Variance estimators in survey sampling

Camelia Goga
Université de Bourgogne
Dijon, France.
email: camelia.goga@u-bourgogne.fr

December 16, 2008

Contents

Chapter 1

Introduction

Variance estimation in survey sampling is of major importance. It gives information on the accuracy of the estimators and allows to build confidence intervals. This report intends to make a review of the major techniques used to derive estimators of the variance of an estimated parameter of interest $\hat{t}$ in the framework of survey sampling. Särndal, Swensson & Wretman (1989) state that a variance estimator $\hat{V}(\hat{t})$ should accomplish at the same time all the following requests: it must have good properties with respect to the sampling design, a simple form and applicability in general. If a system of auxiliary information exists, then it would be advisable that the derived estimator would have good properties with respect to a regression model, including applicability to any linear regression model. At the same time, the variance estimator must have such a form that it can be implemented into a computer software. The principal concern is that such a variance estimator would be able to lead to valid confidence interval for the estimated parameter of study:

$$\hat{t} \pm z_{1-\frac{\alpha}{2}} (\hat{V}(\hat{t}))^{\frac{1}{2}}.$$

There are two main directions for deriving a valid quantity for the unknown variance $V(\hat{t})$:
(a) find an unbiased estimator for the variance when we can calculate it,
(b) find a consistent estimator for the approximative variance.

The choice between the two possibilities depends on the particular features of the survey sampling and on the quantity to be estimated.

There are situations when the minimal value of the variance is desired. Godambe (1955) shows that there exists no linear estimator for the population total with uniformly minimum variance but in more restrictive classes of estimators and certain designs or under a model of superpopulation, we can derive such estimators. An example is the optimum regression estimator obtained by Montanari (1987); unfortunately, it is more of theoretical interest since the obtained value of the optimal variance depends on all the values of the variable of study which are unknown and thus it can not be calculated explicitly in practice. Problems of this kind are discussed in more details in Cassel *et al.* 1977, ch 3.

Chapter 2 deals with the estimation of the fundamental quantities in survey sampling, the total and the mean of a finite population for simple plans with equal probabilities. The Horvitz-Thompson (H-T) estimator is introduced and a general unbiased variance estimator is derived. For the most usual survey designs, we give the precise expressions of the H-T estimator as well as

of the variance estimator. Except in a few cases, this general variance estimator has a complicate expression and it is hard to calculate. This is mainly due to the evaluation of a double sum and to the difficulty of calculating the probabilities of inclusion of second order.

Chapter 3 deals with stratified sampling design and chapter 4 with plans proportional to size. We treat in chapter 5 the multi-stage sampling design with application to two-stage and cluster designs.

In chapter 6, we propose to extend the derivation of a variance estimator when the parameter of interest has a complex form (e.g. non-linear statistics). This will be done with the help of the Taylor expansion. After a presentation of the theoretical results, the technique is used to derive the variance estimator for the ratio estimator, for the mean of the population and for the coefficient of multiple regression.

In the above chapters, no appeal to the auxiliary information was done. Or, it is well-known that the good use of it improves the results. Chapter 7 deals with ways of incorporating auxiliary information in estimating totals and means. We present the *calibration technique* proposed by Deville & Särndal (1992) and Deville (2000) and the *superpopulation approach*. This approach introduces a new structure for our population. Until now, chapter 2-6, we derived results within the context of the *fixed population approach* (Cassel *et al.* 1977) namely each population unit is associated with a fixed and unknown real number which is the value of the variable of interest. For the *superpopulation approach*, each population unit will be the outcome of a random variable for which a stochastic structure is specified.

Notations

Let us consider a finite population $\mathcal{U}$ composed of N elements

$$\mathcal{U} = \{u_1, \ldots, u_k, \ldots, u_N\} = \{1, \ldots, k, \ldots, N\}$$

where for simplicity, we identify the k-th element of $\mathcal{U}$ denoted by u_k with its label k. We will consider in the next that our population is such that each unit u_k can be spotted in a unique way by its label, k. This means that the units have the property of identifiability (Cassel *et al* 1977).

Let consider $\mathcal{Y}$, a variable of interest for which the value for the k-th unit, denoted by y_k, is unknown. We designate by $\mathbf{y} = (y_1, \ldots, y_N)$ the *parameter of the finite population* and any real function of it is called a *parametric function*. The goal of a survey sampling is to make inference about a parametric function such as the total or the mean for example, but more complicated functions may be of interest (the mode, various population quantiles, the population variance).

In the case of a survey sampling, the inference is based on information obtained only from a part of $\mathcal{U}$, called *sample*, obtained from $\mathcal{U}$ by a probabilistic selection scheme. More precisely, let $\mathcal{S}$ be the set of all possible subsets s of $\mathcal{U}$, $s \subset \mathcal{P}(\mathcal{U})$. There are 2^N possible subsets, considering the empty set and the whole population $\mathcal{U}$; a sample s is an element of $\mathcal{S}$. Given $\mathcal{U}$, let $p(s)$ be the probability of selecting $s \in \mathcal{S}$. In other words, the function $p(s)$ which is called *the sampling design* satisfies the following conditions:

1. $p(s) \geq 0$ for all $s \in \mathcal{S}$

2. $\sum_{s\in\mathcal{S}} p(s) = 1$.

In the present work, we will deal only with *noninformative designs*, namely with designs for which the function $p(\cdot)$ does not depend on the values of $\mathcal{Y}$. For this situation, Basu (1958) shows that it is sufficient to consider only the distinct elements of the sample.

We note by n_s the *sample size*, namely the number of elements of s; depending on the chosen scheme, n_s may be fixed or not for all the samples $s \in \mathcal{S}$.

We denote by $I_k = \mathbf{1}_{\{k\in s\}}$ for all $k \in \mathcal{U}$ the *sample membership indicator* (Särndal *et al* 1992). The random variable I_k is a Bernoulli variable indicating if the unit k belongs or not to the sample.

Supposing that a sampling design has been fixed, the probabilities of inclusion are defined as follows :

1. π_k, the first order inclusion probability, is the probability that the element k will be included in a sample . For all $k \in \mathcal{U}$, $\pi_k = \sum_{s\ni k} p(s)$.

2. π_{kl}, the second order inclusion probability, is the probability that the elements k and l will be included in a sample. For all $k, l \in \mathcal{U}$, $\pi_{kl} = \sum_{s\ni k\&l} p(s)$.

Result 1 *: For a given sampling design $p(\cdot)$, the functions I_k have the following properties:*

1. $E(I_k) = \pi_k$,

2. $V(I_k) = \pi_k(1 - \pi_k)$,

3. $Cov\ (I_k, I_l) = \pi_{kl} - \pi_k\pi_l,\ k \neq l$

for all $k, l \in \mathcal{U}$.

Proof : The proof relies on the fact that I_k is a Bernoulli variable of parameter π_k. It results immediately that $E(I_k) = \pi_k$ and $V(I_k) = \pi_k(1 - \pi_k)$. For the last point, we have,

$$\text{Cov } (I_k, I_l) = E(I_kI_l) - E(I_k)E(I_l) = \pi_{kl} - \pi_k\pi_l.$$

□

Consequence 1 *: If the design $p(s)$ has a fixed size, then*

1. $\sum_{\mathcal{U}} \pi_k = n$

2. $\sum_{k\neq l} \sum_{\mathcal{U}} \pi_{kl} = n(n-1)$

3. $\sum_{l\in\mathcal{U}, l\neq k} \pi_{kl}$.

Proof Since we deal with a fixed sample size, $E(n_s) = n$ and $V(n_s) = 0$ with $n_s = \sum_U I_k$. It results for (1) that $E(n_s) = \sum_U \pi_k = n$.
2)

$$\begin{aligned} V(n_s) &= \sum_U V(I_k) + \sum\sum_{k\neq l} \text{Cov } (I_k, I_l) \\ &= \sum_U \pi_k(1-\pi_k) + \sum\sum_{k\neq l}(\pi_{kl} - \pi_k\pi_l) \\ &= \sum_U \pi_k + \sum\sum_{k\neq l}\pi_{kl} - (\sum_U \pi_k)^2 \\ &= n(1-n) + \sum\sum_{k\neq l}\pi_{kl} = 0 \end{aligned}$$

which implies that $\sum\sum_{k\neq l}\pi_{kl} = n(n-1)$.
3) We have

$$\begin{aligned} \sum_{l\in\mathcal{U}, l\neq k}\pi_{kl} &= E(I_k \sum_{l\in\mathcal{U}, l\neq k} I_l) = E(I_k(n - I_k)) \\ &= nE(I_k) - E(I_k^2) = n\pi_k - \pi_k = (n-1)\pi_k \end{aligned}$$

□

For the simplicity of notation, we introduce the Δ-quantities (Särndal *et al* 1992):

$$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$$

$$\check{\Delta}_{kl} = \Delta_{kl}/\pi_{kl}$$

for all $k, l \in \mathcal{U}$.

