965, Section 7.2 .)

In terms of the original metric θ , a locally uniform prior for $\phi = \ell n(\frac{\theta}{1-\theta})$ will be such that

$$p(\theta|M) \propto \theta^{-1}(1-\theta)^{-1} . \qquad (50)$$

Note that, unlike the previous two cases discussed, the admissible range for the present choice of ϕ is infinite. (Specifically, this range is the extended real line, $-\infty \leq \phi \leq \infty$.) Hence it does not make sense to talk about a globally uniform prior for ϕ . Also notice, however, that if terms not involving θ are ignored, then (50) can be obtained as the limiting case of (40) with $b_1 \to 0$ and $b_2 \to 0$ (or, equivalently, with $N_0 \to 0$ for $\theta_0 = \frac{1}{2}$ fixed, using (43)).

It is precisely this limiting argument that Dawid (1979) uses as an illustration of a situation where the appropriateness of model checking via the predictive distribution is, in his view, questionable. Specifically, he notes that the limiting form of the predictive distribution for this example is

$$Pr(Y = 0|M) = Pr(Y = N|M) = \frac{1}{2} , \qquad (51)$$

so that, in Dawid's words, "any value $0 < Y < N$ discredits this 'model-cum-prior'."

It should be noted, however, that although $Y = 0$ and $Y = N$ have a combined predictive probability of unity in the limit, the corresponding limiting posterior distribution of ϕ (or, for that matter, of θ) does not exist (i.e., is improper) for both of these values of Y !*

To better understand what is happening for this case, it is worthwhile to take a closer look at this limiting process. For any fixed $N_0 > 0$, with $\theta_0 = \frac{1}{2}$ also fixed, the prior distribution for θ is equivalent to a globally uniform prior in the metric

$$\int^{\theta} [t(1 - t)]^{\frac{N_0}{2}-1} \, dt \, . \tag{52}$$

(In particular, the choices $N_0 = 2$ and $N_0 = 1$ cause (52) to correspond to θ and $\sin^{-1}\sqrt{\theta}$, respectively; these were the previous two cases discussed.) The result in the Appendix can then be applied to conclude that the predictive cumulative distribution function of the maximum likelihood estimate

*It is interesting to note that Lindley (1965) takes the view that "reliable inferences cannot be made about the ratio of successes to failures until an example of each has occurred", whereas Bernardo (1979) finds using the Case 3 prior to be "less than adequate" in comparison to using the prior discussed in Case 2.

$$\int^{Y/N} [t(1 - t)]^{\frac{N_0}{2}-1} dt \qquad (53)$$

of the metric (52) converges to the uniform prior cumulative distribution function for this metric. Thus the arguments supporting the reasonableness of the predictive distribution for Cases 1 and 2 will also apply to the present general situation where $N_0 > 0$.

However, setting $N_0 = 0$ (rather than N_0 small and positive) causes the metric (52) to correspond to $\ell n\left(\frac{\theta}{1-\theta}\right)$ and thus have an infinite admissible range (rather than the finite range obtained with any $N_0 > 0$). Furthermore, the Appendix result cannot be used as a formal argument supporting the predictive distributional form. Nevertheless, the heuristic argument used in Case 2 to justify the nature of the predictive distribution there applies here, too; since $Y = 0$ and $Y = N$ yield maximum likelihood estimates $\hat{\phi} = -\infty$ and $\hat{\phi} = \infty$, respectively, for $\phi = \ell n\left(\frac{\theta}{1-\theta}\right)$ which are so far removed from the other possible realizations of $\hat{\phi}$ that, in order to appropriately reflect the improper prior for ϕ over the whole extended real line, the predictive distribution for $\hat{\phi}$ assigns probability $\frac{1}{2}$ to each of $\hat{\phi} = \pm\infty$. (In other words, this is an extreme case of a "u-shaped" distribution caused by the discrete nature of $\hat{\phi}$ and the unequal spacing of its realizations.)

6. Further remarks on the binomial example.

The binomial example just discussed illustrates how care should be exercised in the interpretation of predictive distributions which arise from discrete data-generating distributions in situations where the prior information about the parameters is vague.

In this example it was explained how, given a reasonable representation of vague prior knowledge with respect to some metric of θ, the predictive distribution of the maximum likelihood estimate of the metric of interest will appropriately reflect this ignorance. The three most popular candidates for representing the vague prior information were each discussed individually. (For the problem of deciding which of these three choices should be preferred over the others, the reader is referred to Section 3.4 of Bernardo, 1979, and the references in that paper.)

