Available online at www.isas.org.in/jisas
Journal of the Indian Society of
Agricultural Statistics 67(3) 2013 301-304

Panel Rotation with General Sampling Schemes

Arijit Chaudhuri
Indian Statistical Institute, Kolkata

Received 04 August 2012; Revised 06 February 2013; Accepted 06 February 2013

SUMMARY

We consider two consecutive nearby occasions over which a finite survey population changes little in composition. The problem is to estimate the current population total on surveying the population on the previous occasion through a well-designed sampling scheme. Then retaining by probability sampling a part of it as a matched sample for which both the past and the current values are ascertained. A Double-sampling theoretic estimate is first obtained for the current total, a current sample is then gathered independently of what precedes yielding another estimate for the current total.

These two estimates are then appropriately combined into a pooled estimate as an improved one. Horvitz and Thompson's (HT, 1952) and generalized regression methods due to Cassel, Sarndal and Wretman (CSW, 1976) provide basic estimation procedures. As these do not yield a variance explicitly in terms of the sample-size we conclude non-availability of an optimal "Matching Sampling Fraction" (MSF) formula. Estimated coefficient of variation (CV) is derived to guide 'Rotational policy formulation'. For specific unequal probability sampling scheme due for example to Rao, Hartley and Cochran (RHC, 1962) however MSF may be worked out as we have shown in a separate conference paper. In this paper a solution is worked out under a postulated model.

Keywords: Double sampling, Matching sampling fraction, Sampling on successive occasions, Unequal probability sampling.

1. INTRODUCTION

On the first occasion let a survey population U_1 have N_1 units with values x_i, $i \in U_1$ on a variable x of interest. Let this population change into U_2 of N_2 units with values y_i for i in U_2 on the same variate; let this now be denoted as y noting the shift in time. Let $X = \sum_1^{N_1} x_i$, $Y = \sum_1^{N_2} y_i$ be the totals of x and y respectively. Our objective is to suitably estimate Y on surveying well-chosen samples from U_1 and U_2 .

Let us further suppose that some normed size-measures $p_i (0 < p_i < 1, \sum_1^{N_1} p_i = 1)$ well and positively correlated with x_i's and like-wise $q_i (0 < q_i < 1, \sum_1^{N_2} q_i = 1)$ correlated with y_i's be available for utilization. Let from U_1 a sample s_1 of a size n_1 be drawn admitting positive inclusion-probabilities $\pi_i (0 < \pi_i < 1, \sum_1^{N_1} \pi_i = n_1)$ for i in U_i and positive-inclusion-probabilities $\pi_{ij} [\sum_{j=1}^{N_1} \pi_{ij} = (n_1 - 1)\pi_i, \sum_{i \neq j}^{N_1} \sum^{N_1} \pi_{ij} = n_1 (n_1 - 1)]$ for $i, j (i \neq j) = 1, ..., N_1$].

Similarly, let from U_2 be drawn a sample s_2 of size n_2 independently of s_1 with positive inclusion-

E-mail address: arijitchaudhuri1@rediffmail.com probabilities π''_i for i in U_2 and π''_{ij} for i, j in U_2 ($0 < \pi''_i < 1$, $\sum_1^{N_2} \pi''_i = n_2$, $\sum_{j \neq 1}^{N_2} \pi''_{ij} = (n_2 - 1)\pi''_i$ and $\sum_{i \neq j}^{N_2} \sum^{N_2} \pi''_{ij} = n_2(n_2 - 1)$). More importantly, let from s_1 be drawn a sub-sample s_m of size m with positive inclusion-probabilities π'_i ($0 < \pi'_i < 1$, $\sum_1^m \pi'' = m$) for i in s_1 and π'_{ij} ($0 < \pi'_{ij} < 1$, $\sum_{j=1}^{n_1} \pi'_{ij} = (m-1)\pi'_i$ and $\sum_{i \neq j} \sum \pi'_{ij} = m(m-1)$ for i, j ($i \neq j$) in s_1. In section 2 we present suitable estimators.

