VARIANCE REDUCTION AND UNBIASEDNESS FOR SMALL DOMAIN ESTIMATORS

Göran Råbäck, Statistics Sweden, Stockholm and Carl-Erik Särndal, Université de Montréal

1. Introduction

The purpose of this paper is to test empirically the advantages and disadvantages of two kinds of procedures in estimation for small domains: (a) asymptotically design unbiased (ADU) alternatives and (b) design biased alternatives, including the synthetic estimator. The ADU estimators have negligible design bias in moderate to large samples, and confidence intervals are readily obtainable such that the confidence coefficient carries the usual interpretation of coverage rate under repeated draws of samples s by the given sampling design. The design biased estimators, on the other hand, may have a small design variance, but their design bias is often so large that it is virtually certain that an interval centered on the point estimate will fall to contain the true point under estimation.

2. Estimators under study

Let $U=\{1,\ldots,k,\ldots,N\}$ be a finite population from which a sample s of fixed size n is selected by a given sampling design admitting the inclusion probabilities π_k and $\pi_{k\ell}$. Let $U_{.q}$ $(q=1,\ldots,Q)$ be subdomains with known sizes $N_{.q}$ and let $n_{.q}$ be the (random) number of observations contained in $s_{.q}$, the part of s that happens to fall in $U_{.q}$. We seek estimates of the domain means $\bar{Y}_{.q} = \Sigma_{U_{.q}} Y_k/N_{.q}$ $(q=1,..,Q)$.

First, the ADU estimators considered are Design-Model estimators (also known as generalized regression estimators). These are, in a general context, of the form (Särndal, 1981, 1982).

$$\hat{T}_q = \sum_1^N c_{kq}\hat{Y}_k + \sum_s c_{kq}e_k/\pi_k \qquad (2.1)$$

Where $e_k = Y_k - \hat{Y}_k = Y_k - \underset{\sim}{x}_k'\hat{\underset{\sim}{\beta}}$ with

$$\hat{\underset{\sim}{\beta}} = \left(\sum_s w_k \underset{\sim}{x}_k \underset{\sim}{x}_k'\right)^{-1} \sum_s w_k \underset{\sim}{x}_k Y_k \qquad (2.2)$$

is the residual arising in fitting the linear model (ξ, say) such that $E_\xi(Y_k) = \underset{\sim}{x}_k'\underset{\sim}{\beta}$ to the n data points. Here the weights in $\hat{\underset{\sim}{\beta}}$ are chosen as $w_k = 1/\pi_k\sigma_k^2$, where $\sigma_k^2 = V_\xi(Y_k)$. Moreover, $c_{kq} = 1/N_{.q}$ if $k \in U_{.q}$ and $c_{kq} = 0$ otherwise. For a given sampling design, each model fit gives rise to a different ADU estimator.

An approximate variance estimate of $\hat{T}_q$ is given by the Yates-Grundy formula in $c_{kq}e_k$, that is

$$\hat{V}(\hat{T}_q) = \sum\sum_{\substack{k<\ell \\ \in s}} g_{k\ell}\left(\frac{c_{kq}e_k}{\pi_k} - \frac{c_{\ell q}e_\ell}{\pi_\ell}\right)^2 \qquad (2.3)$$

where $g_{k\ell} = (\pi_k\pi_\ell - \pi_{k\ell})/\pi_{k\ell}$. An approximate 100 $(1-\alpha)$% confidence interval for $\bar{Y}_{.q}$ is therefore

$$\hat{T}_q \pm z_{1-\alpha/2}\{\hat{V}(\hat{T}_q)\}^{\frac{1}{2}} \qquad (2.4)$$

where $z_{1-\alpha/2}$ is the unit normal deviate. Under repeated draws of samples s by the given design, roughly $100(1-\alpha)$% of all confidence intervals calculated by (2.4) will cover the true value $\bar{Y}_{.q}$, as confirmed in particular cases by our Monte Carlo study.

In the study reported below, the simple random sampling (srs) design is used, thus $\pi_k = f = n/N$ for all k and $\pi_{k\ell} = n(n-1)/N(N-1)$ for all $k \neq \ell$. The fit of a model usually requires access to a certain amount of auxiliary information.

