COMMUN. STATIST.—THEORY METH., 23(4), 1203-1214 (1994)

VARIANCE ESTIMATION IN MODEL ASSISTED SURVEY SAMPLING

ARIJIT CHAUDHURI AND TAPABRATA MAITI

Indian Statistical Institute
Calcutta - 700035, India,
and
University of Kalyani
Kalyani -741235, India.

Key Words and Phrases: Asymptotic analysis; greg predictor; randomized response; super-population model; survey population; variance estimation.

ABSTRACT

Two versions of Yates-Grundy type variance estimators are usually employed for large samples when estimating a survey population total by a generalized regression (Greg, in brief) predictor motivated by consideration of a linear regression model. Their two alternative modifications are developed so that the limiting values of the design expectations of the model expectations of variance estimators 'match' respectively the (I) model expectations of the Taylor approximation of the design variance of the Greg predictor and the (II) limiting value of the design expectation of the model expectation of the squared difference between the Greg predictor and the population total. The exercise is extended to yield modifications needed when randomized response (RR) is only available rather than direct response (DR) when one encounters sensitive issues demanding protection of privacy. A comparative study based on simulation is presented for illustration.
AMS subject classification: 62 D05.

1. INTRODUCTION

We consider a survey population $U = (1, \ldots, N)$ of N individuals labelled i bearing unknown values y_i and known positive values x_i with respective totals Y and X. The problem is to estimate Y on surveying a sample

s from U chosen according to a suitable design p with probability $p(s)$ having positive inclusion probabilities π_i, π_{ij} respectively for i and (i,j). A model is postulated as plausible for which one may write

$$y_i = \beta x_i + \epsilon_i, \ i \in U \tag{1}$$

Here β is an unknown constant, ϵ_i's are uncorrelated random variables with expectations $E_m(\epsilon_i) = 0$ and variances $V_m(\epsilon_i) = \sigma_i^2$, $i \in U$. By $\sum, \sum\sum$ we denote sums over i, $i,j(i < j)$ in U respectively and by $\sum', \sum'\sum'$ the same over those in s. By $E_p(V_p)$ we shall denote design expectation (variance) operator. Further, $\Delta_{ij} = \pi_i\pi_j - \pi_{ij}$ and $Q_i(> 0)$ are constants to choose at discretion,

$$\begin{aligned} \hat{\beta}_Q &= \frac{\sum' y_i x_i Q_i}{\sum' x_i^2 Q_i}, \quad e_i = y_i - \hat{\beta}_Q x_i, \\ B_Q &= \frac{\sum y_i x_i Q_i}{\sum x_i^2 Q_i}, \quad E_i = y_i - B_Q x_i. \end{aligned}$$

Then Särndal's (1980) Greg predictor for Y is

$$\begin{aligned} t_G &= X\hat{\beta}_Q + \sum{}' \frac{e_i}{\pi_i} & (2) \\ &= \sum{}' \frac{y_i}{\pi_i} g_{si} \ \text{ where } g_{si} = 1 + (X - \sum{}' \frac{x_i}{\pi_i}) \frac{x_i Q_i \pi_i}{\sum' x_i^2 Q_i} & (3) \end{aligned}$$

Two usual choices of Q_i given by Hájek (1971) and Brewer (1979) are respectively $Q_i = \frac{1}{\pi_i x_i}$, $Q_i = \frac{1-\pi_i}{\pi_i x_i}$ and two others are $Q_i = \frac{1}{x_i}$ and $Q_i = \frac{1}{x_i^2}$, $i \in U$. Särndal (1982) considers the Taylor approximation to the variance $V_p(t_G)$ of t_G given by

$$V = \sum\sum \Delta_{ij} (\frac{E_i}{\pi_i} - \frac{E_j}{\pi_j})^2$$

and gave two Yates and Grundy (YG,1953) type variance estimators,

$$\begin{aligned} v_{G1} &= \sum{}'\sum{}' \frac{\Delta_{ij}}{\pi_{ij}} (\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j})^2 \ \text{and} & (4) \\ v_{G2} &= \sum{}'\sum{}' \frac{\Delta_{ij}}{\pi_{ij}} (\frac{e_i g_{si}}{\pi_i} - \frac{e_j g_{sj}}{\pi_j})^2 & (5) \end{aligned}$$

which are discussed in details by Särndal, Swensson and Wretman (1992). Besides having a YG form these do not seem to have any particular properties but are supposed to serve variance estimation purpose well in large samples. Our interest here is to investigate two specific design-cum-model motivated asymptotic properties of them. For this we follow Brewer's (1979) approach to calculate the 'limiting' values of the design expectations of the model expectations of v_{G1} and v_{G2} and compare them to the model expectation of V assuming correctness of (1) and also to the 'limiting' value of the design expectation of the model expectation of the squared error $(t_G - Y)^2$. Since we find 'no match' in either case we proceed to apply 'adjustments' on $v_{Gj}, j = 1, 2$. By 'limiting' expectation we mean the following in accordance with Brewer's (1979) approach.