2. ESTIMATING THE CURRENT TOTAL

For X let us consider the (Horvitz and Thompson's (1952)) HT estimator $e_1 = \sum_{i \in s_1} \frac{x_i}{\pi_i}$ which is unbiased for X. Based on the matched sample s_m, both (x_i, y_i) values for $i \in s_m$ are obviously ascertainable. For $\sum_{i \in s_1} \frac{x_i}{\pi_i}$ the HT estimator based on s_m is $\sum_{i \in s_m} \frac{x'_i}{\pi'_i}$, writing $x'_i = \frac{x_i}{\pi_i}$, $i \in s_1$. Writing $y'_i = \frac{y_i}{\pi_i}$ let the following model be postulated enabling us to write

$$y'_i = \beta x'_i + \epsilon_i, i \in s_1; \tag{2.1}$$

[This is adopted following model (6.4.4) and its special case (6.4.5) given on p. 226 of chapter 6 of Särndal, Swensson and Wretman (SSW, 1992)]

Here β is an unknown constant and ϵ_i's are independently distributed with zero means and unknown variances σ_i^2, $i \in s_1$. Following Cassel, Särndal and Wretman (CSW, 1976) let us take

$$b_Q = (\sum_{i \in s_m} y'_i x'_i Q_i) / (\sum_{i \in s_m} (x'_i)^2 (Q_i) \tag{2.2}$$

with Q_i as suitably chosen positive numbers like $Q_i = \frac{1}{x'_i}$ or $\frac{1}{(x'_i)^2}$ or $\frac{1}{\pi'_i x'_i}$ or $\frac{1 - \pi'_i}{\pi'_i x'_i}$, for example. Then, the generalized regression estimator or the Greg estimator in brief for Y is

$$t_1 = \sum_{i \in s_m} \frac{y'_i}{\pi'_i} + b_Q (\sum_{i \in s_1} x'_i - \sum_{i \in s_m} \frac{x'_i}{\pi'_i}) \tag{2.3}$$

Next let us write E_C, V_C as operators for expectation and variance over the selection of the matched sample s_m conditionally on s_1 and x'_i for $i \in s_1$ remaining fixed. By E_u, V_u we shall mean expectation, variance operators in respect of unconditional sampling of s_2 from U_2. Also by E, V we shall mean expectation, variance operators over the selection of s_1 from U_1. Then, using Brewer's (1979) asymptotic approach utilized by CSW and writing

$$B_Q = (\sum_{i \in s_1} y'_i x'_i Q_i \pi'_i) / (\sum_{i \in s_1} (x'_i)^2 Q_i \pi'_i), \tag{2.4}$$

we have, approximately,

$$E_C(t_1) = \sum_{i \in s_1} y'_i \text{ and } E(t_1) = Y \tag{2.5}$$

Also, approximately,

$$\begin{aligned} t_1 &= \sum_{i \in s_m} \frac{y'_i}{\pi'_i} + B_Q (\sum_{i \in s_1} x'_i - \sum_{i \in s_m} \frac{x'_i}{\pi'_i}) \\ &= \sum_{i \in s_m} (\frac{y'_i - B_Q x'_i}{\pi'_i}) + B_Q \sum_{i \in s_1} x'_i \\ &= \sum_{i \in s_m} \frac{E'_i}{\pi'_i} + B_Q \sum_{i \in s_1} x'_i, \end{aligned}$$

writing

$$E'_i = y'_i - B_Q x'_{ij} \tag{2.6}$$

Then,

$$V_C(t_1) \simeq \sum_{i<j \in s_1} \sum (\pi'_i \pi'_j - \pi'_{ij}) \left(\frac{E'_i}{\pi'_i} - \frac{E'_j}{\pi'_j} \right)^2 \tag{2.7}$$

and writing $e'_i = y'_i - b_Q x'_i$, an estimator for

$$\begin{aligned} V(t_1) \simeq &\sum_{i<j \in U_1} \sum \pi_{ij} \left[(\pi'_i \pi'_j - \pi'_{ij}) \left(\frac{E'_i}{\pi'_i} - \frac{E'_j}{\pi'_j} \right)^2 \right] \\ &+ \sum_{i<j \in U_1} \sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \end{aligned} \tag{2.8}$$

may be taken as

$$v(t_1) = \sum_{i<j \in s_m} \sum \left(\frac{\pi'_i \pi'_j - \pi'_{ij}}{\pi'_{ij}} \right) \left(\frac{e'_i}{\pi'_i} - \frac{e'_j}{\pi'_j} \right)^2$$

$$+\sum_{i<j\in s_m}\left(\frac{\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{1}{\pi'_{ij}} \tag{2.9}$$

