Austral. J. Statist. **38**(1), 1996, 35–42

A NOTE ON COMPETING VARIANCE ESTIMATORS IN RANDOMISED RESPONSE SURVEYS

ARIJIT CHAUDHURI[1], TAPABRATA MAITI[1] & DEBESH ROY[2]

Indian Statistical Institute and Presidency College

Summary

When gathering randomised rather than direct responses on a variable of interest relating to sensitive issues, one may use a modified version of the well-known generalised regression predictor of a finite population total. To construct confidence intervals, this paper proposes four alternative variance estimators — modifications to those usable with direct responses — and examines their relative efficiencies through simulations from simple super-population models.

Key words: Generalised regression predictor; randomised response; variance estimation.

1. Introduction

We consider a sample survey to estimate population totals of several variables including a few that could give a person a bad name, such as amount spent on gambling or alcoholism or number of days of drunken driving etc.

Usually a population is stratified and from each stratum a sample is drawn according to a suitable design, 'independently' across the strata. So, each stratum theoretically may be treated as a population in itself. Accordingly, we present a theory for estimating a 'population' total. For each variable of interest y we assume it is possible to identify a correlated variable x with known population values x_i totalling X. From the population $U = (1, \ldots, i, \ldots, N)$ of size N a sample, s, of n distinct units, is assumed to be drawn with a probability $p(s)$ according to an appropriately chosen design p. For the design p, the probabilities π_i and π_{ij} respectively of including i and i, j $(i \neq j)$ in the sample are assumed to be positive. By E_p, V_p we denote design based operators of expectation and variance.

If y is a non-stigmatizing variable, then its value y_i for a unit i in s may be directly ascertained by survey. On the other hand, if the variable could stigmatize a person, we assume that a randomised experiment may be implemented to

Received February 1993; revised July 1995; accepted October 1995.
[1]Computer Science Unit, Indian Statistical Institute, 203 BT Road, Calcutta–700035, India.
[2]Dept of Statistics, Presidency College, Calcutta–700073, India.
Acknowledgements. The authors thank the referees and an Associate Editor for helpful comments on earlier drafts. The work of the second author is supported by grant no.9/93(35)/95–EMR–I of CSIR, India.

produce a 'randomised response' (RR), say r_i, for i in s, and that a direct response (DR) yielding y_i may not be obtained. As described by Chaudhuri & Mukerjee (1988) a suitable RR technique may be employed so that, writing E_R, V_R as operators of expectation and variance for the 'randomisation' experiment one may have

$$E_R(r_i) = y_i, \quad V_R(r_i) = \alpha_i y_i^2 + \beta_i y_i + \theta_i = V_i \quad \text{say} \quad (i \in U), \tag{1.1}$$

for known α_i, β_i, θ_i. The r_i are assumed to be 'independent' over $i \in U$. In addition,

$$\widehat{V}_i = \frac{1}{1+\alpha_i}(\alpha_i r_i^2 + \beta_i r_i + \theta_i) \qquad (1 + \alpha_i \neq 0), \tag{1.2}$$

satisfies

$$E_R(\widehat{V}_i) = V_i \qquad (i \in U). \tag{1.3}$$

With this set up we postulate a 'super-population' linear regression model $\underline{M}$ permitting us to write

$$y_i = \beta x_i + \epsilon_i \qquad (i \in U). \tag{1.4}$$

Here β is an unknown constant, the ϵ_i are 'random' variates distributed with means and variances respectively as

$$E_m(\epsilon_i) = 0, \qquad V_m(\epsilon_i) = \sigma^2 x_i^g, \tag{1.5}$$

with σ (> 0), g ($0 \leq g \leq 2$) unknown constants. By $\sum$, $\sum\sum$ we denote sums over i in U and i,j ($i < j$) in U; by $\sum'$, $\sum'\sum'$ we denote the same in s. If direct responses y_i are available, then a popular estimator for $Y = \sum y_i$ is

$$t_g = \sum{}' \frac{y_i}{\pi_i} + \hat{\beta}_Q\Big(X - \sum{}' x_i \pi_i\Big) = \sum{}' \frac{y_i}{\pi_i} g_{si},$$
$$g_{si} = 1 + \Big(X - \sum{}' \frac{x_i}{\pi_i}\Big)\frac{x_i Q_i \pi_i}{\sum' x_i^2 Q_i}. \tag{1.6}$$