We suppose from now on that $\pi_k > 0$ for all $k \in \mathcal{U}$, namely that each unit in the population has a chance to be in the sample.

1.1 The Horvitz-Thompson estimator

We consider the class of linear estimators and among these estimators we take the one proposed by Horvitz-Thompson (1952). This estimator is sometimes called the π *estimator* for the total of $\mathcal{Y}$ because the probabilities of first degree appear in its formula:

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}. \tag{1.1}$$

We give now the most important result of this section.

Result 2 *(Horvitz-Thompson 1952). The π estimator for the total of $\mathcal{Y}$, $\hat{t}_\pi$, has the following properties:*

1. *$\hat{t}_\pi$ is unbiased for $t = \sum_\mathcal{U} y_k$.*
2. *The variance of $\hat{t}_\pi$ has the expression :*

$$V(\hat{t}_\pi) = \sum_\mathcal{U}\sum_\mathcal{U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

3. *If $\pi_{kl} > 0$ for all $k\&l \in \mathcal{U}$, an unbiased estimator for $V(\hat{t}_\pi)$ is :*

$$\widehat{V}(\hat{t}_\pi) = \sum_s \sum_s \check{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Proof : The proof relies on the identities:

$$\begin{aligned} \hat{t}_\pi &= \sum_s \frac{y_k}{\pi_k} = \sum_{\mathcal{U}} \frac{y_k}{\pi_k} I_k \\ \widehat{V}(\hat{t}_\pi) &= \sum_{\mathcal{U}} \sum_{\mathcal{U}} \check{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} I_k I_l \end{aligned}$$

and on the fact that the membership indicators satisfy $E(I_k) = \pi_k$ and $\mathrm{Cov}(I_k, I_l) = \Delta_{kl}$ for all $k, l \in U$.

□

Remark 1 *: The H-T estimator is the only unbiased homogeneous linear estimator whose weights do not depend on the sample.*

Remark 2 *: The variance of $\hat{t}_\pi$ can be written as a quadratic form as follows:*

$$V(\hat{t}_\pi) = \boldsymbol{y} \Delta \boldsymbol{y}'$$

where $\Delta = (\frac{\Delta_{kl}}{\pi_k \pi_l})_{k,l \in \mathcal{U}}$ and $\boldsymbol{y} = (y_1, \ldots, y_N)$ the parameter vector.

For a sampling design of fixed size, $n_s = n$, equivalent formulas can be deduced for the variance and variance estimator of $\hat{t}_\pi$, as obtained by Yates and Grundy (1953) and Sen (1953).

Result 3 *(Yates-Grundy-Sen 1953). If $p(s) > 0$ is of fixed size, then $V(\hat{t}_\pi)$ and $\widehat{V}(\hat{t}_\pi)$ have the equivalent expressions:*

1. $V(\hat{t}_\pi) = -\frac{1}{2} \sum \sum_{\mathcal{U}} \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$.
2. *If $\pi_{kl} > 0$ for all $k, l \in \mathcal{U}$,*

$$\widehat{V}(\hat{t}_\pi) = -\frac{1}{2} \sum \sum_s \check{\Delta}_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2.$$

Proof : We use Consequence 1. More exactly,

$$V(\hat{t}_\pi) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} - \sum \sum_{\mathcal{U}} \Delta_{kl} \frac{y_k^2}{\pi_k^2} = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

since $\sum \sum_{\mathcal{U}} \Delta_{kl} \frac{y_k^2}{\pi_k^2} = \sum_{\mathcal{U}} \frac{y_k^2}{\pi_k^2} \sum_{\mathcal{U}} \Delta_{kl} = 0$ by Consequence 1. □

Remark 3 *The YGS variance estimator is not necessarily equal to the HT variance estimator.*

Definition 1 *The coefficient of variation of an estimator $\hat{\theta}$ is given by*

$$cve(\hat{\theta}) = \frac{\sqrt{\widehat{V}(\hat{\theta})}}{\hat{\theta}}.$$

Chapter 2

Equal sampling designs

In this section we describe the properties of estimators of finite population totals,

$$t_y = \sum_{\mathcal{U}} y_k.$$

under equal sampling designs. In chapter 2-6, we place in the context of the fixed population approach so the only randomness is the sampling design, $p(\cdot)$. As a result, the definitions of the expectation, variance and mean square error of an estimator Q for t_y can be formulated for a given design $p(s)$. For example,

$$E(Q) = \sum_{s \in \mathcal{S}} p(s)Q(s).$$

2.1 Sampling without replacement (SI)

We start with the simplest and most used sampling design, the simple random sampling without replacement (SI). We select with equal probability a first element from the population; this element will be kept away during the following selections. Next, we select with equal probability another element among the $N-1$ remaining units of the population and we continue the selection in the same way until the sample has n elements. There are C_N^n samples of size n and the sampling design has the following expression $p(s) = \dfrac{1}{\begin{pmatrix} N \\ n \end{pmatrix}}$ if the the sample s is of size n and zero otherwise. The inclusion probabilities of first and second degree are $\pi_k = \dfrac{n}{N}$ and $\pi_{kl} = \dfrac{n(n-1)}{N(N-1)}$. The Δ-quantities are $\Delta_{kl} = -\frac{f(1-f)}{N-1}$ for any $k \neq l$ and $f = n/N$.

There are several ways to implement a SI sampling. The one described above is called the *draw-sequential scheme*. For a large population of elements stored sequentially, the following *list-sequential scheme* is more recommended.

List-sequential schemes

1. We consider a uniform random variable $\varepsilon \sim Unif(0,1)$ and take independent realizations of it. The SI sample of size n is obtained as follows :

(a) If $\varepsilon_1 < n/N$ then $k = 1$ is selected; otherwise not.

(b) We repeat the operation for $k = 2, 3, \ldots$ Let n_k be the number of elements selected after $k-1$ steps. If $\varepsilon_k < \frac{n-n_k}{N-k+1}$ then k is selected, otherwise not.
The procedure stops when $n_k = n$.

This scheme proposed by Fan, Muller and Rezucha (1962) can be shown to conform to the definition of the SI design. Nevertheless, this scheme requires that the population size is known.

2. Yet another implementation of the SI design is by the following scheme which has the advantage that it permits a simultaneous selection of several nonoverlapping SI samples. The scheme proposes that N independent Unif(0,1) random numbers $\varepsilon_1, \ldots, \varepsilon_N$ are first drawn and then ordered according to size. Now, the n smallest ε-values correspond to the sample.

From **Result 2**, we have :

Result 4 *: Under a sampling **SI**, the π-estimator for the population total becomes*

1. $\hat{t}_{\pi,SI} = N\overline{y}_s = \dfrac{1}{f}\sum_s y_k$.

2. *The variance has the expression* $V_{SI}(\hat{t}_\pi) = N^2\dfrac{1-f}{n}S^2_{y\mathcal{U}}$.

3. *An unbiased estimator of the variance is given by*

$$\widehat{V}_{SI}(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{ys},$$

where $f = \dfrac{n}{N}$,

$$S^2_{y\mathcal{U}} = \frac{1}{N-1}\sum_{\mathcal{U}}(y_k - \overline{y}_{\mathcal{U}})^2$$

with $\overline{y}_{\mathcal{U}} = N^{-1}\sum_{\mathcal{U}} y_k$ *and*

$$S^2_{ys} = \frac{1}{n-1}\sum_s (y_k - \overline{y}_s)^2,$$

for $\overline{y}_s = n^{-1}\sum_s y_k$.

