Impact of Randy Sitter’s Contributions to Survey Sampling Theory and Practice

J. N. K. Rao

Carleton University

Memorial Session for Randy Sitter, SSC 2009
Vancouver, June 1, 2009

Outline

- Some reminiscences
- Early work on bootstrap
- Balanced repeated replication: BOMA
- Jackknife for two-phase sampling
- Jackknife and bootstrap under imputation for missing data
- Constructing combined strata variance estimators
- Empirical likelihood methods for complex surveys

Early work on bootstrap:

Stratified SRSWOR: BWO (Gross 1980), BWR (Rao and Wu 1988: rescaling)

Mirror-match method: Sitter (1992 JASA, 1989 Ph.D. thesis)

Assume $N_h = n_h k_h, n'_h = f_h n_h$ with both k_h and n'_h integers greater than or equal to1. Integer assumption handled by randomization if necessary.

Proposed bootstrap method:

1. Resample n'_h SRSWOR from stratum h sample
2. Repeat step 1 k_h times independently
3. Repeat steps 1 and 2 independently for each stratum
4. Repeat steps 1-3 a large number of times, B, to get bootstrap estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ of θ
5. Estimate the variance of $\hat{\theta}$ as

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

Linear case: second and third moment matching, captures second term of Edgeworth expansion

Modification: Choose $1 \leq n'_h < n_h$ in step 1 and $k_h = [n_h(1 - f_h^*)/[n'_h(1 - f_h)]$ in step 2 where $f_h^* = n'_h / n_h$.

Linear case: second moment matching only. BWO method of McCarthy and Snowden (1985) special case: $n'_h = 1$. Randomization is used to handle non-integer k_h

Extensions: Two-stage cluster sampling: SRSWOR at both stages, Rao-Hartley-Cochran (RHC) method for PPS sampling without replacement

Balanced Repeated Replications based on Orthogonal Multi Arrays (BOMA): Sitter (Biometrika 1993)

Balanced half-sample replication for stratified multi-stage sampling with $n_h = 2$ clusters in each stratum $h = 1,..,L$ proposed by McCarthy (1969). Use Hadamard matrix of order R such that $L + 1 \leq R \leq L + 4$ and R is a multiple of 4 to construct half-sample replicates and associated estimates of θ. BHS ensures second moment matching in the linear case.

General case: Orthogonal array OA (R, $n_1 \times .. n_L$)of strength 2 ensures second moment matching in the linear case (Wu 1989) but number of re-samples R is excessively large.

BOMA $(R; n_1, ..., n_L; \alpha_1, ..., \alpha_L)$ reduces the number of re-samples considerably and yet retains second moment matching property. Here α_h is size of subset of n_h elements in stratum h. The choice $\alpha_h = 1$ gives the usual OA.

Construction of BOMA:

$n_h = p = 4m, L+1 = 4m'$

$B:(L+1)\times L$ Hadamard matrix removing column of +s

$C:(p-1)\times p$ Hadamard matrix removing row of +s

Then BOMA $((p-1)(L+1), p^L, (p/2)^L)$ is given by
$A = B \otimes C$

Example: $L = 7, p = 4$ gives BOMA $(24, 4^7, 2^7)$. Note $R = 24$ compared to $R = 32$ for OA.

$$B = \begin{matrix} + & - & + & - \\ + & - & - & + \\ + & + & - & - \\ + & + & + & + \end{matrix} \qquad +C = \begin{matrix} (1,3) \\ (1,4) \\ (1,2) \end{matrix} \;, \quad -C = \begin{matrix} (2,4) \\ (2,3) \\ (3,4) \end{matrix}$$

A more general method for $n_h = p$ also proposed based on a resolvable BIBD

Jackknife Variance Estimation for Two-phase Sampling (Rao and Sitter, Biometrika 1995):

SRS at first phase and x observed on a sample s' of size n' giving sample mean $\bar{x}'$. A small sub-sample s of size n selected again by SRS and y is observed: $(\bar{y}, \bar{x})$ sub-sample means.

Ratio estimator of the mean $\bar{Y}$: $\bar{y}_r = (\bar{y}/\bar{x})\bar{x}'$

Jackknife: Delete each of the units j in the first-phase sample in turn to get

$$\bar{y}_r(j) = \frac{\bar{y}(j)}{\bar{x}(j)}\bar{x}'(j),\ j \in s \qquad \bar{y}_r(j) = \frac{\bar{y}}{\bar{x}}\bar{x}'(j),\ j \in s' - s$$

Variance estimator: $v_J = \frac{n'-1}{n'} \sum_{j \in s'} \{\bar{y}_r(j) - \bar{y}\}^2$

Jackknife and the corresponding jackknife linearization variance estimators have good conditional properties.