Here $\hat{\beta}_Q = \sum' y_i x_i Q_i / \sum' x_i^2 Q_i$ and Q_i (> 0) are 'arbitrarily' assignable constants. This is called the 'generalised regression' (greg) predictor (Cassel *et al.* 1976). Särndal (1980), following Brewer's (1979) asymptotic approach, showed it to be 'asymptotically design unbiased' (ADU) for Y. Särndal (1982) gave two variance estimators for t_g as

$$v_k = \sum{}'\sum{}' \Delta_{ij}\Big(\frac{e_i}{\pi_i}a_{ki} - \frac{e_j}{\pi_j}a_{kj}\Big)^2 \qquad (k = 1, 2). \tag{1.7}$$

Here $e_i = y_i - \hat{\beta}_Q x_i$; $a_{1i} = 1$; $a_{2i} = g_{si}$ $(i \in U)$; $\Delta_{ij} = (\pi_i \pi_j - \pi_{ij})/\pi_{ij}$. If instead of y_i only r_i is available, we can estimate Y using e_g, the RR version of t_g, which is obtained from t_g by substituting r_i for each y_i $(i \in s)$. We write

$$\hat{\beta}_{Qr} = \frac{\sum' r_i x_i Q_i}{\sum' x_i^2 Q_i} \quad \text{and} \quad e_{ir} = r_i - \hat{\beta}_{Qr} x_i \quad (i \in s).$$

In Section 2 we present alternative formulae for variance estimators of

$$e_g = \sum{}' \frac{r_i}{\pi_i} + \hat{\beta}_{Qr}\Big(X - \sum{}' \frac{x_i}{\pi_i}\Big) = \sum{}' \frac{r_i}{\pi_i} g_{si}.$$

Note that in the same survey, for a given sample, both t_g and e_g may have to be used when dealing with specific (y, x) variables.

2. Variance Estimators and Their Relative Efficacies

A measure of error of e_g as an estimator of Y may be taken as

$$M = E_p E_R (e_g - Y)^2 = E_R E_p (e_g - Y)^2, \tag{2.1}$$

noting that E_p commutes with E_R. Then

$$\begin{aligned} M &= E_p(t_g - Y)^2 + E_p V_R(e_g) \\ &= E_p(t_g - Y)^2 + E_p\Big[\sum{}' \Big(\frac{g_{si}}{\pi_i}\Big)^2 V_i\Big]. \end{aligned} \tag{2.2}$$

Särndal (1982) showed that

$$E_p(v_k) \quad \text{approximates} \quad E_p(t_g - Y)^2. \tag{2.3}$$

Let

$$v_{kr} = \sum{}' \sum{}' \Delta_{ij} \Big(\frac{e_{ir}}{\pi_i} a_{ki} - \frac{e_{jr}}{\pi_j} a_{kj}\Big)^2 \qquad (k = 1, 2),$$

and observe that

$$E_R(e_{ir} - e_i)^2 = V_i + x_i^2 \frac{\sum' x_i^2 Q_i^2 V_i}{(\sum' x_i^2 Q_i)^2} - 2 \frac{x_i^2 Q_i V_i}{\sum' x_i^2 Q_i} = F_i, \quad \text{say;}$$

$$E_R(e_{ir} - e_i)(e_{jr} - e_j) = -x_i x_j \Big[\frac{Q_i V_i + Q_j V_j}{\sum' x_i^2 Q_i} - \frac{\sum' x_i^2 Q_i^2 V_i}{(\sum' x_i^2 Q_i)^2}\Big] = F_{ij}, \quad \text{say.}$$

Then,

$$E_R(v_{kr}) = v_k + \sum{}' \sum{}' \Delta_{ij} \Big[\Big(\frac{a_{ki}}{\pi_i}\Big)^2 F_i + \Big(\frac{a_{kj}}{\pi_j}\Big)^2 F_j - 2 \frac{a_{ki} a_{kj}}{\pi_i \pi_j} F_{ij}\Big].$$

