A comparison of stratified simple random sampling and sampling with probability proportional to size

Edgar Bueno
Dan Hedlin
Per Gösta Andersson

1 Introduction

When planning the *sampling strategy* (i.e. the couple *sampling design* and *estimator*) in a finite population survey setup the statistician is often looking for "the most"efficient strategy. Godambe (1955), Lanke (1973) and Cassel *et. al.* (1977) show that there is no uniformly best estimator, in the sense best for all populations. There is also no best design. Nevertheless, it is often possible to identify a set of strategies that can be considered as candidates. The task is to choose one among this set.

The setup that will be used through this paper is as follows. We are interested in the estimation of the total of a study variable. The values of an auxiliary variable are known from the planning stage for all the elements. We will focus on study variables that are right-skewed and we will assume that ideal survey conditions hold.

The objective is to use the auxiliary variable for obtaining an efficient strategy, where efficiency will be understood in terms of design-based variance. The strategy that couples proportional-to-size sampling with the regression estimator has sometimes been called optimal (see, for example, Särndal *et. al.* (1992); Brewer (1963); Isaki and Fuller (1982). Wright (1983) proposed strong model-based stratification, which couples stratified simple random sampling with the regression estimator.

Both strategies mentioned above rely on the assumption that the finite population of interest can be seen as a realization of a particular super-population model (shown in section 2.2). The aim of this paper is to compare the strategies and try to empirically answer the following questions: **i.** when the super-population model is correctly specified, is in fact πps—reg the best strategy?, and **ii.** if πps—reg was the best strategy under a correctly specified model, is it still the best under a misspecified model?

2 Framework

The aim is to estimate the total $t_y = \sum_U y_k$ of one study variable $\boldsymbol{y}' = (y_1, y_2, \cdots, y_N)$ on a population U with unit labels $\{1, 2, \cdots, N\}$ where N is known. It is assumed that there is one auxiliary variable $\boldsymbol{x}' = (x_1, x_2, \cdots, x_N)$ known for each element in U. A without-replacement sample s of size n is selected and y_k is observed for all units $k \in s$.

In this section we describe five out of the six strategies that are spanned by two designs, stratified simple random sampling —STSI— and proportional-to-size sampling —πps— on the one hand, and three estimators, the Horvitz Thompson estimator —HT—, the post-stratified estimator —pos— and the regression estimator —reg— on the other hand.

The reason behind these strategies is as follows. Regarding the design, simple random sampling does not make any use of the auxiliary information, whereas πps makes, what we call, strong use of it. STSI lies in between, we will say that it makes weak use of the auxiliary information. In a similar way, regarding the estimator, the π-estimator does not make use of the auxiliary information, as opposed to the reg-estimator that makes strong use of it. The pos-estimator lies in between, making weak use of the auxiliary information. Then, the six strategies make use of the auxiliary information at a different degree.

The general regression estimator —GREG—, is described in the first part of this section. The HT, pos and reg estimators are shown as particular cases of it. In the second part of the section, the super-population model that will be considered is described.

2.1 The GREG estimator

The auxiliary vector $\boldsymbol{x}_k = (x_{1k}, x_{2k}, \cdots, x_{Jk})$ is available for every $k \in U$. The General Regression —GREG— estimator of t_y is defined as

$$\hat{t}_{\text{GREG}} \equiv \sum_U \hat{y}_k + \sum_s \frac{e_{ks}}{\pi_k}$$

where $e_{ks} = y_k - \hat{y}_k$ and $\hat{y}_k = \boldsymbol{x}_k \hat{\boldsymbol{B}}$ with $\hat{\boldsymbol{B}} = \left(\sum_s \frac{\boldsymbol{x}_k' \boldsymbol{x}_k}{a_k \pi_k}\right)^{-1} \sum_s \frac{\boldsymbol{x}_k' y_k}{a_k \pi_k}$. The a-values will be defined later.