Proof 2). The SI sampling is of fixed size, so we can use the Yates-Grundy-Sen variance formula (result 3) with $\Delta_{kl} = -\dfrac{f(1-f)}{N-1}$ for any $k \neq l$. We obtain

$$\begin{aligned} V_{SI}(\hat{t}_\pi) &= \frac{1}{2}\frac{f(1-f)}{(N-1)f^2}\sum_U\sum_U (y_k - y_l)^2 \\ &= N^2\frac{1-f}{n}S^2_{yU} \end{aligned}$$

since $\sum\sum_U (y_k - y_l)^2 = N(N-1)S^2_{yU}$.

3). A similar derivation using the Yates-Grundy-Sen variance estimator gives $\widehat{V}_{SI}(\hat{t}_\pi) = N^2\dfrac{1-f}{n}S^2_{ys}$.

□

Remark 4 *The ratio $f = n/N$ is called the **sampling rate** and the difference $1 - f$ the **finite population correction**. We make this correction because with small populations the greater f, the more information we have about the population and thus the smaller the variance. For most samples from large populations, the finite population correction is approximately 1.*

Remark 5 *: For **SI** we have $(\Delta_{k,l})_{k,l\in U} = k(\boldsymbol{I}_N - \boldsymbol{P})$ with $k = f(1-f)\dfrac{N}{N-1}$; $\boldsymbol{I}_N$ is the identity matrix with dimension N and $\boldsymbol{P} = \dfrac{1}{N}\mathbf{1}_N\mathbf{1}'_N$. The variance has the expression $V_{SI} = \dfrac{N}{N-1}\dfrac{1-f}{f}\boldsymbol{y}(\boldsymbol{I}_N - \boldsymbol{P})\boldsymbol{y}'$. As a consequence, $det\ (\Delta_{k,l}) = 0$ and the variance will be minimized for all vector $\boldsymbol{y} = c\ \mathbf{1}'_N$ with c a real constant.*

Remark 6 *The mean $\overline{y}_{\mathcal{U}} = \sum_U y_k/N$ with N known is estimated unbiasedly by dividing by N the π-estimator of the population total from the above result,*

$$\widehat{\overline{y}}_{\mathcal{U}} = \sum_s \frac{y_k}{n} = \overline{y}_s.$$

In this case, the π-estimator coincides with the sample mean which is the most used estimator in classical inferential statistics. To obtain the variance and the variance estimator, we have only to divide by N^2 the corresponding expressions from result 4, namely

$$V_{SI}(\widehat{\overline{y}}_{\mathcal{U}}) = \frac{1-f}{n}S^2_{y\mathcal{U}}, \quad \widehat{V}_{SI}(\widehat{\overline{y}}_{\mathcal{U}}) = \frac{1-f}{n}S^2_{ys}.$$

Remark 7 *Result 4 gives us that the variance $\widehat{V}_{SI}(\hat{t}_\pi)$ is not large if*

1. *the sample size n is large. For large populations, it is n and not the percentage of the population sampled, that determines the precision of the estimator : a sample of size 100 from a population of size $N = 100.000$ has almost the same precision as a sample of size 100 from a 100 millions population units:*

$$V(N\overline{y}_s) = N^2\frac{99.900}{100.000}\frac{S^2}{100} = \frac{S^2}{100}0.999 \quad for \quad N = 100.000$$

$$V(N\overline{y}_s) = N^2\frac{99.999.900}{100.000.000}\frac{S^2}{100} = \frac{S^2}{100}0.999999 \quad for \quad N = 100.000.000$$

2. *the sampling rate $f = n/N$ is large;*

3. *the variance $S^2_{y\mathcal{U}}$ is little.*

2.1.1 Estimation of a proportion

We are interested in estimating the proportion P of individuals from the finite population with an attribute A using a SI sampling design. First of all, we write the proportion P as a finite population mean of the variable $\mathcal{Y}$ defined as follows

$$y_k = \begin{cases} 1 & \text{if k has A} \\ 0 & \text{elsewhere} \end{cases}$$

We have $P = \frac{\sum_U y_k}{N}$ estimated unbiasedly by

$$\hat{P}_{SI} = \frac{\sum_s y_k}{n}$$

with variance

$$V(\hat{P}_{SI}) = \frac{1-f}{n} S^2_{y\mathcal{U}}, \quad \text{where} \quad S^2_{y\mathcal{U}} = \frac{N}{N-1} P(1-P)$$

estimated unbiasedly by

$$\widehat{V}(\hat{P}_{SI}) = \frac{1-f}{n} S^2_{ys}, \quad \text{where} \quad S^2_{ys} = \frac{n}{n-1} \hat{P}_{SI}(1-\hat{P}_{SI}).$$

Let us prove that $S^2_{y\mathcal{U}} = \frac{N}{N-1} P(1-P)$. We have

$$\begin{aligned} S^2_{y\mathcal{U}} &= \frac{1}{N-1} \sum_{\mathcal{U}} (y_k - \overline{y}_{\mathcal{U}})^2 = \frac{1}{N-1} \sum_{\mathcal{U}} (y_k^2 - 2P y_k + P^2) \\ &= \frac{1}{N-1} (\sum_{\mathcal{U}} y_k - 2P \sum_{\mathcal{U}} y_k + NP^2) = \frac{N}{N-1} P(1-P). \end{aligned}$$

We prove in the same way the formula for S^2_{ys}. When $N/(N-1) \simeq 1$, $V(\hat{P}_{SI}) \simeq \frac{1-f}{n} P(1-P)$.

Confidence interval for the finite population total, mean and proportion

Hajek (1960) proves a central limit theorem for simple random sampling without replacement. Under certain technical conditions,we have

$$\frac{\overline{y}_s - \overline{y}_{\mathcal{U}}}{\sqrt{\frac{1-f}{n}} S_{y\mathcal{U}}} \sim \mathcal{N}(0,1).$$

A large-sample $100(1-\alpha)\%$ confidence interval for the population mean is

$$[\overline{y}_s - z_{\alpha/2} \sqrt{\frac{1-f}{n}} S_{y\mathcal{U}}, \overline{y}_s + z_{\alpha/2} \sqrt{\frac{1-f}{n}} S_{y\mathcal{U}}]$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution. Usually, $S_{y\mathcal{U}}$ is unknown and is replaced with the sample estimate S_{ys}. If $n/N \simeq 0$, the confidence interval is the same as the one used in inferential statistics.
One can remark that the length of the above confidence interval is a multiple of $n^{-1/2}$ so, for fixed variance, large sample sizes entails smaller confidence interval. We can see possible consequences of different sample sizes. Figure () shows the value of $1.96 \cdot s/\sqrt{n}$ for a range of sample sizes between 50 and 700 and for two possibles values of the standard deviation s, $s = 500000$ and $s = 700000$. The plot shows that if we ignore the finite population correction and $s = 500000$, a sample of size 300 will give a margin of error of about 60000.

Computation of minimal sample size for having a given precision

We consider the case of a proportion P. Under the normality assumption, the confidence interval for the proportion P is

$$IC_\alpha(P) = [\hat{P}_{SI} - z_{\alpha/2} \sqrt{V(\hat{P}_{SI})}, \hat{P}_{SI} + z_{\alpha/2} \sqrt{V(\hat{P}_{SI})}].$$

We want to find the minimal sample size n such that P should be estimated with a given precision e, namely

$$z_{\alpha/2}\sqrt{V(\hat{P}_{SI})} \quad \leq \quad e$$

and since $V(\hat{P}_{SI}) = \frac{1-f}{n} S^2_{y\mathcal{U}}$, we obtain

$$n \geq \frac{z^2_{\alpha/2} S^2_{y\mathcal{U}}}{e^2 + \frac{z^2_{\alpha/2} S^2_{y\mathcal{U}}}{N}}. \tag{2.1}$$

The problem with the above formula is that $S^2_{y\mathcal{U}} = \frac{N}{N-1} P(1-P)$ is not known. To solve it, we have two possibilities:

1. we have not any information about the proportion P. In this case, we take the worst situation $P = 1/2$ since $P(1-P)$ has a maximum at $P = 1/2$. We obtain

$$n \geq \frac{z^2_{\alpha/2} N}{z^2_{\alpha/2} + 4e^2(N-1)}.$$

   If N large, the sampling rate is approximately zero and $N/(N-1) \simeq 1$. Then,

$$n \geq \frac{z^2_{\alpha/2}}{4e^2}$$

Example (Lohr, 1999): Suppose we want to estimate the proportion of recipes in the Better Homes & Gardens New Cook Book that do not involve animal products. We plan to take an SI sample of the $N = 1251$ test kitchen-tested recipes and we want a 95% CI with margin of error 0.03.
If we ignore the finite population correction, we obtain

$$n \geq n_0 = \frac{1.96^2 \frac{1}{2}\frac{1}{2}}{0.03^2} \simeq 1068$$

which is extremely large comparing with N. On the contrary situation, we obtain

$$n \geq \frac{n_0}{1 + n_0/N} \simeq 577$$

2 we have an estimation $\hat{P}$. This estimation may have been obtained from a **pilot sample**, namely a small sample taken to provide information and guidance for the design of the main survey. Previous studies or surveys can also provide information about P.
   We replace in (2.1) $S^2_{y\mathcal{U}}$ with $S^2_{ys} = \frac{n}{n-1}\hat{P}(1-\hat{P})$ and we obtain

$$n \geq \frac{e^2 + z^2_{\alpha/2}\hat{P}(1-\hat{P})}{e^2 + \frac{z^2_{\alpha/2}\hat{P}(1-\hat{P})}{N}}.$$

   If N is large,

2.1.2 The design effect

The *design effect* is the ratio between the variance of the π-estimator $N\overline{y}_s$ under the SI sampling and the variance of the π-estimator under another sampling procedure $p(\cdot)$,

$$\text{deff}(p, \hat{t}_\pi) = \frac{V_p(\hat{t}_\pi)}{V_{SI}(N\overline{y}_s)}. \tag{2.2}$$

This ratio expresses how well the strategy $(p, \hat{t}_\pi)$ conducts in comparison with $(SI, N\overline{y}_s)$. When $\text{deff}(p, \hat{t}_\pi) > 1$, precision is lost by not using the SI design and in the contrary case, precision is gained compared to SI design.

2.1.3 SI with SAS

A SI sample can be obtained using PROC SURVEYSELECT with the option METHOD=SRS. One may use SAMPSIZE=n to specify the sample size SAMPRATE=n/N to specify the sampling fraction. The file of "sorties" gives information about the sampled individuals and the first-order inclusion probabilities. One example is given below.

```
/*  Plan de sondage simple   */

title1 'Logement Hte Gne (rec99) : plan simple';
proc surveyselect data=sondage.rec99htegne method=srs n=70 stats
   seed=47279 out=sondage.logsi1;
run;
```

One estimation of the finite population total can be obtained using PROC SURVEYMEANS with option SUM. If the finite population correction factor is desired, one can use TOTAL= or RATE= options.

```
title1 'Logement Hte Gne (rec99)';
title2 'Total estim log vacants Plan SI';
proc surveymeans data=sondage.logsi1 total=554 sum varsum cvsum clsum;
    var logvac;
weight Samplingweight;
ods output Statistics = sondage.rlog99;
 run;
```

SI with R

In order to get a simple random sampling without replacement, one may use the R function SAMPLE(N,n) which returns the vector of sampled individuals. The package Sampling contains the function SRSWR(n,N) which returns a 0,1 vector of size N.

```
 rec99<-read.csv("C:/Documents and Settings/Administrateur/Bureau/cours_sondages_Besancon/pr
library(sampling)
#slection de 70 individus parmi 554
si.rec99<-srswor(70,554)
t_ht<-554*mean(rec99$LOGVAC[which(si.rec99==1)])
```

2.2 Bernoulli sampling (BE)

In this case the sample is composed of all elements k from $\mathcal{U}$ which satisfy $\varepsilon_k < \pi$, where ε_k for all $k \in \mathcal{U}$ are independent realizations of a random variable with uniform distribution in the interval (0, 1) and $0 < \pi < 1$ is a fixed constant. As a result all the units from the population have the same probability of inclusion of first degree, $\pi_k = \pi$ for all $k \in \mathcal{U}$. Besides, the selection of the unit k is made independently from the selection of unit l with $l \neq k$; as a consequence, $\pi_{kl} = \pi^2$ for all $k \neq l \in \mathcal{U}$ and $\Delta_{kl} = 0$ for all $k \neq l \in \mathcal{U}$. Then, the variance-covariance matrix has the expression $\Delta = \text{diag}\pi(1-\pi)$. In BE sampling, the sample membership indicators I_k are independent.
We have 2^N possible samples and the probability of selecting one is

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s}$$

where n_s is the sample size. In BE sampling, the sample size is a binomially distributed random variable with parameters N and π, $n_s \sim \mathcal{B}(N, \pi)$.

Example :

Result 5 *: Under a **BE** sampling, the π-estimator for the population total t_y can be written*

1. $\hat{t}_{\pi,BE} = \dfrac{1}{\pi}\sum_s y_k$;

2. *The variance has the expression* $V_{BE}(\hat{t}_\pi) = \left(\dfrac{1}{\pi} - 1\right)\sum_{\mathcal{U}} y_k^2$.

3. *An unbiased estimator of the variance is given by*

$$\hat{V}_{BE}(\hat{t}_\pi) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_s y_k^2.$$

Proof 1) $\hat{t}_{\pi,BE} = \sum_s \dfrac{y_k}{\pi_k} = \dfrac{1}{\pi}\sum_s y_k$.
2) We have $\Delta_{kl} = 0$ for $k \neq l$ and $\pi(1-\pi)$ for $k = l$. So,

$$V_{BE}(\hat{t}_\pi) \quad = \quad \sum_U \sum_U \Delta_{kl}\check{y}_k\check{y}_l = \sum_U \pi(1-\pi)\frac{y_k^2}{\pi^2} = \left(\frac{1}{\pi} - 1\right)\sum_U y_k^2.$$

3) $\widehat{V}_{BE}(\hat{t}_\pi) = \sum_s \sum_s \check{\Delta}_{kl}\check{y}_k\check{y}_l = \sum_U \pi(1-\pi)\dfrac{y_k^2}{\pi^2} = \dfrac{1}{\pi}\left(\dfrac{1}{\pi} - 1\right)\sum_s y_k^2$.

□

The π estimator in the case of Bernoulli sampling is often inefficient because of the variable sample size. Nevertheless, the Bernoulli sampling conditioned to the sample size n_s is a **SI** sample. That's why, once selected the **BE** sample, we can consider the conditional frame.
Although designs with fixed size are desired, there are situations in which variable sample size conducts better or it can not be avoided. Two examples are relevant. The first one is the selection in a domain of a finite population, situation not treated here and the second one is the selection in the presence of nonresponse. In this case, the response behaviour in different population subgroups is often modeled as a **BE** sample selection.

Table 2.1: Population and samples

Sample	Elements of sample:				
s_1	$1,$	$a+1,$	$2a+1,$	$\cdots$	$(n-1)a+1$
$\cdots$		$\cdots$			
s_r	$r,$	$a+r,$	$2a+r,$	$\cdots$	$(n-1)a+r$
$\cdots$		$\cdots$			
s_a	$a,$	$2a,$	$3a,$	$\cdots$	$na=N$

2.2.1 BE sampling with SAS and R

2.3 Systematic sampling (SY)

The systematic sampling is very easy to implement and it is used when no list of population exists or when the list is in roughly random order.
To select a SY sample of size n, we determine a as being the ratio N/n supposed to be integer (the contrary situation will be treated in next section). The quantity a is called *the sampling interval*. Next, we choose at random and with equal probability a first element r between the first a elements in the population list; r is called *random start*. The sample will be composed of the ath element from the population list,

$$s_r = \{k : k = r + (j-1)a \leq N, j = 1, \ldots, n\}$$

For example (Lohr, 1999), to select a sample of 45 students from a list of 45 000 students at Arizona State University, the sampling interval a is 1000. Suppose the random integer we choose is 597. Then the students numbered 597, 1597, ..., 44 597 would be in the sample.
There are situations when it is more appropriate to choose the sampling interval and not the sample size. This happens when the population size is not known and the SY sampling is very useful in such situations. The resulting sample will be of random size. Let us consider an example.