It should be noted that, although these three choices give rise to quite different predictive distributions for Y, the corresponding posterior distributions for θ will not differ greatly in most cases, when N is not too small. Specifically, Good (1965) comments that "when there are more than three successes and three failures, there is little difference between the three methods. . .for many practical purposes."

Finally, note that from the beginning it has been assumed that only the number of successes Y that occur in the N trials is observed by the investigator. Suppose, instead, that each of the N actual Bernoulli trial results is observed individually, so that a vector $\underset{\sim}{y} = (y_1,\ldots,y_N)'$ of zeros and ones corresponding to failures and successes, respectively, represents the experimental results. Then, in terms of the previous model $p(Y,\theta|M)$ given by (38), the relevant model now becomes

$$\begin{aligned} p(\underset{\sim}{y},\theta|M) &= p(\underset{\sim}{y}|\theta,M)p(\theta|M) \\ &= p(\underset{\sim}{y}|Y,M)p(Y,\theta|M) , \end{aligned} \tag{54}$$

where the additional factor

$$p(\underset{\sim}{y}|Y,M) = \begin{cases} \binom{N}{Y}^{-1} & \text{for } \underset{\sim}{y}'\underset{\sim}{y} = Y \\ 0 & \text{otherwise} \end{cases} , \tag{55}$$

can be viewed as the conditional predictive distribution of $\underset{\sim}{y}$ given Y and can thus be used in obtaining diagnostic checks for departure from the assumed form of the data-generating distribution caused by, say, serial dependence.

7. Summary.

This paper has dealt with an approach to model building whereby a sampling theory argument is used in criticizing a tentative model by referring diagnostic functions of the observed data to their appropriate reference predictive distributions, while a Bayesian argument is used in estimating the model parameters via the posterior distribution. Examples involving the normal and the binomial distributions illustrated the main points of this approach. Particular attention was given to situations in which prior knowledge about the parameters was vague and, for the binomial case, particular difficulties in predictive distribution interpretation were discussed.

Appendix. Proof that the predictive distribution of $\hat{\theta}$ converges to the prior distribution of θ for the binomial model.

The prior cumulative distribution function of θ and the predictive cumulative distribution of $\hat{\theta}$ are, respectively,

$$F_\theta(t) = \begin{cases} 0 & , \ t < 0 \\ \int_0^t \frac{\Gamma(b_1 + b_2)}{\Gamma(b_1)\Gamma(b_2)} u^{b_1-1}(1-u)^{b_2-1} du & , \ 0 \le t \le 1 \\ 1 & , \ t > 1 \end{cases} \tag{A1}$$

$$F_{\hat{\theta}}(t) = \begin{cases} 0 & , \ t < 0 \\ \sum_{i=0}^{[Nt]} \binom{N}{i} \frac{\Gamma(b_1+b_2)}{\Gamma(b_1)\Gamma(b_2)} \frac{\Gamma(i+b_1)\Gamma(N-i+b_2)}{\Gamma(N+b_1+b_2)} & , \ 0 \le t \le 1 \\ 1 & , \ t > 1 \ . \end{cases} \tag{A2}$$

These agree exactly on the intervals $t < 0$ and $t \ge 1$. For $t = 0$ $F_\theta(0) = 0$, while $F_{\hat{\theta}}(0) = \frac{\Gamma(b_1+b_2)}{\Gamma(b_1)} \frac{\Gamma(N+b_2)}{\Gamma(N+b_1+b_2)} \to 0$ as $N \to \infty$, since $\frac{\Gamma(N+b_2)}{\Gamma(N+b_1+b_2)}$ behaves like N^{-b_2} for N large, as can be verified from Stirling's series. (See, for example, Box and Tiao, 1973, Appendix A2.2.) It thus suffices to show that

$$\sum_{i=0}^{[Nt]} \binom{N}{i} \frac{\Gamma(i + b_1)\Gamma(N - i + b_2)}{\Gamma(N + b_1 + b_2)} \longrightarrow \int_0^t u^{b_1-1} (i - u)^{b_2-1} du \tag{A3}$$

as $N \to \infty$ for $0 < t < 1$.

