DISCUSSION

I. GUTTMAN (*University of Toronto*):

Firstly, I would like to congratulate Professor Geisser for his article and presentation -this is a very stimulating piece of work, and I am honoured to be asked to discuss this paper.

Now I have to report that I have gone through several phases since accepting the invitation to be a discussant of this paper. The first phase said -glance through the paper, get the flavour, report on the flavour, and then, like all good discussants, do the *fungible* thing, talk about my work in the area of reliability, censored data, etc. The second phase, however, intruded, because as I read the paper I became more and more stimulated, interested, frustrated, etc. (underline all of these words) by the underlying ideas. For example, the Sample Re-Use Method is *not* Bayesian and the use of words prior and predictive at various points in the paper are somewhat misleading. (The point that Sample Re-Use is not Bayesian is indeed admitted by Geisser). Indeed, if Geisser had given me permission (how presumptuous can one be) to construct a title, I would perhaps have suggested "Smoothing and Approximations to Properties of Predictive Distributions". Allow me, in the ensuing time to say why.

Suppose the population being sampled has distribution $f(\cdot\,|\,\theta)$, and that n independent observations from this population have been taken, say $\mathbf{y}' = (y_1,\ldots,y_n)$. Then, if additionally we are to observe a (future) observation from this population, the conditional distribution h of y, given $\mathbf{y}$, is, using the rules of probability, given by

$$h(y\,|\,\mathbf{y}) = \int_{\theta\in\Omega} f(y\,|\,\theta)p(\theta\,|\,\mathbf{y})d\,\theta \tag{1}$$

where the posterior $p(\theta\,|\,\mathbf{y})$ is such that

$$p(\theta\,|\,\mathbf{y}) = c(\mathbf{y})q(\theta)\ell(\theta\,|\,\mathbf{y}) \tag{2}$$

with

$$\ell(\theta\,|\,\mathbf{y}) = \Pi_{i=1}^{n} f(y_i\,|\,\theta) \text{ and } [c(\mathbf{y})]^{-1} = \int_{\theta\in\Omega} q(\theta)p(\theta\,|\,\mathbf{y})d\,\theta\,. \tag{3}$$

Here, of course, $q(\theta)$ is the prior distribution of θ and summarizes all the information available to the experimenter *prior* to the taking of the data $\mathbf{y}$. Now suppose the prior q itself depends on certain constants α, i.e.,

$$q(\theta) = q(\theta \mid \alpha) = q(\theta \mid \alpha_1, \alpha_2) \tag{4}$$

Then we should write (1) as

$$h(y \mid \mathbf{y}; \alpha) = \int f(y \mid \theta) p(\theta \mid \mathbf{y}; \alpha) d\theta \tag{5}$$

Now Geisser proceeds as follows: Suppose indeed that the distribution function $q(\theta \mid \alpha_1, \alpha_2)$ is the "prior" for θ, *and* that α_1 is known, but α_2 is *unknown*, and that $\mathbf{y}$ is observed. Consider the discrepancy function $D(\mathbf{y}; \alpha_1; \alpha_2)$. Select for α_2 *that value* $\hat{\alpha}_2$ which is such that

$$D(\mathbf{y}; \alpha_1, \hat{\alpha}_2) = \min_{\alpha_2} D(\mathbf{y}; \alpha_1, \alpha_2) \tag{6}$$

Note that

$$\hat{\alpha}_2 = \hat{\alpha}_2(\mathbf{y}; \alpha_1) \tag{7}$$

(For an example of D, see (2.7) of Section 2 of Geisser's paper). Then, use this to construct a function

$$h^{(\cdot)}(y \mid \mathbf{y}) = [c(\mathbf{y})] \int f(y \mid \theta) q(\theta \mid \alpha_1, \hat{\alpha}_2) \ell(\theta \mid \mathbf{y}) d\theta = h^{(\cdot)}(y \mid \mathbf{y}; \alpha_1; \hat{\alpha}_2) \tag{8}$$

and needless to say

$$h^{(s)}(y \mid \mathbf{y}) = \int f(y \mid \theta) p(\theta \mid \mathbf{y}; \alpha_1, \hat{\alpha}_2) d\theta \neq h(y \mid \mathbf{y}), \tag{8a}$$

where

$$h(y \mid \mathbf{y}) = \int f(y \mid \theta) p(\theta \mid \mathbf{y}; \alpha_1, \alpha_2) d\theta \tag{8b}$$

(The superscript (s) in (8a) stands for smoothing $f(y \mid \theta)$ with $p(\theta \mid \mathbf{y}; \alpha_1, \hat{\alpha}_2)$). Now in (8b), the Bayesian and/or his client is using that value of α_1 and α_2 that arises due to prior information about θ, and in general, this choice will be different than $(\alpha_1, \hat{\alpha}_2)$ — in fact, the $\hat{\alpha}_2$'*s* themselves may vary according to the different nature of the choice of D. This point aside, we are now asked to make the step that regards $h^{(s)}$ as an approximation to h, i.e.,

$$h^{(s)}(y \mid \mathbf{y}) \approx h(y \mid \mathbf{y}) \tag{8c}$$

This, it seems to me, must be justified.

