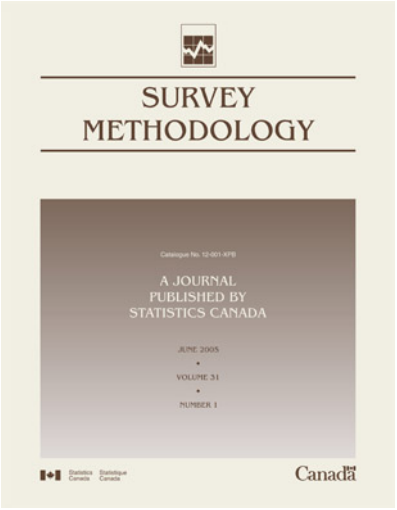
Catalogue no. 12-001-XIE

Survey Methodology

June 2002



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-800-263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Depository Services Program inquiries	1-800-700-1033
Fax line for Depository Services Program	1-800-889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About us > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2002

Published by authority of the Minister responsible for Statistics Canada

October 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 77-85
Statistics Canada, Catalogue No. 12-001-XIE

Unbiased Estimation by Calibration on Distribution in Simple Sampling Designs Without Replacement

Yves Tillé [1]

Abstract

The post-stratified estimator sometimes has empty strata. To address this problem, we construct a post-stratified estimator with post-strata sizes set in the sample. The post-strata sizes are then random in the population. The next step is to construct a smoothed estimator by calculating a moving average of the post-stratified estimators. Using this technique it is possible to construct an exact theory of calibration on distribution. The estimator obtained is not only calibrated on distribution, it is linear and completely unbiased. We then compare the calibrated estimator with the regression estimator. Lastly, we propose an approximate variance estimator that we validate using simulations.

Key Words: Unbiased estimation; Calibration on a distribution function; Conditional inclusion probabilities; Weighting.

1. Introduction

It is possible during a survey by sampling to identify the values of an auxiliary character for all population units. This information may be available when the units are selected in a database containing other variables of interest. The temptation is then to calibrate the results of a survey on this auxiliary information. The decision is made either to retain from this auxiliary variable only certain functions (moments, sizes) for the purpose of using a calibration method (see for example Deville and Särndal 1992 or Estevao, Hidiroglou and Särndal 1995), or this variable can be divided into classes with the view to using a post-stratified estimator.

If the decision is to opt for the post-stratified estimator, deciding on the strata divisions can be delicate. Theoretically, the strata must be defined prior to the selection of the sample. Where should the post-strata boundaries be placed? What size should the post-strata be? This latter question is the most embarrassing because the main problem with post-stratification is the possibility of obtaining empty post-strata. This means that the post-strata have to be large enough so that the probability of obtaining an empty post-stratum is negligible. These problems are not limited to post-stratified estimators. Indeed, it is also possible to have no regression or calibrated estimators for some samples.

Our objective is to define a new method of using auxiliary information in the population. This method is based on the definition of post-strata for which the number of units is set in the sample and not in the population. In this way, it is possible to import into the estimator complex auxiliary information resulting from knowledge of all of the values taken by the auxiliary variable, while avoiding both the problem of defining post-strata borders and the problem of empty post-strata.

This article is organized as follows. In section 2, the notation is defined and in section 3, we describe the principle of rank conditioning, which is used to define the unbiased estimators in section 4. In section 5, the smoothed estimator is defined, and a specific case is examined in detail in section 6. Section 7 contains an application of the estimation of a distribution function. In section 8, this new estimator is compared with the regression estimator and the estimator for a simple design without replacement. Computation of variance is discussed in section 9. As a result of the impossibility of providing an exact solution, an approximation is provided in section 10, which is tested by simulations in section 11. Lastly, general conclusions are presented in section 12.

2. Notation

We assume a population composed of N observation units, with the labelling being denoted as $\{1, ..., k, ..., N\}$. In this population, we are interested in a character of interest $Y_k, k \in U$. The objective is to estimate the total $Y = \sum_{k \in U} Y_k$. We select a random sample S of fixed size n by means of a simple random design without replacement. We denote I_k the random indicator variable, which takes the value 1 if the unit k is in the sample and 0 if not. The inclusion probabilities first order are therefore defined by $\Pr(k \in S) = \pi_k = n/N, k \in U,$ and the second order inclusion probabilities by $\Pr(k, l \in S) = \pi_{kl} = n(n-1)/(N(N-1)), k \neq l \in U.$

We will be interested in the class of linear estimators of Y, which is written as

$$\hat{Y}_w = \sum_{k \in S} w_k Y_k,$$

1. Yves Tillé, Groupe de Statistique, Université de Neuchâtel, Espace de l'Europe 4, Case postale 827, 2002 Neuchâtel, Suisse. E-mail : yves.tille@unine.ch.

where the weights w_k may depend on the sample S and therefore be random.

One of the possibilities is to take $w_k = 1/\pi_k = n/N$, which gives the Horvitz-Thompson estimator,

$$\hat{Y}_{\mathrm{HT}} = \sum_{k \in S} \frac{Y_k}{\pi_k} = \frac{N}{n} \sum_{k \in S} Y_k,$$

which is unbiased.