Let $\widehat{F}_i$, $\widehat{F}_{ij}$ stand for F_i, F_{ij} with V_i replaced by $\widehat{V}_i$ $(i \in U)$ in the latter and for $k = 1, 2$

$$\hat{v}_{kg} = v_{kr} - \sum{}' \sum{}' \Delta_{ij} \Big[\Big(\frac{a_{ki}}{\pi_i}\Big)^2 \widehat{F}_i + \Big(\frac{a_{kj}}{\pi_j}\Big)^2 \widehat{F}_j - 2 \frac{a_{ki} a_{kj}}{\pi_i \pi_j} \widehat{F}_{ij}\Big] + \sum{}' \Big(\frac{g_{si}}{\pi_i}\Big)^2 \widehat{V}_i. \tag{2.4}$$

Then,

$$E_R(\hat{v}_{kg}) = v_k + \sum{}' \Big(\frac{g_{si}}{\pi_i}\Big)^2 V_i, \tag{2.5}$$

and hence $E_p E_R(\hat{v}_{kg})$ approximates M, see equations (2.2)–(2.3). So we propose $\hat{v}_{kg}$ $(k = 1, 2)$, as two variance estimators of e_g.

Alternatively, writing $R = \sum r_i$ we neglect the error in equating $E_p(e_g)$ to R. Recalling that E_p commutes with E_R we approximate M by

$$E_R V_p(e_g) + E_R(R - Y)^2 = E_R V_p(e_g) + \sum V_i. \tag{2.6}$$

To find an elegant approximation for $V_p(e_g)$ by a first order Taylor series expansion we proceed as follows.

Let

$$w_{1i} = r_i, \quad w_{2i} = x_i, \quad w_{3i} = r_i x_i Q_i \pi_i, \quad w_{4i} = x_i^2 Q_i \pi_i \quad \text{for } i \in U;$$

$$\widehat{T}_j = \sum{}' \frac{w_{ji}}{\pi_i}, \quad T_j = \sum w_{ji} \quad (j = 1, \ldots, 4),$$
$$\underline{\widehat{T}} = (\widehat{T}_1, \ldots, \widehat{T}_4), \quad \underline{T} = (T_1, \ldots, T_4).$$

Noting

$$e_g = \widehat{T}_1 + \frac{\widehat{T}_3}{\widehat{T}_4}(X - \widehat{T}_2) = f(\underline{\widehat{T}}), \quad \text{say},$$

e_g approximates $f(\underline{T}) + \sum_{j=1}^4 \lambda_j(\widehat{T}_j - T_j)$, where $\lambda_j = \delta f(\underline{\widehat{T}})/\delta \widehat{T}_j\Big|_{\underline{\widehat{T}} = \underline{T}}$. Writing

$$\phi_i = \sum_{j=1}^4 \lambda_j w_{ji}, \quad \hat{\lambda}_j = \lambda_j\Big|_{\underline{T} = \underline{\widehat{T}}}, \qquad \hat{\phi}_i = \sum_{j=1}^4 \hat{\lambda}_j w_{ji},$$

we approximate

$$V_p(e_g) \quad \text{by} \quad V_p\Big(\sum{}' \frac{\phi_i}{\pi_i}\Big).$$

So, using (2.6), (2.7) we propose the following additional variance estimators of e_g,

$$m_{1g} = \sum{}' \sum{}' \Delta_{ij}\Big(\frac{\hat{\phi}_i}{\pi_i} - \frac{\hat{\phi}_j}{\pi_j}\Big)^2 + \sum{}' \frac{\widehat{V}_i}{\pi_i}, \tag{2.8}$$

and

$$m_{2g} = \sum{}' \frac{\hat{\phi}_i^2}{\pi_i}\Big(\frac{1}{\pi_i} - 1\Big) + \sum{}' \sum{}' \frac{\hat{\phi}_i \hat{\phi}_j}{\pi_{ij}}\Big(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\Big) + \sum{}' \frac{\widehat{V}_i}{\pi_i}, \tag{2.9}$$

observing that $E_R E_p(m_{kg})$ $(k = 1, 2)$ approximate M.