We assume that $x_1,\ldots, x_N$ are known positive values of an auxiliary variable correlated with Y. The N values of x are size ordred and the population units are grouped by nonoverlapping intervals of x into H groups $U_{h.}$ of known size $N_{h.}$ $(h=1,..,H)$. The population is thereby crossclassified into HQ cells U_{hq} of known size N_{hq}. At the sample level, the corresponding notation is $s_{h.}$, $n_{h.}$, s_{hq} and n_{hq}. Thus, $n = \Sigma_{h=1}^H n_{h.} = \Sigma_{q=1}^Q n_{.q} = \Sigma_{h=1}^H \Sigma_{q=1}^Q n_{hq}$, relations which also hold for the big N's.

We consider the following ADU estimators obtained from the general formula (2.1) for the srs design

$$\hat{T}_{1q} = N_{.q}^{-1}\sum_1^H (N_{hq}\bar{Y}_{s_{h.}} + f^{-1}\sum_{s_{hq}} e_{1k}) \qquad (2.5)$$

$$\hat{T}_{2q} = N_{.q}^{-1}\sum_1^H (\hat{\beta}_h \sum_{U_{hq}} x_k + f^{-1}\sum_{s_{hq}} e_{2k}) \qquad (2.6)$$

where, for $k \in s_{h.}$, $e_{1k} = Y_k - \bar{Y}_{sh.} = Y_k - \Sigma_{sh.}Y_k/n_{h.}$ and $e_{2k} = Y_k - \hat{\beta}_h x_k$ with $\hat{\beta}_h = \Sigma_{sh.}Y_k/\Sigma_{sh.}x_k$. In (2.3), we use $e_k = e_{1k}$ and $e_k = e_{2k}$, respectively.

The estimators $\hat{T}_{1q}$ and $\hat{T}_{2q}$ arise for the models of independent Y_k such that, respectively,

$$E_\xi(Y_k) = \beta_h \; ; \; V_\xi(Y_k) = \sigma_h^2 \quad \text{all } k \in U_{h.} \qquad (2.7)$$

and

$$E_\xi(Y_k) = \beta_h x_k \; ; \; V_\xi(Y_k) = \sigma_h^2 x_k \quad \text{all } k \in U_{h.} \qquad (2.8)$$

Other ADU estimators under srs studied below are the "simple direct estimator"

$$\hat{T}_{3q} = \sum_{s_{.q}} Y_k/n_{.q} = \bar{Y}_{s_{.q}} \qquad (2.9)$$

and the "combined cell mean estimator"

$$\hat{T}_{4q} = \sum_1^H N_{hq}\bar{Y}_{s_{hq}}/N_{.q} \qquad (2.10)$$

The latter is frequently undefined for subdomain q, namely, as soon as the cell count n_{hq} in the denominator of $\bar{Y}_{s_{hq}} = \Sigma_{s_{hq}} Y_k/n_{hq}$ is zero for some h. (To make $\hat{T}_{4q}$ operational, a rule for collapsing cells would first have to be defined.)

Now $\hat{T}_{3q}$ and $\hat{T}_{4q}$ are also obtained from (2.1) under obvious model formulations; their respective variance estimates are given by (2.3) with

$e_k = e_{3k} = Y_k - \bar{Y}_{s.q}$ and $e_k = e_{4k} = Y_k - \bar{Y}_{shq}$.

Turning now to the design biased estimators, we consider the much debated synthetic estimator

$$\hat{T}_{5q} = \sum_1^H N_{hq} \bar{Y}_{s_h.} / N_{.q} \qquad (2.11)$$

It requires the same auxiliary information as (2.5) above and (2.12) below, that is, the population cell counts N_{hq}.