We will be focusing instead on the more general class of conditionally weighted estimators (Tillé 1998, 1999a) where the units are weighted by inverses of conditional inclusion probabilities. If Z is some statistic, then the conditionally weighted estimator

$$\hat{Y}_Z = \sum_{k \in S} \frac{Y_k}{E(I_k \mid Z)} \tag{1}$$

is strictly unbiased if and only if $E(I_k \mid Z) > 0$, for all $k \in U$. In effect,

$$E(\hat{Y} \mid Z) = \sum_{k \in U} \frac{E(I_k \mid Z) Y_k}{E(I_k \mid Z)} = Y.$$

Since the estimator is conditionally unbiased, it is also unconditionally unbiased. Depending on which statistic Z is used, estimator (1) generalizes the stratified estimator as well as (a close approximation) the regression estimator (see Tillé 1998).

3. Conditioning on Ranks

Let us now assume that the N values $X_1, ..., X_k, ..., X_N$ of an auxiliary character x are known for N units of the population. First, we assume that all of the X_k take separate values (this hypothesis will be removed in section 5). The rank R_k of unit k is

$$R_k = \#\{l \in U \mid X_l \le X_k\}.$$

Moreover, we denote r_j, $j = 1, ..., n$, the ordered population ranks of the n selected units in the sample, thus $r_1 < r_2 < ... < r_{n-1} < r_n$. The r_j are random variables with a negative hypergeometric distribution (see Tillé 1999b).

The statistic used to define the conditional probabilities of inclusion is a subset of $\{r_1, ..., r_j, ..., r_n\}$. First, we define

- an integer q such that $2 \le q \le n$, defining the period,
- an integer b such that $2 \le b$, defining the border,
- an integer l such that $b \le l \le b + q - 1$, defining the interval.

The quantities $q, b,$ and l serve to define a subset of indices:

$$E_l = \{r_l, r_{l+q}, r_{l+2q}, ..., r_{l+hq}, ..., r_{l+Hq}\},$$

$$\text{for } l = b, ..., b + q - 1.$$

For example, if $n = 18$, $q = 4$, $b = 3$, then

$$E_3 = \{r_3, r_7, r_{11}, r_{15}\},$$

$$E_4 = \{r_4, r_8, r_{12}, r_{16}\},$$

$$E_5 = \{r_5, r_9, r_{13}\},$$

$$E_6 = \{r_6, r_{10}, r_{14}\}.$$

The conditional inclusion probability is computed in relation to one of the E_l.

The value of H is defined in such a way that $l + Hq \le n - b + 1$ and thus H is the largest integer such that $H \le (n - b - l + 1)/q$. It is clear that H depends on l.

The next step is to compute the inclusion probabilities:

$$E(I_k \mid E_l) = \begin{cases} 1 & \text{if } k \in E_l \\ \dfrac{q-1}{r_{l+hq} - r_{l+(h-1)q} - 1} & \text{if } r_{l+(h-1)q} < k < r_{l+hq},\ h = 1, ..., H \\ \dfrac{l-1}{r_l - 1} & \text{if } k < r_l \\ \dfrac{n - (l + Hq)}{N - r_{l+Hq}} & \text{if } k > r_{l+Hq}. \end{cases}$$

These inclusion probabilities are thus relatively uneven. However, they are all positive, including the borders. It is important to use a border $b \ge 2$ so that the first and the last post-stratum are not empty.

4. Class of Unbiased Estimators

Since $E(I_k \mid E_l) > 0$, we can construct an estimator that is unbiased and even conditionally unbiased with respect to E_l. By denoting $y_1, ..., y_j, ..., y_n$ the n values taken by the units in the sample ordered according to the R_k, we obtain

$$\hat{Y}_l = \sum_{k \in S} \frac{Y_k}{E(I_k \mid E_l)}$$

$$= \frac{r_l - 1}{l - 1} \sum_{j=1}^{l-1} y_j + y_l$$

$$+ \sum_{h=1}^{H} \left(\frac{r_{l+hq} - r_{l+(h-1)q} - 1}{q-1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j} + y_{l+hq} \right)$$

$$+ \frac{N - r_{l+Hq}}{n - (l + Hq)} \sum_{j=l+Hq+1}^{n} y_j$$

$$= N_{0|l}\, \hat{\bar{y}}_{0|l} + y_l + \sum_{h=1}^{H} (N_{h|l} \hat{\bar{y}}_{h|l} + y_{l+hq}) + N_{H+1|l}\, \hat{\bar{y}}_{H+1|l}$$

where

$$N_{0|l} = r_l - 1,$$

$$N_{h|l} = r_{l+hq} - r_{l+(h-1)q} - 1,\ h = 1, \ldots, H,$$

$$N_{H+1|l} = N - r_{l+Hq},$$

$$\hat{\bar{y}}_{0|l} = \frac{1}{l-1} \sum_{j=l}^{l-1} y_j,$$

$$\hat{\bar{y}}_{h|l} = \frac{1}{q-1} \sum_{j=1}^{q-1} y_{l+(h-1)q+j},\ h = 1, \ldots, H,$$

and

$$\hat{\bar{y}}_{H+1|l} = \frac{1}{n - (l + Hq)} \sum_{j=l+Hq+1}^{n} y_j .$$

This estimator is in reality a post-stratified estimator where the sizes of the post-strata are set in the sample. Since $E(I_k \mid E_l) > 0$, $\hat{Y}_l$ is strictly unbiased unconditionally and conditionally to E_l, which is clearly not the case for the traditional post-stratified estimator, because the latter has a non-zero probability of having an empty post-stratum. By setting the size of the post-strata in the sample, creating empty post-strata becomes impossible. The corresponding size of the post-stratum in the population is a random variable $N_{h|l}$.