Systematic sample is much compared with a SI sample. There are situations when it behaves like a SI sample and other when it does not. If the population is in random order, than SY sampling will be much like an SI sample.
Systematic sampling does not necessarily give a representative sample if the listing of population units is in some periodic or cyclical order. If male and female names alternate in the list and a is even, the SY sample will contain either all man or all women. In this case, the sample will not behave as a SI sample.
On the other hand, some populations are in increasing or decreasing order. A SY sample from a population of accounts ordered will contain some large amounts and some small amounts. With a SI sample, we may have only small or only large amounts which is not very desirable.

The set of all possible samples $\mathcal{S}_{SY} = \{s_1, \ldots, s_a\}$ consists of the a different and non-overlapping sets corresponding to the a possible random starts. This is represented in the table (2.1):

The probability of selecting $s \in \mathcal{S}_{SY}$ is

$$p(s) = 1/a \quad \text{if} \quad s \in \mathcal{S}_{SY} \quad \text{and zero elsewhere.}$$

The probability of first degree is

$$\pi_k = 1/a, k \in U$$

and of second degree,

$$\pi_{kl} = 1/a \quad \text{if} \quad k \neq l \in s \quad \text{and zero otherwise}$$

So, the condition that $\pi_{kl} > 0$ for all k, l it is not satisfied which means that we can not determine the Horvitz-Thompson variance estimate.
We have $U = \cup_{r=1}^{a} s_r$ and the finite population total of $\mathcal{Y}$ may be written as

$$t_y = \sum_U y_k = \sum_{r=1}^{a} t_{s_r}$$

with $t_{s_r} = \sum_{s_r} y_k$.

Result 6 *For a SY sampling, the π-estimator is*

1. $\hat{t}_\pi = at_s$ *for* $s \in \mathcal{S}_{SY}$.
2. *The variance of* $\hat{t}_\pi$ *is* $V_{SY}(\hat{t}_\pi) = a\sum_{r=1}^{a}(t_{s_r} - \bar{t})^2$ *with* $\bar{t} = \sum_{r=1}^{a} t_{s_r}/a$.

Proof 1) We have $\pi_k = 1/a$ and $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k} = at_s$.
2) The $\Delta_{kl} = \dfrac{1}{a} - \dfrac{1}{a^2}$ if $k \neq l \in s$ and $\Delta_{kl} = -\frac{1}{a^2}$ otherwise.

$$\begin{aligned} V_{SY}(\hat{t}_\pi) &= \sum_U \sum_U \frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l} y_k y_l = \sum_U \sum_U \frac{\pi_{kl}}{\pi_k\pi_l} y_k y_l - \left(\sum_U y_k\right)^2 \\ &= a\sum_{r=1}^{a} \sum_{s_r} \sum_{s_r} y_k y_l - t^2 \\ &= a\sum_{r=1}^{a} \left(\sum_{s_r} y_k\right)^2 - t^2 = a\sum_{r=1}^{a}(t_{s_r} - \bar{t})^2. \end{aligned}$$

□

2.3.1 Controlling the sample size

The fractional interval method

Let $a = N/n$ where n is the desired sample size.

1. Draw a random number ξ from the uniform distribution on the interval $(0, a)$.
2. The sample will consist of elements k for which

$$k - 1 < \xi + (j-1)a \leq k, \quad j = 1, 2, \ldots, n$$

Any element k has the probability $\pi_k = 1/a = n/N$ to be chosen and every possible sample is of size n.

Circular Systematic Sampling Method

The frame is laid out circularly : the last element is followed by the first and so on.

1. Draw a random number r between 1 and N with equal probability.
2. The sample will consist of elements k for which, for $j = 1, \ldots, n$ we have

$$k = r + (j-1)a \quad \text{if} \quad r + (j-1)a \leq N$$

or

$$k = r + (j-1)a - N \quad \text{if} \quad r + (j-1)a > N$$

Any element k has the probability $\pi_k = 1/a = n/N$ to be chosen and every possible sample is of size n.

2.3.2 The efficiency of systematic sampling

One can remark that the variance

$$V_{SY}(\hat{t}_\pi) = a \sum_{r=1}^{a} (t_{s_r} - \bar{t})^2$$

is zero if all the finite population totals are the same, $t_{s_r} = \bar{t}$. So, the efficiency of the SY sampling depends on the particular ordering of the N elements on which the systematic selection is applied. We study in the next the efficiency of the SY sampling as a function of the population ordering. We recall that we consider the case $N = a \cdot n$. Then, $\hat{t}_\pi = N \sum_{s_r} y_k/n = N\overline{y}_{s_r}$ with variance

$$V_{SY}(\hat{t}_\pi) = N^2 \frac{1}{a} \sum_{r=1}^{a} (\overline{y}_{s_r} - \overline{y}_U)^2$$

with $\overline{y}_{s_r} = \sum_{s_r} y_k/n$. Consider the ANOVA decomposition:

$$\underbrace{\sum_{U} (y_k - \overline{y}_U)^2}_{SST} = \underbrace{\sum_{r=1}^{a} \sum_{s_r} (y_k - \overline{y}_{s_r})^2}_{SSW} + \underbrace{\sum_{r=1}^{a} n(\overline{y}_{s_r} - \overline{y}_U)^2}_{SSB}$$

which means that the total variation (SST) is the sum of the variation within systematic samples (SSW) and the variation between systematic samples (SSB). The total variation is fixed and SSB determines the variance under SY sampling,

$$V_{SY}(\hat{t}_\pi) = N \cdot SSB.$$

In other words, the more homogeneous the elements within systematic samples are, the less efficient the SY sampling is. Homogeneous is used here to connote the tendency to have equal y-values. Thus, to achieve a favorable population ordering for SY sampling, we should strive for an ordering that entails a low degree of homogeneity among the elements within the same

systematic sample.
We discuss two measures of homogeneity.
1. We define *the intraclass correlation coefficient* ρ,

$$\rho = 1 - \frac{n}{n-1}\frac{SSW}{SST}$$

- $\rho > 0$ if the elements from the same sample tend to have similar y-values.
- $\rho = 1$ if $SSW = 0$ or complete homogeneity,
- $\rho = -1/(n-1)$ if $SSB = 0$ or complete heterogeneity.

2. The second homogeneity measure is defined as follows

$$\delta = 1 - \frac{N-1}{N-a}\frac{SSW}{SST}$$

The advantage with δ is that it can be used when the systematic samples are not of equal sizes. The extreme values of δ are

$$\delta_{min} = -\frac{a-1}{N-a}$$

for $SSB = 0$ and $\delta_{max} = 1$ which occurs if $SSW = 0$ or complete heterogeneity.

Comparison with SI sampling

We have

$$deff = \frac{V_{SY}(\hat{t}_\pi)}{V_{SI}(\hat{t}_\pi)} \simeq 1 + (n-1)\rho.$$

Then,
- if $\rho \simeq 1$ we have $deff \simeq n$ and SY is less efficient that SI;
- if $\rho = 0$ then $deff \simeq 1$ and SY =SI;
- if $\rho < 0$ then SY is more efficient than SI.

2.3.3 Variance estimation

There is no unbiased estimator of the variance $V_{SY}(\hat{t}_\pi)$. Several biased variance estimators have been proposed in the literature. We give below one.
Under the assumption that SY sampling is close to SI sampling, one can use as variance estimator

$$\widehat{V}(\hat{t}_\pi) = \frac{N^2(1-f)}{n} S^2_{ys_r}$$

with $S^2_{ys_r} = \frac{1}{n-1}\sum_{s_r}(y_k - \overline{y}_{s_r})^2$ if s_r is the selected sample.

2.3.4 Implementation of SY sampling in SAS and R

Chapter 3

Stratified sampling

In stratified sampling, the population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ is divided into H subpopulations denoted U_h of size N_h for $h = 1, \ldots, H$ such that $U_h \cap U_{h'} = \phi$ for $h \neq h'$.

$$\mathcal{U} = \{1, \ldots, k, \ldots, N\} = \bigcup_{h=1}^{H} U_h$$

$$|U| = N,\ |U_h| = N_h,\ N = \sum_{h=1}^{H} N_h$$

From each subpopulation $\mathcal{U}_h, h = 1, \ldots, H$ we select a sample s_h, $s_h \subset U_h$, of size n_h according to a sample design $p_h(\cdot)$. The selection in each subpopulation is made independently. The resulting sample s is :

$$s = s_1 \cup s_2 \cup \ldots \cup s_H$$

$$|s_h| = n_h \Rightarrow |s| = \sum_{h=1}^{H} n_h.$$

Then the probability of selecting s is:

$$p(s) = p_1(s_1) \ldots p_H(s_H).$$

Let π_k^h for $k \in \mathcal{U}$ and π_{kl}^h for $k \neq l \in \mathcal{U}$ be the first and second order inclusion probabilities according to p_h, $h = 1, \ldots, H$. Then, π_k, the first order inclusion probability according to $p(\cdot)$, is equal to π_k^h if $k \in s_h$ and π_{kl}, the second order inclusion probability according to $p(\cdot)$, is equal to $\pi_k^h \pi_l^{h'}$ if k, l belong to different strata h, h' and equal to π_{kl}^h if k, l belong to the same stratum h.