The population U and $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$, $\underline{X} = (x_1, \ldots, x_i, \ldots, x_N)$, $\underline{Q} = (Q_1, \ldots, Q_i, \ldots, Q_N)$ are supposed to produce themselves $T(> 1)$ times so as to yield the following entities:
$U_T = (U(1), \ldots, U(j), \ldots, U(T))$, $\underline{Y}_T = (Y(i), \ldots, Y(j), \ldots, Y(T))$,
$U(j) = ((j-1)N + 1, \ldots, (j-1)N + i, \ldots, (j-1)N + N)$,
$\underline{Y}(j) = (y_{(j-1)N+1}, \ldots, y_{(j-1)N+i}, \ldots, y_{(j-1)N+N})$,
$j = 1, \ldots, T$ where $(j-1)N + i$ for each $j = 1, \ldots, T$ stands for the same unit i for each respective $i (= 1, \ldots, N)$. Similarly for $\underline{X}_T$ and $\underline{Q}_T$. From each $U(j)$ a sample $s(j)$ is 'independently' chosen according to the same p as noted earlier. The T such samples are amalgamated into a sample s_T, say, which consequently is selected according to a design p_T such that

$$p_T(s_T) = p(s(1)) \ldots p(s(T)).$$

If t_G is based on s_T, then $t_G(s_T)$ is purported to estimate TY. The limiting value

$$\lim_{T \to \infty} E_p \ \left(\frac{1}{T} t_G(s_T)\right)$$

denoted as $\lim E_p(t_G)$ then equals Y as one may check - this property of t_G is known as its 'asymptotic design unbiasedness' (ADU, in brief). In calculating similar limiting expectation of other functions of survey data $d = (s, y_j, j \in s)$ an easy and fruitful way is to apply Slutzky's (cf, Crame'r 1966) theorem available in particular for continuous, especially rational functions and we shall profitably use it throughout below to derive convenient results of interest in section 2. Finally, in section 3 we shall extend this approach to cover situations when y_i's relate to stigmatizing issues and so they are not directly available and only RR's relevant to them may only be procured. It is now well-known, especially from recent books by Särndal, Swensson and Wretman (SSW, in brief, 1992) and Chaudhuri and Stenger (1992), why one need not insist on design- unbiased estimators like Horvitz and Thompson's (1952) for a survey population total and should rather explore improved alternatives with controlled mean square errors utilizing available auxiliary data. Särndal's (1980) greg predictor is such an alternative even when only one regressor is available. To construct confidence intervals one has of course Särndal's (1982) two variance estimators for it though with no known theoretical properties. Our motivation here is to seek further improvements and if possible extend the investigation to cover 'randomized responses'. The extent of our success is revealed below.

2. ALTERNATIVE VARIANCE ESTIMATORS

$$\begin{aligned}
E_m(V) &= \sum \sigma_i^2(\frac{1}{\pi_i}-1)+\frac{\sum \sigma_i^2 x_i^2 Q_i^2 \pi_i^2}{(\sum x_i^2 Q_i \pi_i)^2} V_p(\sum{}' \frac{x_i}{\pi_i}) \\
&\quad -2\sum\sum \Delta_{ij}(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j})\frac{(\sigma_i^2 x_i Q_i-\sigma_j^2 x_j Q_j)}{\sum x_i^2 Q_i \pi_i} \\
&= A_G, \text{ say.}
\end{aligned}$$

$$\begin{aligned}
E_m(v_{G2}) &= \sum{}'\sum{}' \frac{\Delta_{ij}}{\pi_{ij}}[(\frac{\sigma_i^2 g_{si}^2}{\pi_i^2}+\frac{\sigma_j^2 g_{sj}^2}{\pi_j^2})+\frac{\sum' \sigma_i^2 x_i^2 Q_i^2}{(\sum' x_i^2 Q_i)^2}(\frac{x_i g_{si}}{\pi_i}-\frac{x_j g_{sj}}{\pi_j})^2 \\
&\quad -\frac{2}{\sum' x_i^2 Q_i}(\frac{x_i g_{si}}{\pi_i}-\frac{x_j g_{sj}}{\pi_j})(\frac{x_i Q_i g_{si}\sigma_i^2}{\pi_i}-\frac{x_j Q_j g_{sj}\sigma_j^2}{\pi_j})].
\end{aligned}$$