To judge the relative merits of $\hat{v}_{kg}$, m_{kg} $k = 1, 2$ in terms of their variances is difficult. So, to discriminate among them we consider their respective efficacies in yielding confidence intervals for Y of the form

$$e_g \pm \tau_{\alpha/2}\sqrt{v}, \tag{2.10}$$

with v standing for $\hat{v}_{kg}$, m_{kg} $(k = 1, 2)$. Here, for a chosen α in (0,1), $\tau_{\alpha/2}$ stands for the $100\frac{\alpha}{2}$ point on the right tail of the distribution of τ which is the standard normal deviate. This confidence interval has a nominal confidence coefficient of $100(1-\alpha)$ and is constructed on the basis of usual convention of approximating the distribution of $(e_g - Y)/\sqrt{v}$ by that of τ. In Section 3 we present a numerical comparison of the relative performances, based on a simulation study, of the four alternative confidence intervals above.

3. A Numerical Exercise by Simulation

Take $N = 70$. To generate $\underline{Y} = (y_i, \ldots, y_i, \ldots, y_N)$, $\underline{X} = (x_i, \ldots, x_i, \ldots, x_N)$ subject to model $\underline{M}$ we generate u_i $(i = 1, \ldots, N)$ as random samples from the density

$$f(u) = \mu e^{-\mu u} \qquad (u > 0),$$

with several choices of μ $(\mu > 0)$, and generate τ_i $(i = 1, \ldots, N)$ from $N(0,1)$ and take $x_i = 10 + u_i$, $\epsilon_i = \sigma\tau_i x_i^{g/2}$ with various choices of σ (> 0) and g $(0 \leq g \leq 2)$. With several choices of β (> 0), we then generate $y_i = \beta x_i + \epsilon_i$ $(i = 1, \ldots, N)$. In order to draw samples from U, we take $n = 15$ and apply two well-known schemes, one due to Lahiri (1951) and the other due to Hartley & Rao (HR) (1962). Both require use of size-measures, z_i say $(i = 1, \ldots, N)$, positively well correlated with y_i. We generate $\underline{Z} = (z_i, \ldots, z_i, \ldots, z_N)$ taking $z_i = 8.2 + 0.65x_i^{\gamma}$ choosing $\gamma = 0.78$. To apply Lahiri's scheme we equivalently follow Midzuno (1952) and select a unit i of U with a probability proportional to z_i on the first draw and take a simple random sample without replacement (SRSWOR) of size $(n-1)$ from the remaining population. Formulae for π_i, π_{ij} are easily found. Hartley & Rao's scheme chooses a circular systematic sample of size n with probability proportional to z_i from U after randomly permuting the elements of U, ensuring an inclusion probability $nz_i/\sum z_i$ for i in U. These authors give formulae for π_{ij}. We apply Chaudhuri & Mukerjee's (1988) method to generate randomised responses in the following way. We choose two vectors of suitable real numbers $\underline{A} = (A_i, \ldots, A_h, \ldots, A_H)$, $\underline{B} = (B_i, \ldots, B_j, \ldots, B_J)$ with means μ_A $(\neq 0)$, μ_B and variances σ_A^2, σ_B^2. For a sampled individual i, an element, A_m say, is chosen randomly from $\underline{A}$ and 'independently' an element, B_l say, is chosen randomly from $\underline{B}$ and a 'randomised' response is elicited as ψ_i which is

$$\psi_i = y_i A_m + B_h. \tag{3.1}$$

This is repeated 'independently' for every i in s. Then, $r_i = (\psi_i - \mu_B)/\mu_A$ is generated. For such r_i the relation (1.1) is satisfied with $\alpha_i = \sigma_A^2/\mu_A^2$, $\beta_i = 0$ and $\theta_i = \sigma_B^2/\mu_A^2$ $(i \in U)$. In our numerical exercise presented in the table below we choose (μ, β, σ, g) as

(i) $(0.6, 2.0, 5.0, 1.4)$,
(ii) $(0.2, 2.0, 4.0, 1.5)$,
(iii) $(0.2, 1.5, 3.5, 1.6)$,
(iv) $(0.6, 2.5, 4.5, 1.7)$,
(v) $(0.2, 1.5, 3.5, 1.8)$,
(vi) $(0.4, 2.0, 4.0, 1.9)$.

