

## RESAMPLING METHODS IN COMPLEX SURVEYS: AN OVERVIEW IN HONOUR OF J.N.K. RAO'S RETIREMENT

Randy R. Sitter<sup>1</sup>

### ABSTRACT

This article attempts to review some of the developments in resampling methods in the context of sample surveys. This will include the jackknife, balanced repeated replications and the bootstrap, and will focus primarily on variance estimation and the contributions of J.N.K. Rao. Complex sampling structures and issues such as stratification, multistage cluster sampling, missing data and imputation etc., will be considered and the resulting impact these have on the structure and complexity of resampling methods will be briefly discussed. I will finish with discussion of some recent developments that were partially motivated by J.N.K. Rao's work.

KEY WORDS: Balanced Bootstrap; Balanced Repeated Replications; Bootstrap; Jackknife; Stratified Multi-stage Sampling.

### RÉSUMÉ

Cet article essaiera de passer en revue quelques-uns des développements en méthodes de rééchantillonnage dans le contexte de sondages. Ces méthodes comprennent la méthode du jackknife, les réplifications répétées balancées et le bootstrap. L'accent sera mis sur l'estimation de la variance et les contributions de J.N.K. Rao. Des structures d'échantillonnage complexes et des problématiques telles que la stratification, l'échantillonnage en grappe à plusieurs niveaux, les données manquantes et imputation seront discutés. Nous finirons par une discussion sur les plus récents développements, qui étaient motivé by le travail de J.N.K. Rao.

MOTS CLÉS : Bootstrap; bootstrap équilibré; échantillonnage à plusieurs phase stratifié; le jackknife; réplifications répétées équilibré.

### 1. INTRODUCTION

J.N.K. (Jon) Rao has had a major and positive influence on my life. First through his work, some of which I was taught in courses and read from the literature as a graduate student. Indeed some of his work on resampling methods became the basis from which my PhD thesis began. Then, as a colleague in the Department of Mathematics and Statistics at Carleton University in Ottawa, Canada (now the School of Mathematics and Statistics), where he encouraged me, supported me, taught me, and became a good friend. I miss the days at Carleton, where I learned more over coffee breaks with Jon than in any other setting....about statistics, probability, mathematics...about professionalism, scholarship, excellence...about life as an academic.

Thus, I would like to take this opportunity to thank Jon Rao, for all he has done for me personally. Thanks Jon, and I congratulate you on your *reduced teaching load*.

I thought I was fairly intelligent to pick as my focus a narrow area of Jon's work and the area I am most familiar with. I thought it far too daunting to try to give an overview of Jon's entire catalogue of contributions in a 30 minute presentation and now in a short proceedings paper. I would be doomed to drown under the weight of it and equally doomed to fail to do it justice. But, as I began to attempt to formulate my talk, I realized I had not been cagey enough. The amount of research that Jon has produced even in this relatively small sub-area of his work is still very large, a career for many. Thus, I will begin by saying that I have not done the body of work justice. I have not even been thorough. Therefore, I took the liberty of placing papers of Jon's in the references which are not referred to in this article.

In Section 2, I narrow my scope further by restricting attention to stratified multi-stage sampling. Sections 3-5 then respectively describe the jackknife, balanced repeated replications and the bootstrap in this setting. Section 5 first discusses imputation and non-response

<sup>1</sup> Randy R. Sitter, Department of Statistics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, sitter@stat.sfu.ca

and the adjusted jackknife, and then describes some new research on balanced resampling and the bootstrap in this context (Saigo, Shao and Sitter, 2001). I conclude with a brief description, in Section 6, of some of Jon's more important works which were not discussed here.

## 2. STRATIFIED MULTI-STAGE SAMPLING

Though some of the results presented can be more generally applied, for the purposes of brevity, I restrict attention to the commonly used stratified multistage sampling design. Suppose that the population contains  $L$  strata and in stratum  $h$ ,  $n_h$  clusters are selected with probabilities  $p_{hi}$ ,  $i = 1, \dots, n_h$ . Samples are taken independently across strata. In the case of complete response on item  $y$ , let

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} / (n_h p_{hi})$$

be a linear unbiased estimator of the stratum total  $Y_h$ , where  $\hat{Y}_{hi}$  is a linear unbiased estimator of the cluster total  $Y_{hi}$  for a selected cluster based on sampling at the second and subsequent stages. A linear unbiased estimator of the total,  $Y = \sum Y_h$ , is given by  $\hat{Y} = \sum \hat{Y}_h$ , which may be written as

$$\hat{Y} = \sum_{(hik) \in s} w_{hik} y_{hik}, \quad (1)$$

where  $s$  is the complete sample of elements, and  $w_{hik}$  and  $y_{hik}$  denote the sampling weight and the item value attached to the  $(hik)$ -th sampled element, respectively. It is common practice to select psu's without replacement with probability proportional to size. However, at the variance estimation stage, we often treat the sample as if the psu's were selected with-replacement. This generally leads to overestimation of the variance of  $\hat{Y}$ , but the bias is small if the first-stage sampling fractions are small. If so,

$$Var(\hat{Y}) = \sum_{h=1}^L \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 = v(y_{hi}),$$

where  $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$  and  $\bar{y}_h = \sum_i y_{hi} / n_h$ . Jon Rao's operator notation,  $v(y_{hi})$ , emphasizes that  $Var(\hat{Y})$  depends only on the weighted cluster totals,  $y_{hi}$ .