We suppose that N_h is known for all $h = 1, \ldots, H$. The population total can be written as:

$$t_y = \sum_{\mathcal{U}} y_k = \sum_{h=1}^{H} t_h = \sum_{h=1}^{H} N_h \overline{y}_{U_h}$$

where $t_h = \sum_{\mathcal{U}_h} y_k$ is the total of the stratum h and $\overline{y}_{U_h} = \frac{1}{N_h} t_h$ is the mean of the stratum h.

We can formulate the following result for the π-estimator in the case of stratified sampling:

Result 7 *: Under a stratified sampling design, the π-estimator for the total of the population is:*

1. $\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi}$ *where* $\hat{t}_{h\pi}$ *is the* π*-estimator for the stratum h.*
2. *The variance has the expression* $V(\hat{t}_\pi) = \sum_{h=1}^{H} V_h(\hat{t}_{h\pi})$*, where* $V_h(\hat{t}_{h\pi})$ *is the variance of* $\hat{t}_{h\pi}$ *for all h.*
3. *An unbiased variance estimator is given by* $\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}_h(\hat{t}_{h\pi})$*, where* $\hat{V}_h(\hat{t}_{h\pi})$ *is the variance estimator for* V_h *for all h.*

The selection of the sample s_h can be done differently or in the same way in all strata. For example, we can choose all s_h by sampling without replacement or by Bernoulli sampling. In each case, formulas for variance and for an estimation for variance can be obtained using the results derived for the designs **SI**, **BE**.

Stratified sampling with SI in each strata (STSI)

For stratified sampling with sampling without replacement in each stratum, we obtain:

1. $\hat{t}_\pi = \sum_{h=1}^{H} N_h \left(\sum_{s_h} \frac{y_h}{n_h} \right)$.
2. The variance of the π-estimator is

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{yU_h}^2 \tag{3.1}$$

   where $S_{yU_h}^2 = \frac{1}{N_h - 1} \sum_{\mathcal{U}_h} (y_k - \overline{y}_{U_h})^2$ is the stratum variance and $f_h = \frac{n_h}{N_h}$ the sampling fraction in stratum h.
3. $\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{ys_h}^2$ where $S_{ys_h}^2 = \frac{1}{n_h - 1} \sum_{s_h} (y_k - \overline{y}_{s_h})^2$ is the sample variance in stratum h.

Choice of strata

3.0.5 Optimal sample allocation under STSI sampling

Consider a population with fixed strata and we want to determine the sample sizes n_h, $h = 1, \ldots, H$ which minimize the variance $V(\hat{t}_\pi)$ for a fixed cost

$$C = c_0 + \sum_{h=1}^{H} n_h c_h \tag{3.2}$$

where c_0 is a fixed overhead cost and c_h is the cost of surveying one element in stratum h.

Result 8 *Consider the STSI sampling. Minimizing the variance* $V(\hat{t}_\pi)$ *for a fixed cost C gives the solution*

$$n_h = \frac{(C - c_0) N_h S_{yU_h} / (c_h)^{1/2}}{\sum_{h=1}^{H} N_h S_{yU_h} (c_h)^{1/2}}, \quad h = 1, \ldots, H$$

and the minimum variance is

$$V_{opt} = \frac{1}{C - c_0} \left[\sum_{h=1}^{H} N_h S_{yU_h} c_h^{1/2} \right]^2 - \sum_{h=1}^{H} N_h S_{yU_h}^2$$

Proof We have

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{yU_h}^2 = \sum_{h=1}^{H} \frac{N_h^2 S_{yU_h}^2}{n_h} - \sum_{h=1}^{H} N_h S_{yU_h}^2 = \sum_{h=1}^{H} \frac{A_h}{n_h} + B$$

with $A_h = N_h^2 S_{yU_h}^2$ and $B = -\sum_{h=1}^{H} N_h S_{yU_h}^2$. Then, minimizing the variance under a cost constraint is equivalent to minimizing the product

$$\left(\sum_{h=1}^{H} \frac{A_h}{n_h} \right) (C - c_0).$$

Using the Cauchy inequality, we have that the above product is always superior to $\left(\sum_{h=1}^{H} (A_h c_h)^{1/2} \right)^2$ with equality for

$$n_h \propto (A_h / c_h)^{1/2}.$$

Using relation (3.2), we obtain that

$$n_h = (C - c_0) \frac{\sqrt{\frac{A_h}{c_h}}}{\sum_{h=1}^{H} (A_h c_h)^{1/2}}.$$

Next, we replace A_h by $N_h^2 S_{yU_h}^2$ and we obtain the desired relation. □

We consider in the next that all c_h are equal and let n be the total sample size, so that

$$n = \sum_{h=1}^{H} n_h$$

Optimum allocation

The optimum allocation is obtained from result 8 with c_h constant. It results

$$n_h = n \frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}} \tag{3.3}$$

which gives the optimal variance

$$V_{STSI,o}(\hat{t}_\pi) = \frac{N^2}{n} \left(\sum_{h=1}^{H} W_h S_{yU_h} \right)^2 - N \sum_{h=1}^{H} W_h S_{yU_h}^2. \tag{3.4}$$

This result is also called the *Neyman allocation*. For deriving the above n_h, one needs the standard deviations S_{yU_h} which are unknown. As a consequence, this result can not be used in practice. But in repeated surveys, one may be able to use past experience to state close approximations to the true S_{yU_h}.
We can see that n_h is proportional to S_{yU_h} which means that the sample size will be larger if the variation of $\mathcal{Y}$ is larger in the stratum h.

x-optimal allocation

Suppose that $\mathcal{X}$ is an auxiliary information available highly correlated with $\mathcal{Y}$ and that we know S_{xU_h} for every $h = 1, \ldots, H$. The x-optimal allocation is given by

$$n_h = n \frac{N_h S_{xU_h}}{\sum_{h=1}^{H} N_h S_{xU_h}} \tag{3.5}$$

If the relation between $\mathcal{X}$ and $\mathcal{Y}$ is perfectly linear, than the x-optimal allocation is in fact optimal.

Proportional allocation

The proportional allocation is defined by

$$n_h = n \frac{N_h}{N} \tag{3.6}$$

Remark that if S_{yU_h} are all equal, then the proportional allocation is optimal.
The variance of the π-estimator and the proportional allocation is equal to

$$V_{STSI,p}(\hat{t}_\pi) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{h=1}^{H} W_h S^2_{y,U_h}. \tag{3.7}$$

This formula can be obtained by using the variance formula (3.1), relation (3.6) and the fact that $W_h = N_h/N$.

Allocation proportional to the y-total

Allocation proportional to the y-total is defined as follows

$$n_h = n \frac{\sum_{U_h} y_k}{\sum_U y_k} \tag{3.8}$$

which can not be used in practice since the stratum y-totals are unknown.

Allocation proportional to the x-total

Allocation proportional to the x-total is defined by

$$n_h = n \frac{\sum_{U_h} x_k}{\sum_U x_k} \tag{3.9}$$

3.0.6 Comparison between SI and STSI

We use the ANOVA decomposition :

$$(N-1) S^2_{yU} = \sum_{h=1}^{H} N_h (\overline{y}_{U_h} - \overline{y}_U)^2 + \sum_{h=1}^{H} (N_h - 1) S^2_{yU_h}$$

or

$$SST = SSB + SSW$$

3.0.7 Implementation of stratified sampling with R and SAS

Chapter 4

Unequal probabilities sampling designs

Up to now, we have only discussed equal sampling designs, namely the probabilities of selecting the sampling units are equal. Equal probabilities give schemes that are often easy to design, explain and implement. Such designs are not, however, always possible or, if practicable, as efficient as schemes using unequal probabilities. An intermediate case is the stratified sampling where the units from *different strata* have *different* probabilities to be chosen while units from the *same strata* have *equal* selection probabilities. We have seen at the time that the stratified sampling design reduced the variance of the Horvitz-Thompson estimator of the population total if the strata were very homogeneous with respect to the interest variable or with respect to an auxiliary variable roughly correlated to the study variable. An optimal allocation scheme would sample a very high fraction (perhaps 100%) in strata with high variability and small fraction in strata with little variability. Stratified sampling design is a particular case of sampling with unequal probabilities of selection.