Putting $g_{si}=1$ in $E_m(v_{G2})$ we get an expression for $E_m(v_{G1})$. Noting that $\lim E_p(g_{si})=1$ and $\lim E_p(g_{si}^2)=1+\frac{(x_i Q_i \pi_i)^2}{E_p(\sum' x_i^2 Q_i)^2}V_p(\sum' \frac{x_i}{\pi_i})$, we have

$$\begin{aligned}
\lim E_p E_m(v_{G2}) &= \sum \sigma_i^2(\frac{1}{\pi_i}-1)+\frac{V_p(\sum' \frac{x_i}{\pi_i})}{E_p(\sum' x_i^2 Q_i)^2}[\sum\sum \Delta_{ij}(\sigma_i^2 x_i^2 Q_i^2+\sigma_j^2 x_j^2 Q_j^2 \\
&\quad +(1+\frac{\sum\sum \Delta_{ij}(x_i^4 Q_i^2+x_j^4 Q_j^2)}{E_p(\sum' x_i^2 Q_i)^2}\sum \sigma_i^2 x_i^2 Q_i^2 \pi_i)] \\
&\quad -\frac{2}{E_p(\sum' x_i^2 Q_i)}\sum\sum \Delta_{ij}(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j})(\frac{x_i Q_i \sigma_i^2}{\pi_i}-\frac{x_j Q_j \sigma_j^2}{\pi_j}) \\
&= V_{G2}, \text{ say .}
\end{aligned}$$

Replacing only the expression in the square brackets by $\sum \sigma_i^2 x_i^2 Q_i^2 \pi_i$ and keeping the rest intact give a formula for $V_{G1}=\lim E_p E_m(v_{G1})$.

$$\begin{aligned}
\lim E_p E_m(t_G-Y)^2 &= \lim E_p E_m[(t_G-E_m(t_G))+(E_m(t_G)-E_m(Y)) \\
&\quad -(Y-E_m(Y))]^2 \\
&= \lim E_p V_m(t_G)-V_m(Y) \\
&= L_G, \text{ say ,}
\end{aligned}$$

following Godambe and Thompson (1977), noting that $(i)\lim E_p$ and E_m commute, $(ii) E_m(t_G-Y)=0$ and $(iii)\lim E_p(t_G)=Y$.

$$\begin{aligned}
V_m(t_G) &= \sum{}' \sigma_i^2[\frac{1}{\pi_i^2}+\frac{x_i^2 Q_i^2}{(\sum' x_i Q_i)^2}(X^2+(\sum{}' \frac{x_i}{\pi_i})^2-2X\sum{}' \frac{x_i}{\pi_i}) \\
&\quad +\frac{2x_i Q_i}{\pi_i(\sum' x_i^2 Q_i)}(X-\sum{}' \frac{x_i}{\pi_i})]
\end{aligned}$$

So,

$$L_G=\sum \sigma_i^2(\frac{1}{\pi_i}-1)+\frac{\sum \sigma_i^2 x_i^2 Q_i^2 \pi_i}{E_p(\sum' x_i^2 Q_i)^2}V_p(\sum{}' \frac{x_i}{\pi_i}).$$

For practical purposes we assume from now on

$$\sigma_i^2 = \sigma^2 f_i, i \in U \tag{6}$$

with $\sigma(>0)$ unknown, but $f_i(>0)$ known, $i = 1,\ldots,N$. For example, following Smith (1938), Brewer, Foreman, Mellor and Trewin (1977) it is useful to take $f_i = x_i^g, 0 \le g \le 2, i = 1,\ldots,N$. In practice g is not known beyond this. But we treat below a special case where g is fully known as g_0 in [0,2] and it is of interest to examine the consequence if g is in [0,2] but different from g_0. Such a study of robustness is not yet undertaken. Writing

$$A_{QG} = E_p(\sum{}' x_i^2 Q_i)^2,\ B = V_p(\sum{}' \frac{x_i}{\pi_i}),\ C_{QG} = E_p(\sum{}' x_i^2 Q_i)$$

