Some Uses of Order Statistics in Bayesian Analysis

by

Seymour Geisser
University of Minnesota

Technical Report No. 604
February 1995

*Research supported in part by NIGMS 25271

Some Uses of Order Statistics in Bayesian Analysis

Seymour Geisser
University of Minnesota

1 Introduction

Order statistics, although playing a prominent role in frequentist methodology, especially in nonparametric inference, are not often featured in Bayesian analysis. One area, however, where order statistics can be of interest to Bayesians is in the detection of outliers or discordant observations. These situations are such that there is an observation that appears to be somewhat removed from the remaining ones but no discernible alternative can readily be specified for the potential discordancy. Alternatives, although sometimes considered, Dixit (1994), require additional distributional assumptions and prior probabilities over and above the original model assumptions that initially an investigator may not be prepared to contemplate. In these situations a simple Bayesian test of significance may be appropriate in determining whether an observation or several of them are discordant or if it is necessary to contemplate alternative model assumptions for the entire data set. Sections 2 through 6 will consider Bayesian discordancy testing. Another area involves situations where the calculation of the probability that the minimum (maximum) of a set of future observables is greater (smaller) than some critical threshold. More generally we shall be interested in the chance that R out of M future values are in some interval or some set. This essentially involves the Rth future order statistic. This will be the subject of sections 7 and 8.

2 Discordancy Testing

The notion of a Bayesian significance test was introduced by Box (1980) for goodness-of-fit problems. This view was adopted for discordancy testing by Geisser (1980, 1989, 1990). In this paper we delineate some approaches for the use of Bayesian significance testing to the detection of potentially discordant observations. These tests can be useful in situations where no distributional alternative is readily contemplated or easily modelled. These cases

presumably may have arisen from errors in transcribing data, numbers misread, digits transposed, an incorrect sign before a number, a stipulated experimental condition that did not obtain or any of a host of possibilities that would serve to flaw one or more observations in an experiment. Therefore the surprise spawned by a small P-value of an appropriate significance test is useful in detecting potentially flawed observations. If the "discordant" observation(s) make(s) an appreciable difference in a potential inference or decision then a determination needs to be made as to whether the apparently discordant observables are really incompatible with the rest of the observations or whether the modeling requires revision or both. In univariate situations these discordancies often take the form of outliers in that one or several observations appear to be distantly removed from the others.

We shall present a framework for such discordancy tests that depend on (a) the identification of a potentially discordant observation because of the intrusion of some untoward event connected with the particular observation (b) taking into account a diagnostic ransacking of the data in search of potentially discordant observations. Approaches are discussed that depend on the differing circumstances in identifying the suspect observations.

3 Suspicious Circumstances

Assume $Y_1, \ldots, Y_N$ are independently distributed with known covariates $x_1, \ldots, x_N$ such that the distribution function of Y_i is specified as $F(y_i|x_i, \theta)$. In addition, we assume a prior density $p(\theta)$ for θ. Hence based on this model we can compute the predictive distribution of a future value or set of such values $z^{(n)} = (z_1, \ldots, z_n)$

$$F\left(z^{(n)}|y^{(N)}\right) = EF\left(y^{(N)}|x^{(N)}, \theta\right)$$

where $y^{(N)} = (y_1, \ldots, y_N)$ is the observed set of values of $Y^{(N)} = (Y_1, \ldots, Y_N)$ and $x^{(N)} = (x_1, \ldots, x_N)$. In the process of sampling a particular value Y_i say some untoward event or suspicious circumstance occurred that may have affected the observable y_i. A determination can be made as to whether the observation was discordant and if so its effect on the inference or decision. Discordancy can be assessed using the predictive distribution of the particular observable $Y_i = y_i$ given the rest of the sample $y_{(i)}$ which denotes all of the observations in $y^{(N)}$

with y_i deleted. If it is determined that y_i is likely to have been flawed then a comparison of either posterior distributions of particular parameters of interest or the predictive distribution of future observations with and without the potentially flawed observable is in order.

Since we are dealing essentially with "potentiality" it is clear that there is little concern with observables well within the ambit of those for which no doubt is manifest. Hence we would restrict our attention to those that appear extreme in some sense. When dealing with independent and identically univariate observables, our attention is directed to those that are possibly extreme, i.e. the largest and the smallest observables. For an extreme single value for which an untoward event occurred, we can construct a significance test by calculating the predictive probability

$$P_M = \Pr\left[Z_M \geq y_M | y_{(M)}\right] \tag{3.1}$$

for the largest observation and

$$P_m = \Pr\left[Z_m \leq y_m | y_{(m)}\right], \tag{3.2}$$

where $y_{(i)}$ is $y^{(n)}$ with y_i deleted. If both are to be tested simultaneously, then

$$P_{m,M} = \Pr\left[Z_m \leq y_m, Z_M \geq y_M | y_{(m,M)}\right] \tag{3.3}$$

when $y_{(i,j)}$ is $y^{(N)}$ with y_i and y_j deleted.

As long as the largest or smallest y_M and y_m were tagged because of a prior potential problem there is no need to concern ourselves with the distribution of order statistics. We have restricted ourselves to extreme points because non-extreme points are not likely to be of concern, unless values more extreme are also of concern because of their removal from the bulk of observations.

More generally for non-identically distributed variables (usually because of known covariates) we would calculate

$$P_i = P[Z_i \in R | y_{(i)}, x^{(N)}] \tag{3.4}$$

where R is a region indicated for discordancy by some diagnostic procedure.