The vectors $\underline{A}, \underline{B}$ are chosen as

(I) $\underline{A} = (1,1,1,1,1,1,1)$, $\underline{B} = (0,0,0,0,0,0)$,
(II) $\underline{A} = (42,36,50,30,45,28,52)$, $\underline{B} = (15,12,18,9,11,8)$,
(III) $\underline{A} = (100,102,99,105,101,98,103)$, $\underline{B} = (75,72,71,69,72,70)$,

where case I corresponds to DR (direct response). To calculate e_g we choose $Q_i = 1/\pi_i x_i$, take $F = 1000$ replicates of the samples for both the schemes and choose $\alpha = 0.05$ to construct 95% confidence intervals. By $\sum_r$ we denote summation over the replicates, and by $\widehat{M}$ we denote the approximations of M by (2.6). To evaluate the performances of (e_g, v) we consider the following three usual criteria:

1. ACP(actual coverage percentage) $\equiv$ the percent of replicated samples for which the confidence intervals actually cover Y. The closer it is to 95, with everything else at par, the better.
2. ACV (average coefficient of variation) $\equiv$ the average, over the replicates, of the values of $\sqrt{v}/e_g$. This reflects the length of confidence interval: the shorter the better.
3. ARB (absolute pseudo relative bias) $\equiv 1/F \sum_r |v - \widehat{M}|/\widehat{M}$: the smaller the better.

In Table 1 we present the values of $(ACP, 10^3 ACV, 10^3 ARB)$ for (e_g, v) with v as $\hat{v}_{kg}$, m_{kg} $(k = 1,2)$, corresponding to several combinations of choices of (μ, β, σ, g) as in (i)–(vi) and $\underline{A}, \underline{B}$ as in (I)–(III) noted above. The values based on Lahiri's scheme are given below those for the HR scheme.

Concluding comments: Like v_k $(k = 1,2)$, the variance estimators $\hat{v}_{kg}$ $(k = 1,2)$ and m_{1g} are suggested by Yates & Grundy's (1953) form while m_{2g} is suggested by Horvitz & Thompson's (1952). However m_{2g} seems to outperform its competitors in terms of ACP and ARB though not ACV. For DR as well as RR cases all the procedures seem quite acceptable and competitive. Midzuno's scheme is simpler than HR's and performs better. For Midzuno's scheme, π_i equals $[n - 1 + (N - n)z_i/\sum z_i]/(N - 1)$ which is rather close to n/N, the π_i-value for SRSWOR. Yet it is preferable in employing e_g, at least in the present context, over HR's with wider variation in π_i.

TABLE 1

Values of $(ACP, 10^3ACV, 10^3ARB)$ for (e_g, v) under several alternative situations