No closed expression for the variance of the GREG-estimator is available, but it can be approximated by (see Särndal *et. al.*, 1992)

$$\text{AV}_p\left(\hat{t}_{\text{GREG}}\right) = \sum_U \sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \qquad \text{with } E_k = y_k - \boldsymbol{x}_k \boldsymbol{B} \tag{1}$$

where $\boldsymbol{B} = \left(\sum_U \frac{\boldsymbol{x}_k' \boldsymbol{x}_k}{a_k}\right)^{-1} \sum_U \frac{\boldsymbol{x}_k' y_k}{a_k}$..

This is the same expression as the variance of the HT-estimator with E_k instead of y_k. From now on we will write $\text{V}_p\left(\hat{t}_{\text{GREG}}\right)$ instead of $\text{AV}_p\left(\hat{t}_{\text{GREG}}\right)$.

Note that the following are sufficient (but not necessary) conditions for (1) being equal to zero:

i. $E_k = 0$ for all $k \in U$. A GREG-estimator that correctly explains the study variable will lead to small residuals and therefore a small variance.

ii. $\pi_k = n\, E_k / t_E$. Even if the E_k were known, this condition cannot be fulfilled, as some residuals will be smaller than zero and some will be larger than zero, leading to negative probabilities. Also t_E is often very close to zero, leading to many $\pi_k > 1$.

iii. $\pi_k = n \frac{|E_k|}{t_{|E|}}$ together with $\pi_{kl} = \pi_k \pi_l$ if $k \in U^+$ and $l \in U^-$. One method for satisfying the second part of the condition would be to stratify the population U with respect to the sign of E_k, which, however, requires a knowledge about the finite population at a level of detail that is seldom available. We will assume that this

knowledge is not available and we will settle for the next condition.
iii'. $\pi_k = n\frac{|E_k|}{t_{|E|}}$, which is obtained if we drop the $\pi_{kl} = \pi_k\pi_l$ part of condition **iii**. Note that **iii'** does not yield a zero variance. Why to consider condition **iii'** then? First, as will be shown below, the HT-estimator can be seen as a particular case of the GREG-estimator, then if we have $y_k > 0$, it is equivalent to condtion **ii** above, leading to a zero variance. Second, it will be useful for defining the so-called optimal strategy and model-based stratification in a more intuitive way, without explicitly defining concepts like anticipated variance.

As can be seen, in the context of the GREG-estimator, conditions **i** and **iii'** suggest the specific role of the design and the estimator in the sampling strategy. The estimator must explain the trend of the study variable with respect to the auxiliary variable, leading to small residuals. The design, on the other hand, must explain the residuals, in other words, how the study variable is spread around the trend.

The HT-estimator as a particular case of the GREG-estimator Let $\boldsymbol{x}_k = 0$ for all $k \in U$. If we allow $0/0 = 0$ (this terrible blasphemy is justified by using a generalized inverse in $\hat{\boldsymbol{B}}$ instead of the inverse, and noting that 0 is a generalized inverse of itself) we have that $\hat{\boldsymbol{B}} = \left(\sum_s \frac{\boldsymbol{x}_k'\boldsymbol{x}_k}{a_k\pi_k}\right)^- \sum_s \frac{\boldsymbol{x}_k'y_k}{a_k\pi_k} = 0$. Then $\hat{y}_k = \boldsymbol{x}_k\hat{\boldsymbol{B}} = 0$ and $e_{ks} = y_k - \hat{y}_k = y_k - 0 = y_k$. The GREG-estimator becomes

$$\hat{t}_{\text{GREG}} = \sum_U \hat{y}_k + \sum_s \frac{e_{ks}}{\pi_k} = \sum_U 0 + \sum_s \frac{y_k}{\pi_k} = \hat{t}_\pi$$

which explicitly shows that the π-estimator can be seen as the case where no auxiliary information is used into the GREG-estimator. Note also that $E_k = y_k - \boldsymbol{x}_k\boldsymbol{B} = y_k$.