4 Examples

As an example we consider the exponential distribution. Let $Y_1, \ldots, Y_N$ be a random sample from

$$f(y|\alpha,\gamma) = \alpha e^{-\alpha(y-\gamma)}, \quad y > \gamma, \ \alpha > 0.$$

Let $y_1, \ldots, y_d$ represent fully observed values, and $Y_{d+1}, \ldots, Y_N$ be censored at $y_{d+1}, \ldots, y_N$ respectively. Further let $m = \min(y_1, \ldots, y_d)$, and for reasons previously discussed (Geisser 1984), assume that $m < \min(y_{d+1}, \ldots, y_N)$. Let the conjugate prior density be $p(\gamma,\alpha) = p(\gamma|\alpha)p(\alpha)$ where

$$p(\gamma|\alpha) = N_0\alpha e^{\alpha N_0(\gamma - m_0)}, \quad \gamma < m_0,$$

and

$$p(\alpha) \propto \alpha^{d_0-2}e^{-\alpha N_0(\overline{y}_0 - m_0)} \quad \alpha > 0, \ \overline{y}_0 > m_0,$$

where $1 < d_0 \leq N_0$. Then the posterior densities are

$$p\left(\gamma|\alpha, y^{(N)}\right) \propto e^{\alpha N^*(\gamma - m^*)}, \qquad \gamma < m^*,$$

$$p\left(\alpha|y^{(N)}\right) \propto \alpha^{d^*-2}e^{-\alpha N^*(\overline{y}^* - m^*)}, \quad \overline{y}^* > m^*, \ \alpha > 0,$$

for

$$1 < d^* \leq N^*, \quad d^* = d_0 + d, \quad N^* = N_0 + N, \quad m^* = \min(m_0, m), \tag{4.1}$$

$$\overline{y}^* = (N_0 + N)^{-1}(N_0\overline{y}_0 + N\overline{y}), \text{ and } N\overline{y}_i = \sum_1^N y_i.$$

The predictive distribution of a future observable Z is

$$F(z) = \begin{cases} \frac{1}{N^*+1}\left(\frac{\overline{y}^*-m^*}{\overline{y}^*-z}\right)^{d^*+1}, & z \leq m^*, \\ 1 - \frac{N^{*d}(\overline{y}^*-m)^{d-1}}{(N^*+1)\{z-m^*+N(\overline{y}^*-m^*)\}^{d^*-1}}, & z > m^*. \end{cases} \tag{4.2}$$

Note that for the noninformative prior $p(\gamma,\alpha) \propto \alpha^{-1}$,

$$m^* \to m, \quad \overline{y}^* \to \overline{y}, \quad d^* \to d, \quad N^* \to N.$$

In what follows we shall remove the stars although it is clear that when the proper prior is available we can replace the unstarred values with starred values.

Clearly, then, for (3.1) we would calculate a P-value for the largest,

$$P_M = \frac{(N-1)^c(\overline{y}_{(M)} - m)^{c-1}}{N(M - m + (N-1)(\overline{y}_{(M)} - m))^{c-1}} \tag{4.3}$$

where

$$\begin{aligned} c &= d-1 \quad \text{if } y_M \text{ is fully observed} \\ c &= d \quad \text{if } y_M \text{ is censored} \end{aligned}$$

and for the smallest,

$$P_m = 1 - \frac{1}{N}\left(\frac{\overline{y}_{(m)} - m_2}{\overline{y}_{(m)} - m}\right)^{d-2} \tag{4.4}$$

where m_2 is smaller than the censored observations when m is excluded.

For combinations of the largest and smallest we first calculate (bereft of stars) the predictive probability of a pair of future observables Z_1 and Z_2

$$\begin{aligned} \Pr[Z_1 \le z_1, Z_2 \le z_2 | y^{(N)}] &= \left(\frac{\overline{y} - m}{\overline{y} - v}\right)^{d-1} + \frac{N}{N+2}\left(\frac{N(\overline{y} - m)}{N\overline{y} + z_1 + z_2 - (N+2)v}\right)^{d-1} \\ &\quad - \frac{N}{N+1}\left(\frac{N(\overline{y} - m)}{N(\overline{y} - v) + z_2 - v}\right)^{d-1} \\ &\quad - \frac{N}{N+1}\left(\frac{N(\overline{y} - m)}{N(\overline{y} - v) + z_1 - v}\right)^{d-1} \end{aligned} \tag{4.5}$$

for $v = \min(z_1, z_2)$ and $\max(z_1, z_2) \le m$,

$$\Pr[Z_1 \le z_1, Z_2 > z_2 | y^{(N)}] = \frac{N}{(N+1)(N+2)}\left(\frac{N(\overline{y} - m)}{N(\overline{y} - z_1) + z_2 - z_1}\right)^{d-1} \tag{4.6}$$

for $z_1 \le m \le z_2$, and

$$\Pr[Z_1 > z_1, Z_2 > z_2 | y^{(N)}] = \frac{N}{(N+2)}\left(\frac{N(\overline{y} - m)}{N(\overline{y} - m) + z_1 + z_2 - 2m}\right)^{d-1} \tag{4.7}$$

for $\min(z_1, z_2) \ge m$.

Hence, for the two smallest we can calculate P_{m,m_2} by substituting in (4.5) $N-2$, $\overline{y}_{(m,m_2)}$, $d-2$ and m_3 for N, $\overline{y}$, d and m. For z_1 and z_2 we use m and m_2.

For the smallest and the largest, $P_{m,M}$ can be calculated from (4.6) with $N-2$, $\overline{y}_{(m,M)}$, m_2, c substituted for N, $\overline{y}$, m, d where

$$c = \begin{cases} d-1 & \text{if } M \text{ is censored} \\ d-2 & \text{if } M \text{ is uncensored,} \end{cases}$$

and substitute m for z_1 and M for z_2.