Note that we are using a "prior" in (8) which is a function of the data $\mathbf{y}$, a violation of the cannon of Bayesianism which loosely speaking says that if we are to be coherent then the *prior cannot depend on the data.* In fact when we use (8), it seems to me that we can ligitimately ask is there a better way of smoothing $f(y \mid \theta)$ than by use of $p(\theta \mid \alpha_1, \hat{\alpha}_2)$? Incidentally, I gather that the use of the term "Predictive Sample Reuse" in the title comes up here because we are using the data not only in the functional form

of (2), but also through the use of (6) to use $p(\theta \mid \mathbf{y}; \alpha_1, \hat{\alpha}_2)$ as the smoothing function in $h^{(\cdot)}$, so that even here in the uncomplicated case of no censored data, all observations recorded, we have used the data twice.

But this is compounded in the censored case. In this case the data has a certain structure, viz:

$\mathbf{y}_1' = (y_1, \ldots, y_{n-d})$ are recorded (9)
$\mathbf{y}_2' = (y_{n-d+1}, \ldots, y_n)$ is such that it is known only that $y_j > a_{2j}$ where the a_{2j}, $j = n-d+1, \ldots, n$ are known constants.

Using the superscript (c) to denote the presence of censored data as in (9), we have that the predictive distribution of y, given censored data is

$$h^{(c)}(y \mid \mathbf{y}_1, \mathbf{a}_2) = \int f(y \mid \theta) p(\theta \mid \mathbf{y}_1, \mathbf{a}_2; \alpha_1, \alpha_2) d\theta \tag{10}$$

where here the posterior p is given by

$$p(\theta \mid \mathbf{y}_1, \mathbf{a}_2; \alpha_1, \alpha_2) = k(y) q(\theta \mid \alpha_1, \alpha_2) \ell(\theta \mid \mathbf{y}_1; \mathbf{a}_2) \tag{11}$$

with

$$\ell(\theta \mid \mathbf{y}_1; \mathbf{a}_2) = \{\Pi_{i=1}^{n-d} f(y_i \mid \theta)\} \{\Pi_{j=n-d+1}^{n} [1 - F(a_{2j} \mid \theta)]\} \tag{12}$$

where F is the cumulative of the distribution $f(. \mid \theta)$. Geisser now proceeds as follows. Remove a_{2j} from the likelihood and let $\mathbf{a}_2^{(j)}$ be the $(d\text{-}1)$ vector obtained from $\mathbf{a}_2$ by deleting a_{2j} from it. Find $h^{(c)}(y \mid \mathbf{y}_1, \mathbf{a}_2^{(j)})$ using the prescription (10) etc. From this, find the conditional $h^{(c)}(y \mid y > a_{2j}; \mathbf{y}_1, \mathbf{a}_2^{(j)})$ and obtain

$$\mathrm{E}^{(c)}(y \mid y > a_{2j}) = y_j^*, \tag{13}$$

the conditional expectation of the predictive variable y, given that $y > a_{2j}$, and where we are given the structure (9) with $(d\text{-}1)$ censored observations etc.
The set $(y_{n-d+1}^*, \ldots, y_n^*)' = \mathbf{y}_2^*$ is then used along with $\mathbf{y}_1$ in a discrepancy function D^*, to produce a value α_2^* that is such that

$$D^*(\mathbf{y}_1, \mathbf{y}_2^*; \alpha_1, \alpha_2^*) = \min_{\alpha_2} D^*(\mathbf{y}_1, \mathbf{y}_2^*, \alpha_1, \alpha_2) \tag{14}$$

(An example of D^* is the weighted discrepancy function defined at (3.10) of which I will have something to say below.) The value α_2^* so obtained is then used, as before, to obtain (see (10))

$$h^{(c,s)}(y \mid \mathbf{y}_1, \mathbf{y}_2^*; \alpha_1, \hat{\alpha}_2^*) = \int f(y \mid \theta) p(\theta \mid \mathbf{y}_1, \mathbf{y}_2^*, \alpha_1, \hat{\alpha}_2^*) d\theta \tag{15}$$

Note that we are re-using the data $(d+1+1)$ times — d times to find the set $\mathbf{y}_2^*$

from (13), one additional time in (14) and of course a further additional time in $p(\theta \mid \mathbf{y}_1, \mathbf{y}_2; \alpha_1, \hat{\alpha}_2)$.

To proceed further with this discussion, allow me to start at the beginning of Geisser's paper. We are considering the case of addressing an exponential process, whose distribution is given by

$$f(y \mid \sigma) = \sigma^{-1} \exp[-y/\sigma], \; y, \sigma > 0 \tag{16}$$

We are assuming that before the taking of sample information taken from the above (single) exponential, the applicable prior of σ is such that

$$p(\sigma) \propto \sigma^{-(\delta+1)} \exp[-\gamma/\sigma], \; \sigma, \gamma, \delta > 0 \tag{17}$$

This of course implies that, a-priori

$$2\gamma/\sigma = \chi^2_{2\delta} \text{ or } \sigma = 2\gamma/\chi^2_{2\delta} \tag{18}$$

and that the prior expectation and variance are

$$\begin{aligned} &\text{(i)} \; \mathrm{E}(\sigma) = \gamma/(\delta\text{-}1) = g, \; \delta > 1 \\ &\text{(ii)} \; \mathrm{V}(\sigma) = g^2/(\delta\text{-}2), \; \delta > 2 \end{aligned} \tag{19}$$

In practice, the prior comes "armed" with fixed values of γ and δ, or g and δ, fixed by prior sample information or "experimenter's expertise", and this does touch on the problem of determining whether (17) is applicable, and if so, how to use prior information or experimenter's expertise to arrive at a suitable choice of (g,δ) or (γ,δ) etc. I do not go into this here, but assume that we do have (17) available and that (γ,δ) represent the values chosen (wisely) by the experimenter.