and assuming (6) with f_i known, we get

$$\begin{aligned}
V_{G1} &= \sigma^2[\sum f_i(\frac{1}{\pi_i} - 1) + \frac{B}{A_{QG}}\sum f_i x_i^2 Q_i^2 \pi_i \\
&\quad -\frac{2}{C_{QG}}\sum\sum \Delta_{ij}(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j})(\frac{f_i x_i Qi}{\pi_i} - \frac{f_j x_j Qj}{\pi_j})] \\
&= \sigma^2 a_{G1f}, \text{ say },\\
V_{G2} &= \sigma^2[\sum f_i(\frac{1}{\pi_i} - 1) + \frac{B}{A_{QG}}\{\sum\sum \Delta_{ij}(f_i x_i^2 Q_i^2 + f_j x_j^2 Q_j^2) \\
&\quad +\sum f_i x_i^2 Q_i^2 \pi_i (1 + \frac{1}{A_{QG}}\sum\sum \Delta_{ij}(x_i^4 Q_i^2 + x_j^4 Q_j^2))\} \\
&\quad -\frac{2}{C_{QG}}\sum\sum \Delta_{ij}(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j})(\frac{f_i x_i Qi}{\pi_i} - \frac{f_j x_j Qj}{\pi_j})] \\
&= \sigma^2 a_{G2f}, \text{ say },\\
A_G &= \sigma^2[\sum f_i(\frac{1}{\pi_i} - 1) + \frac{B}{C_{QG}^2}\sum f_i^2 x_i^2 Q_i^2 \pi_i \\
&\quad -\frac{2}{C_{QG}}\sum\sum \Delta_{ij}(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j})(f_i x_i Qi - f_j x_j Qj)] \\
&= \sigma^2 b_{Gf}, \text{ say.}
\end{aligned} \tag{7}$$

So, two proposed alternatives to v_{G1}, v_{G2} are

$$v'_{G1} = v_{G1}\frac{b_{Gf}}{a_{G1f}},\ \text{and } v'_{G2} = v_{G2}\frac{b_{Gf}}{a_{G2f}}.$$

Noting

$$\begin{aligned}
L_G &= \sigma^2[\sum f_i(\tfrac{1}{\pi_i} - 1) + \tfrac{B}{A_{QG}}\sum f_i x_i^2 Q_i^2 \pi_i] \\
&= \sigma^2 C_{Gf}, \text{ say },
\end{aligned}$$

two more alternatives to v_{G1}, v_{G2} follow as

$$v''_{G1} = v_{G1}\frac{C_{Gf}}{a_{G1f}},\ v''_{G2} = v_{G2}\frac{C_{Gf}}{a_{G2f}}$$

3. RANDOMIZED RESPONSE

In case y_i's relate to sensitive issues like amount spent on gambling, amount of tax evaded etc., often instead of 'direct responses' (DR), 'randomized responses' (RR) are gathered. The above developments may extend as follows to cover them.

As described by Chaudhuri (1987) and Chaudhuri and Mukerjee (1988) it is conceivably possible to elicit RR from sampled individuals i of U as r_i, say, independently of one another, such that, writing $E_R(V_R)$ as operator for expectation (variance) with respect to 'randomization', one may have $(i) E_R(r_i) = y_i, (ii) V_R(r_i) = \alpha_i y_i^2 + \beta_i y_i + \theta_i = V_i$, say, with $\alpha_i, \beta_i, \theta_i$ as pre-assigned constants, $(iii) \hat{V}_i = (\alpha_i r_i^2 + \beta_i r_i + \theta_i)/(1+\alpha_i)$, provided $(1+\alpha_i) \neq 0$, satisfying $E_R(\hat{V}_i) = V_i, i \in U$.

Granting availability of r_i with $(i) - (iii)$, we define and write

$$B_Q(r), \hat{\beta}_Q(r), t_G(r), v_G(r)$$

etc. to denote $B_Q, \hat{\beta}_Q, t_G, v_G$ etc. with y_i in the latter just replaced by r_i throughout in the former keeping everything else in tact. As a measure of error of $t_G(r)$ in estimating Y we may take $E_p E_R(t_G(r) - Y)^2$ or $E_m E_p E_R(t_G(r) - Y)^2$ using the extra operator E_R. Noting that $E_R(t_G(r)) = t_G$, we obtain

$$\begin{aligned} E_p E_R(t_G(r) - Y)^2 &= E_p E_R[(t_G(r) - t_G) + (t_G - Y)]^2 \\ &= E_p(t_G - Y)^2 + \sum \tfrac{V_i}{\pi_i} + E_p[\tfrac{\sum' V_i x_i^2 Q_i^2}{(\sum' x_i^2 Q_i)^2}(X - \sum' \tfrac{x_i}{\pi_i})^2 \\ &\quad + \tfrac{2}{\sum' x_i^2 Q_i}(X - \sum' \tfrac{x_i}{\pi_i}) \sum' \tfrac{V_i x_i Q_i}{\pi_i}] \\ &= E_p(t_G - Y)^2 + D_Q(V), \text{ say.} \end{aligned}$$