For the two largest, say M and M_2, we can use (4.7) with $N-2$, $\overline{y}_{(M,M_2)}$ and c substituted for N, $\overline{y}$, d where

$$c = \begin{cases} d-2 & \text{if neither of } M \text{ or } M_2 \text{ are censored} \\ d-1 & \text{if only one is uncensored} \\ d & \text{if both are censored,} \end{cases}$$

and set $z_1 = M$, $z_2 = M_2$.

For $Y_1, \ldots, Y_N$ independent in normal linear regression the setup is as follows:

$$Y = X\beta + U$$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} & \ldots, & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \ldots, & x_{Np} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

where X is known, β is unknown, and

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_N \end{pmatrix} \sim N(0, \sigma^2 I_N).$$

Hence,

$$f_Y(y) = \frac{e^{-(1/2\sigma^2)(y-X\beta)'(y-X\beta)}}{(2\pi)^{N/2}\sigma^N} \tag{4.8}$$

where y is the realization of the vector Y.

Although there is no more difficulty with the usual normal-gamma prior for (β, σ^2), we shall illustrate this with the noninformative prior

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2},$$

so that

$$p(\beta, \sigma^2|y) \propto \frac{1}{(\sigma^2)^{(N/2)+1}} e^{-(1/2\sigma^2)(y-X\beta)'(y-X\beta)}.$$

Suppose we are interested in predicting a set of M new variates

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_M \end{pmatrix}$$

at known design matrix

$$W = \begin{pmatrix} w_{11} & , \ldots , & w_{1p} \\ \vdots & & \vdots \\ w_{M1} & , \ldots , & w_{Mp} \end{pmatrix}$$

such that

$$Z = W\beta + U$$

where

$$U \sim N(0, \sigma^2 I_M).$$

Then for the future set Z

$$f(z|y) = \int f(z|\sigma^2, \beta, W)\, p(\sigma^2, \beta|y, X)\, d\sigma^2 d\beta.$$

Let

$$A = I + W(X'X)^{-1}W'$$

$$(N-p)s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})$$

then we can calculate

$$\begin{aligned} f(z|y) &= \frac{\Gamma((N+M-p)/2)}{\pi^{M/2}\Gamma((N-p)/2)(N-p)^{M/2}|s^2 A|^{1/2}} \\ &\quad \times \left[1 + \frac{(z - W\hat{\beta})'A^{-1}(z - W\hat{\beta})}{(N-p)s^2}\right]^{-(N+M-p)/2} \end{aligned} \tag{4.9}$$

an M-dimensional student density. Further, it is easy to show that the predictive distribution of

$$V = \frac{(Z - W\hat{\beta})'A^{-1}(Z - W\hat{\beta})}{(N-p)s^2} = F_{M,N-p}. \tag{4.10}$$

Hence, to test whether a tagged subset $y^{(n)}$ of $y^{(N)}$ is an outlier group (usually n will be 1 or 2), we can calculate

$$P^{(n)} = \Pr\left[F_{n,N-n-p} \geq \frac{\left(y^{(n)} - X_{N-n}\hat{\beta}_{N-n}\right)' A_{N-n}^{-1} \left(y^{(n)} - X_{N-n}\hat{\beta}_{N-n}\right)}{(N-n-p)s_{N-n}^2}\right] \tag{4.11}$$

where the subscript $N-n$ indicates that the values s_{N-n}^2, $\hat{\beta}_{N-n}$, A_{N-n}, X_{N-n} are based on the undeleted $N-n$ observations and $F_{n,N-n-p}$ is an F variate with n and $N-n-p$ degrees

of freedom. For the special case where $p = 1$ and $x_{i1} = 1$ for $i = 1, \ldots, N$ and $n = 1$ we have

$$\frac{(y_i - \overline{y}_{(i)})\sqrt{N-1}}{s_{(i)}\sqrt{N}} \sim t_{N-2} \tag{4.12}$$

a "Student" with $N-2$ degrees of freedom where $(N-2)s^2_{(i)} = \sum_{j \neq i}(y_j - \overline{y})^2$. If a direction (too large or too small) is considered an outcome of the untoward event than one-sided P_i can be computed. If the untoward event does not imply a direction then the two-sided significance value is appropriate. These methods can easily be extended for a normal-gamma conjugate prior.

For Poisson regression we assume

$$\Pr[Y = y|x, \theta] = \frac{e^{-x\theta}(\theta x)^y}{y!} \quad y = 0, 1, \ldots$$

and obtain independent values $Y_1, \ldots, Y_N$ with known covariates $x_1, \ldots, x_N$. Again the conjugate prior is a gamma but we shall illustrate with the conventional improper prior $p(\theta) \propto \frac{1}{\theta}$. A future value Y_{N+1} has predictive probability function

$$\Pr\left[Y_{N+1} = z|x_{N+1}, y^{(N)}, x^{(N)}\right] = \binom{t+z-1}{t-1}\left(\frac{x_{N+1}}{u+x_{N+1}}\right)^z\left(\frac{u}{u+x_{N+1}}\right)^t \tag{4.13}$$

$z = 0, 1, \ldots, t = \sum_1^N y_i \geq 1$, and $u = \sum_1^N x_i$. Hence if y_i were tagged we would calculate the significance level of y_i by replacing (z, t, x_{N+1}, u) with $(y_i, t_{(i)}, x_i, u_{(i)})$ where $t_{(i)}$ and $u_{(i)}$ are computed with y_i and x_i deleted. Call this q_i and if this is larger than the prescribed P_i for rejection then one can stop, otherwise if this is smaller than the prescribed value, then one could calculate tails by probability ordering for the values of z.

5 Ransacked Data

Here the situation is such that at the time the data were generated, no known untoward event occurred to influence the observables. However the data were ransacked, whether graphically or numerically, to determine whether the set of observations are consonant with the assumed model.

We shall present methods that are appropriate in these circumstances. In univariate situations where the values are generated by an i.i.d. process, potentially discordant values are generally extremes, i.e. particularly large or small values.

