

DISCUSSION OF PAPERS BY CHEN/SHAO AND BELLHOUSE/STAFFORD

M.E. Thompson¹

It is a pleasure to discuss two papers on very different aspects of methodology for complex surveys.

data, the median can be estimated as the μ root of the estimation equation

PAPER BY CHEN AND SHAO

$$\phi_1(\mu) = \left[\sum_A w_{hij} (I_{y_{hij}}(\mu) - \frac{1}{2}) \right] = 0, \quad (1)$$

Professors Chen and Shao have obtained an interesting set of results, which show that there is a simple and extendible solution to the problem of confidence interval construction when (i) hot deck imputation uses selection probabilities proportional to the weights of the respondents and (ii) there is a consistent estimator of non-response probability - and this solution is available even when non-respondents are not identifiable in the data set.

where $I_y(\mu) = 1$ if $y \leq \mu$, $= 0$ otherwise. If we keep μ fixed, we can estimate the variance of $\phi_1(\mu)$ from the sample, and form a studentized ratio which is approximately $N(0, 1)$ or t distributed when μ is the true median. Confidence interval construction follows by inverse testing, with more computational effort, but less delicate distribution theory. It seems to me that the authors' method can be applied to this approach when some data are missing, in a straightforward way.

Condition (ii) seems to require essentially that there *exists* a constant response probability p_y for a given item and a given imputation class, and that we can estimate it because we know the *number* of respondents r for the item and class. If the identification of non-respondents has really been thrown away, we will not be able to estimate p_y consistently. Furthermore, it seems to me to be important for statistical and scientific purposes always to flag imputed values, in spite of the inconvenience of the storage requirements. Thus I would be tempted to modify the context for these potentially very useful results: we know which values are imputed, and thus can estimate response probabilities; however, we do not wish to use the detailed information - the flags - otherwise in constructing confidence intervals.

For the low income proportion, say λ , we would add to (1) a second equation

$$\phi_2(\mu, \lambda) = \sum_A w_{hij} (I_{y_{hij}}(\mu/2) - \lambda) = 0,$$

as suggested by Binder and Kovačević (1993). A suitable combination of $\phi_2(\mu, \lambda)$ and $\phi_1(\mu)$ for the estimation of λ is then given by

$$\hat{F}^* \left(\frac{1}{2} \mu \right) - F \left(\frac{1}{2} \mu \right) + \frac{f \left(\frac{1}{2} \mu \right)}{2f(\mu)} [F(\mu) - \hat{F}^*(\mu)] + o_p \left(\frac{1}{\sqrt{n}} \right)$$

(the right hand side of an equation provided in an earlier draft of Chen and Shao's paper).

It would have been interesting to see simulation results also for estimation of a ratio under marginal imputation. Is it necessary to assume independence of response/non-response and (y, z) for v_s^* to be consistent?

PAPER BY BELLHOUSE AND STAFFORD

In connection with the estimation of the median and the low income proportion, the estimating function formulation is very natural. When there are no missing

It is also interesting to see the beginning of the application of powerful symbolic computation methods to the automation of survey sampling calculations. I

¹ M. E. Thompson, Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1.

have some doubts about the pedagogical value of introducing the package early on, but I think I can see the value for researchers and consultants in the generation of useful asymptotic and finite sample results.

I found the section on the expression of finite population moments and k statistics intriguing, and wonder whether in this connection, it would be possible to make use of the interesting identity of I.J. Good (1977) that

$$K_r = \frac{1}{rN^{(r)}} \sum \dots \sum (y_{j_1} + \omega y_{j_2} + \omega^2 y_{j_3} + \dots + \omega^{r-1} y_{j_r})^r,$$

where ω is the r -th root of unity $e^{2\pi i/r}$, and the r -fold sum is taken over all sequences of r distinct subscripts between 1 and N .

Note that from this expression, writing down an unbiased estimator of K_r in terms of joint inclusion probabilities of the sampling design is immediate.

For exploitation of the estimation function approach to sampling estimation, I would be interested to see the authors consider incorporating algebraic or numerical inversion of probability intervals for approximate pivots derived from estimating functions. This would diminish the reliance on Taylor linearization, and to some extent reduce the complexity of asymptotic calculations. For example, in the very simple case of estimation of a ratio R from a simple random sample, instead of beginning with inversion of

$$\left(\frac{\bar{y}}{\bar{x}} - R \right) / A \approx N(0, 1)$$

where

$$A = \frac{1}{\bar{x}} \sqrt{\frac{1}{n} \left(1 - \frac{n}{N} \right) \frac{\sum_{j \in s} (y_j - \hat{R}x_j)^2}{n-1}},$$

which involves one linearization operation to arrive at the denominator, we would consider finding an approximate confidence interval by inverting

$$\left(\sum_{j \in s} (y_j - Rx_j) \right) / B \approx N(0, 1)$$

where

$$B = \sqrt{n \left(1 - \frac{n}{N} \right) \frac{\sum_{j \in s} (y_j - Rx_j)^2}{n-1}}.$$

To begin to assess the adequacy of these approximations, we could then expand the pivots to first order. This gives for the first a bias of approximately

$$-\frac{1}{2} \sqrt{1-\lambda} \gamma n^{-1/2} - \sqrt{1-\lambda} (\mu_{xx} / \mu_x \sigma) (R - \beta) n^{-1/2} + O(n^{-1}),$$

where $\lambda = n/N$, $\beta = \sum_{j=1}^N x_j y_j / \sum_{j=1}^N x_j^2$, $\mu_{xx} = (\sum_{j=1}^N x_j^2) / N$, $\sigma^2 = (\sum_{j=1}^N z_j^2) / N$, $\gamma = \sum_{j=1}^N (y_j - Rx_j)^3 / N \sigma^{3/2}$, and $z_j = y_j - Rx_j$; and a simpler bias expression for the second of

$$-\frac{1}{2} \sqrt{1-\lambda} \gamma n^{-1/2} + O(n^{-1}).$$

REFERENCES

- Binder, D.A., and Kovačević, M.S. (1993). "Estimating some measures of income inequality from survey data: an application of the estimating equation approach", *Proceedings of the American Statistical Association*, Survey Research Methods Section, 551-556.
- Good, I.J. (1977). "A new formula for k -statistics", *Annals of Statistics*, 5, 1, 224-228.