Approximating $E_p(t_G - Y)^2$ by V, we approximate $E_m E_p E_R(t_G(r) - Y)^2$ by

$$M = A_G + E_m D_Q(V) \tag{8}$$

Now,

$$v_{G1}(r) = \sum' \sum' \frac{\Delta_{ij}}{\pi_{ij}} \left(\frac{e_i(r)}{\pi_i} - \frac{e_j(r)}{\pi_j} \right)^2$$

So,

$$\begin{aligned} E_R v_{G1}(r) &= v_{G1} + \sum' \sum' \tfrac{\Delta_{ij}}{\pi_{ij}} [(\tfrac{V_i}{\pi_i^2} + \tfrac{V_j}{\pi_j^2}) + \tfrac{\sum' x_i^2 Q_i^2 V_i}{(\sum' x_i^2 Q_i)^2} (\tfrac{x_i}{\pi_i} - \tfrac{x_j}{\pi_j})^2 \\ &\quad - \tfrac{2}{\sum' x_i^2 Q_i} (\tfrac{x_i}{\pi_i} - \tfrac{x_j}{\pi_j})(\tfrac{x_i V_i Q_i}{\pi_i} - \tfrac{x_j V_j Q_j}{\pi_j})]. \end{aligned}$$

So,

$$\lim E_p E_m E_R[v_{G1}(r) - \sum{}'\sum{}'\frac{\Delta_{ij}}{\pi_{ij}}\{(\frac{\hat{V}_i}{\pi_i^2}+\frac{\hat{V}_j}{\pi_j^2})+\frac{\sum' x_i^2Q_i^2\hat{V}_i}{(\sum' x_i^2Q_i)^2}(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j})^2$$
$$- \frac{2}{\sum' x_i^2Q_i}(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j})(\frac{x_i\hat{V}_iQ_i}{\pi_i}-\frac{x_j\hat{V}_jQ_j}{\pi_j})\}]$$
$$= \lim E_pE_mE_R(v'_{G1}(r)), \text{ say },$$
$$= \lim E_pE_m(v_{G1})$$
$$= \sigma^2 a_{G1f}. \tag{9}$$

So, combining (7), (8), and (9) it follows that

$$v''_{G1}(r) = b_{Gf}\frac{v'_{G1}(r)}{a_{G1f}}+\sum{}'\frac{\hat{V}_i}{\pi_i^2}+[\frac{\sum'\hat{V}_ix_i^2Q_i^2}{(\sum' x_i^2Q_i)^2}(X-\sum{}'\frac{x_i}{\pi_i})^2$$
$$+\frac{2}{\sum' x_i^2Q_i}(X-\sum{}'\frac{x_i}{\pi_i})\sum{}'\frac{\hat{V}_ix_iQ_i}{\pi_i}]$$
$$= b_{Gf}\frac{v'_{G1}(r)}{a_{G1f}}+D_Q(\hat{V}), \text{ say }, \tag{10}$$

may be taken as an estimator for a measure of error of $t_G(r)$ as an estimator of Y because $\lim E_pE_mE_R(v''_{G1}(r))$ equals M. Again,

$$v_{G2}(r)=\sum{}'\sum{}'\frac{\Delta_{ij}}{\pi_{ij}}(\frac{e_i(r)g_{si}}{\pi_i}-\frac{e_j(r)g_{sj}}{\pi_j})^2$$

$$E_Rv_{G2}(r) = v_{G2}+\sum{}'\sum{}'\frac{\Delta_{ij}}{\pi_{ij}}[(\frac{V_ig_{si}^2}{\pi_i^2}+\frac{V_jg_{sj}^2}{\pi_j^2})+\frac{\sum' x_i^2Q_i^2V_i}{(\sum' x_i^2Q_i)^2}(\frac{x_ig_{si}}{\pi_i}-\frac{x_jg_{sj}}{\pi_j})^2$$
$$-\frac{2}{\sum' x_i^2Q_i}(\frac{x_ig_{si}}{\pi_i}-\frac{x_jg_{sj}}{\pi_j})(\frac{x_iV_iQ_ig_{si}}{\pi_i}-\frac{x_jV_jQ_jg_{sj}}{\pi_j})].$$

