DISCUSSION

J.N.K. Rao, Carleton University
Department of Mathematics & Statistics, Carleton University, Ottawa, Canada

The four papers presented in this session "Estimation problems in complex surveys" cover three important topics in sample survey theory and methods. The papers by Binder (and Kovacevic) and Shao deal with variance estimation for complex statistics, while the paper by Singh, Stukel and Pfeffermann studies the problem of measuring uncertainty associated with model-based small area estimators. Godambe and Thompson in their paper consider optimal estimation in a causal framework.

Binder and Kovacevic consider several descriptive measures of income inequality including ordinates of the Lorenz curve, income shares, the family of Gini coefficients, and a low income measure. A design-based estimator, $\hat{\theta}$, is obtained as the solution of an estimating equation of the form

$$\int \hat{u}(y,\hat{\theta})\, d\hat{F}(y) = 0,$$

where $\hat{F}(y)$ is design-unbiased for the population distribution function $F(y)$ and $\hat{u}(y,\hat{\theta})$ is an estimator of $u(y,\theta)$ that defines the population parameter θ as the solution of $\int u(y,\theta)\, dF(y) = 0$. This class covers measures of income inequality as well as customary estimators like ratios, linear regression and logistic regression coefficients which correspond to the special case $\hat{u}(y,\theta) = u(y,\theta)$. Binder's (1983) well-known paper considered the latter special case and obtained Taylor linearization variance estimators. The present paper extends Binder's paper to the general case of $\hat{u}(y,\hat{\theta})$, using clever techniques but ignoring the remainder terms. As shown in Shao and Rao (1993; unpublished manuscript) for the low income measure, the justification for ignoring the remainder terms involves non-trivial technical arguments, however.

In their illustration under a stratified multistage design, Binder and Kovacevic compared the standard errors obtained using the proposed linearization method and the "delete-one *cluster*" jackknife method. The jackknife variance estimator is known to be design-consistent for smooth estimators like quintile shares, Gini coefficients and Lorenz curve ordinates (Shao, 1992). In a limited simulation study (Rao, Wu and Yu, 1992), the delete-*cluster* jackknife also performed quite well for the median, a nonsmooth estimator, although the "delete-one *element*" jackknife variance estimator is known to be inconsistent for nonsmooth estimators in the case of simple random sampling. It is somewhat surprising that Binder and Kovacevic obtained much smaller Taylor standard errors compared to jackknife standard errors for a smooth estimator like quantile share (e.g., 0.119 vs 0.0337 for $Q(0.6, 0.8)$). They also obtained much smaller Taylor standard errors for nonsmooth estimators (median and low income measure). It would be useful to conduct a simulation study to throw light on the two methods in the context of stratified multistage designs, particularly when the contribution from several of the sample clusters is zero.

Shao gives an excellent account of the asymptotic properties of balanced repeated replication (BRR) variance estimators under stratified multistage sampling. An advantage of BRR is that the variance estima-

tors are design-consistent for both smooth and nonsmooth statistics. However, the construction of balanced replicates for arbitrary n_h is not always easy or the number of replicates to achieve balance can be large or such replicates may not exist as in the case of $n_h = 6$, where n_h is the number of sample clusters in stratum h. In view of these difficulties, one often uses grouped BRR with two random groups of clusters in each stratum h. This method requires only a Hadamard matrix as in the case of $n_h = 2$, but the resulting variance estimator is very unstable, as shown by Krewski (1978). The proposed repeatedly grouped BRR (RGBRR) method looks promising since it is simple to implement and leads to a stable variance estimator. When the number of strata, L, is small and n_h is large, the customary grouped BRR variance estimator is in fact inconsistent, as shown by Rao and Shao (1993; unpublished manuscript). On the other hand, RGBRR variance estimator performs well even when the number of repetitions, G, is as small as 10 or 15. By choosing the groups in a balanced manner, Sitter (1993) overcomes the difficulties associated with the BRR, but it is less flexible compared to RGBRR in terms of the number of replicates although his variance estimator is slightly more efficient.

The proposed BRR variance estimator for a total under imputation for missing data is also useful. It complements the jackknife variance estimator of Rao and Shao (1992), and both are design-consistent under uniform response within imputation classes that may cut across sample clusters.

Model-based estimators for small areas have received considerable attention in recent years. Such estimators "borrow strength" from related areas to increase the efficiency of estimators. Prasad and Rao (1990) and others used a frequentist approach to obtain an empirical best linear unbiased prediction (EBLUP) estimator and a second-order approximation (PR) to the estimator of its MSE, under some random effect models. Kass and Steffey (1989), on the other hand, employed a Bayesian framework to obtain a first-order approximation (KS-I) to the posterior variance, while Hamilton (1986) used a Monte Carlo integration method (H) of approximating the posterior variance. Singh, Stukel and Pfefferman have investigated, under a simplified model, some frequentist properties of KS-I and H approximation. They also suggest modification to improve their accuracy; in particular, a simplified version of the second-order approximation (KS-II) of Kass and Steffey. It would be useful to provide similar improved approximations for more complex random effect models and to study their frequentist properties. We agree with the authors that the proposed approximations to the posterior variance have the advantage of dual interpretation in both frequentist and Bayesian contexts, although PR-approximation to estimator of MSE performed better with respect to frequentist properties, as one would expect.

Godambe and Thompson define a finite population parameter $\Delta = \Sigma(y_i - y_i')/N$ as a measure of causal effect (difference between mean responses under two treatments). This parameter is hypothetical unlike the customary finite population parameters since each individual can receive only one of the two treatments. Using an estimating functions approach, Godambe and Thompson obtain an "approximately optimal" estimator of Δ, under a semi-parameter superpopulation model and the assumption $P\{z_i = 1|(y_i, y_i')\} = \alpha_i$, where $z_i = 1$ if the individual i receives treatment 1 and $z_i = 0$ if treatment 2 is assigned. The proposed estimator $\hat{\Delta}$ depends on α_i which is estimated using a logistic regression model.

It may be more natural to assume the generation of z_i as part of the design rather than as part of the superpopulation model, as in the case of nonresponse situations. If so, the proposed estimator $\hat{\Delta}$ may not be design-consistent. It would be useful to develop "model-assisted" estimators that are both design-consistent and (approximately) model-unbiased, and compare their performances relative to $\hat{\Delta}$. Also, it would be useful to provide suitable variance estimators for $\hat{\Delta}$.

Finally, I would like to congratulate all the authors for their excellent papers.