Another group of design biased alternatives that deserve attention for comparative purposes are those arising from the predictive approach, (Holt, Smith and Tomberlin, 1979), that is, from the general "predictor formula".

$$\hat{T}_q = \sum_s c_{kq} Y_k + \sum_{U-s} c_{kq} \hat{Y}_k$$

where $\hat{Y}_k$ is the predicted Y_k-value under the model. For the models (2.7) and (2.8) we get, respectively (with $w_k = \sigma_k^{-2}$ in (2.2))

$$\hat{T}_{6q} = N_{.q}^{-1} \sum_1^H \left\{ N_{hq} \bar{Y}_{s_h.} + \sum_{s_{hq}} e_{1k} \right\} \qquad (2.12)$$

$$\hat{T}_{7q} = N_{.q}^{-1} \sum_1^H \left\{ \hat{\beta}_h \sum_{U_{hq}} x_k + \sum_{s_{hq}} e_{2k} \right\} \qquad (2.13)$$

with e_{1k} and e_{2k} as in (2.5) and (2.6), respectively.

It is unclear what variance estimate to use for (2.11)-(2.13); suggestions made for (2.11) (Gonzalez and Waksberg, 1973) do not seem to have the usual repeated sampling interpretation.

For (2.12) we use the variance estimator pooled over all groups h and suggested by Holt, Smith and Tomberlin (1979),

$$\hat{V}(\hat{T}_{6q}) = N_{.q}^{-2} d_q \sum_{h=1}^{H} \sum_{q=1}^{Q} \sum_{s_{hq}} (Y_k - \bar{Y}_{s_h.})^2 / (n-H) \qquad (2.14)$$

where $d_q = \sum_{h=1}^{H} (N_{hq} - n_{hq})(N_{hq} - n_{hq} + n_{h.})/n_{h.}$. This is a model based variance estimate, and although we use it below in our Monte Carlo study for the construction of confidence intervals, we expect a priori that the coverage rate achieved in repeated samples by these intervals will not come near the theoretical confidence level aimed at.

3 Description of the Monte Carlo Simulation

We identified a population consisting of N=1287 Swedish households classified into Q=12 subdomains defined by household type (number of household members, age and accupation of head of household). Let Y_k and x_k denote, respectively, disposable income and net income (as per the income tax return), both calculated as totals for all members of the k:th household. For every unit k=1,.., N assume that the following information is avaliable (1) the value of x_k (from the income tax return) and (2) the identity of the subdomain to which k belongs.

The subdomain sizes $N_{.q}$ are thus known. The relative subdomain sizes $N_{.q}/N$ varied between 1.3% and 29.5%. In the discussion below we distinguish the "six larger subdomains" varying in size from 7.1% to 29.5%, and the "six smaller subdomains" varying is size from 1.3% to 4.7%.

The target of estimation is the mean disposable income in each domain,

$$\bar{Y}_{.q} = \sum_{U_{.q}} Y_k / N_{.q} \quad (q=1,.., Q).$$

The x-variable was used to create H=5 income classes. For this population the correlation between x and Y is 0.80. It expected, therefore, that the x-dimension will have a strong variance reducing effect. That is, estimators such as $\hat{T}_{1q}$ and $\hat{T}_{2q}$ which capitalize on the x-variable will be better than estimators (such as the simple direct formula $\hat{T}_{3q}$) which do not.

Our Monte Carlo simulation consisted in the selection of J=1000 samples, each of size n=200, from the population of 1287. Thus, the expected number of observations in a subdomain ranged from 2.6 for the smallest to 59 for the largest subdomain.

For each sample from 1 to 1000, and for i=1,.., 7; q=1,.., 12, we calcualted the estimator value $\hat{T}_{iq}$, its estimated variance $\hat{V}_{iq} = \hat{V}_{srs}(\hat{T}_{iq})$, and the cofidence interval $\hat{T}_{iq} \pm z_{1-\alpha/2} \hat{V}_{iq}^{1/2}$ for $1-\alpha$ = 80%, 90% and 95%. In summary of the whole experiment we calculated the mean, ME_{iq}, and the variance VA_{iq}, of the 1000 estimator values, as well as the coverage rate CR_{iq} = the fraction of the 1000 intervals that covered the true point $\bar{Y}_{.q}$ (at nominal confidence levels of 80%, 90% and 95%). Here ME_{iq}, VA_{iq} and CR_{iq} can be taken as sufficiently good approximations of the true mean, variance and coverage rate of the procedure.