For the translated exponential, if the largest is chosen by ransacking, then we first calculate $F_M(u|\theta)$ the distribution of the largest M conditional on $\theta = (\alpha, \gamma)$. Then, for $p(\theta)$ the proper gamma prior of section 4,

$$P_M = 1 - \int F_M(u|\theta)\, p(\theta)\, d\theta$$

so that

$$P_M = \sum_{j=1}^{N} \binom{N}{j} (-1)^{j-1} \frac{N_0}{N_0 + j} \left(\frac{N_0(\overline{y}_0 - m_0)}{N_0(\overline{y} - m_0) + (M - m_0)j} \right)^{d_0 - 1}. \tag{5.1}$$

Since this represents the probability that the maximum is at least as large as its observed value, the result is appropriate for the maximum observation whether fully observed or censored.

Similarly, for the smallest observation m, we obtain

$$P_m = \begin{cases} 1 - \left(\frac{N_0}{N_0+N}\right) \frac{\{N_0(\overline{y}_0 - m_0)\}^{d_0-1}}{\{N(m-m_0)+N_0(y_0-m_0)\}^{d_0-1}}, & m \geq m_0, \\ \frac{N}{N_0+N} \left(\frac{\overline{y}_0 - m_0}{\overline{y}_0 - m}\right)^{d_0-1}, & m \leq m_0. \end{cases} \tag{5.2}$$

Of course, the above tests exist only for a proper prior distribution and that this prior will have a rather considerable effect on the significance test. Unless these prior parameters $\overline{y}_0$, m_0, N_0, d_0 can be specified or perhaps approximated with some precision this may be impractical in many situations.

In contrast the frequency approach attempts to obtain a statistic that reflects in some way whether the maximum (or minimum) is discrepant and has a sampling distribution independent of the parameters. For example in the non-censored case $d = N$, a statistic used is

$$T = \frac{M - M_2}{M - m} \tag{5.3}$$

where as before M, M_2, m are largest, second largest and smallest values, Dixon (1951), Likes (1966), Kabe (1970). Here it is easily shown that

$$\Pr[T \geq t] = 1 - (N-1)(N-2)B\left(\frac{2-t}{1-t}, N-2\right) \tag{5.4}$$

where $B(u, v)$ is the beta function. Similarly for the smallest a statistic used is

$$T = \frac{m_2 - m}{M - m}$$

whose sampling distribution yields for a P-value,

$$\Pr[T \leq t] = (N-2)B\left(\frac{1+(n-2)t}{1-t}, n-2\right). \tag{5.5}$$

When $d < N$, then straightforward frequentist solutions are not available.

In general then assume some diagnostic, say D, is used in ransacking to order the y_i. Then the transformation $D(Y_i) = D_i$ yields random variables $D_1, D_2, \ldots, D_N$. Hence we need to find the distribution of D_M the transform which yields the observed y which is most discrepant, namely

$$F_{D_M}(d|\theta),$$

the conditional distribution of D_M associated with the most removed Y_i given θ. Then

$$P_M = 1 - \int F_{D_M}(d|\theta)\, p(\theta)\, d\theta \tag{5.6}$$

where $p(\theta)$ is a proper prior. Tests of this sort were termed Unconditional Predictive Discordancy (UPD), Geisser (1989). They allow the prior to play a substantial role in determining the outlier status of an observation. One can continue by finding the joint distribution of the ordered D_i's given θ and test whether jointly some most discrepant subset in terms of the diagnostic's discrepancy ordering is a discordant subset.

For a simple normal case we assume Y_i, $i = 1, \ldots, N$ are i.i.d. $N(\mu, \sigma^2)$ with σ^2 known and $\mu \sim N(\theta, \tau^2)$. Now the unconditionally Y_i, $i = 1, \ldots, N$ are an exchangeable set of normal variates with mean θ, variance $\sigma^2 + \tau^2$ and covariance τ^2. This might imply that $V_i = \frac{(Y_i - \theta)^2}{\sigma^2 + \tau^2}$ is the appropriate diagnostic with the $\max_i V_i = V_0$ being used to construct the significance test for the largest deviation, namely

$$P_M = \Pr[V_0 \geq v]. \tag{5.7}$$

It is clear that $V_1, \ldots, V_N$ are all exchangeable and marginally distributed as χ^2 with one degree of freedom. Although the distribution of V_0 is not analytically explicit, P_M can be calculated by either numerical approximation or Monte Carlo techniques, see also tables by Pearson and Hartley (1966). However, this is not the critical issue. The question is whether V_i is an appropriate discrepancy measure because V_i only reflects distance from the prior mean and this could cause some discomfort as it need not reflect sufficient distance from

the rest of the observations. The latter is often the implied definition of an outlier or a discordant observation. One could also use

$$\max_i \frac{(Y_i - \overline{Y})^2 N}{(\sigma^2 + \tau^2)(N+1)} = \max Z_i = Z_0 \tag{5.8}$$

again a joint distribution of exchangeable normal random variables, each marginally χ^2 with one degree of freedom, and though slightly more complex, it is still calculable. Again, this is essentially the frequentist approach for $\tau^2 = 0$ which in essence is independent of the prior assumptions. Perhaps this goes too far in the other direction, i.e. disregarding the prior information. Some compromise may be needed and the one that suggests itself is

$$W_i = \left(Y_i - \frac{N\tau^2\overline{Y} + \theta\sigma^2}{N\tau^2 + \sigma^2}\right)^2 \tag{5.9}$$

where the deviation is from the posterior mean—an appropriate convex combination of sample mean and prior mean. Although unconditionally $W_1, \ldots, W_N$ are still exchangeable the marginal distribution of W_i is essentially proportional to a non-central χ^2, thus complicating matters still further for $W_0 = \max W_i$. However, deviations such as W_i seem more sensible in that both prior and likelihood play a part in contrast to only either prior or likelihood. Further distributional complications ensue when the usual conjugate gamma prior is assumed for σ^{-2}. In addition, the two hyperparameters of the gamma prior also must be known.