The estimator $\hat{Y}_l$ has another interesting property. By using the definition of the $E(I_k \mid E_l)$, we can quite easily show that

$$\sum_{k \in S} \frac{1}{E(I_k \mid E_l)} = N.$$

The estimator is thus calibrated on the size of the population. This property, which is also held by the Horvitz-Thompson estimator in simple designs, is therefore not lost.

Units where the ranks are in E_l are called pivot units, and are assigned a weight equal to 1, which makes the weights very unequal. A downside to $\hat{Y}_l$ is the use of widely dispersed weights. This problem can be resolved by smoothing the estimators.

5. Smoothing Estimators

To resolve the problem of the dispersion of the weights, we compute a moving average for the estimators as follows:

$$\hat{Y}_c = \frac{1}{q} \sum_{l=b}^{b+q-1} \hat{Y}_l.$$

$\hat{Y}_c$ retains all of the properties of the $\hat{Y}_l$. This means that it is unbiased, calibrated on N and linear and can therefore be written as

$$\hat{Y}_c = \sum_{j=1}^{n} w_j\, y_j,$$

where

$$w_j = \begin{cases} \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{r_l - 1}{l-1}, & j < b, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{m^-(j+l-b-q)} - 1}{j+l-b-m^-(j+l-b-q) - 1} + 1 \right), & b \le j < b+q-1, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right), & b+q-1 \le j \le n-b+2-q, \\ \frac{1}{q} \left(\sum_{l=b}^{b+q-1} \frac{r_m + (j+l-b) - r_{j+l-b-q} - 1}{m + (j+l-b) - j + l - b - q - 1} + 1 \right), & n-b+2-q < j \le n-b+1, \\ \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N + 1 - r_{n+1-l} - 1}{n + 1 - (n + 1 - l) - 1} = & \\ \qquad \frac{1}{q} \sum_{l=b}^{b+q-1} \frac{N - r_{n+1-l}}{l-1}, & n-b+1 < j, \end{cases}$$

$$m^-(x) = \begin{cases} 0 & \text{if } x < b \\ x & \text{if not,} \end{cases}$$

$$m^+(x) = \begin{cases} n+1 & \text{if } x > n-b+1 \\ x & \text{if not,} \end{cases} \qquad (2)$$

$$r_0 = 0, \text{ and } r_{n+1} = N+1.$$

Under the apparent complexity arising from the specific treatment of the borders, the weighting system is relatively simple. In the case where we are not too close to the borders, it takes the value

$$w_j = \frac{1}{q}\left(\sum_{l=b}^{b+q-1} \frac{r_{j+l-b} - r_{j+l-b-q} - 1}{q-1} + 1 \right)$$

$$= \frac{1}{q(q-1)} \sum_{\alpha=0}^{q-1} (r_{j+\alpha} - r_{j+\alpha-q}).$$

If all of the values of the auxiliary variable are not distinct, we can assign the unit ranks with common values randomly. For example, if $X_1 = 2, X_2 = 5, X_3 = 5, X_3 = 5, X_4 = 7, X_5 = 8,$ we select with a probability ½, between, ranks $R_1 = 1, R_2 = 2, R_3 = 3, R_4 = 4, R_5 = 5,$ or $R_1 = 1, R_2 = 3, R_3 = 2, R_4 = 4, R_5 = 5.$ We the compute the smoothed estimator for each permutation, and we calculate their mean. The advantage of this method is that it preserves an unbiased estimator. In effect, for each possible permutation, the estimator is unbiased. In practice, it is not necessary to compute estimators for all of the permutations. We can calculate the estimator for one permutation and then simply equalize the weights of the units having the same values for the variable x.

6. Case where $q = 2, b = 2$

When $q = 2,$ and $b = 2,$ we obtain after a few calculations

$$\hat{Y}_c = \frac{1}{2}\left\{ \sum_{j=3}^{n-2} y_j (r_{j+1} - r_{j-1}) + \frac{r_3 + 2r_2 - 3}{2} y_1 + \frac{r_3 + 1}{2} y_2 + \frac{r_{n+1} - r_{n-2} + 1}{2} y_{n-1} + \frac{3r_{n+1} - 2r_{n-1} - r_{n-2} - 3}{2} y_n \right\}$$

$$= \frac{1}{2}\left\{ \begin{array}{l} \displaystyle\sum_{j=1}^{n} y_j (r_{j+1} - r_{j-1}) \\ + y_1 \dfrac{r_3 - 3}{2} + y_2 \dfrac{2r_1 + 1 - r_3}{2} \\ + y_{n-1} \dfrac{r_{n+1} + r_{n-2} + 1 - 2r_n}{2} + y_n \dfrac{r_{n+1} - r_{n-2} - 3}{2} \end{array} \right\},$$

where $r_0 = 0$ and $r_{n+1} = N + 1.$ This brings us to an estimator proposed by Ren (2000, page 140) and obtained using a calibration argument. The way in which the borders are managed is the only slight difference.