So,

$$\lim E_pE_mE_R[v_{G2}(r) - \sum{}'\sum{}'\frac{\Delta_{ij}}{\pi_{ij}}\{(\frac{\hat{V}_ig_{si}^2}{\pi_i^2}+\frac{\hat{V}_jg_{sj}^2}{\pi_j^2})+\frac{\sum' x_i^2Q_i^2\hat{V}_i}{(\sum' x_i^2Q_i)^2}(\frac{x_ig_{si}}{\pi_i}-\frac{x_jg_{sj}}{\pi_j})^2$$
$$- \frac{2}{\sum' x_i^2Q_i}(\frac{x_ig_{si}}{\pi_i}-\frac{x_jg_{sj}}{\pi_j})(\frac{x_i\hat{V}_iQ_ig_{si}}{\pi_i}-\frac{x_j\hat{V}_jQ_jg_{sj}}{\pi_j})\}]$$
$$= \lim E_pE_mE_R(v'_{G2}(r)), \text{ say },$$
$$= \lim E_pE_m(v_{G2})$$
$$= \sigma^2 a_{G2f}. \tag{11}$$

So, combining (7), (8), (10) and (11) it follows that

$$v''_{G2}(r)=b_{Gf}\frac{v'_{G2}(r)}{a_{G2f}}+D_Q(\hat{V})$$

may be taken as another estimator for a measure of error of $t_G(r)$ as an estimator of Y because $\lim E_p E_m E_R(v''_{G2}(r))$ equals M. Again

$$\lim E_p E_m E_R(t_G(r) - Y)^2 = F_G,$$

say, which equals

$$L_G + \lim E_p E_m[\sum{}' \frac{V_i}{\pi_i^2} + \frac{\sum' V_i x_i^2 Q_i^2}{(\sum' x_i^2 Q_i)^2}(X - \sum{}' \frac{x_i}{\pi_i})^2 + \frac{2}{\sum' x_i^2 Q_i}(X - \sum{}' \frac{x_i}{\pi_i})\sum{}' \frac{V_i x_i Q_i}{\pi_i}].$$

So, $v'''_{G1}(r) = v'_{G1}(r)\frac{C_{Gf}}{a_{G1f}} + D_Q(\hat{V})$
and, $v'''_{G2}(r) = v'_{G2}(r)\frac{C_{Gf}}{a_{G2f}} + D_Q(\hat{V})$
may be taken as alternative estimators for F_G because it is easily checked that

$$\lim E_p E_m E_R(v'''_{G1}(r)) = F_G = \lim E_p R_M E_R(v'''_{G2}(r)).$$

4. KOTT'S ESTIMATOR

Finally we consider Kott's (1990, a,b) variance estimators

$$v_{kj} = \frac{v_{Gj}}{E_m(v_{Gj})} E_m(t_G - Y)^2, j = 1, 2,$$

which are 'free' of model parameters under (6). Noting

$$E_m(t_G - Y)^2 = \sigma^2[\frac{\sum' f_i x_i^2 Q_i^2}{(\sum' x_i^2 Q_i)^2}(X - \sum{}' \frac{x_i}{\pi_i})^2 + \sum{}' \frac{f_i}{\pi_i^2} + \sum f_i + \frac{2}{\sum' x_i^2 Q_i}(X - \sum{}' \frac{x_i}{\pi_i})\sum{}' \frac{f_i x_i Q_i}{\pi_i}(1 - \pi_i) - 2\sum{}' \frac{f_i}{\pi_i}],$$

formulae for v_{k1} and v_{k2} easily follow with DR but not with RR.

5. A SIMULATION STUDY

For a comparative study of the alternative procedures with DR we resort to simulation.

Treating the model (1) as valid, we take $(i)\epsilon_i$'s as $N(0, \sigma_i^2), \sigma_i^2 = \sigma^2 x_i^g, \sigma = 1.0, g = 1.5, \beta = 5.5, (ii) x_i$'s as independently identically exponentially distributed with a density

$$f(x) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda}), x > 0. \quad (12)$$

Taking $N = 50, \lambda = 7.0$, using these we first generate two vectors $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$ and $\underline{X} = (x_1, \ldots, x_i, \ldots, x_N)$. To draw a sample s of size $n = 11$, we apply two separate sampling schemes namely (1) due to Lahiri, Midzuno and Sen (LMS, in brief, 1951, 1952, 1953) and (2) Hartley and Rao (HR, say, 1962). For this we generate a vector of real numbers $\underline{Z} = (z_1, \ldots, z_i, \ldots, z_N)$, z_i's independently identically distributed with a common density (12) with $\lambda = 15.0$. We take $w_i = 5 + z_i, i \in U$, as the size-measure needed for the sample selection. We draw $R = 100$ replicates of the sample chosen by each of these two methods. To study the relative performances of $v_{Gj}, v'_{Gj}, v''_{Gj}$ and $v_{kj}, j = 1,2$, we proceed as explained below.