4.1 Poisson sampling (PO)

The **BE** sampling is not a fixed size design. Another example is the Poisson sampling (**PO**) when the selection of an element is decided by $\varepsilon_k < \pi_k$ where $\{\pi_1, \ldots, \pi_n\}$ is a set of fixed constants between 0 and 1. We give the first and second order inclusion probabilities. Based on the same arguments as in the case of **BE** sampling, we have that π_k for all $k \in \mathcal{U}$ are the set of first order inclusion probabilities; as for the second degree inclusion probabilities, we have for $k \neq l \in \mathcal{U}$, $\pi_{kl} = \pi_k \pi_l$. Because of these particular expressions of the inclusion probabilities, the **Result** 2 will become:

Result 9 *: Under a **PO** sampling, the π-estimator for the population total has the following expression:*

1. $\hat{t}_{\pi,PO} = \sum_s \frac{y_k}{\pi_k}$.
2. *The variance is given by* $V_{PO}(\hat{t}_\pi) = \sum_{\mathcal{U}} \pi_k(1-\pi_k)\breve{y}_k^2 = \sum_{\mathcal{U}} \frac{(1-\pi_k)}{\pi_k} y_k^2$.
3. *An unbiased variance estimator is* $\hat{V}_{PO}(\hat{t}_\pi) = \sum_s (1-\pi_k)\breve{y}_k^2$.

As in the case of **BE** sampling, the **PO** sampling is of variable sample size n_s, fact which entails a great value of $V_{PO}(\hat{t}_\pi)$. By conditioning, we may remove this drawback. Hajek (1981) proves that every conditional **PO** sampling maximizes the entropy in the class of designs having the same first degree inclusion probability and taking part to the same class Ω of samples satisfying $\sum_{s\in\Omega} p(s) = 1$. The entropy is a measure of spread of the sample design defined by

$$e = -\sum_s p(s) \ln p(s).$$

The main role of **PO** sampling is to help to define and analyse other sampling method (Hajek 1981). By conditioning, we can obtain **SI** sampling, two-stage sampling described in sections 2.2 and 2.9, etc.

In the case of a **BE** sampling we have:

$$E_{BE}(n_s) = N\pi \text{ and } \pi_k = \pi \quad \text{for all } k = 1, \ldots, N.$$

Thus, if we fix the expected sample size and we suppose that N is known, then π_k for $k = 1, \ldots, N$ are completely specified. In **PO** sampling, we do not have the same thing:

$$E_{PO}(n_s) = \sum_{k=1}^{N} \pi_k.$$

Thus, for fixed $E_{PO}(n_s)$ we have to make a choice for π_k. We will choose the π_k that minimize the variance. We get the following expression for the first order inclusion probabilities:

$$\pi_k = \frac{ny_k}{\sum_{\mathcal{U}} y_k} \quad \text{for all } k = 1, \ldots, N$$

assuming that $y_k < \dfrac{\sum_{\mathcal{U}} y_k}{n}$ for all $k = 1, \ldots, N$. Because the expression of π_k requires y_k for all $k = 1, \ldots, N$ which in general are unknown, the inclusion probabilities so obtained can not be used. But, if we have auxiliary information $\mathcal{X}$ which takes the value x_k for the k-th element of the population and the variable of interest $\mathcal{Y}$ is approximately proportional to $\mathcal{X}$, then we can consider:

$$\pi_k = \frac{nx_k}{\sum_{\mathcal{U}} y_k} \quad \text{for all } k = 1, \ldots, N$$

and

$$x_k < \frac{\sum_{\mathcal{U}} x_k}{n} \quad \text{for all } k = 1, \ldots, N.$$

The inclusion probabilities π_k given by these expressions are called *probability proportional to size*. If $\mathcal{Y}$ is proportional to $\mathcal{X}$, then the associated π-estimator $\hat{t}_{\pi,PO}$ has a small variance.
This discussion states a more general problem: how we can use available auxiliary information at the sampling stage? One possibility is selecting units with unequal probabilities, such as *probability proportional to size* described above. We will describe briefly in the next subsection the most important methods of units selection with unequal probabilities. Till (2006) gives a detailed description of these methods as well as the associated algorithms.

4.2 πps estimator

We need auxiliary information $\mathcal{X}$ for selecting unequal probabilities sampling designs. The derivation of first order inclusion probabilities of a fixed size design is the first step of all the methods. Conditions on the second order inclusion probabilities are formulated, fact which makes possible the resolution of this problem. Hanif & Brewer (1980) and Brewer & Hanif (1983) present all the existing methods for selecting unequal probabilities sampling designs until that moment.
We can have designs with or without replacement. We start with the study of the πps sampling design when $\pi_k \propto x_k$ for all k in the population, called πps sampling. In this case, several schemes for selecting elements such that $\pi_k \propto x_k$ have been proposed. All these schemes have been devised in order to accomplish the following five requests (Särndal *et al.* 1992):

1. The selection of the sample is simple;
2. the π_k are strictly proportional to x_k for all $k = 1, \ldots, N$;
3. $\pi_{kl} > 0$ for all $k \neq l$;
4. the π_{kl} can be computed simply;
5. $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l < 0$ for all $k \neq l$ to guarantee that the Yates-Grundy variance estimator is nonnegative.

Let n be the sample size. We will study different schemes depending on the values of n.

For $n = 1$, we have the total cumulative method which is strictly a πps design, but $\pi_{kl} = 0$ for all $k \neq l$ so an unbiased variance estimator can not be obtained.

For $n = 2$, we mention only the design proposed by Brewer (1963, 1975) which ensures $\pi_k = \frac{2x_k}{\sum_{\mathcal{U}} x_k}$ for $k = 1, \ldots, N$ and $\pi_{kl} > 0$ for all $k \neq l$. This scheme satisfies also the condition that $\Delta_{kl} < 0$ for all $k \neq l$ that allows the Yates-Grundy variance estimator to be always nonnegative. So, all the above requirements are satisfied.

For $n > 2$, several schemes exit, most of them are complicated because of the requirement that π_k must be proportional to x_k for all $k \in \mathcal{U}$. If we relax this condition, Sunter (1977) proposes a schema which gives π_k strictly proportional to x_k except for a small portion of the population, corresponding to the smallest values of x_k. We have also $\pi_{kl} > 0$ and $\Delta_{kl} < 0$ for $k \neq l$. Sunter (1986) presents a list-sequential scheme which achieves a strictly proportionality between π_k and x_k. Madow (1949) proposes a systematic unequal probabilities sampling design, which is one of the best because of its simplicity.

The method of Rao, Hartley and Cochran (1962) gives an unbiased estimator for the population mean and at the same time a variance estimator.