Extension to multiple linear regression with normally distributed errors, though clear for all 3 approaches, involves further unknown hyperparameters.

For Poisson regression we also would require a discordancy ordering perhaps based on the least probable Y_i as the potential outlier. As this becomes quite complicated we shall merely illustrate for i.i.d. Poisson variates with a gamma prior, for θ

$$p(\theta|\gamma, \delta) = \frac{\gamma^\delta \theta^{\delta-1} e^{-\gamma\theta}}{\Gamma(\delta)}.$$

If the maximum Y_i has the smallest probability we let $Z = \max_i Y_i$, assuming this is the potential outlier. Then

$$\begin{aligned}
\Pr[Z \geq z|\theta] &= 1 - \left[\sum_{y=0}^{z-1} \frac{e^{-\theta}\theta^y}{y!}\right]^N \\
\Pr[Z \geq z] &= 1 - \int_0^\infty \left[\sum_{y=0}^{z-1} \frac{e^{-\theta}\theta^y}{y!}\right]^N p(\theta)\, d\theta \qquad (5.10) \\
&= 1 - \frac{\gamma^\delta}{\Gamma(\delta)} \int_0^\infty e^{-(N+\gamma)\theta}\theta^{\delta-1} \left(1 + \theta + \frac{\theta^2}{2} + \cdots + \frac{\theta^{z-1}}{(z-1)!}\right)^N d\theta.
\end{aligned}$$

Clearly one can write the multinomial expansion for the term raised to the Nth power in the integrand and integrate termwise and obtain a complex but finite and explicit solution involving gamma functions. If $\min Y_i = W$ has the smallest probability, then

$$\begin{aligned}
\Pr[W \leq w|\theta] &= 1 - \left[1 - \sum_{y=0}^{w} \frac{e^{-\theta}\theta^y}{y!}\right]^N \\
\Pr[W \leq w] &= 1 - \int p(\theta|\gamma,\delta) \left[1 - \sum_{y=0}^{w} \frac{e^{-\theta}\theta^y}{y!}\right]^N d\theta. \qquad (5.11)
\end{aligned}$$

Again this is complex but explicitly computable in terms of a finite series involving gamma functions.

Although simple analytic expressions, except when dealing with the exponential distribution, are rare, Monte Carlo methods are generally available to handle such situations. However, the major difficulty is of course the assignment of the proper prior distribution and the ensuing set of hyperparameters. Because of these difficulties we shall present another way of handling these situations which can be used with proper or improper priors.

6 Conditional Predictive Discordancy (CPD) Tests

We shall now present a method which (a) turns out to be much easier to calculate, (b) can be used for the usual improper priors, (c) depends on a proper prior and its hyperparameters when a proper prior is used, (d) is seamless in its transition from a proper prior to an improper prior and to censoring, and (e) in certain instances when an improper prior is used it will yield a result identical to a frequency significance test.

The idea is relatively simple if $D(Y_i)$ represents the scalar diagnostic which characterizes the discrepancy of the observation from the model and orders the observables from most to least discrepant $D_1, D_2, \ldots, D_N$, then a significance test

$$P - \Pr[D_1 > d_1 | D_1 > d_2, d_{(1)}] \tag{6.1}$$

where $d_{(1)}$ refers to $d^{(N)}$ with d_1 deleted. Here we assume only D_1 is random and conditioning is on $D_{(1)}$, i.e. all but the largest value.

Alternatively, we could consider conditioning $D_{(1,2)}$, i.e. all but the largest and second largest discrepant values which would result in

$$P = \Pr[D_1 > d_1 | D_2 > d_2, d_{(1,2)}]. \tag{6.2}$$

As an example we consider the exponential case of section 4. For testing the largest for discordancy using (6.1) we obtain

$$\begin{aligned} P_M &= \Pr[Z \geq M | Z > M_2, y_{(M)}] = \left[\frac{N(\overline{y} - m) - (M - M_2)}{N(\overline{y} - m)}\right]^c \\ P_M &= (1 - t)^c \end{aligned} \tag{6.3}$$

where

$$\begin{aligned} t &= \frac{M - M_2}{N(\overline{y} - m)} \\ c &= d - 1 \quad \text{if } M \text{ were censored} \\ c &= d - 2 \quad \text{if } M \text{ were uncensored} \end{aligned}$$

when the non-informative prior is used. For the conjugate prior we need only to affix stars to $\overline{y}$, m, d, and N, using the previous definitions of (4.1).

Using (6.2) we obtain

$$\begin{aligned} P'_M &= P[Z_1 \geq M | Z_2 > M_2, y_{(M,M_2)}] \\ &= \frac{P[Z_1 \geq M, Z_2 \geq M_2 | y_{(M,M_2)}]}{P[Z_2 \geq M_2 | y_{(M,M_2)}]} \\ &= \frac{N - 1}{N}\left(\frac{N(\overline{y} - m) - (M - m)}{N(\overline{y} - m)}\right)^c \\ &= \frac{N - 1}{N}(1 - t)^c \end{aligned} \tag{6.4}$$

where

$$t = \frac{M - m}{N(\overline{y} - m)}$$

and

$$c = \begin{cases} d-1 & \text{if } M \text{ and } M_2 \text{ are censored} \\ d-2 & \text{if one of } M \text{ or } M_2 \text{ is censored} \\ d-3 & \text{if } M \text{ and } M_2 \text{ are uncensored.} \end{cases}$$

We know that if $d = N$, the uncensored case, the sampling distribution of the statistic

$$T = \frac{M - M_2}{N(\overline{y} - m)} \tag{6.5}$$

which can be used to test for the largest being an outlier is such that

$$\Pr[T \geq t] = (1-t)^{N-1} = P_M$$

i.e. the same value as (6.3). Hence we have a seamless transition from proper prior with censoring to the usual non-informative prior without censoring yielding the sampling distribution statistic.