## 3. JACKKNIFE

To construct a jackknife variance estimator, obtain  $\hat{Y}_{(gj)}$  by deleting the  $j$ -th sampled cluster from the  $g$ -th stratum ( $g = 1, \dots, L; j = 1, \dots, n_g$ ). Alternately, one can accomplish this by adjusting the weights as follows:

$$w_{hik(gj)} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ n_g w_{gjk} / (n_g - 1) & \text{if } h = g \text{ and } i \neq j \\ w_{hik} & \text{if } h \neq g \end{cases}$$

and using  $\hat{Y}_{(gj)} = \sum_{(hik) \in s} w_{hik(gj)} y_{hik}$ . The jackknife variance estimator is then

$$v_j(\hat{Y}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{(gj)} - \hat{Y})^2.$$

For more general smooth statistics,  $\hat{\theta} = g(\hat{Y})$ , merely replace  $\hat{Y}_{(gj)}$  and  $\hat{Y}$  by  $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$  and  $\hat{\theta} = g(\hat{Y})$ , respectively. See Krewski and Rao (1981) and Rao and Wu (1985) for results on consistency.

## 4. BALANCED REPEATED REPLICATIONS (BRR)

McCarthy (1969) proposed a BRR for  $n_h = 2$ . Balanced half-samples are obtained by deleting 1 unit from each stratum using an orthogonal array. For

$$B = \begin{pmatrix} & & & & & & & h \\ & t & & & & & & \\ & 1 & + & + & + & + & + & + \\ & 2 & - & + & - & + & - & + \\ & 3 & - & - & + & + & - & - \\ & 4 & + & - & - & + & + & - \\ & 5 & + & + & + & - & - & - \\ & 6 & - & + & - & - & + & - \\ & 7 & - & - & + & - & + & + \\ & 8 & + & - & - & - & - & + \end{pmatrix}$$

example, let  $L = 7$  and  $n_h = 2$  and use

to obtain half-sample  $t$  via adjusting weights such that

$$(w_{h1k(t)}, w_{h2k(t)}) = \begin{cases} (0, 2w_{h2k}) & \text{if } (th) = - \\ (2w_{h1k}, 0) & \text{if } (th) = +. \end{cases}$$

**Table 1: A BOMA (24, 4<sup>7</sup>; 2<sup>7</sup>)**

h=1	2	3	4	5	6	7
(1,3)	(1,3)	(1,3)	(1,3)	(1,3)	(1,3)	(1,3)
(1,4)	(1,4)	(1,4)	(1,4)	(1,4)	(1,4)	(1,4)
(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)
(2,4)	(1,3)	(2,4)	(1,3)	(2,4)	(1,3)	(2,4)
(2,3)	(1,4)	(2,3)	(1,4)	(2,3)	(1,4)	(2,3)
(3,4)	(1,2)	(3,4)	(1,2)	(3,4)	(1,2)	(3,4)
(2,4)	(2,4)	(1,3)	(1,3)	(2,4)	(2,4)	(1,3)
(2,3)	(2,3)	(1,4)	(1,4)	(2,3)	(2,3)	(1,4)
(3,4)	(3,4)	(1,2)	(1,2)	(3,4)	(3,4)	(1,2)
(1,3)	(2,4)	(2,4)	(1,3)	(1,3)	(2,4)	(2,4)
(1,4)	(2,3)	(2,3)	(1,4)	(1,4)	(2,3)	(2,3)
(1,2)	(3,4)	(3,4)	(1,2)	(1,2)	(3,4)	(3,4)
(1,3)	(1,3)	(1,3)	(2,4)	(2,4)	(2,4)	(2,4)
(1,4)	(1,4)	(1,4)	(2,3)	(2,3)	(2,3)	(2,3)
(1,2)	(1,2)	(1,2)	(3,4)	(3,4)	(3,4)	(3,4)
(2,4)	(1,3)	(2,4)	(2,4)	(1,3)	(2,4)	(1,3)
(2,3)	(1,4)	(2,3)	(2,3)	(1,4)	(2,3)	(1,4)
(3,4)	(1,2)	(3,4)	(3,4)	(1,2)	(3,4)	(1,2)
(2,4)	(2,4)	(1,3)	(2,4)	(1,3)	(1,3)	(2,4)
(2,3)	(2,3)	(1,4)	(2,3)	(1,4)	(1,4)	(2,3)
(3,4)	(3,4)	(1,2)	(3,4)	(1,2)	(1,2)	(3,4)
(1,3)	(2,4)	(2,4)	(2,4)	(2,4)	(1,3)	(1,3)
(1,4)	(2,3)	(2,3)	(2,3)	(2,3)	(1,4)	(1,4)
(1,2)	(3,4)	(3,4)	(3,4)	(3,4)	(1,2)	(1,2)