(e_g, v)	(I,i)			(II,i)			(III,i)		
$(e_g, \hat{v}_{1g})$	81.4	200	486	75.7	219	558	81.6	200	592
	91.1	249	377	85.9	268	480	91.1	249	383
$(e_g, \hat{v}_{2g})$	83.7	215	429	78.6	233	515	83.1	215	434
	91.3	250	386	85.4	203	483	91.4	250	391
(e_g, m_{1g})	83.7	215	429	77.3	228	531	83.1	215	434
	91.3	250	386	85.4	263	483	91.4	250	391
(e_g, m_{2g})	90.9	267	375	85.1	281	425	90.3	266	373
	91.3	250	386	85.4	263	483	91.4	250	391
(e_g, v)	(I,ii)			(II,ii)			(III,ii)		
$(e_g, \hat{v}_{1g})$	83.9	166	702	77.9	185	715	83.9	166	705
	92.4	199	791	84.7	214	813	92.3	199	794
$(e_g, \hat{v}_{2g})$	86.5	180	646	79.8	197	685	85.8	180	650
	92.9	201	780	85.5	216	805	92.6	201	784
(e_g, m_{1g})	86.5	180	646	78.8	192	699	85.8	180	651
	92.9	201	780	85.2	215	806	92.6	201	784
(e_g, m_{2g})	91.9	216	505	84.4	230	582	91.9	216	511
	92.9	201	780	85.2	215	806	92.6	201	784
(e_g, v)	(I,iii)			(II,iii)			(III,iii)		
$(e_g, \hat{v}_{1g})$	82.5	228	506	78.3	249	576	82.1	228	512
	91.4	279	438	86.3	294	536	91.5	279	446
$(e_g, \hat{v}_{2g})$	83.9	245	448	80.1	265	531	83.6	246	453
	91.8	280	443	86.0	295	535	91.6	280	450
(e_g, m_{1g})	83.9	245	448	79.6	260	545	83.6	246	453
	91.8	280	443	86.0	295	535	91.6	280	450
(e_g, m_{2g})	90.7	301	467	86.2	318	435	90.2	301	366
	91.8	280	443	86.0	285	535	91.6	280	450
(e_g, v)	(I,iv)			(II,iv)			(III,iv)		
$(e_g, \hat{v}_{1g})$	81.2	208	483	76.4	228	556	81.6	207	481
	91.1	255	373	85.9	268	476	91.1	255	378
$(e_g, \hat{v}_{2g})$	83.8	224	426	79.0	243	513	83.2	222	431
	91.4	255	381	85.6	269	478	91.5	256	386
(e_g, m_{1g})	83.8	224	426	77.5	238	528	83.1	222	431
	91.4	255	381	85.6	269	468	91.5	256	386
(e_g, m_{2g})	90.9	278	373	85.3	293	422	90.4	275	371
	91.4	255	381	85.6	269	478	91.5	256	386
(e_g, v)	(I,v)			(II,v)			(III,v)		
$(e_g, \hat{v}_{1g})$	84.4	266	564	79.9	302	627	84.0	266	569
	92.7	317	622	88.1	340	682	92.6	317	629
$(e_g, \hat{v}_{2g})$	86.7	287	500	83.3	323	579	86.0	288	505
	93.2	321	606	87.5	344	669	93.2	321	612
(e_g, m_{1g})	86.7	287	500	82.1	318	590	86.0	288	505
	93.2	321	606	87.5	344	670	93.2	321	612
(e_g, m_{2g})	92.2	345	388	87.6	379	472	91.9	345	391
	93.2	321	606	87.5	344	670	93.2	321	612
(e_g, v)	(I,vi)			(II,vi)			(III,vi)		
$(e_g, \hat{v}_{1g})$	82.6	287	478	79.2	299	557	81.1	289	480
	91.6	359	390	87.5	361	495	91.7	371	396
$(e_g, \hat{v}_{2g})$	84.2	309	423	81.4	319	513	83.7	311	428
	91.9	361	401	87.3	363	496	91.7	374	407
(e_g, m_{1g})	84.2	309	423	80.9	314	525	83.7	311	428
	91.9	361	401	87.3	363	496	91.7	374	407
(e_g, m_{2g})	90.9	378	370	87.1	383	422	90.6	381	368
	91.9	361	400	87.3	363	496	91.7	374	406

References

BREWER, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *J. Amer. Statist. Assoc.* **74**, 911–915.

CASSEL, C.M., SÄRNDAL, C.E. & WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620.

CHAUDHURI, A. & MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques.* New York: Marcel Dekker Inc.

HARTLEY, H.O. & RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Statist.* **33**, 350–374.

HORVITZ, D.G & THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.

LAHIRI, D.B. (1951). A method of sample selection providing unbiased ratio estimators. *Bull. Internat. Statist. Inst.* **32**, 451–454

MIDZUNO, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Statist. Math.* **3**, 99–107.

SÄRNDAL, C.E. (1980). On π-inverse weighting versus best linear weighting in probability sampling. *Biometrika* **67**, 639–650.

— (1982). Implications of survey design for generalized regression estimation of linear functions. *J. Statist. Plann. Inf.* **7**, 155–170.

YATES, F. & GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc. Ser. B* **15**, 253–261.