The post-stratified estimator Let $a_k = c_j$ and $\boldsymbol{x}_k = (x_{1k}, x_{2k}, \cdots, x_{Jk})$ with x_{jk} defined as

$$x_{jk} = \begin{cases} 1 & \text{if } k \in U_j' \\ 0 & \text{if not} \end{cases}$$

where the U_j' $(j = 1, \cdots, J)$ form a partition of U. The post-stratified estimator, or simply pos-estimator, is obtained when this particular type of auxiliary information is used in the GREG-estimator. The residuals become $E_k = y_k - \bar{y}_{U_j'}$ $(k \in U_j')$ where $\bar{y}_{U_j'} = \frac{1}{N_j}\sum_{U_j'} y_k$ and N_j is the size of the j-th post-stratum.

The regression estimator Let $a_k = c$ and $\boldsymbol{x}_k = (1, z_k)$, with z_k the result of a known function applied on the known x_k. The regression estimator, or simply, reg-estimator, is obtained when this $\boldsymbol{x}_k$ is used in the GREG-estimator. The residuals become

$$E_k = y_k + B_2\frac{t_z}{N} - \frac{t_y}{N} - B_2 z_k \qquad \text{with } B_2 = \frac{Nt_{zy} - t_z t_y}{Nt_{z^2} - t_z^2} \tag{2}$$

where $t_y = \sum_U y_k$, $t_z = \sum_U z_k$, $t_{z^2} = \sum_U z_k^2$ and $t_{zy} = \sum_U z_k y_k$.

2.2 The super-population model and the strategies under comparison

We will assume that the statistician is willing to admit that the following model *adequately describes* the relation between the study variable, $\boldsymbol{y}$, and the auxiliary

variable, $\boldsymbol{x}$. The values of $\boldsymbol{y}$ are realizations of the model ξ_0

$$Y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k \tag{3}$$

where the ϵ_k are random variables satisfying $\mathrm{E}_{\xi_0}(\epsilon_k) = 0$, $\mathrm{V}_{\xi_0}(\epsilon_k) = \delta_3 x_k^{2\delta_4}$ and $\mathrm{E}_{\xi_0}(\epsilon_k \epsilon_l) = 0$ $(k \neq l)$. Moments are taken with respect to the model ξ_0 and δ_i are constant parameters. The term $\delta_0 + \delta_1 x_k^{\delta_2}$ in model ξ_0 will be called *trend*. The term $\delta_3 x_k^{2\delta_4}$ will be called *spread*. Brewer (1963; 2002, p. 111 and p. 200-201) shows rather heuristically that for most survey data $1/2 \leq \delta_4 \leq 1$ when $\delta_2 = 1$.

Cassel *et. al.* (1977) prove that if model (3) holds, the minimum variance strategy in the class of linear, design-unbiased estimators and fixed size design is πps with $\pi_k = n\frac{x_k^{\delta_4}}{t_{x^{\delta_4}}}$ where $t_{x^{\delta_4}} = \sum_U x_k^{\delta_4}$ and GREG with known parameters. This is a theoretical strategy that cannot be implemented in practice, as it assumes that the model is correct and its parameters known. Model ξ_0 as defined above may be used for assisting the definition of the sampling strategy as follows.

Revisiting (πps—reg) If model ξ_0 is assumed, it is natural to consider the GREG-estimator with $\boldsymbol{x}_k = (1, x_k^{\delta_2})$ at the estimation stage. In this case, we have $y_k = B_0 + B_1 x_k^{\delta_2} + E_k$ but also $y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k^*$, where E_k is the residual resulting from fitting the regression underlying the GREG-estimator and ϵ_k^* is a realization of the random variable ϵ_k. Then

$$E_k = (\delta_0 - B_0) + (\delta_1 - B_1) x_k^{\delta_2} + \epsilon_k^* \approx \epsilon_k^*$$