Now, the summand on the left-hand side of (A3) can be written

$$\frac{\Gamma(N + 1)}{\Gamma(N + b_1 + b_2)} \frac{\Gamma(i + b_1)}{\Gamma(i + 1)} \frac{\Gamma(N - i + b_2)}{\Gamma(N - i + 1)} , \tag{A4}$$

which behaves like

$$N^{1-b_1-b_2} \, i^{b_1-1} \, (N - i)^{b_2-1} \tag{A5}$$

for N , i and $N-i$ all large; so that summation from $i = [\ell n\ N] + 1$ to $i = [Nt]$ for N large gives

$$\frac{1}{N} \sum_{i=[\ell n\ N]+1}^{[Nt]} \left(\frac{i}{N}\right)^{b_1-1} \left(1 - \frac{i}{N}\right)^{b_2-1} , \tag{A6}$$

which is recognized as a Riemann sum representation of the integral

$$\int_{[\ell n\ N]/N}^{[Nt]/N} u^{b_1-1} (i - u)^{b_2-1} du . \tag{A7}$$

Furthermore, when N is large and i is small in comparison, (A4) behaves like

$$N^{1-b_1-b_2} \frac{\Gamma(i + b_1)}{\Gamma(i + 1)} N^{b_2-1} ; \qquad (A8)$$

so that summation from $i = 0$ to $i = [\ell n\ N]$ for N large gives

$$N^{-b_1} \sum_{i=0}^{[\ell n\ N]} \frac{\Gamma(i + b_1)}{\Gamma(i + 1)} = N^{-b_1} \frac{\Gamma([\ell n\ N] + b_1 + 1)}{\Gamma([\ell n\ N] + 1)b_1} , \qquad (A9)$$

which behaves like

$$\frac{1}{b_1} \left(\frac{[\ell n\ N]}{N}\right)^{b_1} \qquad (A10)$$

and thus approaches zero as $N \to \infty$. Hence, letting $N \to \infty$ in (A7) gives the desired result (A3).

References.

Andrews, D. F. (1971). Significance tests based on residuals. Biometrika, 58, 139-48.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). J. Roy. Statist. Soc. B, 41, 113-47.

Box, G. E. P. (1960). Fitting empirical data. Ann. New York Academy of Sciences, 86, 792-816.

Box, G. E. P. (1979a). Robustness in the strategy of scientific model building. In Robustness in Statistics, 201-36, eds. Launer, R. L. and Wilkinson, G. N. New York, Academic Press.

Box, G. E. P. (1979b). Sampling and Bayes' inference in the advancement of learning. Tech. Report 1969, Math. Research Center, Univ. of Wisconsin, Madison, Wis.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York, Wiley.

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. San Francisco, Holden-Day.

Box, G. E. P. and Tiao, G. C. (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.

Dawid, A. P. (1979). Discussion of the paper by G. E. P. Box (1979b) read at the International Meeting on Bayesian Statistics, May 28-June 2, 1979 at Valencia, Spain.

DeGroot, M. H. (1970). Optimal Statistical Decisions. New York, McGraw-Hill.

Good, I. J. (1965). The Estimation of Probabilities. Cambridge, M.I.T. Press.

Kendall, M. G. and Stuart, A. (1969). The Advanced Theory of Statistics. Volume 1: Distribution Theory. (3rd edition.) New York, Hafner.

Lindley, D. V. (1965). Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2, Inference. London, Cambridge University Press.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2084	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Some Aspects of Model Estimation and Model Criticism		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Steven P. Bailey George E. P. Box		8. CONTRACT OR GRANT NUMBER(s) DAAG29-75-C-0024 DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE May 1980 13. NUMBER OF PAGES 47
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Model building, Bayesian inference, sampling theory inference, diagnostic checks, predictive distribution, vague prior knowledge

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The recently advanced philosophy of model building is developed further. It is stressed how Bayesian inferences based on the posterior distribution of the model parameters are appropriate only after sampling theory inferences based on the predictive distribution of the data fail to discredit the model. An example involving the normal distribution is discussed in detail. Diagnostic checking functions are developed which can be applied in an intuitive sequential manner. Careful attention is also given to the nature of the predictive

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT (Cont'd.)

distribution for the extreme situation where information about the parameters is very precise or very vague. For the latter case, it is illustrated how the predictive distribution can simultaneously (i) reflect this vague information in an appropriate manner and (ii) allow for the checking of the adequacy of the basic distributional assumptions such as normality and independence.

A particular problem in the interpretation of predictive distributions arises in situations involving a discrete data-generating distribution with vague prior knowledge about the parameter(s). This problem is explored in depth for the case of the binomial distribution.