Based on s_2 and (s_2, y_i, $i \in s_2$) let for Y the HT unbiased estimator be considered, namely t_2 = $\sum_{i\in s_2}\frac{y_i}{\pi''_i}$. Its variance is

$$V_U(t_2)=\sum i<j\in U_2\sum(\pi''_i\pi''_j-\pi''_{ij})\left(\frac{y_i}{\pi''_i}-\frac{y_j}{\pi''_j}\right)^2 \tag{2.10}$$

of which an unbiased estimator is

$$v(t_2) = \sum_{i<j\in s_2}\sum\left(\frac{\pi''_i\pi''_j-\pi''_{ij})}{\pi''_{ij}}\right)\left(\frac{y_i}{\pi''_i}-\frac{y_j}{\pi''_j}\right)^2 \tag{2.11}$$

Now, let us take for the approximately unbiased pooled estimator

$$\overline{t}=\frac{(t_1)/v(t_1)+(t_2)/v(t_2)}{\dfrac{1}{v(t_1)}+\dfrac{1}{v(t_2)}} \tag{2.12}$$

Then,

$$V(\overline{t})=1\bigg/\left(\frac{1}{v(t_1)}+\frac{1}{v(t_2)}\right)=\frac{V_1V_2}{V_1+V_2} \tag{2.13}$$

writing $V_1 = v(t_1)$, $v_2 = v(t_2)$, is the approximate variance of $\overline{t}$ and we may take $v = \frac{\hat{V}_1\hat{V}_2}{\hat{V}_1+\hat{V}_2}$ as an estimator for $V(\overline{t})$.

Of course, $\hat{V}_1=v(t_1)$ and $\hat{V}_2=v(t_2)$

So,

$$v=\frac{v(t_1)v(t_2)}{v(t_1)+v(t_2)}. \tag{2.14}$$

Finally $CV=100\frac{\sqrt{v}}{\overline{t}}$ is to be taken as an estimated coefficient of variation for $\overline{t}$.

3. PANEL ROTATION POLICY DETERMINATION

With this general approach unless a "sampling scheme" is specified, it may not be possible to minimize $V(t_1)$ in respect of m. As a consequence, an appropriate choice of the "Matching Sampling Fraction" (MSF), namely $\frac{m}{n_1}$ cannot be worked out. Brewer and Hanif (1983) and Chaudhuri and Vos (1988) have narrated many sampling schemes to be employed while estimating a population total by Horvitz and Thompson's (HT, 1952) method.

It is however clear that $\frac{m}{n_1}$ adapted from SSW (1992) should be such that $V_C(t_1)$ is controlled to the extent possible. With this objective in mind let us modify the model (2.1) by postulating that the model-based variance $V_m(\in_i)$ be taken as $V_m(\in_i) = \sigma^2 x_i'$, $i \in s_1$ for every sample s_1.

Postulating this model, say, (3.1), let us minimize $E_mV_C(t_1)$, denoting by E_m the expectation calculated under the model (3.1). Observing under Model (3.1) that $B_Q=\beta+\frac{\Sigma_{i\in s_1}\in_i x'_iQ_i\pi'_i}{\Sigma_{i\in s_1}(x'_i)^2Q_i\pi'_i}$, $E'_i = y'_i - B_Q x'_i$ = $\in_i\left[1-\frac{\Sigma_{i\in s_1}\in_i x_iQ_i\pi'_i}{\Sigma_{i\in s_1}(x'_i)^2Q_i\pi'_i}\right]$ it follows that we may write

$$E_m\left(\frac{E'_1}{\pi'_i}-\frac{E'_j}{\pi'_j}\right)^2$$

$$=E_m\left[\left(\frac{\in_i}{\pi'_i}-\frac{\in_j}{\pi'_j}\right)\left(1-\frac{\Sigma_{i\in s_1}\in_i x'_iQ_i\pi'_i}{\Sigma_{i\in s_1}(x'_i)^2Q_i\pi'_i}\right)\right]^2 \tag{3.2}$$