In order to minimize the variance in the sense of condition **iii'** one would like to use a design having $\pi_k = n\frac{|E_k|}{t_{|E|}}$. Using the approximation above we get

$$|E_k| \approx |\epsilon_k^*| = \sqrt{\epsilon_k^{*2}} \approx \sqrt{\mathrm{E}_{\xi_0}(\epsilon_k^2)} = \sqrt{\delta_3 x_k^{2\delta_4}} = \delta_3^{1/2} x_k^{\delta_4}$$

Therefore the design must satisfy $\pi_k = n\frac{x_k^{\delta_4}}{t_{x^{\delta_4}}}$.

This strategy is often found in the literature and referred as "optimal", in the sense that it minimizes the anticipated variance, a model dependent statistic. A comprehensive definition of the strategy can be found in, for example, Särndal *et. al.* (1992). We have decided to introduce the strategy in a more intuitive form, without explicitly defining concepts like anticipated variance. The sampling strategy defined in this way will be denoted by $\pi\mathrm{ps}(\delta_4)$—$\mathrm{reg}(\delta_2)$.

Revisiting (STSI—reg) There are many ways for implementing STSI. We focus on one version of Wright's (1983) model-based stratification, as described in Särndal et. al. (1992, sec. 12.4). Assuming the model ξ_0, the GREG estimator with $\boldsymbol{x}_k = (1, x_k^{\delta_2})$ is used again. Regarding the design, STSI is considered. In order to minimize the variance (1) we want the variances $S_{EU_h}^2$ to be as small as possible. The model is used in the same way as above, defining $|E_k| \approx \delta_3^{1/2} x_k^{\delta_4}$.

Ignoring the scale factor δ_3, the resulting "residuals"are stratified using the approximation to the cum$\sqrt{f}$-rule together with Neyman allocation. Wright (1983) showed a lower bound for the efficiency of this strategy compared to the optimal one. The sampling strategy defined in this way will be denoted by $\mathrm{STSI}(\delta_4)$—$\mathrm{reg}(\delta_2)$.

Revisiting STSI—HT As mentioned before, the HT-estimator can be seen as the case when null-auxiliary information is used in the GREG-estimator. In this case the residuals are $E_k = y_k$ and in order to have a small variance (1) we look for strata leading to a small sum-of-squares-within $SSW_y = \sum_{h=1}^{H} \sum_{U_h} (y_k - \bar{y}_{U_h})^2$.

Using the model, the y_k may be approximated by $y_k \approx \delta_0 + \delta_1 x_k^{\delta_2}$, which leads to

$$\sum_{h=1}^{H} \sum_{U_h} (y_k - \bar{y}_{U_h})^2 \approx \delta_1^2 \sum_{h=1}^{H} \sum_{U_h} \left(x_k^{\delta_2} - x_{U_h}^{\bar{\delta}_2} \right)^2$$

So we have to look for strata leading to small SSW of $x_k^{\delta_2}$. The strata are then created using the approximation to the cum$\sqrt{f}$-rule on $x_k^{\delta_2}$ together with Neyman allocation. The strategy defined in this way will be denoted by STSI(δ_2)—HT and will be considered as a benchmark.

Revisiting πps—pos Regarding the definition of the post-strata, recall that the residuals of the pos-estimator can be written as $E_k = y_k - \bar{y}_{U'_j}$ for all $k \in U'_j$. When looking for post-strata that minimize these E_k, a natural criterion would be to minimize its square sum $\sum_U E_k^2$, but note that $\sum_U E_k^2 = \sum_{j=1}^{J} \sum_{U'_j} E_k^2 = \sum_{j=1}^{J} \sum_{U'_j} \left(y_k - \bar{y}_{U'_j} \right)^2$, which is the SSW shown in STSI(δ_2)—HT above. Therefore the post-strata will be created using the approximation to the cum$\sqrt{f}$-rule on $x_k^{\delta_2}$.