Then recalculate  $\hat{\theta}$  using these BRR weights for  $t = 1, \dots, T$  and let

$$v_{BRR} = T^{-1} \sum_{t=1}^T (\hat{\theta}_{(t)} - \hat{\theta})^2.$$

BRR variance estimates are consistent for smooth and non-smooth estimators. For extensions and related work, see Rao and Shao (1996, 1999). The difficulty is the construction of balanced half-samples for arbitrary  $n_h$ .

For  $n_h \geq 2$ , one can use balanced orthogonal multi-arrays (BOMAs) to obtain balanced half-samples (Sitter, 1993). For example, if  $L=7$  and  $n_h = 4$  for  $h=1, \dots, 7$ , one can use the BOMA in Table 1 to construct half-samples.

#### 4. BOOTSTRAP

Rao and Wu (1988) (see also: Kovar, Rao and Wu, 1988; Rao, Wu and Yue, 1992) propose the a rescaling bootstrap variance estimator obtained as follows: Draw a simple random sample with-replacement of  $m_h = n_h - 1$  psu's from the  $n_h$  sampled psu's; Repeat a

large number of times,  $B$ , and let  $m_{hi}(b)$  = the number of times cluster ( $hi$ ) appears in the  $b$ -th bootstrap resample; Let the boot strap weights be  $w_{hik(b)} = w_{hik} n_h m_{hi}(b) / (n_h - 1)$  and recalculate  $\hat{\theta}$  using these bootstrap weights to get  $\hat{\theta}_{(b)}$  for  $b = 1, \dots, B$ ; Define the bootstrap variance estimator to be

$$v_B = B^{-1} \sum_{b=1}^B (\hat{\theta}_{(b)} - \bar{\theta}_{(.)})^2,$$

where  $\bar{\theta}_{(.)} = B^{-1} \sum_b \hat{\theta}_{(b)}$ . The bootstrap variance estimator is consistent for both smooth and nonsmooth estimators.

#### 5. IMPUTATION FOR NONRESPONSE

Item non-response is typically handled by imputation to fill in missing values. Marginal imputation methods include, regression or ratio imputation using auxiliary variables, and random imputation within imputation classes. Rao and Shao (1992) considered a weighted random imputation within imputation class which cut across sampled clusters. They derive the following estimator which is design-unbiased under uniform response within imputation class:

$$\hat{Y}_I = \sum_v \sum_{hik \in s_{vr}} w_{hik} y_{hik} + \sum_v \sum_{hik \in s_{vm}} w_{hik} y_{hik}^*$$

where  $y_{hik}^*$  is an imputed value,  $v$  represents imputation class, and  $s_{vr}$  and  $s_{vm}$  denote the sample of respondents and nonrespondents within imputation class  $v$ .

##### 5.1 Jackknife

Rao and Shao (1992) demonstrate the need to adjust  $y_{hik}^*$  by the amount

$$E_*^{(gj)} y_{hik}^* - E_* y_{hik}^*,$$

where  $E_*$  and  $E_*^{(gj)}$  are the expectation under random imputation given the donor set and the same quantity given the donor set with the  $gj$ -th cluster removed. Letting  $\hat{Y}_I^{(gj)}$  be the imputed estimator recalculated with the  $gj$ -th cluster removed and using the adjusted imputed values, a correct jackknife variance estimator is

$$v_j(\hat{y}_j) = \sum_{g=1}^L \frac{(n_g - 1)}{n_g} \sum_{j=1}^{n_g} (\hat{y}_{j(g)}^a - \hat{y}_j)^2.$$

Rao and Shao (1992) establish the design-consistency of  $v_j$ .

## 5.2. Balanced Half-samples and Repeated Half-Sample Bootstrap Using Re-imputation

A simple idea is to resample from the imputed data set, and then use the obtained respondents in the resample to re-impute the obtained imputed values using the original random imputation method. One would hope to thus capture both the sampling variation and the extra variation due to the random imputation. This works well provided: the  $n_h/(n_h - 1)$  goes to 1, or the resampling method is approximately design-unbiased, has resample size  $n_h$  within each stratum, and requires no rescaling. Thus, if  $n_h/(n_h - 1)$  is small, one could just use the Rao-Wu bootstrap or the BRR applied to the imputed data set and then re-impute the imputed values using the resampled respondents (see Shao and Sitter, 1996).