Note that I use the parametrization given in (16), but of course if we let $\mu = 1/\sigma$, we obtain Geisser's formulation. I prefer using σ as in (16) since $E_f(y \mid \sigma) = \sigma$. Note again that $p(\sigma)$ exists for $\delta > 0$, $E_p(\sigma)$ exists for $\delta > 1$, and that $V_p(\sigma)$ exists for $\delta > 2$. Also, we make note of the fact that if δ and γ tend to zero such that $\gamma/(\delta\text{-}1)$ tends to g, then $p(\sigma)$ tends to $p_{ni}(\sigma)$ which is such that

$$p_{ni}(\sigma) \propto 1/\sigma$$

the so called and much maligned *non*-informative prior (*ni*) for σ.

Now an interesting and somewhat novel theme of the author intrudes at this point, and that is the calculation and fitting of the predictive $h\left(y \mid {}^{no}_{data}\right)$, *based* on the prior alone, where

$$h\left(y \mid {}^{no}_{data}\right) = \int f(y \mid \sigma) p(\sigma) d\sigma \tag{21}$$

which after using (16) and (17) yields

$$h\left(y \mid \overset{no}{data}\right) = \delta\gamma^{\delta}(\gamma + y)^{-(\delta+1)}, \tag{21a}$$

that is, a priori, we predict y to behave as a scaled Snedecor-F variable, i.e., *a priori*

$$y = (\gamma/\delta)\ F_{2,2}^{\ \delta}. \tag{21b}$$

Note that the mean and variance of this distribution are

$$\text{(i)}\ E_h(y) = \frac{\gamma}{\delta\text{-}1} = g = E_p(\sigma), \qquad \delta > 1 \tag{22}$$

and if $\delta > 2$,

$$\text{(ii)}\ V_h(y) = g^2\ \frac{\delta}{\delta\text{-}2} > g^2\ \frac{1}{\delta\text{-}2} = V_p(\sigma).$$

Now the moments show that fitting the parameter (γ,δ) or (g,δ) using (21) is associated with a distribution that is located at the same place as the prior $p(\sigma)$, but has *larger variance* (by a factor δ, which could be considerable), and the moments of h are functions of the parameter of the prior. A person who would want to nail down information about prior parameters by fitting his information about (γ,δ) through h (which has larger variance) rather than through the prior, must believe in putting the cart before the horse, and notice too that the experimenter is asked to examine his experience and relate it to future y's based on h, which is not based on current experimental data - I doubt that many experimenters would do this.

Now what is going on can be summarized by the following tableau (We shall let $\delta = (1/\alpha) + 1$ or $\alpha = (\delta\text{-}1)^{-1}$.)

A-Priori

Predictive $y = \frac{\gamma}{\delta}\ F_{2,2\delta}$

Prior on σ: $\sigma = \dfrac{2\gamma}{\chi^2_{2\delta}} = \dfrac{\gamma}{\delta}\ \dfrac{2\delta}{\chi^2_{2\delta}}$

or $\quad \sigma = \dfrac{\gamma}{\delta}\ \lim_{m\to\infty}\ \dfrac{\chi^2_m/m}{\chi^2_{2\delta}/2\delta}$

$$\text{or} \qquad \sigma = \frac{\gamma}{\delta} \lim_{m \to \infty} F_{m, 2\delta}$$

$$E_h(y) = \frac{\gamma}{\delta\text{-}1} = \gamma\alpha = g \qquad E_p(\sigma) = \frac{\gamma}{\delta} \lim_{m \to \infty} \frac{2\delta}{2\delta\text{-}2} = g$$

$$V_h(y) = g^2 \frac{\delta}{\delta\text{-}2} \qquad V_p(\sigma) = \frac{\gamma^2}{\delta^2} \lim_{m \to \infty} (F_{m, 2\delta})$$

$$\vdots$$

$$= g^2 \frac{1+\alpha}{1\text{-}\alpha} \qquad = g^2 \frac{\alpha}{1\text{-}\alpha}$$

Note again that we may write

$$V_p(\sigma) = \frac{\gamma^2}{\delta^2} \lim_{m \to \infty} V(F_{m, 2\delta}) \tag{23}$$

as

$$V_p(\sigma) = \frac{\gamma^2}{(\delta\text{-}1)^2} \frac{1}{\delta\text{-}2} \lim_{m \to \infty} u(m) \tag{23a}$$

$$= g^2 \frac{1}{\delta\text{-}2} \lim_{m \to \infty} u(m)$$

where $u(m)$ is such that

$$u(m) = 1 + \frac{2(\delta\text{-}1)}{m}$$

so that

$$\lim u(m) = 1 \tag{23c}$$

Now Geisser's method of fitting using h amounts to saying replace $\lim_{m \to \infty}$

$u(m)$, which equals 1, by $u(2) = \delta$, while of course, the Bayesian, who is using that $p(\sigma)$ given by (17), that is, σ is *a-priori*, the scales inverted Chi-Square variable given in (18), is using $u(\infty) = 1$.