Example 1: With a population of size $N = 20.$ Let us assume that the values of the variable of interest are found in Table 1. We also assume that the sample of size $n = 7$ is composed of the units with ranks {3, 7, 8, 11, 12, 15, 17}. If we take $q = 2, l = 2, b = 2$ we obtain $E_2 = \{r_2, r_4, r_6\} = \{7, 11, 15\}.$ We can then calculate $E(I_k \, E_2 = \{7, 11, 15\}).$ The conditional inclusion probabilities are as follows:

$$E(I_3 \mid E_2 = \{7, 11, 15\}) = 1/6,$$
$$E(I_7 \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_8 \mid E_2 = \{7, 11, 15\}) = 1/3,$$
$$E(I_{11} \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_{12} \mid E_2 = \{7, 11, 15\}) = 1/3,$$
$$E(I_{15} \mid E_2 = \{7, 11, 15\}) = 1,$$
$$E(I_{17} \mid E_2 = \{7, 11, 15\}) = 1/5.$$

Table 1
Example of a Population of Size $N = 20$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_k	9	71	72	35	91	14	3	36	64	38	81	52	78	62	86	16	20	59	84	55
R_k	2	14	15	6	20	3	1	7	14	8	17	9	16	12	19	4	5	11	18	10

The estimator

$$\hat{Y}_0 = \sum \frac{y_k}{E(I_k \mid E_2 = \{7, 11, 15\})}$$

is therefore unbiased and conditionally unbiased. Further, it is linear and

$$\sum_{k \in S} \frac{1}{E(I_k \mid E_2 = \{7, 11, 15\})} = N.$$

However, if we take $q = 2, l = 3, b = 2,$ we obtain $E_3 = \{r_3, r_5\} = \{8, 12\}.$ Using the same example, we then compute $E(I_k \mid E_3 = \{8, 12\}),$ and we obtain

$$E(I_3 \mid E_3 = \{8, 12\}) = 2/7,$$
$$E(I_7 \mid E_3 = \{8, 12\}) = 2/7,$$
$$E(I_8 \mid E_3 = \{8, 12\}) = 1,$$
$$E(I_{11} \mid E_3 = \{8, 12\}) = 1/3,$$
$$E(I_{12} \mid E_3 = \{8, 12\}) = 1,$$
$$E(I_{15} \mid E_3 = \{8, 12\}) = 2/8 = 1/4,$$
$$E(I_{17} \mid E_3 = \{8, 12\}) = 2/8 = 1/4.$$

The estimator

$$\hat{Y}_1 = \sum \frac{y_k}{E(I_k \mid E_3 = \{8, 12\})}$$

is also unbiased and linear.

Lastly, we compute the mean of the two estimators:

$$\hat{Y}_c = \frac{\hat{Y}_0 + \hat{Y}_1}{2}.$$

The weights are obtained simply by calculating the mean of the weights of estimators $\hat{Y}_0$ and $\hat{Y}_1$, and have the values

$$w_3 = (6+7/2)/2 = 19/4,$$
$$w_7 = (1+7/2)/2 = 9/4,$$
$$w_8 = (3+1)/2 = 2,$$
$$w_{11} = (1+3)/2 = 2,$$
$$w_{12} = (3+1)/2 = 2,$$
$$w_{15} = (1+4)/2 = 5/2,$$
$$w_{17} = (5+4)/2 = 9/2.$$

Estimator $\hat{Y}_c$ is linear and strictly unbiased.

7. Application to the Estimation of the Distribution

There are several ways to appropriately use auxiliary information to estimate a distribution function. A description of these techniques can be found in Ren (2000) and in Wu and Sitter (2001). The method that we suggest also makes it possible to estimate the distribution. The distribution in the population is defined by

$$F_1(y) = \frac{1}{N}\sum_{k\in U} I\{y_k \le y\},$$

and can be estimated by

$$\hat{F}_1(y) = \frac{\sum_{k\in S} w_k I\{y_k \le y\}}{\sum_{k\in S} w_k},$$

where $I\{y \le y_k\}$ is the indicator function, and the w_k are the weights allocated to the units k which have the value $1/\pi_k = N/n$ for the Horvitz-Thompson estimator, and which are given in (2) for the calibrated estimator.

Note that the two functions are discrete, but that there are far fewer jumps in S than in U. To offset the differences in the distributions between the sample and the population, we have smoothed the distribution functions by using, as Deville (1995) did, a linear interpolation of the centres of the risers, which involves defining $F_2(y)$ by linking the points

$$\frac{1}{2}\{F_1(y_k) - F_1(y_k - \varepsilon)\},$$

for $k \in U$, where ε is a strictly positive, arbitrarily small real number. We then define $\hat{F}_2(y)$ by linking the points

$$\frac{1}{2}\{\hat{F}_1(y_k) - \hat{F}_1(y_k - \varepsilon)\},$$

for the sample.

Example 2: A population of size $N = 1,000$ was generated using independent log-normal variables that are equally distributed. A sample of size $n = 16$ was then selected and we set $h = 5$. Figure 1 gives $F_2(x)$ in the population.