- Extension to regression estimators: Sitter, JASA 1997
- Application to estimating measurement error bias using two-phase sampling: jackknife and bootstrap variance estimation using modification of re-scaling method (Rao and Sitter, 1997).
- Efficient replication variance estimation for two-phase sampling (Kim and Sitter, Statistica Sinica 2003): Replicate variance estimators with number of replications slightly larger than the size of the second-phase sample.

Jackknife variance estimation under imputation for missing data:

- Mass imputation of y under two-phase sampling (Rao and Sitter, Biometrika 1995): ratio imputation $\hat{y}_i = (\bar{y}/\bar{x})x_i$.

  Let $y_i^* = y_i, i \in s'; y_i^* = \hat{y}_i, i \in s' - s$.

  Imputed estimator: $\bar{y}_I = \frac{1}{n'}\sum_{i=1}^{n'} y_i^* = \bar{y}_r$ = ratio estimator
- Jackknife variance estimation: use Rao-Shao (1992) adjusted imputed values $\{\bar{y}(j)/\bar{x}(j)\}x_i$ when $j \in s$ is deleted and imputed values remain unchanged otherwise.

- Use the same idea under non-response: x always observed but y is missing completely at random or a ratio imputation model holds and probability of missing can depend on x (MAR assumption). Jackknife remains design-consistent under uniform response as well as design-model unbiased without actually specifying the response mechanism (doubly protective).
- Extension to the case of different imputations when x also may be missing: Sitter and Rao (CJS, 1997).

- Chen, Rao and Sitter (Statistica Sinica 2000): Elimination of imputation variance under random imputation and still preserve the distribution. Jackknife variance estimation based on adjusted imputed values (Rao and Shao 1992) is used.

Bootstrap variance estimation under imputation for missing survey data

1. Shao and Sitter (JASA 1996) propose a bootstrap method to handle variance estimation for stratified multi-stage designs under imputation for missing data. It avoids adjusted imputed values used in the jackknife method of Rao and Shao (1992).
2. Idea is to re-impute the bootstrap data set in the same way as the original data set is imputed. Under random imputation, this method requires that $n_h/(n_h-1)$ goes to 1 which is not valid when n_h is small.

3. Saigo, Shao and Sitter (SMJ 2001) get around this difficulty by proposing a modified bootstrap method. If $n_h = 2m_h$, draw m_h PSUs by SRSWOR and repeat each obtained unit twice to get bootstrap sample of size n_h. Use this bootstrap in step 1 and follow step 2. For variance estimation using B bootstrap samples, one should not take deviations from the full sample imputed estimate.

Instead the mean of the bootstrap imputed estimates should be used.

Constructing combined strata variance estimators under stratified multi-stage sampling

- Combined variance strata method is a way to reduce the number of replications with the Jackknife or the BRR. Each deletion is done simultaneously in a combined stratum.
- Lu, Brick and Sitter (JASA 2006): “Optimal” grouping using algorithm from scheduling theory.
- Lu and Sitter (Statistica Sinica 2007): Application to minimizing disclosure risk associated with a replicate weights data file.

Empirical Likelihood Methods for Complex Surveys:

- Chen and Sitter (Statistica Sinica 1999): Pseudo empirical likelihood (PEL) approach to inference from survey data.
- Wu and Sitter (JASA 2001), Sitter and Wu (JASA 2002): Model-calibration approach; Model-calibrated pseudo empirical likelihood approach.
- Chen, Sitter and Wu (Biometrika 2002): Efficient algorithms for empirical likelihood; Obtain range-restricted survey weights through PEL.
- Empirical likelihood method to raking: Talk at JSM 2006 (work in progress with Changbao Wu)

Concluding remarks:

- Truly amazing that Randy Sitter made so many major contributions to sample survey theory and methods within 15 years after receiving his PhD degree.
- His contributions to design of experiments (particularly to industrial problems) are equally fundamental.
- Editorial contributions: Editor of Technometrics, Associate editor of Biometrics, Survey Methodology and Canadian Journal of Statistics.
- Randy Sitter’s tragic death at such young age is a great loss to statistical community.