We discuss the main findings with respect to the concepts (a) design bias; (b) design variance; (c) coverage rate. In Table 1 to 3, results are given seperately for "six larger domains" and "six smaller domains". For domains q in each of these two categories, "max" and "min" gives, respectively, the largest and smallest value of the concept in question. This gives some indication of the deterioration of the quality of an estimator as domain size decreases.

It turned out that $\hat{T}_{4q}$ was undefined for many of the 1000 samples (because of at least one zero cell count n_{hq}) and T_{4q} is therefore excluded from the comparisons below.

(a) The design bias. For all subdomains, including the very small ones, E_{iq} should according to theory be close to the true point $\bar{Y}_{.q}$ for the ADU estimators (i=1,2,3), but not likely so for the design biased ones (i=5,6,7). This was confirmed by the Monte Carlo results reported in Table 1, which gives values of the relative bias $100(E_{iq} - \bar{Y}_{.q})/\bar{Y}_{.q}$. It should be noted that the sample size here is to be thought of as n=200. This is large enough to count on approximate design unbiasedness for ADU estimators, even though the expected sample size in the smallest domains is only about 3.

Table 1 reveals a striking contrast between the three ADU estimators and the tree design biased ones.

The relative design bias in the ADU group is never important; the max and min values tend, however, to be slightly larger in the smaller domains. Deviations from zero could be a result of the insufficiency of the 1000 repetitions.

Earlier claims that the design bias of the synthetic estimator $\hat{T}_{5q}$ can not be estimated from the sample are unfounded, considering the results of this study. The term $f^{-1}\Sigma_{s_{hq}} e_{1k}$ of $\hat{T}_{1q}$ given by (2.5) effectively removes the design bias of the synthetic estimator. For the design biased group, the direction and magnitude of the relative design bias is unpredictable. It is to be expected that particularly large biases arise if the underlying model fits the population badly.

Table 1. Relative design bias $100(ME_{iq}-Y_{.q})/Y_{.q} = BI_{iq}$.

Estimator	Six larger domains Max BI_{iq} q	Min BI_{iq} q	Six smaller domains Max BI_{iq} q	Min BI_{iq} q
ADU:				
$\hat{T}_{1q}$	0.73	-0.41	1.63	-1.62
$\hat{T}_{2q}$	0.59	-0.39	2.00	-1.59
$\hat{T}_{3q}$	0.04	-0.48	2.70	-1.16
Design biased:				
$\hat{T}_{5q}$	26.0	-15.0	14.6	-34.1
$\hat{T}_{6q}$	22.1	-12.7	12.1	-30.0
$\hat{T}_{7q}$	20.9	-13.1	2.0	-24.7

(b) Design variance. Among the ADU methods, $\hat{T}_{1q}$ and $\hat{T}_{2q}$ capitalize on the information contained in the x-variable and are expected to give smaller variances than the simple direct method $\hat{T}_{3q}$.

On the other hand, we expect the design biased methods $\hat{T}_{5q}$, $\hat{T}_{6q}$ and $\hat{T}_{7q}$ to profit even more strongly from the variance reducing effect of the x-dimension, unhampered as they are by the ADU-ness constraint.

These expectations are confirmed by Table 2, which, using the standard error of $\hat{T}_{1q}$ as the point of reference, shows max, min and average values of the efficiency measure $\{VA_{1q}/VA_{2q}\}^{1/2} = EF_{iq}$, seperately for larger domains and for smaller domains. Within the ADU group, $\hat{T}_{1q}$ and $\hat{T}_{2q}$ perform about equally well. For example, $\hat{T}_{1q}$ is better than $\hat{T}_{3q}$ in all domains, and the tendency is especially pronounced in the smaller domains.

The variance reduction achieved by the design biased methods is of more dramatic proportions and is especially apparent in the smaller domains, where the synthetic estimator $\hat{T}_{5q}$ performs particularly well with an average efficience measure of 7.0.