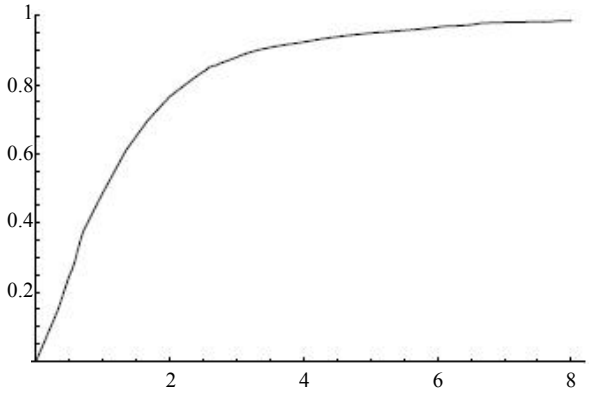


Figure 1. Population distribution function.

Figure 2 shows $F_2(x)$ and the distribution estimated by the Horvitz-Thompson estimator. Lastly, Figure 3 shows $F_2(x)$ and the distribution estimated by the calibrated estimator

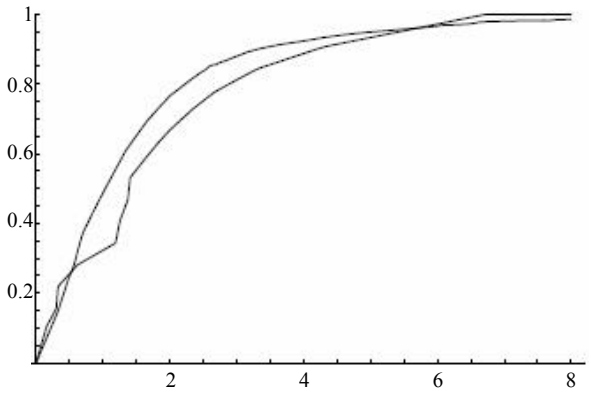


Figure 2. Population distribution function and Horvitz-Thomson distribution estimator.

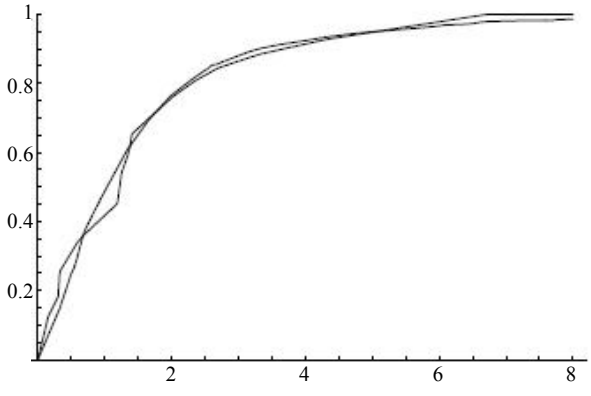


Figure 3. Population distribution function and calibrated distribution estimator.

8. Comparison with the Regression Estimator

In order to compare the qualities of the proposed estimator, a series of simulations was conducted to compare the estimator calibrated on distribution with the Horvitz-Thompson estimator and the regression estimator. Three populations of size 1,000 were generated by means of the following models.

- *Model A Linear dependence*: The population is generated using the model $X_k \sim N(0,1)$ and $Y_k = X_k + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0,1)$. The coefficient of correlation obtained in the population is $\rho = 0.616154$.
- *Model B Non-linear dependence* 1: The population is generated using the model $X_k \sim N(0,1)$ and $Y_k = (0.2 + X_k)^2 + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0,1)$. The coefficient of correlation obtained in the population is $\rho = 0.28975$.
- *Model C Non-linear dependence* 2: The population is generated using the model $X_k \sim N(0,1)$ and $Y_k = (0.4 + X_k)^2 + 1.33333 \times \varepsilon_k$ where $\varepsilon_k \sim N(0,1)$. The coefficient of correlation obtained in the population is $\rho = 0.476158$.

In each population, 100,000 samples of size 100 were selected. Three weighting systems were computed for each sample.

1. the weights associated with the simple design $w_k = N/n$,
2. the weights of the distribution calibrated estimator given in (2) using the window $q = 10$ and border $b = 6$,
3. the weights of the regression estimator given by

$$w_k = \frac{N}{n} + (X - \hat{X}_{\mathrm{HT}}) \frac{(X_k - \hat{\bar{X}})}{\sum_{k \in S} (X_k - \hat{\bar{X}})^2},$$

where X is the total of the X_k in the population, $\hat{X}_{\mathrm{HT}}$ is the Horvitz-Thompson estimator of X, and $\hat{\bar{X}} = \hat{X}_{\mathrm{HT}} / N$.

Using these weights, the estimator of the mean and of the nine deciles were calculated for each sample. We then estimate the variance of these estimators by means of the simulations.

The results are given in Tables 2, 3 and 4. The variances were brought to 1 for the simple design. For the linear model, the regression estimator is slightly preferable. However, in the non-linear case, the distribution calibrated estimator significantly increases the precision on the mean and on the quantiles. This means that our proposed estimator is robust when there is a non-linear relationship between the auxiliary variable and the variable of interest.