The second method illustrated by (6.4) does not provide a frequentist analogue for the sampling distribution of $T = \frac{M-m}{N(\overline{y}-m)}$ and this cannot be reconciled with (6.4).

For the smallest observation we obtain, basically using (6.1),

$$P_m = \Pr[Z \leq m | Z \leq m_2, y_{(m_2)}]$$

where

$$P_m = \begin{cases} A(m)/A(m_2), & m_0 \leq m, \\ B(m)/A(m_2), & m \leq m_0 \leq m_2, \\ B(m)/B(m_2), & m \leq m_2 \leq m_0, \end{cases}$$

where

$$\begin{aligned} A(z) &= 1 - \frac{N^* - 1}{N^*} \left(\frac{(N^*-1)(\overline{y}_{(m)} - m_0)}{(N^*-1)(\overline{y}^*_{(m)} - m_0) + z - m_0} \right)^{d^*-2}, \\ B(z) &= \frac{1}{N^*} \left(\frac{\overline{y}^*_{(m)} - m_0}{\overline{y}^*_{(m)} - z} \right)^{d^*-2}. \end{aligned}$$

The noninformative prior, however, yields the simple form

$$P_m = \left(\frac{\overline{y}_{(m)} - m_2}{\overline{y}_{(m)} - m}\right)^{d-2} = (1-(N-1)t)^{d-2},$$

where

$$t = \frac{m_2 - m}{N(\overline{y} - m)}.$$

For $D = N$ in the uncensored case, the sampling distribution of T yields

$$\Pr[T \geq t] = P_m$$

so again we have a seamless transition, Geisser (1989).

For normal linear regression, a CPD test for the ransacked potential outlier we suggest using as a diagnostic

$$V_i = \frac{(Y_i - x_i'\hat{\beta}_{(i)})' A_{(i)}^{-1} (Y_i - x_i'\hat{\beta}_{(i)})}{(N-1-p)s_{(i)}^2}$$

where the notation (i) refers to the entire set of $y = (y_1, \ldots, y_N)$ with y_i and associated $x_i' = (x_{i1}, \ldots, x_{ip})$ deleted.

Once the V_i's are ordered into the largest U_c and the second largest U_{c-1} with the values y_c and y_{c-1} corresponding to U_c and U_{c-1}, then we can compute as the significance level

$$P_c = \Pr[U_c > u_c | U_c > u_{c-1}, y_{(c)}]$$

where u_c and u_{c-1} are the realized values of U_c and U_{c-1}. We suggest that the significance computation be made as follows:

$$P_c = \frac{\Pr[U > u_c]}{\Pr[U > u_{c-1}]} \tag{6.6}$$

where U is distributed as an F-variate with 1 and $N-1-p$ degrees of freedom.

Similarly for Poisson regression we can order (3.1) using

$$\begin{aligned} p_i &= \Pr[Y_i = y_i | x_i, y_{(i)}, x_{(i)}] \\ &= \binom{t_{(i)} + y_i - 1}{t_{(i)} - 1} \left(\frac{x_i}{u_{(i)} + x_i}\right)^{y_i} \left(\frac{u_{(i)}}{u_{(i)} + x_i}\right)^{t_{(i)}} \end{aligned} \tag{6.7}$$

where $t_{(i)} = \sum_{j \neq i} y_j$, $u_{(i)} = \sum_{j \neq i} x_j$ and p_c and p_{c-1} are the smallest and second smallest probabilities corresponding to say y_c and y_{c-1}. At this point one could use as significance level the simple computation

$$P_c = \frac{p_c}{p_{c-1} + p_c}.$$

7 Ordering Future Values

In certain problems where there is data regarding an observable such as the yearly high water mark on the banks of a river or a dam there is an interest in calculating the probability that no flood will occur in the next M years, i.e. the maximum high water mark does not exceed a given value. Conversely, another situation is the use of possibly harmful drugs serving a limited number of patients given to alleviate very serious but rather rare diseases. Here the drug may be lethal or severely damage some important bodily function if some measured physiological variable falls below (or above) some defined value. A more mundane situation is where a buyer of say M light bulbs, whether connected in series or not, wants to calculate the probability of no failures in a given time based on previous information of bulb lifetimes. In the last two situations we require the chance that the minimum physiological value (or failure time) exceeds a certain threshold.

In the first case we are interested in calculating the maximum Z of future values say $Y_{N+1}, \ldots, Y_{N+M}$ not exceeding a given value z, i.e. the distribution function of the maximum Z,

$$\Pr[Z \leq z|y^{(N)}] = \int \Pr[Z \leq z|\theta, y^{(N)}] p(\theta|y^{(N)})\, d\theta \tag{7.1}$$

where

$$p(\theta|y^{(N)}) \propto f(y^{(N)}|\theta)\, p(\theta).$$

In the second case we are interested in calculating $1 - F_W(w)$ where W is the minimum of the future set $Y_{N+1}, \ldots, Y_{N+M}$ or

$$\Pr[W > w|y^{(N)}] = \int \Pr[W > w|y^{(N)}, \theta] p(\theta|y^{(N)})\, d\theta. \tag{7.2}$$