Having advocated the fitting of the *no-data* predictive, there is what amounts to some backtracking from this position by Geisser, because he now assumes that $g = E_h(y)$ $\left(= E_p(\sigma)\right)$ is assumed known $\left(\text{i.e. picked by soliciting from the experimenter information, sample or otherwise, about } E_h(y)\right)$ and then, rather than continuing with the fitting of h, chooses $\delta = \frac{1}{\alpha} + 1$ by employing the discrepancy function D or D^* and finally the value of α that minimizes the chosen discrepancy function. Here $D = D(\mathbf{y}; g, \alpha)$ is used it all observations recarded, while $D^* = D^*(\mathbf{y}_1; \mathbf{y}_2^*; g, \alpha^*)$ is used if there is censored observations -see the previous discussion here and Geisser's paper, relations (2.7) and (3.10). We again note that doing this amount to choosing a value for a parameter of the prior which depends on the data, a cannon of Bayesianism thus being violated.

Indeed, what would a "Strict Bayesian" do in this problem? (I am indebted to George Barnard for pointing out that the definite article "a" instead of "the" should be used before the words "Strict Bayesian"). We suppose that the process being sampled is as given in (16), that the appropriate prior based on the experimenter's experience and knowledge is given by (17) with δ and γ fixed. Now suppose n units are put on test, and that

(i) n_1 observations, say $y_j^{(c)}$, $j = 1,\ldots,n_1$, unrecorded, but known that lifetimes are less than a_1, that is, $y_i^{(c)} < a_1$, $i = 1,\ldots,a_1$;

(ii) $n\text{-}n_1\text{-}n_2$ observations recorded, say y_j, $j = n_1 + 1,\ldots,n\text{-}n_2$; (24)

(iii) n_2 observations, say $y_t^{(c)}$, $t = n\text{-}n_2 + 1,\ldots,n$, unrecorded, but known that lifetimes are greater than a_2, that is, $y_t^{(c)} > a_2$, $t = n\text{-}n_2 + 1,\ldots,n$

(in our previous discussion, $n_1 = 0$ and $y_t^{(c)} > a_{2_t}$, where $a_{2_t} \equiv a_2$; Geisser's illustrative example involves the case $a_{2_t} \equiv a_2$ and that is why I have decide to look at this case at this point).

From (24) we have that the likelihood is such that

$$\ell(\sigma|\mathbf{y}_2; a_1, a_2) \propto [1\text{-}\exp(-a_1/\sigma)]^{n_1} \times \sigma^{-(n-n_1-n_2)} \exp(-t^{(0)}/\sigma) \times (\exp(-a_2/\sigma)^{n_2} \tag{25}$$

where $t^{(0)} = \sum_{i=1}^{n-n_1-n_2} y_i$ is the sum of the recorded observations. We can thus use all the above ingredients and find

$$p(\sigma|\text{data}) = K \sum_{j=0}^{n_1} \binom{n_1}{j} (-1)^{\sigma^{-(n-n_1-n_2+\delta+1)}} \times \exp[-(t^{(0)} + n_2 a_2 + j a_1 + \gamma)/\sigma] . \tag{26}$$

Using (26) the results for ultimate calculation of the predictive distributions are as follows:

I. **Uncensored Case:** $(n_1 = n_2 = 0)$.

Using the previous definitions we find for this case that:

$$p(\sigma|\mathbf{y}) = \frac{(t+\gamma)^{n+\delta}}{\Gamma(n+\delta)} \frac{1}{\sigma^{n+\delta+1}} \exp - \frac{t+\gamma}{\sigma} \tag{27}$$

where $t = t^{(0)} = \Sigma_i^n y_i$, so that, a posteriori,

$$2(t+\gamma)/\sigma = \chi^2_{2(n+\delta)}. \tag{27a}$$

This in turn implies that

$$h(y|\mathbf{y}) = \int_0^\infty f(y|\sigma)p(\sigma|\mathbf{y})d\sigma \tag{28}$$

$$= \frac{n+\delta}{t+\gamma}\left(1 + \frac{y}{t+\gamma}\right)^{-(n+\delta+1)}$$

that is, the predictive distribution is such that

$$y = \frac{t+\gamma}{n+\delta} F_{2,2(n+\delta)}. \tag{28a}$$

We find

$$\text{(i)} \quad E(y|\mathbf{y}) = \frac{t+\gamma}{n+\delta\text{-}1} = \frac{\alpha n\bar{y}+g}{n\alpha+1} = f \tag{28b}$$

$$\text{(ii)} \quad V(y|\mathbf{y}) = f^2 \frac{n+\delta}{(n+\delta\text{-}1)^2(n+\delta\text{-}2)}$$

and it is to be recalled that Geisser assumes g known and picks α to minimize D given by his (2.7), viz.

$$D(\alpha) = n^{-1}\Sigma_1^n [f_i - y_i]^2 \tag{28c}$$

where f_i has the same form as f in (28b), but leaves y_i out, that is,

$$f_i = \frac{\alpha(n\text{-}1)\bar{y}_i + g}{\alpha(n\text{-}1) + 1}$$

and $\bar{y}_i = (n\text{-}1)^{-1} \sum_{j \neq i} y_j$.