Table 2
Model A: Estimator Variance
(Reference: Horvitz-Thompson = 1)

Parameter	Distribution calibration	Regression estim.
Mean	0.674422	0.632608
1st decile	0.905273	0.893876
2nd decile	0.815403	0.802082
3rd decile	0.842681	0.815071
4th decile	0.806749	0.768283
5th decile	0.783731	0.740765
6th decile	0.818051	0.782549
7th decile	0.794411	0.773794
8th decile	0.857114	0.844874
9th decile	0.884424	0.884032

Table 3
Model B: Estimator Variance
(Reference: Horvitz-Thompson = 1)

Parameter	Distribution calibration	Regression estim.
Mean	0.429689	0.953025
1st decile	0.913598	0.958656
2nd decile	0.919394	1.009270
3rd decile	0.829860	0.987950
4th decile	0.792094	0.989114
5th decile	0.703908	0.992023
6th decile	0.622705	1.009830
7th decile	0.550028	0.981249
8th decile	0.443828	1.010340
9th decile	0.549615	1.029120

Table 4
Model C: Estimator Variance
(Reference: Horvitz-Thompson = 1)

Parameter	Distribution calibration	Regression estim.
Mean	0.30768	0.808114
1st decile	0.95560	0.983582
2nd decile	0.85920	0.970913
3rd decile	0.73854	0.930401
4th decile	0.65728	0.950651
5th decile	0.60500	0.956807
6th decile	0.52139	0.930514
7th decile	0.45709	0.907537
8th decile	0.40752	0.903593
9th decile	0.39820	0.860050

9. Variance and Estimation of Variance

To compute the variance of $\hat{Y}_c$, we begin by computing the variance of $\hat{Y}_l$. Since $\hat{Y}_l$ is unbiased conditionally to E_l, we have

$$V(\hat{Y}_l) = EV(\hat{Y}_l \mid E_l).$$

As with each of the post-strata, conditionally to E_l the design is a fixed-size simple sampling without replacement, we have

$$V(\hat{Y}_l \mid E_l) = \sum_{h=0}^{H+1} N_{h|l}^2 V(\hat{\bar{y}}_{h|l})$$

$$= \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{n|l}} \frac{S_{h|l}^2}{n_{h|l}}, \qquad (3)$$

where

$$n_{0|l} = l - 1,$$

$$n_{h|l} = q - 1,\ h = 1, \ldots, H,$$

$$n_{H+1|l} = n - (l + Hq),$$

$$\bar{Y}_{0|l} = \frac{1}{N_{0|l}} \sum_{k=1}^{r_{l-1}} Y_{(k)},$$

$$\bar{Y}_{h|l} = \frac{1}{N_{h|l}} \sum_{k=r_{l+(h-1)q}+1}^{r_{l+hq-1}} Y_{(k)},\ h = 1, \ldots, H,$$

$$\bar{Y}_{H+1|l} = \frac{1}{N_{H|l}} \sum_{k=N-r_{l+Hq}+1}^{N} Y_{(k)},$$

$$S_{0|l}^2 = \frac{1}{N_{0|l} - 1} \sum_{k=1}^{r_{l-1}} (Y_{(k)} - \bar{Y}_{0|l})^2,$$

$$S_{h|l}^2 = \frac{1}{N_{h|l} - 1} \sum_{k=r_l+(h-1)q+1}^{r_l+h_q-1} (Y_{(k)} - \bar{Y}_{h|l})^2,\ h = 1, \ldots, H,$$

and

$$S_{H+1|l}^2 = \frac{1}{N_{H+1|l} - 1} \sum_{k=N-r_{l+Hq}+1}^{N} (Y_{(k)} - \bar{Y}_{H+1|0})^2,$$

where the $Y_{(k)}$ represent the values of Y_k sorted by increasing order of the X_k.

Note that it is very difficult to calculate the unconditional variance of $\hat{Y}_l$, that is, the expectation of $V(\hat{Y}_l \mid E_l)$. In effect, $N_{h|l}$ and $S_{h|l}^2$ are random. However, we can estimate $V(\hat{Y}_l \mid E_l)$ simply and obtain an unbiased estimator of the conditional variance (and thus of the variance) by simply estimating (3), by

$$\hat{V}(\hat{Y}_l \mid E_l) = \sum_{h=0}^{H+1} N_{h|l}^2 \frac{N_{h|l} - n_{h|l}}{N_{h|l}\, n_{h|l}} s_{h|l}^2, \qquad (4)$$

where

$$s_{0|l}^2 = \frac{1}{n_{0|l} - 1} \sum_{j=1}^{l-1} (y_j - \hat{\bar{y}}_{0|l})^2,$$

$$s_{h|l}^2 = \frac{1}{n_{h|l} - 1} \sum_{j=1}^{q-1} (y_{l+(h-1)q+j} - \hat{\bar{y}}_{h|l})^2,\ h = 1, \ldots, H,$$

and

$$s_{H+1|l}^2 = \frac{1}{n_{h|l} - 1} \sum_{j=l+Hq+1}^{n} (y_j - \hat{\bar{y}}_{H+1|l})^2.$$

The estimator $\hat{V}(\hat{Y}_l \mid E_l)$ is not only unbiased for $\hat{V}(\hat{Y}_l \mid E_l)$ but also for $\hat{V}(\hat{Y}_l)$.