For the exponential case we can obtain explicit results for the previously discussed exponential sampling with a gamma prior. Here we obtain the predictive probability that the maximum Z will not exceed z, to be

$$P[Z \leq z|y^{(N)}] = \begin{cases} \sum_{j=0}^{M} \binom{M}{j} (-1)^j \frac{N^*}{N^*+j} \left[\frac{N^*(\overline{y}^*-m^*)}{N^*(\overline{y}^*-m^*)+j(z-m^*)}\right]^{d^*-1} & \text{for } z > m^* \\ \sum_{j=0}^{M} \binom{M}{j} (-1)^j \frac{N^*}{N^*+j} \left[\frac{N^*(\overline{y}^*-m^*)}{N^*(\overline{y}^*-m^*)+N^*(m^*-z)}\right]^{d^*-1} & \text{for } z \leq m^*. \end{cases} \tag{7.3}$$

For the problem where the minimum W should exceed a value w we obtain

$$\Pr[W > w|y^{(N)}] = \begin{cases} \frac{N^*}{N^*+M}\left[\frac{N^*(\overline{y}^*-m^*)}{N^*(\overline{y}^*-m^*)+M(w-m^*)}\right]^{d^*-1} & \text{for } w \geq m^* \\ 1-\frac{M}{N^*+M}\left[\frac{\overline{y}^*-m^*}{\overline{y}^*-w}\right]^{d^*-1} & \text{for } w < m^*, \end{cases} \tag{7.4}$$

c.f. Dunsmore (1974).

Sometimes the situation is such that we are interested more generally in the chance that at least the rth largest will not exceed a given value. We first obtain the probability that exactly r out of M will not exceed the threshold w, Geisser (1984). Let

$$\begin{aligned} V_i &= 1 \quad \text{if } Y_{N+i} \leq w \quad i = 1, \ldots, M \\ &= 0, \qquad \text{otherwise} \end{aligned}$$

and set $R = \sum_1^N V_i$. Then after some algebra we obtain

$$\begin{aligned} \Pr[R = r|y^{(N)}] &= \left(\frac{\overline{y}^*-m^*}{\overline{y}^*-w}\right)^{d^*+1}\binom{N^*+M-r-1}{M-r}\Big/\binom{N^*+M}{M} \\ &\quad \text{if } r > 0, \quad w < m^*, \\ &= 1-\frac{M}{N^*+M}\left(\frac{\overline{y}^*-m^*}{\overline{y}^*-w}\right)^{d^*-1} \quad \text{if } r = 0, \quad w \leq m^*, \\ &= N^*\binom{M}{r}\sum_{j=0}^{r}\binom{r}{j}\frac{(-1)^j}{(N^*+M-r+j}} \\ &\quad \times\left(1+\frac{(M-r+j)(w-m^*)}{N^*(\overline{y}^*-m^*)}\right)^{-(d^*-1)} \quad \text{if } w > m^*. \end{aligned} \tag{7.5}$$

Thus

$$\begin{aligned} \Pr[R \leq r_0|y^{(N)}] &= 1-\left(\frac{\overline{y}^*-m^*}{\overline{y}^*-w}\right)^{d^*-1}+\left(\frac{\overline{y}^*-m^*}{\overline{y}^*-w}\right)^{d^*-1}\sum_{r=0}^{r_0}\binom{N^*+M-r-1}{M-r}\Big/\binom{N^*+M}{M} \\ &\quad \text{if } w \leq m^* \\ &= N^*\sum_{r=0}^{r_0}\binom{M}{r}\sum_{j=0}^{r}\binom{r}{j}\frac{(-1)^j}{(N^*+M-r+j)}\left(1+\frac{(M-r+j)(w-m^*)}{N^*(\overline{y}^*-m^*)}\right)^{1-d^*} \\ &\quad \text{if } w > m^*, \end{aligned} \tag{7.6}$$

and $1-\Pr[R \leq r_0|y^{(N)}]$ is also the distribution function of the rth order statistic of the future random variables Y_{N+i}, $i = 1, \ldots, M$. For further ramifications on interval estimation of the rth order statistics, see Geisser (1985).

Other sampling distributions cum conjugate priors are generally not amenable to explicit results but numerical approximations or Monte Carlo simulations are often capable of yielding appropriate numerical answers for the most complicated situations.

Now for $Y_1, \ldots, Y_N, Y_{N+1}, \ldots, Y_{N+M}$ i.i.d. $N(\mu, \sigma^2)$,

$$\Pr[Y_{N+i} \leq w | \mu, \sigma^2] = \Phi\left(\frac{w-\mu}{\sigma}\right) = \Theta \tag{7.7}$$

and assuming that $p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$,

$$\begin{aligned} \Pr[\Theta \leq \theta] &= P\left[\Phi\left(\frac{w-\mu}{\sigma}\right) \leq \theta\right] \\ &= P\left[\frac{w-\mu}{\sigma} \leq \Phi^{-1}(\theta)\right] \\ &= P[\beta \leq \Phi^{-1}(\theta)] \end{aligned}$$

where $\frac{w-\mu}{\sigma} = \beta$. For fixed w using the posterior density for μ and σ^2 given $y^{(N)}$ we obtain the density of β to be

$$f(\beta | y^{(N)}) = \frac{\sqrt{N} e^{-N\beta^2/2}}{\sqrt{2\pi}\Gamma\left(\frac{(N-1)}{2}\right)} \sum_{j=0}^{\infty} \frac{(\sqrt{2}Nd\beta)^j \Gamma\left(\frac{N+j-1}{2}\right)}{j!(1+Nd^2)^{\frac{N+j-1}{2}}} \tag{7.8}$$

where $d = (w - \overline{y})/[(N-1)s^2]^{1/2}$ and $\sum_1^N (y_i - \overline{y})^2 = (N-1)s^2$. Thence

$$\Pr[R = r | y^{(N)}] = \binom{M}{r} \int_{-\infty}^{p} \Phi^r(\beta)[1 - \Phi(\beta)]^{M-r} f(\beta | y^{(N)})\, d\beta \tag{7.9}$$

which can be approximated numerically or by simulation techniques, Geisser (1987). From (7.9) one can obtain the distribution function of the rth order statistic among the future set $Y_{N+1}, \ldots, Y_{N+M}$. Although the presentation here is for the number in a semi-infinite interval, it is easily extended to finite intervals and for a set of exchangeable normal variates, Geisser (1987).