II. **Case of Censoring on the right only** $(n_1 = 0; n_2 > 0)$

Using previous definitions, for this case we find:

$$p(\sigma \mid \mathbf{y}_2 ; a_2) = \frac{(t^{(0)} + \gamma + n_2 a_2)^{n-n_2+\delta}}{\Gamma(n\text{-}n_2 + \delta)} \frac{\exp -\left(\frac{(t^{(0)} + \gamma + n_2 a_2)}{\sigma}\right)}{\sigma^{(n-n_2+\sigma+1)}}$$

that is, a posteriori

$$2(t^{(0)} + n_2 a_2 + \gamma)/\sigma = \chi^2_{2(n-n_2+\delta)}. \tag{29a}$$

Further, the predictive distribution $h^{(c)}$ is such that

$$y = \frac{(t^{(0)} + n_2 a_2 + \gamma)}{n\text{-}n_2 + \delta} F_{2,2(n-n_2+\delta)}. \tag{30}$$

Recall that $\gamma = g(\delta\text{-}1) = g/\alpha$. Note too that (γ,δ) $\left(\text{or } (g,\delta) \text{ or } (g,\alpha)\right)$ is specified at the outset by the experimenter. So a Strict Bayesian who wants to do some predicting in this situation uses (30) which is completely specified. Note too, that using (30) and letting $\alpha \to 0$ implies that

$$y = g\,\chi^2_2/2. \tag{31}$$

Hence, in particular, we would estimate the 90th percentile of future y's, say $\tilde{y}_{\cdot 10}$, that is, the point exceeded with probability $\cdot 10$ when using the predictive as

$$\tilde{y}_{\cdot 10} = \begin{cases} \dfrac{t^{(0)} + n_2 a_2 + (g/\alpha)}{n\text{-}n_2 + 1 + (1/\alpha)} F_{2,2(n-n_2+1)+\frac{2}{\alpha}; \cdot 10} & \text{if } \alpha \neq 0, \\ g\chi^2_{2,\cdot 10}/2 = g\,(2{\cdot}3026) & \text{if } \alpha = 0. \end{cases} \tag{32}$$

TABLE 1

$\tilde{y}_{.10}$ (to nearest integer)

MLE of σ to the nearest integer is 4290.

		g							
α	δ	60	3550	3700	4280	4290	5000	5150	15,000
0	∞	1138	8174	8520	9855	9878	11,513	11,858	34,539
1/10	11	5273	9124	9290	9929	9940	10,724	10,889	21,759
2/10	6	6888	9420	9528	9949	9956	10,471	10,580	17,724
3/10	$4\frac{1}{3}$	7678	9563	9644	9957	9963	10,346	10,427	15,747
4/10	$3\frac{1}{2}$	8147	9648	9713	9963	9967	10,272	10,367	14,575
5/10	3	8457	9705	9759	9966	9970	10,223	10,277	13,799
6/10	$2\frac{2}{3}$	8678	9745	9791	9968	9971	10,188	10,234	13,247
7/10	$2\frac{3}{7}$	8842	9775	9815	9970	9973	10,162	10,202	12,834
8/10	$2\frac{1}{4}$	8970	9798	9834	9971	9974	10,142	10,178	12,514
9/10	$2\frac{1}{9}$	9072	9817	9849	9972	9974	10,126	10,158	12,259
1	2	9155	9832	9861	9973	9975	10,113	10,142	12,050
Geisser - - -		9890	9226	8520	9855	9878	11,513	10,151	10,018

[$F_{m_1,\, m_2,\beta}$ denotes the point exceeded with probability β when using the Snedecor F with (m_1,m_2) degrees of freedom and $\chi^2_{2;.10}$ is the point exceeded with probability $\cdot 10$ when using the Chi-Square distribution with 2 degrees of freedom and is equal to $4\cdot 6052$.]

We illustrate the different types of results that emerge using Geisser's data. Consulting Geisser's paper, we find

$$\begin{aligned} n = 100;\ n_2 = 89;\ n - n_2 = 11;\ a_2 = 500; \qquad (33)\\ n_2\, a_2 = 89\ (500) = 44{,}500;\ t^{(0)} + n_2\, a_2 = 47{,}187; \\ \text{MLE} = 47{,}187/11 = 4290\,. \end{aligned}$$

Recall that Geisser uses (32) with α replaced by $\hat{\alpha}$, and $\hat{\alpha}$ is obtained by following the procedure described starting at (12). The results are given in Table 1. We are assuming that $\alpha > 0$ (so that the mean g exists) and we have cut off the table at the line $\alpha = 1$, but it could continue indefinitely in principle. As the last line, we have inserted Geisser's results. (At this writing Geisser did not supply the values of $\hat{\alpha}$ found, but of course, it is easy to see that for his entries for the cases $g = 3700, 4280, 4290$ and 5000 that his $\hat{\alpha} = 0$.)