10. Approximations for Computing the Variance

Unfortunately, computing the variance of $\hat{Y}_c$ becomes more complex because of covariances. In effect, we have

$$V(\hat{Y}_c) = \frac{1}{q^2} \sum_{l=b}^{b+q-1} \sum_{i=b}^{b+q-1} \mathrm{Cov}(\hat{Y}_l, \hat{Y}_i).$$

When $l = i$, the problem is to estimate $V(\hat{Y}_i)$, which is done easily. When $l \neq i$, it is necessary to compute

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) = E\,\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i \mid E_l) + \mathrm{Cov}(E(\hat{Y}_l \mid E_l),\ E(\hat{Y}_i \mid E_l)).$$

Since $E(\hat{Y}_l \mid E_l) = Y$, we obtain

$$\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i) = E\,\mathrm{Cov}(\hat{Y}_l, \hat{Y}_i \mid E_l) = EE(\hat{Y}_l, \hat{Y}_i \mid E_l) - Y^2.$$

Unfortunately, it does not appear possible to extricate the computation of $E(\hat{Y}_l,\ \hat{Y}_i \mid E_l)$ and therefore we must find an approximation.

One way is to find a value that is greater than the variance. Since

$$\mathrm{Cov}(\hat{Y}_l,\ \hat{Y}_i) \leq \sqrt{V(\hat{Y}_l)\, V(\hat{Y}_i)},$$

we have a greater value given by

$$V(\hat{Y}_c) \le \frac{1}{q^2} \sum_{l=b}^{b+q-1} \sum_{i=b}^{b+q-1} \sqrt{V(\hat{Y}_l)\, V(\hat{Y}_i)}$$

$$= \frac{1}{q^2} \left(\sum_{l=b}^{b+q-1} \sqrt{V(\hat{Y}_l)} \right)^2,$$

which can be estimated by

$$\hat{V}_1(\hat{Y}_c) = \frac{1}{q^2} \left(\sum_{l=b}^{b+q-1} \sqrt{\hat{V}(\hat{Y}_l \mid E_l)} \right)^2,$$

which comes back to estimating the standard deviation of the means by the mean of the standard deviations.

Lastly, a second possibility involves using a residuals technique. Generally, when an estimator is corrected using a calibration technique, the variance is estimated by means of a residuals technique (see Deville and Särndal 1992 and Deville 1999 on this topic). When computing the variance of $\hat{Y}_l$, it is possible to use a residuals technique to obtain the exact variance. Consider the variable

$$v_k(l) = \begin{cases} \left(\dfrac{N^2(N-n)}{Nn(n-1)} \right)^{-1/2} \left(\dfrac{N_{h|l}^2 (N_{h|l} - n_{h|l})}{N_{h|l} n_{h|l} (N_{h|l} - 1)} \right)^{1/2} (Y_k - \overline{Y}_{h|l}) \\ \qquad \text{if } k = r_{l+(h-1)q+1}, \ldots, r_{l+hq-1} \\ 0 \qquad \text{if } k = r_{l+(h-1)q} \quad \text{or} \quad k = r_{l+hq} \end{cases}$$

which can appear as a residual associated with the estimator $\hat{Y}_l$. The variable $v_k(l)$ inserted in the traditional expression of the fixed-size simple sampling design without replacement is exactly equal to the conditional variance $\hat{Y}_l$ given in (3). In effect,

$$N^2 \frac{N-n}{nN} \frac{1}{N-1} \sum_{k \in U} \left(v_k - \frac{\sum_{k \in U} v_k}{N} \right)^2 = V(\hat{Y}_l \mid E_l).$$

This variable, however, depends on the $\overline{Y}_{h|l}$ which are unknown. We can estimate $v_k(l)$ by

$$\hat{v}_j(l) = \begin{cases} \left(\dfrac{N^2(N-n)}{Nn(n-1)} \right)^{-1/2} \left(\dfrac{N_{h|l}^2 (N_{h|l} - n_{h|l})}{N_{h|l} n_{h|l} (n_{h|l} - 1)} \right)^{1/2} (y_j - \hat{\overline{y}}_{h|l}) \\ \qquad \text{if} \quad j = l + (h-1)\, q + 1, \ \ldots, \ l + hq - 1 \\ 0 \qquad \text{if} \quad j = l + (h-1)\, q \quad \text{or} \quad j = l + hq \end{cases}$$

If we insert $\hat{v}_k(l)$ in the estimator of the variance for the simple design without replacement, we obtain an unbiased estimator of the conditional variance, and therefore of the variance.

$$N^2 \frac{N-n}{nN} \frac{1}{n-1} \sum_{j=1}^{n} \left(\hat{v}_j - \frac{\sum_{j=1}^{n} \hat{v}_j}{n} \right)^2 = \hat{V}(\hat{Y}_l \mid E_l).$$

Deville (1999) shows that the variance of a sum of estimators can be determined by adding the residuals associated with these estimators, the residuals having been computed by linearization. To estimate the variance of $\hat{Y}_c$, we could therefore simply take the mean of the residuals $\hat{v}_k(l)$, which is written

$$\hat{v}_k = \frac{1}{q} \sum_{l=b}^{b+q-1} \hat{v}_k(l).$$

In this way, it would be possible to estimate the variance by

$$\hat{V}_2(\hat{Y}_c) = \frac{N^2(N-n)}{nN} \frac{1}{n-1} \sum_{k \in S} \left(\hat{v}_k - \frac{\sum_{k \in S} \hat{v}_k}{n} \right)^2.$$

These two variance estimators need to be tested by simulations.