8 Multivariate Problems

In situations where $Y_1, \ldots, Y_N, Y_{N+1}, \ldots, Y_{N+M}$ are q-dimensional vector variables such that

$$\begin{aligned} V_i &= 1 \qquad \text{if } Y_{N+i} \in G \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

where G is some specified region, again interest is in $R = \sum_{i=1}^{M} V_i$, the number of future variables in G. In the medical arena, interest would be on the future number of patients who would be administered a therapeutic agent, exchangeable with past patients who had received the agent. Hence if

$$P[Y \in G|\theta] = \beta, \tag{8.1}$$

then symbolically

$$P(R = r|y^{(N)} = \binom{M}{r} \int \beta^r (1-\beta)^{M-r} f(\beta|y^{(N)})\, d\beta. \tag{8.2}$$

In essence, this generalizes the problem of order statistics.

For $Y_i \sim N(\mu, \Sigma)$, and using

$$p(\mu, \Sigma^{-1}) \propto |\Sigma|^{\frac{q+1}{2}} \tag{8.3}$$

suggested by Geisser and Cornfield (1963)

$$P(R = r|y^{(N)}) = \int\int G^r(\mu, \Sigma^{-1})[1 - G(\mu, \Sigma^{-1})]^{M-r} p(\mu, \Sigma^{-1}|y^{(N)})\, d\mu d\Sigma^{-1} \tag{8.4}$$

where

$$P(Y \in G|\mu, \Sigma^{-1}) = \int_G f(y|\mu, \Sigma^{-1})\, dy = G(\mu, \Sigma^{-1}). \tag{8.5}$$

Although in any practical application G will not be an arbitrary region but one in which each component of the vector would be in an interval (finite or semi-finite) and here $G = I_1 \times I_2 \times \cdots \times I_q$, or a hyperrectangle. This obviously precludes any explicit analytical solution for the problem. Although an alternate form for $P(R = r|y^{(N)})$ is available, it is not clear which would be more susceptible to numerical approximation and/or simulation, if either. Here one finds the joint predictive distribution of $Y_{N+1}, \ldots, Y_{N+M}$ directly,

$$\begin{aligned} \Pr[R = r|y^{(N)}] &= \int_{I_1} \cdots \int_{I_q} f(y_{N+1}, \ldots, y_{N+M}|y^{(N)})\, dy_{N+1} \ldots dy_{N+M} \qquad (8.6) \\ &= \int_{I_1} \cdots \int_{I_q} \frac{(2\pi)^{-qM/2} K(q, N-1)|(N-1)S|^{\frac{N-1}{2}} |\Omega|^{q/2} d\underline{y}}{K(q, N-1+M)|(N-1)S + (\underline{y} - \overline{y}e)\Omega(\underline{y} - \overline{y}e)'|^{(N-1+M)/2}} \end{aligned}$$

where

$$K^{-1}(q, \nu) = 2^{\nu q/2} \pi^{q(q-1)/4} \prod_{j=1}^{q} \Gamma\left(\frac{\nu + 1 - j}{2}\right),$$

$$\overline{y} = M^{-1}\sum_{i=1}^{M} y_{N+i},\ (N-1)S = \sum_{j=1}^{N}(y_j - \overline{y})(y_j - \overline{y})'$$

$\Omega = I + ee'$, $e' = (1, \ldots, 1)$ an M-dimensional vector and $\underline{y} = (y_{N+1}, \ldots, y_{N+M})'$, c.f. Geisser (1993, p. 207).

9 Acknowledgement

This work was supported in part by NIGMS 25271.

References

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series B* **143** 383–340.

Dixit, U.J. (1994). Bayesian approach to prediction in the presence of outliers for Weibull distribution. *Metrika* **41** 127–136.

Dixon, W. J. (1994). Ratios involving extreme values. *Annals of Mathematical Statistics* **22** 68–78.

Dunsmore, I.R. (1974). The Bayesian predictive distribution in life testing models. *Technometrics* **3** 455–460.

Geisser, S. (1980). Contribution to Discussion. *Journal of the Royal Statistical Society, Series A* **143** 416–417.

Geisser, S. (1984). Predicting Pareto and exponential observables. *Canadian Journal of Statistics* **12** 143–152.

Geisser, S. (1985). Interval prediction for Pareto and exponential observables. *Journal of Econometrics* **29** 173–185.

Geisser, S. (1987). Some remarks on exchangeable normal variables with applications. *Contributions to the Theory and Applications of Statistics*, Academic Press, 127–153.

Geisser, S. (1989). Predictive discordancy testing for exponential observations. *Canadian Journal of Statistics* **17** (2) 19–26. Correction, (1991) **19** (4) 453.

Geisser, S. (1990). Predictive approaches to discordancy testing. *Bayesian and Likelihood Methods in Statistics and Econometrics*. S. Geisser et al. eds., Amsterdam: North-Holland, 321–335.

Geisser, S. (1993). *Predictive Inference*. Chapman and Hall, New York.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society B* **25** 368–376.

Kabe, D.G. (1970). Testing outliers from an exponential population. *Metrika* **15** 15–18.

Likes, J. (1966). Distribution of Dixon's statistics in the case of an exponential population. *Metrika* **11** 46–54.

Pearson, E.S. and Hartley, H.O. (1966). Biometrika Tables for Statisticians, Volume I, Cambridge.