Regarding the inclusion probabilities, we use an approach analogous to the one considered for πps—reg. Note that $y_k = \bar{y}_{U'_j} + E_k$ but also $y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k^*$. Then $E_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k^* - \bar{y}_{U'_j}$. In order to minimize the variance in the sense of condition **iii'** one would like to use a design having $\pi_k = n \frac{|E_k|}{t_{|E|}}$. As the E_k are unknown, we use

$$|E_k| \approx \sqrt{\mathrm{E}_{\xi_0} \left[\left(\delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k - \bar{Y}_{U'_j} \right)^2 \right]} \approx \delta_3^{1/2} \sqrt{\left(1 + \frac{2}{N_j} \right) x_k^{2\delta_4} + \frac{t_{x^{2\delta_4}, U'_j}}{N_j^2}} \equiv \delta_3^{1/2} v_k$$

where the approximation $x_k^{\delta 2} \approx \bar{x}_{U'_j}^{\delta_2}$ was used in order to obtain the last expression, N_j is the size of the j-th post-stratum and $t_{x^{2\delta_4}, U'_j} = \sum_{U'_j} x_k^{2\delta_4}$. Using condition **iii'** and these proxies for the residuals, we have that the design must satisfy $\pi_k = n \frac{v_k}{t_v}$ where $t_v = \sum_U v_k$. The sampling strategy defined in this way will be denoted by πps—pos(δ_2).

Revisiting STSI—pos In this case the post-stratified estimator is used again in the same way as in the strategy above, this means that post-strata are created using the approximation to the cum$\sqrt{f}$-rule on $x_k^{\delta_2}$. The same approximated residuals are then obtained. The strata are defined by applying the approximation to the cum$\sqrt{f}$ on the v_k defined above and the sample is allocated using Neyman allocation. The sampling strategy defined in this way will be denoted by STSI—pos(δ_2).

References

Brewer, K.R.W. (1963). *A Model of Systematic Sampling with Unequal Probabilities.* Australian Journal of Statistics, **5**, 5-13.

Brewer, K.R.W. (2002). *Combined Survey Sampling Inference: Weighing Basu?s Elephants*. London: Arnold.

Cassel, C.M., Särndal, C. E. and Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.

Dalenius, T. and Hodges, J.L. (1959) *Minimum variance stratification*. Journal of the American Statistical Association, **54**, 88-101.

Godambe, V.P. (1955). *A unified theory of sampling from finite populations*. Journal of the Royal Statistical Society, Series B **17**, 269-278.

Hanif, M. and Brewer K. R. W. (1980). *Sampling with Unequal Probabilities without Replacement: A Review*. International Statistical Review **48**, 317-335.

Holmberg, A. and Swensson, B. (2001). *On Pareto πps Sampling: Reflections on Unequal Probability Sampling Strategies*. Theory of Stochastic Processes, **7(23)**, 142-155.

Isaki, C.T. and Fuller, W.A. (1982) *Survey design under the regression superpopulation model*. Journal of the American Statistical Association **77**, 89-96.

Kozak, M. and Wieczorkowski, R. (2005). *πps Sampling versus Stratified Sampling ? Comparison of Efficiency in Agricultural Surveys*. Statistics in Transition, **7**, 5-12.

Lanke, J. (1973). *On UMV-estimators in Survey Sampling*. Metrika **20**, 196 202.

Rosén, B. (1997). *On sampling with probability proportional to size*. Journal of statistical planning and inference **62**, 159-191.

Rosén, B. (2000a). *Generalized Regression Estimation and Pareto πps*. R&D Report 2000:5. Statistics Sweden.

Rosén, B. (2000b). *On inclusion probabilities for order πps sampling*. Journal of statistical planning and inference **90**, 117-143.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Wright, R.L. (1983). *Finite Population Sampling with Multivariate Auxiliary Information*. Journal of the American Statistical Association, **78**, 879 884.