From now on we shall take $Q_i=\frac{1}{x'_i}$

Then,

$$E_mV_C(t_1) = \sigma^2(\sum\frac{x'_i}{\pi'_i}-\sum_{s_1}x'_i)$$

$$+\sigma^4(\sum_{s_1}\frac{x'_i}{\pi'_i}-\sum_{s_1}x'_i)\frac{\sum_{s_1}x'_i{\theta^2}_i}{(\sum_{s_1}x'_i\theta_i)^2}. \tag{3.3}$$

Let for simplicity,

$$\pi'_i = m\theta_i \tag{3.4}$$

with known θ_i such that $0 < \theta_i < 1$ $\forall\, i\in s_1$ and $\Sigma_{i\in s_1}\theta_i=1$ for every s_1 with a positive selection-probability.

Then, under Model (3.1) and design (3.4),

$$E_m V_C(t_1) \simeq \frac{A}{m} - B$$

$$\text{With} \quad A = \sigma^2 \left(\sum_{s_1} \frac{x_i'}{\theta_i}\right)\left[1 + \sigma^2 \frac{\sum_{s_1} x_i'\theta_i^2}{\left(\sum_{s_1} x_i'\theta_i\right)^2}\right]$$

$$\text{and} \quad B = -\sigma^2 \left[\sum_{s_1} x_i + \sigma^2 \frac{\sum_{s_1} x_i'\theta_i^2}{\left(\sum_{s_1} x_i'\theta_i\right)}\right] \qquad (3.5)$$

Then, (3.5) diminishes with increasing m.

By a panel in a survey we mean the set of units to be surveyed on a certain occasion. By its rotation we mean dropping a part of a panel for a survey on another occasion and allowing their re-emergence on another.

Our motivation in this paper is to start with an initial panel of n_1 units, retain a part of this, namely a sub-sample of m units of it, dropping $(n_1 - m)$ the remaining units initially sampled and adding a fresh set of n_2 units, the total panel to be surveyed on the second occasion is to consist of $n_2 + m$ units instead of the initial n_1 units, though m of the units are to remain common on both the occasions. Questions are which m to be surveyed on both the occasions and how many are they in number. We presume we have provided answers to both.

4. A SIMULATION-BASED EMPIRICAL EXERCISE TO CHOOSE MSF

Data Source: The Clustered MU 284 Population

Särndal-Swensson-Wretman (pp—660-661)

N = 50 clusters

Z --- size measure variable = Number of municipalities in the clusters

y: total 1985 population

x: total 1975 population

Y = 8339.00

X = 7992.88

Z = 284

The original and the matched samples are chosen following Hartley and Rao's (1962) method using the same size-measures.

Sample size	Estimate = t_1	Variance of t_1
Original Sample size = 19 Matched sample size = 5	8105.218	7805.53
Original Sample size = 19 Matched sample size = 7	8451.674	5307.29
Original Sample size = 19 Matched sample size = 11	8176.660	5023.92

Obviously, increasing m one achieves enhanced efficiency.

ACKNOWLEDGEMENTS

The author is grateful to the referee for help in revising an earlier submission with probable improvements. Also Dr. Sanghamitra Pal's help in implementing the Section 4 for the revision is deeply appreciated.

REFERENCES

Brewer, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Jour. Amer. Stat. Assoc.,* **74(368),** 911-915.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer-Verlag, New York, U.S.A.

Chaudhuri, A. (2010). *Essentials of Survey Sampling*. Prentice Hall of India, Delhi, India.

Chaudhuri, A. and Vos, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. North-Holland, NL.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika,* **63**, 615-620.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Stat. Math.* **33**, 350-374.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a universe. *Jour. Amer. Stat. Assoc.,* **47**, 663-685.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Jour. Roy. Stat. Soc.,* **B24**, 482-491.

Särndal, C.E, Swensson, D.E. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.