Note that unlike Geisser's 1 line table, there are no reversals along rows in the main body of the table. Further, the columns to the right of the MLE are decreasing and viceversa. And it is interesting to note again that priors do produce the different results indicated by the Table, different from the 1-line table of Geisser's, which after all could be very different itself depending on the type of D^* function used. Note that if Geisser were to use the weighted function given by his (3.10), then a minor quarrel could be picked. In our notation, the weights used are

(i) $$\left[E_\sigma\{V(y\,|\,\sigma)\}\right]^{-1} = \left[E_\sigma\{\sigma^2\}\right]^{-1}$$

for the uncensored variables, where the expectation is taken with respect to the posterior of σ given in (29), and

(ii) $$\left[V(y\,|\,y > a_2)\right]^{-1}$$

for the censored variables (recall these are all censored at a_2), where here the variance is taken with respect to the conditional predictive $h^{(c)}(y\,|\,y > a_2)$ where the unconditional $h^{(c)}$ is specified in (30). The minor quarrel is with (i) -in most applications $a_{2i} \equiv a_2$ (in this example $a_2 = 500$) because of a time constraint, or a gauge calibrated between (0, a_2) only, etc., and realistically then, once we see the data the recorded observations are known to be less than a_2. Hence the recommendation would be, not to use (i), but $[V(y\,|\,y < a_2)]^{-1}$.

III. Censored observations on the left and right $(n_1 > 0, n_2 > 0)$.

For this case it is easy to show that we may write the posterior as

$$p(\sigma \mid \text{data}) = K \sum_{j=0}^{n_1} \binom{n_1}{j} (-1)^j \sigma^{-(n-n_1-n_2+\delta+1)} \exp(-\nu_j/\sigma) \tag{34}$$

where

$$\nu_j = t^{(0)} + n_2 a_2 + j a_1 + \gamma , \tag{34a}$$

and K is such that

$$K^{-1} = \sum_{j=0}^{n_1} (-1)^j \binom{n_1}{j} \frac{\Gamma(n-n_1-n_2+\delta)}{\nu_j^{n-n_1-n_2+\delta}} \tag{34b}$$

Using this we may in turn find that the predictive density is given by

$$h^c(y \mid \mathbf{y}_1^{(c)}; y_2; \mathbf{y}_2^{(c)}) = \sum_{j=0}^{n_1} c_j q(y; \nu_j; n-n_1-n_2+\delta) \tag{35}$$

where in general

$$q(y;b;c) = \frac{c}{b}\left[1 + \frac{y}{b}\right]^{-(c+1)}, b, c > 0 \tag{35a}$$

and where

$$c_j = \binom{n_1}{j} (-1)^j \nu_j^{-(n-n_1-n_2+\delta)} \Big/ \sum_{l=1}^{n_1} \binom{n_1}{l} (-1)^l \nu_l^{-(n-n_1-n_2+\delta)} . \tag{35b}$$

To illustrate what happens in this case, we have censored Geisser's data on the right at $a_1 = 60$, so that we are pretending that we have the following sample information:

$n_1 = 2$ observations less than $a_1 = 60$;
$n_2 = 89$ observations greater than $a_2 = 500$ (36)
$n-n_1-n_2 = 9$ observations recorded, and observed to be 90, 135, 161, 249, 323, 353, 833, 436, 477.

Note that $t^{(0)} = 2{,}607$. Using the above data in (35), and dealing with the case where $\alpha = \cdot 5$, that is, $\delta = 3$, we find that the 90^{th} percentile of this distribution for various values of g are as given in Table II.

TABLE II 90^{th} percentile for the predictive density (35) based on the data set (36), for $\alpha = .5$ and g as tabled

g	60	3,550	3700	4280	4290	5000	5150	15,000
90^{th} percentile	8,414	9,666	9720	9927	9930	10,185	10,239	13,782

A comparison with the $\alpha = .5$ line of Table 1 yields the (expected!) fact that the corresponding entries are all less than the corresponding entries of Table I. (A program that tabulates the cumulative of (35) is available from the Department of Statistics, University of Toronto).

Finally, I want to congratulate Geisser again. His paper is very throught provoking and has proven to be very stimulating (to this person at least) and some very subtle issues are raised in this paper. To the data analyst and to those who worry about foundations, this work raises some profound questions to which some clear answers are deserved. In the meantime, the methods proposed by Geisser are of great interest, and he is to be congratulated for the inventive procedures that he has developed.

S.J. PRESS (*University of California, Riverside*):

Professor Geisser has provided us with yet another illustration of the versatility of the predictive sample reuse method that he and Professor Mervyn Stone introduced in different forms, independently, in 1974, in their now well-known and celebrated papers that both appeared in England, in *Biometrika*, and in *JRSS* (B), respectively. Professor Geisser has now shown us how to apply this methodology to the prediction of future observations, when some of the sample data are *censored*. The problem here, of course, that makes this application different from his earlier applications is that not all of the data are inmediately available as candidates for deletion, in the basic discrepancy function, because of the censoring.