For large samples, with e as an estimator for Y having v as a variance estimator,

$$d = (e - Y)/\sqrt{v}$$

is usually supposed to be distributed as τ, the standardized normal distribution $N(0,1)$. As a result with $\tau_{\alpha/2}$ as the $100\alpha/2\%$ point in the right tail of the distribution of τ, the interval $(e \pm \tau_{\alpha/2}\sqrt{v})$ is taken to provide the $100(1-\alpha)\%$ confidence interval (CI,in brief) for Y. Here α is a number in (0,1)- we take it only as 0.05. We take e as t_G and v as the various variance estimators mentioned so far. As measure of performances of (t_G, v) we consider the following as recommended by Rao and Wu (1983), among others.

1. 'Actual coverage probability' (ACP, in brief): This is the proportion of the $R(= 100)$ replicated samples for which $(t_G \pm \tau_{.025}\sqrt{v})$ covers Y. The closer it is to 0.95, which is the 'nominal confidence coefficient', the better for (t_G, v).
2. 'Average coefficient of variation' (ACV, in brief): This is the average over the above R samples, of the values of $\sqrt{v}/t_G$, which reflects the length of the CI relative to t_G and as such the smaller the ACV, the better the (t_G, v).

Numerical findings are given in the table below presenting the values based on HR scheme within parentheses just below those for LMS scheme.

6. A SUMMARY OF NUMERICAL FINDINGS

Even though the sample and population sizes are small, for LMS scheme all the variance estimators seem to fare well and advantages in using our modified estimators are discernible. For $v_{Gj}, v'_{Gj}, v''_{Gj}$, the three choices of Q_i excluding $1/x_i^2$ which is bad seem equally effective. For v_{kj} the choice

Table
ACP and ACV for (v, Q)

(v, Q)	ACP	ACV	(v, Q)	ACP	ACV
$(v_{G1}, \frac{1}{x_i})$	.90	.016	$(v_{G2}, \frac{1}{x_i})$	.93	.018
	(.82)	(.018)		(.85)	(.020)
$(v'_{G1}, \frac{1}{x_i})$	.93	.019	$(v'_{G2}, \frac{1}{x_i})$	.96	.019
	(.84)	(.021)		(.87)	(.021)
$(v''_{G1}, \frac{1}{x_i})$	.94	.020	$(v''_{G2}, \frac{1}{x_i})$	.97	.020
	(.86)	(.022)		(.87)	(.023)
$(v_{G1}, \frac{1}{x_i^2})$	.93	.079	$(v_{G2}, \frac{1}{x_i^2})$	1.00	.097
	(.84)	(.075)		(.97)	(.092)
$(v'_{G1}, \frac{1}{x_i^2})$	.90	.067	$(v'_{G2}, \frac{1}{x_i^2})$	.98	.071
	(.82)	(.068)		(.92)	(.076)
$(v''_{G1}, \frac{1}{x_i^2})$	.94	.081	$(v''_{G2}, \frac{1}{x_i^2})$	1.00	.085
	(.84)	(.077)		(.96)	(.086)
$(v_{G1}, \frac{1}{\pi_i x_i})$	.90	.016	$(v_{G2}, \frac{1}{\pi_i x_i})$	.93	.018
	(.80)	(.018)		(.85)	(.021)
$(v'_{G1}, \frac{1}{\pi_i x_i})$	.93	.019	$(v'_{G2}, \frac{1}{\pi_i x_i})$	.95	.019
	(.84)	(.023)		(.85)	(.023)
$(v''_{G1}, \frac{1}{\pi_i x_i})$	.94	.020	$(v''_{G2}, \frac{1}{\pi_i x_i})$	.97	.020
	(.87)	(.025)		(.85)	(.025)
$(v_{G1}, \frac{1-\pi_i}{\pi_i x_i})$	.89	.016	$(v_{G2}, \frac{1-\pi_i}{\pi_i x_i})$	.93	.018
	(.80)	(.019)		(.84)	(.021)
$(v'_{G1}, \frac{1-\pi_i}{\pi_i x_i})$	.93	.019	$(v'_{G2}, \frac{1-\pi_i}{\pi_i x_i})$	.95	.019
	(.85)	(.024)		(.85)	(.023)
$(v''_{G1}, \frac{1-\pi_i}{\pi_i x_i})$	.94	.020	$(v''_{G2}, \frac{1-\pi_i}{\pi_i x_i})$	.97	.021
	(.88)	(.027)		(.85)	(.026)
$(v_{k1}, \frac{1}{x_i})$	.97	.020	$(v_{k2}, \frac{1}{x_i})$	.97	.020
	(.70)	(.022)		(.66)	(.022)
$(v_{k1}, \frac{1}{x_i^2})$	.99	.075	$(v_{k2}, \frac{1}{x_i^2})$	1.00	.079
	(.84)	(.074)		(.85)	(.145)
$(v_{k1}, \frac{1}{\pi_i x_i})$	.97	.020	$(v_{k2}, \frac{1}{\pi_i x_i})$	.97	.020
	(.73)	(.021)		(.73)	(.021)
$(v_{k1}, \frac{1-\pi_i}{\pi_i x_i})$	.97	.020	$(v_{k2}, \frac{1-\pi_i}{\pi_i x_i})$	.97	.020
	(.73)	(.021)		(.73)	(.021)