In practice,  $n_h \geq 4$  seems to be enough since this aspect only induces a bias in the component of the variance due to imputation, which is typically a small portion of the overall variance. However, the case of small  $n_h$  is an important special case. Two resampling schemes that meet the stated requirements and thus can handle small  $n_h$  but remain valid for arbitrary  $n_h$  are proposed in Saigo, Shao and Sitter (2001):

### A.A Repeated Half-sample Bootstrap:

Assume  $n_h = 2m_h$ . For each bootstrap sample take a simple random sample without-replacement of size  $m_h$  from  $n_h$  for  $h=1, \dots, L$  and repeat each obtained unit twice. Recalculate the estimator using the twice repeated obtained units and the original weights, repeat a large number of times and obtain the bootstrap variance estimator as described previously. One can handle odd  $n_h$  by a randomization strategy.

Using this method in stratified multi-stage sampling with no imputation yields valid bootstrap variance estimates, has a resample size equal to the original sample size and requires no rescaling. If one applies this with the strategy of re-imputation to imputed data the bootstrap is valid as  $L$  gets large for bounded  $n_h$ .

### B.A Repeated BRR or a type of balanced bootstrap

For  $n_h = 2$ , instead of obtaining a half-sample and re-weighting, obtain a half-sample and repeat each obtained unit twice. The BRR remains valid with no imputation and now has a resample size equal to the original sample size and requires no reweighting. It can also be viewed as a type of balanced bootstrap. Similarly for balanced half-samples obtained using BOMA's.

If applied to a randomly imputed data set and combined with re-imputation, a valid variance estimate results. Note that Nigam and Rao (1996) derive balanced bootstraps for stratified multi-stage sampling that require rescaling.

One cautionary note on using re-imputation bootstrap and BRR methods. The bootstrap variance estimator

$$v_B = B^{-1} \sum_{b=1}^B (\hat{\theta}_{1(b)} - \bar{\theta}_{1(\cdot)})^2$$

is valid, however the more typically used

$$v_B = B^{-1} \sum_{b=1}^B (\hat{\theta}_{1(b)} - \hat{\theta}_1)^2$$

is not. The reason is that  $\hat{\theta}_1$  is a single realization of a random imputation, whereas  $\bar{\theta}_{1(\cdot)} \approx E_1(\hat{\theta}_1)$ . These are not close for random imputation.

## 6. CONCLUDING REMARKS

I conclude this expository by listing a few important topics in Jon Rao's body of work on resampling methods in surveys which I have neglected. One is resampling methods for two-phase sampling in Rao and Sitter (1995, 1997) and Sitter and Rao (1997). A second is the work on resampling methods for poststratification (see Yung and Rao, 1996). Finally, Jon's work on the linearized jackknife (see Rao, 1997).

## REFERENCES

- Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, Supplement, 25-45.
- Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

- McCarthy, P.J. (1969). Pseudoreplication half samples. *Review International Statistical Institute*, 37, 239-264.
- Nigam, A.K. and Rao, J.N.K. (1996). On balanced bootstrap for stratified multistage samples. *Statistica Sinica*, 6, 199-214.
- Rao, J.N.K. (1997). Developments in sample survey theory: an appraisal. *The Canadian Journal of Statistics*, 25, 1-21.
- Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rao, J.N.K. and Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.
- Rao, J.N.K. and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.
- Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Rao, J.N.K. and Sitter, R.R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. *Survey Measurement and Process Quality: Wiley Series in Probability and Statistics*, (Ed. L. Lyberg, P. Biemer, M Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), pages 753-768, John Wiley and Sons Inc.
- Rao, J.N.K. and Wu, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.
- Rao, J.N.K. and Wu, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data: some recent work. *Bull. Internat. Statist. Inst.* Proceedings of the 46th session, Book 3, 5-19.
- Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- Rust, K.F. and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Saigo, H., Shao, J. and Sitter, R.R. (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology*, to appear.
- Shao, J. and Rao, J.N.K. (1993a). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyà A*, 55, 393-414.
- Shao, J. and Rao, J.N.K. (1993b). Jackknife inference for heteroscedastic linear regression models. *The Canadian Journal of Statistics*, 21, 377-395.
- Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Sitter, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.
- Sitter, R.R. and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation in survey data. *The Canadian Journal of Statistics*, 25, 61-73.
- Yung, W. and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multistage sampling. *Survey Methodology*, 22, 23-31.