As a solution to this inherent difficulty, the author proposes that we introduce pseudo observations, obtained by using the expectation of the predictive distribution of a censored observation given that the observation (that is, the observed failure time) exceeds a preassigned value, namely, the censored value. To obtain this predictive distribution we must introduce substantial structure into the problem. It seems that, we must have a likelihood function, and a *bona fide* prior distribution on the unknown parameters. From this structure we obtain a posterior, and subsequently, a predictive distribution. Taking expectations in the latter yields a "pseudo observation".

The author suggests that when we turn the sample reuse crank, we should utilize the pseudo observations as well as the uncensored ones, and he suggests two methods for doing so. He also suggests that the discrepancy function should be formed as a weighted average of the individual discrepancies obtained by deletion of observations, the weights being assigned according to some specific suggestions.

Finally, Professor Geisser has applied his paradigm to some actual failure time data.

I would like now to make some comments and to raise some questions.

1. My first question concerns the parametric structure imposed on the problem. The recommended approach requires that we make parametric assumptions about both

the sample data and about the parameters of the sampling distribution. If we must impose such structure anyhow, as we would do in a conventional frequentist approach, or in a conventional Bayesian approach, why should we utilize the sample reuse method at all, in this application? To do so, we must introduce some ad hockery regarding the form of our discrepancy function, our predictor function, etc. In other applications, we would presumably be trading off some precision of results, as a result of this ad hockery, in order to gain robustness of prediction with respect to distributional assumptions. In this case, what do we gain?

2. I would like now to question the assignment of weights. Isn't the assignment of weights to the discrepancies quite arbitrary? Certainly the assignment is no less arbitrary that the assumptions made about the form of the discrepancy function, the form of the predictor function, etc. On what basis has the author selected the weights? It seems to me that using precisions as weights is motivated by a normal distribution assumption. But in the case where the data are more likely to be some member of a family of non-normal waiting-time distributions (exponential is what Professor Geisser used as an illustration), why use precision weights?

3. The author combines subjective information with sample information, in a more or less Bayesian way, but violates Bayes' theorem by using sample data to assess the parameters of the prior distribution. It seems to me that there has been ample precedent in the literature for this kind of approach, called empirical Bayes. But this raises the natural question, should we use a moment matching assessment technique or perhaps we should use some other method of getting at the parameters, and then do maximum likelihood estimation of the hyperparameters by maximizing the marginal distribution of the data given the hyperparameters? We would of course need to adopt a likelihood function to do this. Perhaps a smaller risk would be obtained, an important consideration for an empirical Bayesian.

4. My last question involves the underlying parameters of the prior distribution again. A gamma prior is suggested in the paper, for the mean of the sampling distribution of failure time. This is a two parameter prior. But the ensuing analysis really involves only the shape parameter and assumes we know the scale parameter. Perhaps the analysis could be carried out for both parameters simultaneously? Perhaps the mathematics is too intractable.

In conclusion, I would like to thank Professor Geisser for an extremely stimulating and thought provoking paper that clearly extends his earlier research in this area, into new and important fields. But given the methodology we have heard about today, as it relates to *censored* data, it seems to me that there is another problem that could probably be treated in an analogous way - this is the problem of *missing data*. We could generate pseudo observations for the missing data and carry out the analysis in like fashion.

Perhaps Professor Geisser will tell us how to do this in one of his future papers on the subjct.

REPLY TO THE DISCUSSION

S. GEISSER (*University of Minnesota*):

In the introduction to my paper I outlined how sample reuse procedures could be executed in the presence of censored data. Two such procedures were suggested, neither requiring distributional assumptions.

Believing that I. J. Good is essentially correct in his view that most reasonable Bayesian applications are inherently compromises with other methods and also that the predictive sample reuse method can be an attractive empirical Bayes procedure - I offered such an application. It was described first for full data sets and then for incomplete data sets with censoring as a particular application at varying levels of inferential structure running the gamut from low to high.

Even in any real subjective application of Bayesian procedures, there comes a point at some level in the possibly infinite hierarchy of hyperparameters and hyperdistributions where one is no longer willing to continue regressing. Among the several alternatives are: (a) assign precise values to some final set of hyperparameters, (b) introduce a so-called non-informative distribution for them, (c) devise an empirical Bayes procedure for their estimation. Given that certain conditions obtain, coherence is guaranteed for (a), problematic for (b), and inevitably vitiated for (c).

My paper, in part, sets forth a new procedure that can be substituted for others useful in (c) and one which has the robust quality of simulating to a large degree on the available data what it requires from a predictor. Although originally the predictive sample reuse method was introduced to provide point predictors for low structure paradigms, here its effectiveness is amply demonstrated as a useful empirical Bayes estimator of a hyperparameter, an intermediate step towards prediction for a high structure paradigm. In particular, a situation is described where it turns out to be easiest and most convenient to apply amongst the usual estimators of this type. For example, in the uncensored situation, we easily obtain the marginal density of $X_1,\ldots X_N$ to be

$$f(x_1,\ldots,x_N|\delta,\gamma) = \frac{\Gamma(N+\delta)\gamma^{\delta}}{\Gamma(\delta)[Nx+\gamma]^{N+\delta}} \tag{1}$$

where x is the mean of the observations.