Table 2. Features of $EF_{iq} = \{VA_{1q}/VA_{iq}\}^{1/2}$

Estimator	Six larger domains Max EF_{iq} q	Min EF_{iq} q	$\overline{EF}_{i.}$
ADU:			
$\hat{T}_{2q}$	1.16	0.77	1.00
$\hat{T}_{3q}$	0.99	0.69	0.84
Design biased:			
$\hat{T}_{5q}$	2.53	0.98	1.64
$\hat{T}_{6q}$	2.34	1.08	1.66
$\hat{T}_{7q}$	2.41	1.09	1.64

Estimator	Six smaller domains Max EF_{iq} q	Min EF_{iq} q	$\overline{EF}_{i.}$
ADU:			
$\hat{T}_{2q}$	1.09	0.55	0.93
$\hat{T}_{3q}$	0.91	0.65	0.75
Design biased:			
$\hat{T}_{5q}$	17.26	3.03	7.00
$\hat{T}_{6q}$	5.43	2.90	4.17
$\hat{T}_{7q}$	5.68	2.72	3.95

(c) Coverage rate. Table 3 shows coverage rates (at nominal values 1-α of 90% and 95%) achieved by the confidence intervals centered on $\hat{T}_{1q}$, $\hat{T}_{2q}$, $\hat{T}_{3q}$ and $\hat{T}_{6q}$.

The intervals are constructed according to (2.4), where $\hat{V}(\hat{T}_{iq})$ is given by the Yates-Grundy formula (2.3) in the case of $\hat{T}_{1q}$, $\hat{T}_{2q}$ and $\hat{T}_{3q}$, and by (2.14) in the case of $\hat{T}_{6q}$.

The coverage rates achieved by the ADU methods have a tendency to fall short of the intended nominal rate, a feature which becomes particularly apparent for the smaller domains, where the achieved CR drops, in the worst case, to about 60 % (for an intended 90%) and 65% (for an intended 95%). This is not surprising, since the smallest domains yield extremely few non-zero values $c_{kq}e_k/\pi_k$ on which to base the calculation of the estimated variance.

Table 3. Features of the coverage rate CR_{iq}. The upper (lower) portion of the table refers to the nominal rate of $1-\alpha = 90\%$ (95%).

Estimator	Six larger domains: Max CR_{iq}	Six larger domains: Min CR_{iq}	Six smaller domains: Max CR_{iq}	Six smaller domains: Min CR_{iq}
ADU:				
$\hat{T}_{1q}$	89.7	84.4	89.5	59.0
$\hat{T}_{2q}$	92.5	83.3	90.9	67.7
$\hat{T}_{3q}$	88.3	77.1	82.0	61.8
Design biased:				
$\hat{T}_{6q}$	99.7	0.0	99.7	0.0
ADU:				
$\hat{T}_{1q}$	94.2	89.9	94.3	66.6
$\hat{T}_{2q}$	96.9	88.3	96.5	72.1
$\hat{T}_{3q}$	94.6	83.1	87.0	64.4
Design biased:				
$\hat{T}_{6q}$	100.0	0	100.0	0.3

The coverage rates of the design biased method $\hat{T}_{6q}$ bear no resemblance to the nominal rates aimed at. We conclude that a model based variance estimate such as (2.14) is of little use in attempting to construct confidence intervals with the customary randomization theory interpretation, that is, one that appeals to repeated draws of samples s under the given sampling design. (The design biased estimators $\hat{T}_{5q}$ and $\hat{T}_{7q}$ were omitted from the coverage rate study because of the apparent lack of a confidence interval procedure that carries the randomization theory interpretation.

REFERENCES

Gonzalez, M.E. and Waksberg, J. (1973). Estimation of the error of synthetic estimates. Paper presented at the 1st meeting of the International Association of Survey Statisticians, Vienna.

Holt, D., Smith, T.M.F. and Tomberlin, T.J. (1979). A model-based approach to estimation for small subgroups of a population. Journal of the American Statistical Association, 74, 405-410.

Särndal, C.E. (1982). When robust estimation is not an obvious answer: The case of the synthetic estimator versus alternatives for small areas. To appear in the Journal of the American Statistical Association.

Särndal, C.E. (1982). Implications of survey design for generalized regression estimation of linear functions. To appear in the Journal of Statistical Planning and Inference.