DISCUSSION

M.E. Thompson, University of Waterloo

The authors of the papers presented here today have provided some new insights into the area of robustness in sampling estimation. Although the approaches and the results are different, I feel there is at least the nucleus of a common understanding present, and I hope that a recognition of this will lead to fruitful research in the area. I am particularly struck by the final comments of Royall and Pfeffermann on the role of randomization, which is of course fundamental in Godambe's paper.

As in much of the recent literature, the estimation of a population total $T(\underset{\sim}{y}) = \sum_{i=1}^{N} y_i$ in the presence of a regression model for y in both papers provides a framework for the discussion of robustness. The regression model is simple and mathematically tractable, and amply illustrates the implications of conflicting approaches. A special case of the "working model" in both cases would be given by

$$y_i = \beta x_i + e_i,$$

$$Ee_i = 0,\ Ee_ie_j = 0,\ \text{Var } e_i = \sigma^2 f(x_i).$$

For Godambe, β is unknown and the e_i are independent with essentially arbitrary distributions; for Royall and Pfeffermann the e_i are normal and β has a uniform (improper) prior.

To me the Royall and Pfeffermann paper represents a first step in the Bayesian analysis of robustness, inasmuch as it examines primarily conditions for total insensitivity of the posterio distribution to certain modifications of the working model. The authors find that the posterior mean is $N\bar{y}_s$ as long as the sample is balanced on x and other possible variables z. Less obviously, they find that the variance of the posterior is insensitive to the presence of the other possible regressor variables z if Condition L holds: In our special case, this means $f(x_i) \propto x_i$, and this is a condition which cannot be counted upon in practice. If σ^2 does not have a degenerate prior, even Condition L cannot guarantee that the posterior distribution remains inviolate by the intrusion of z. It seems clear that, although the point estimation problem is taken care of by balancing, the interval estimation problem remains to be solved. For example, it ought to be possible to come up with a Bayesian interpretation of the robust variance estimators of Royall, Eberhardt and Cumberland (1975, 1978).

Godambe quite frankly concentrates on point estimation. He shows how the design can be chosen so that a point estimator which is optimal under the working model is still unbiased and near optimal (in a very carefully defined sense) under certain departures from it. What emerges is a stratified design with, in general, unequal inclusion probabilities within strata. These are, in fact, proportional to $\sqrt{f(x_i)}$, and will allow more efficient point estimation than will balancing, at least in some situations.

For example, consider the somewhat artificial case of a stratified population in which the 'true' model is

$$y_i = \beta_h x_i + e_i,\ \text{Var}(y_i) = \sigma^2 x_i.$$

As shown by Royall and Pfeffermann, balancing (and invariance of the posterior for fixed σ^2) is achieved if the sample is balanced on x within each stratum and proportionally allocated among strata. Godambe's prescription, if the same stratification is used, uses a non-proportional allocation; namely, n_h is proportional to

$$\sum_{i\in str.h} \sqrt{f(x_i)} = \sigma \sum_{i\in str.h} \sqrt{x_i},$$

giving more representation to more variable strata. If $\beta_h = \beta + \gamma_h$, then $a_i = \gamma_h x_i$, and

$$E_p[\varepsilon_\zeta(e(\underset{\sim}{Q})-T)]^2 = E_p[(\frac{\sum_{i=1}^{N}\sqrt{x_i}}{n} \sum_h \sum_{i\in S_h} \beta_h \frac{x_i}{\sqrt{x_i}} - \sum_h \sum_{i\in S_h} \beta_h x_i]^2.$$

Thus as long as the p-variance of $e(\underset{\sim}{Q})$ is small at the point $\underset{\sim}{\chi} = (\beta_1 x_1, \beta_1 x_2, \ldots, \beta_h x_{N-1}, \beta_h x_N)$ (it will be 0 if the x's are constant within strata), then $e(\underset{\sim}{Q})$ is nearly ζ-unbiased as well as being exactly p-unbiased, and all is well. Clearly there are operational details to be worked out, such as the size of strata to take when 'sharp stratification' is not possible and the true model is unknown. The specification of interval estimation procedures must also be tackled. But Godambe's claim that the use of unequal probability sampling designs can enhance efficiency without the sacrifice of robustness is clearly worthy of serious consideration.

Finally, a few words about robustness of regression model techniques in general. The kinds of departures from the working model considered by both sets of authors are mainly simply expressed

perturbations of the linear relationship. That is, Royall and Pfeffermann in their example emphasize the case where the z variables correspond to constant or quadratic terms or x times indicator functions of strata. Godambe requires that his mean departure vector $\underset{\sim}{a}$ should be such that

$$E_p(e(\underset{\sim}{0})-T)^2$$

is small at the point $(a_1+\beta x_1,\ldots,a_N+\beta x_N)$. But if we look at populations encountered in practice, we often find a close to linear relationship cluttered up with 'outliers'. (See, for example, the populations displayed by Royall and Cumberland, 1981). If these are set aside, most sensible methods of estimating $T(\underset{\sim}{y})$ will produce comparable results, while if they are treated as ordinary sample points when sampled, the outcomes may be in greater error than they need to be. Is it possible to devise a theory of estimators of $T(\underset{\sim}{y})$ which are robust to this kind of departure from the linear model? Already much work has been done on robust regression estimation, and possibly some of this will apply. And of course, the key problem, the real test of such a theory, will be its implications for interval as well as point estimation.

REFERENCES

ROYALL, R.M., and EBERHARDT, K.R. (1975), "Variance Estimates for the Ratio Estimator," Sankhya, Ser. C., 37, 43-52.

ROYALL, R.M., and CUMBERLAND, W.G. (1978), "Variance Estimation in Finite Population Sampling". Journal of the American Statistical Association, 73, 351-358.