Assuming $g = \gamma/(\delta\text{-}1)$ is known and transforming to $Y_i = g^{-1}X_i$, the marginal density of $Y_1, \ldots, Y_N$ is

$$f(y_1,\ldots,y_N|\delta) = \frac{\Gamma(N+\delta)(\delta-1)^{\delta}}{\Gamma(\delta)[Ny+\delta-1]^{N+\delta}} \tag{2}$$

Here $S = \Sigma_{i=1}^{N} Y_i$ is sufficient for δ and clearly $(\delta - 1)^{-1}S$ is distributed as $\beta_2(N,\delta)$, a beta distribution of the second kind. The method of moments fails because $E(S) = N$ and using the second moment restricts the range of δ. Hence it would require that $\hat{\delta} > 2$, which results in the estimating method imposing a restriction unassumed by the model. The maximum likelihood estimator is a solution to the unwieldy equation

$$\log \frac{\delta\text{-}1}{s+\delta\text{-}1} + \frac{\delta}{\delta\text{-}1} - \frac{N+\delta}{s+\delta\text{-}1} + \Sigma_{j=0}^{N-1} \frac{1}{\delta + j} = 0 \qquad (3)$$

the number of whose terms increases with the sample size. So much for competitors in terms of ease in getting a sensible estimator.

Professor Press wonders if the analysis could be carried out for both g and δ unknown. For that case, there is no apparent relief for method of moments and maximum likelihood procedures when applied to (1). The PSR method requires solving a cubic equation in the uncensored case and is somwhat more complicated in the censored case. When a given value for g is specified (which is much more likely to be specifiable than δ), the PSR solution as described in the paper is explicit for the uncensored case and easy to achieve in the censored case using the recursive algorithms of section 4 and has the appealing property of being similar to a "testimator".

It is a rare event indeed when a discussion comes perilously close to exceeding the length of the paper at issue. Even rarer when the discussant begins and ends with the same litany of praise and yet the author must disagree with most of the views expressed. I refer, of course, to Professor Guttman's critique. First an exception: I applaud his use of the term *fungible* which I introduced in an attempt to extend exchangeable. It has indicated that a prediction made when a colleague expressed his aversion to such a singularly unattractive word, may yet take hold. Mustering my most somber demeanor, I portentously responded, "It will grow on you."

First we address some minor details. Professor Guttman's equation (6) is meaningless unless a predictive function is specified. A demonstration of consistency; i.e., that the two sides of (8c) approach the same density as the sample size increases, does not present any difficulty. Although I already responded in part to the great to-do about violating primordial Bayesian canons, still permit me to take this opportunity to expose a further serious transgression on my part. To assume that a prior depends on the likelihood is, of course, original sin itself in this theology. Apparently undetected by Professor Guttman, who usually performs yeoman service as a sort of Bayesian superego, was my use of a conjugate prior density - *mea culpa*. Professor Guttman admonishes me for a prior that comes only "one-armed" instead of what he considers to be appropriate - the investigator determining exact values for both hyperparameters. We obviously describe different situations.

Now to more serious questions. I must take very strong issue with his horse-cart analogy. It derives, I believe, from a fundamental misunderstanding of the practical value of parametric infusions into statistical paradigms. Parameters are basically artifices introduced by the statistician to lubricate the modeling procedure, and of course, hyperparameters even more so. In most instances, they are completely alien to the experimenter's thinking who works with and thinks about observables. Hence, if properly questioned he can respond in those terms directly. If you want to elicit more than just a curious stare, try explaining a hyperparameter to an investigator; it is a sure ticket to non-communication. Further, the exercise on predictive and prior variances which has exercised Professor Guttman invites exorcism. They are irrelevant calculations devoid of purpose and meaning in regard to the issues.

Professor Guttman has taken the trouble to calculate tables of the 90th percentile points of the predictive distribution for varying but known α and g and claims to have uncovered the fatal flaw (certain reversals in the probabilities) in using an empirical Bayes procedure - the fault being that it is not "Bayesian." He could have saved himself the trouble by discerning from the table and graphs in the original paper that this had to be the case. On the one hand, these reversals actually demonstrate the fact that when the guessed value of g is very far from the experimental data, the sample reuse procedures wisely discount the value to a greater and greater extent as if g were the product of a demented prior opinion. On the other hand, when the mean of the sample values is within a certain small interval of g, the procedure behaves as if μ were known to be g^{-1} from the start. This is the "testimator" quality of the procedure - it makes every effort to temper the rigidity of coherence with the facts embodied in the data.

Professor Press complains about my weight functions. If he has a better scheme, I would be happy to entertain it because the plethora I presented complicate the procedure far too much. In fact, the more information used, the greater the computational complexity. Even if a set of weights, indisputably appropriate and yielding a reasonably computable solution, were adduced, which is unlikely, I believe the algorithmic method would still be preferable. This was fully described in section 4 and illustrated for the data set. Hence, I echo his complaint but for different reasons.

On the other hand, Professor Guttmam insists that weights be based on a predictive variance conditioned on the observable being less than a given value when in fact it is known that it exceeds that value. This logical inversion indeed makes even a cart-horse analogy pale by comparison.