$Q_i = 1/x_i^2$ seems decidedly poor. For HR scheme there is definite reduction in ACP though the relative performances of the variance estimators follow a similar pattern as in LMS scheme. Again $1/x_i^2$ is a bad choice. So, we conclude that LMS scheme should be preffered to HR in situations similar to the one considered here and our alternative variance estimators are worth consideration as good competitors against the traditional ones, both in theory and practice.

ACKNOWLEDGEMENT

Thanks are due to the referee for helpful comments.

BIBLIOGRAPHY

Brewer, K.R.W. (1979). " A class of robust sampling designs for large-scale surveys," *Jour. Amer. Stat. Assoc.,* 74, 911-15.

Brewer, K.R.W., Foreman, E.K., Mellor, R.W. and Trewin, D.J. (1977). " Use of experimental design and population modelling in survey sampling," *Bull. Int. Stat. Inst.*, 47:3,173-190.

Chaudhuri, A. (1987). " Randomized response surveys of finite populations: a unified approach with quantitative data," *Jour. Stat. Plan. Inf.*, 15, 157-165.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*, Marcel Dekker, Inc. N.Y.

Chaudhuri, A. and Stenger, H. (1992). *Survey sampling: Theory and methods*, Marcel Dekker. Inc. N.Y.

Crame'r, H. (1966). *Mathematical methods of statistics*, Princeton Univ. Press.

Godambe, V.P. and Thompson, M.E. (1977). " Robust near optimal estimation in survey practice," *Bull.Int. Stat. Inst.*, 97:3,129-146.

Hájek, J. (1971). " Comment on a paper by Basu, D.", *Foundations of Statistical Inference* , (V.P. Godambe, and D.A. Sprott, Eds.). Holt, Rinehart, Winston. Toronto,203-242.

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

Horvitz, D.G. and Thompson, D.J. (1952). " A generalization of sampling without replacement from a finite universe," *Jour. Amer. Stat. Assoc.*, 47, 663-85.

Kott, P.S. (1990a). " Estimating the conditional variance of design consistent regression estimator", *Jour. Stat. Plan. Inf.*, 24, 287-296.

—(1990b). " Recently proposed variance estimators for the simple regression estimator", *Jour. Official Stat.*, 6, 451-454.

Lahiri, D.B. (1951). " A method of sample selection providing unbiased ratio estimators", *Bull. Int. Stat. Inst.*, 32,350-374.

Midzuno, H. (1952). " On the sampling system with probabilities proportional to sum of sizes", *Ann. Inst. Stat. Math.*, 3, 99-107.

Rao, J.N.K. and Wu, C.F.J. (1983). " Methods for standard errors and confidence intervals from sample survey data: Some recent work", *Invited paper in 46-th session of Int. Stat. Inst.*, 1-16.

Särndal, C.E. (1980). " On II inverse weighing versus best linear weighing in probability sampling", *Biometrika*, 67, 639-650.

—(1982). " Implications of survey design for generalized regression estimation of linear functions", *Jour. Stat. Plan. Inf.*, 7, 155-170.

Särndal, C.E. Swensson, B.E. and Wretman, J.H. (1992). *Model assisted survey sampling*, Springer-Verlag, N.Y. Inc.

Sen, A.R. (1953). " On the estimator of the variance in sampling with varying probabilities", *Jour. Ind. Soc. Agr. Stat.*, 5:2, 119-127.

Smith, H.F. (1938). " An empirical law describing heterogeneity in the yields of agricultural crops", *Jour. Agri. Sc.*,28, 1-23.

Yates, F. and Grundy, P.M. (1953). " Selection without replacement from within strata with probability proportional to size", *Jour. Roy. Stat. Soc. B*,15, 253-261.

Received September 1992; Revised August 1993.