11. Simulations for Variance Estimators

The simulations shown in Table (5) are based on populations of size $N = 100$, that are generated by means of normal independent random variables. For each case studied, we give the value of q and the coefficient of correlation between the variable of interest Y_k and the rank R_k of the auxiliary variable X_k. The border b is defined by taking the integer of $q/2 + 1$. Since our purpose is to validate the variance estimator, we use 3,000 samples of size $n = 20$ for each simulation and we compare the variance estimated by the simulations of the calibrated estimator $V_{si}(\hat{Y}_c)$ with the expectations under the simulations of the two variance estimators denoted $E_{si}(\hat{V}_\alpha(\hat{Y}_c))$, $\alpha = 1, 2$. The last two columns of the tables show the relative bias defined by

$$\mathrm{RB}_{si}\, \hat{V}_\alpha(\hat{Y}_c) = \frac{E_{si}\hat{V}_\alpha(\hat{Y}_c) - V_{si}(\hat{Y}_c)}{V_{si}(\hat{Y}_c)}, \ \alpha = 1, \ 2.$$

The simulations show that the two proposed estimators overestimate the variance. The overestimation appears to diminish as q increases. The estimator $\hat{V}_2(\hat{Y}_c)$ definitely has the smallest bias. We will therefore prefer to use $\hat{V}_2(\hat{Y}_c)$.

Table 5
Simulation Results

	Correlation:	0.802			
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$\mathrm{RB}_{si}\hat{V}_1(\hat{Y}_c)$	$\mathrm{RB}_{si}\hat{V}_2(\hat{Y}_c)$
4	0.045	0.065	0.054	0.444	0.200
5	0.045	0.066	0.057	0.467	0.267
6	0.056	0.076	0.070	0.357	0.250
7	0.058	0.079	0.059	0.362	0.017
8	0.063	0.088	0.087	0.397	0.381
	Correlation:	0.481			
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$\mathrm{RB}_{si}\hat{V}_1(\hat{Y}_c)$	$\mathrm{RB}_{si}\hat{V}_2(\hat{Y}_c)$
4	0.048	0.066	0.059	0.375	0.229
5	0.045	0.060	0.054	0.333	0.200
6	0.044	0.056	0.051	0.273	0.159
7	0.044	0.054	0.051	0.227	0.159
8	0.045	0.052	0.048	0.156	0.067
	Correlation:	0.111			
q	$V_{si}(\hat{Y}_c)$	$E_{si}\hat{V}_1(\hat{Y}_c)$	$E_{si}\hat{V}_2(\hat{Y}_c)$	$\mathrm{RB}_{si}\hat{V}_1(\hat{Y}_c)$	$\mathrm{RB}_{si}\hat{V}_2(\hat{Y}_c)$
4	0.281	0.471	0.363	0.676	0.292
5	0.297	0.420	0.356	0.414	0.199
6	0.279	0.363	0.316	0.301	0.133
7	0.267	0.342	0.324	0.281	0.213
8	0.282	0.327	0.281	0.160	– 0.004

12. Conclusions

Our proposed estimator is one of the rare estimators that is both unbiased and linear, that uses auxiliary information and that is calibrated on the size of the population. It can be parameterized using the width of window q. This new estimator is robust compared with the regression estimator. It can take into account auxiliary information with a non-linear relationship with the variable of interest. Most simulations appear to show that the width of the window does not significantly impact the preciseness of the mean estimator. However, it also appears that a small window width is not penalizing, even if there is no dependence between the auxiliary variable and the variable of interest. The smaller q is, the tighter the calibration, and the variance estimator will then be significantly penalized because the degree of freedom is lost in each post-stratum. The choice of q must therefore reflect this problem.

There are many other methods that allow for the use of the information given by a distribution function (see Ren 2000) to improve an estimator. The results that we have presented are limited to simple sampling designs, but we think they are important just as post-stratification is important as a specific case of calibration techniques. Post-stratification is one of the few examples where it is possible to show with accuracy that calibration corresponds to a conditional approach. Further, our approach can be seen as a calibration on a distribution function providing an unbiased estimator. A good general distribution calibration technique should therefore include in simple sampling designs the method we have presented.

Acknowledgements

We would like to thank Jean-Claude Deville and Anne-Catherine Favre, two referees and an associate editor for their constructive comments, which considerably improved this article.

References

Deville, J.-C. (1995). *Estimation de la variance du coefficient de Gini mesuré par sondage*. INSEE Méthode, working paper, Methodology F9510.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Estevao, V., Hidiroglou, M.A. and Särndal, C.-E. (1995). Methodological principle for a generalized estimation system in Statistics Canada. *Journal of Official Statistics,* 11, 181-204.

Ren, R. (2000). *Estimation par calage sur la répartition*. Thèse de Doctorat en préparation, Paris, Université Paris Dauphine.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.

Tillé, Y. (1999a). Sur la détermination a posteriori des bornes des post-strates. In *Les Sondages* (Eds. G. Brossier and A.-M. Dussaix). Dunod, 202-208.

Tillé, Y. (1999b). Estimation in surveys using conditional inclusion probabilities: Complex design. *Survey Methodology*, 25, 57-66.

Wu, C., and Sitter, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statistics*, 29, 289-307.