4.2.1 Pwr sampling

For sampling with replacement, denoted by **SIR**, Hansen and Hurwitz (1943) proposed the *pwr* estimator: p-expanded with replacement, corresponding to a generalization of simple random sampling with replacement. The method consists in drawing with replacement m different elements with unequal probabilities $p_1, \ldots, p_k, \ldots, p_N$ retaining the independence of the draws.

$$Pr(\text{selecting } k) = p_k \text{ for } k = 1, \ldots, N.$$

Then the sets of $\{p_k\}_{k\in\mathcal{U}}$ satisfy the properties :

1. $p_k > 0$ for all $k \in \mathcal{U}$
2. $\sum_{\mathcal{U}} p_k = 1$

The proposed *pwr* estimator for the total t_y is:

$$\hat{t}_{pwr} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}}$$

We have the following result for the *pwr* estimator:

Result 10 *:*

1. *The variance of* $\hat{t}_{pwr}$ *is :*

$$V(\hat{t}_{pwr}) = \frac{V_1}{m} \text{ where } V_1 = \sum_{\mathcal{U}}\left(\frac{y_k}{p_k} - t\right)^2 p_k;$$

2. *An unbiased estimator for* $V(\hat{t}_{pwr})$ *is:*

$$\widehat{V}(\hat{t}_{pwr}) = \frac{\hat{V}_1}{m}; \quad \hat{V}_1 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{pwr}\right)^2.$$

Proof

□

One can see from above that the variance of the *pwr*-estimator is zero if the study variable is proportional to *p*-values, namely

$$y_k/p_k = c, \quad \text{for all} \quad k = 1, \ldots, N$$

where c is a constant. Unfortunately, this proportionality can not be attained. A solution is to take

$$p_k \propto x_k$$

where x_k is an auxiliary variable value roughly proportional to y_k. We obtain then

$$p_k = \frac{x_k}{\sum_U x_k}, \quad \text{for} \quad kk = 1, \ldots, N.$$

If we want to select only one element, the cumulative method can be used, otherwise we repeat the cumulative method m times independently. We start with $p_k = \frac{x_k}{\sum_{\mathcal{U}} x_k}, k \in U$ and let v_k be defined as follows:

$$v_0 = 0 \text{ and } v_k = \sum_{l=1}^{k} p_l \quad \text{for } k = 1, \ldots, N.$$

Then we generate u an Unif(0, 1) variable and the selection or not selection of the element k will be decided as follows: the element k is selected if $v_{k-1} \leq u < v_k$. We repeat this operation m times until the sample s is obtained.

$$P(\text{ element k is selected }) = P(v_{k-1} \leq u < v_k) = v_k - v_{k-1} = p_k$$

and $\hat{t}_{yHH} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}}$ with the same variance and estimate variance as $\hat{t}_{pwr}$ deduced above. As it can be noticed, the expressions for $V(\hat{t}_{pwr})$ and $\hat{V}(\hat{t}_{pwr})$ are simpler than the corresponding ones if we had used the π-estimator, when the cross-products $\check{\Delta}_{kl}\check{y}_k\check{y}_l$ must be computed. However, $\hat{t}_{pwr}$ is less efficient than the π estimator. The following strategy gives an estimator for the variance that combines the two estimators:

1. a fixed size m πps sampling design is used such that $\pi_k = mp_k = \frac{mx_k}{\sum_{\mathcal{U}} x_k}$;
2. the π-estimator is used to estimate the population total t_y;
3. the variance of the π-estimator is estimated by the *pps*-sampling formula:

$$\hat{V} = \frac{1}{m(m-1)}\sum_{s}\left(\frac{y_k}{p_k} - \frac{1}{m}\sum_{s}\frac{y_k}{p_k}\right)^2$$

The variance estimator $\hat{V}$ will not be unbiased for $V(\hat{t}_{\pi})$.

The R and SAS

Chapter 5

Multi-stage sampling

All of the designs presented before suppose that direct element sampling is possible, namely there is a sampling frame describing rigorously the target population and one can use it to select the sample. But in practice, it is very often that we do not possess such a sampling frame or it could be too expensive to have one. Secondly, the population could be spread over a wide area entailing very high travel expenses for the personal interviewers. How can we select a sample in such situations? A variety of sampling designs are available for surveys in which direct element sampling is impossible or impractical. One possibility is to have a *multistage sampling* which consists in several stages of sampling (three, four stages) and the last-stage sampling is one of direct elements. Let consider an example.

Example (Särndal *et al.*, 1992) : The Swedish Board of Education sponsors annual surveys in Sweden to measure drug among ninth-grade students. In this survey, data on drug use is collected through anonymous questionnaires from every student in a sample of ninth-grade classes. The sampling frame consists of a list of all ninth-grade classes. This is an example of *cluster sampling* or one-stage sampling.
Alternatively, we could select a sample of schools and for each selected school, a sample of ninth-grade classes is drawn next. Finally, a sample of students is selected in each sampled ninth-grade classes. We have *three-stage sampling.*

We are interested in this chapter with the estimation of the finite population total t_y knowing that the finite population size N is unknown. As a consequence, the mean $\overline{y}$ is a non-linear parameter being the ratio between two unknown totals.

1. The population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ is now partitioned into N_I primary sampling units called PSUs, $\mathcal{U}_1, \ldots, \mathcal{U}_i, \ldots, \mathcal{U}_{N_I}$ of size $|\mathcal{U}_i| = N_i$ for $i = 1, \ldots, N_I$ with N_i often unknown. For simplicity, we note:

   $$\mathcal{U}_I = \{1, \ldots, i, \ldots, N_I\}.$$

   In the first stage a sample $s_I, s_I \subset \mathcal{U}_I$ of PSUs is drawn according to a sampling design $p_I(\cdot)$.

2. For each $i \in \mathcal{U}_I, \mathcal{U}_i$ is partitioned into N_{IIi} secondary sampling units, SSUs, $\mathcal{U}_{i1}, \ldots, \mathcal{U}_{iq}, \ldots, \mathcal{U}_{iN_{IIi}}$

symbolically represented by

$$\mathcal{U}_{IIi} = \{1, \ldots, q, \ldots, N_{IIi}\}.$$

In the second stage, for each $i \in s_I$, a sample s_{IIi} is selected from $\mathcal{U}_{IIi}$ according to a sampling design $p_{IIi}(\cdot)$.

3. We repeat the two previous steps until the r-th stage when the r-th sampling units are the population elements.

The general procedure will be referred to the $r-1$ subsequent stages. We suppose that we have invariance and independence with respect to the $r-1$ stages of selection.
In multi-stage sampling we have the inclusion probabilities according to the first stage:

$$\pi_{Ii}, \pi_{Iij}, \text{ for } i \neq j \text{ and } i, j \in s_I$$

$$\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}, \quad \check{\Delta}_{Iij} = \frac{\Delta_{Iij}}{\pi_{Iij}}.$$

Let t_y be the population total:

$$t_y = \sum_{\mathcal{U}} y_k = \sum_{\mathcal{U}_\mathcal{I}} t_i;$$

where $t_i = \sum_{\mathcal{U}_i} y_k$ is the total of $\mathcal{U}_i$. We assume that we can build the π-estimator $\hat{t}_{i\pi}$ for t_i with respect to the last $r-1$ stages of selection, $E(\hat{t}_{i\pi}|s_I) = t_i$, and let $V_i = V(\hat{t}_{i\pi}|s_I)$ be the variance of $\hat{t}_{i\pi}$; let $\hat{V}_i$ be an unbiased estimator of V_i, given s_I, namely $E(\hat{V}_i|s_I) = V_i$. With these notations, we can give the result for the π-estimator for the population total and also the expressions for the variance and variance estimator in multi-stage sampling.

Result 11 *In r-stage sampling, $r \geq 2$, we have:*

1. *the estimator $\hat{t}_\pi = \sum_{s_I} \frac{\hat{t}_{i\pi}}{\pi_{Ii}}$ is unbiased for t_y;*

2. *the variance of $\hat{t}_\pi$ is $V(\hat{t}_\pi) = \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j + \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}}$;*

3. *an unbiased variance estimator is $\hat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} \frac{\hat{t}_{j\pi}}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}$.*

Proof : 1.

$$E(\hat{t}_\pi) = E_I E(\hat{t}_\pi|s_I) = E_I \left(\sum_{s_I} \frac{1}{\pi_{Ii}} E(\hat{t}_{i\pi}|s_I) \right) = E_I \left(\sum_{s_I} \frac{t_i}{\pi_{Ii}} \right) = \sum_{\mathcal{U}_\mathcal{I}} t_i = t.$$

2.

$$\begin{aligned} V(\hat{t}_\pi) &= V_I\{E(\hat{t}_\pi|s_I)\} + E_I\{V(\hat{t}_\pi|s_I)\} = V_I \left\{ \sum_{s_I} \frac{t_i}{\pi_{Ii}} \right\} + E_I \left\{ \sum_{s_I} \frac{V_i}{\pi_{Ii}^2} \right\} \\ &= \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j + \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}}. \end{aligned}$$

3.

$$\begin{aligned} E(\hat{V}(\hat{t}_\pi)) &= E_I E(\hat{V}(\hat{t}_\pi)|s_I) = E_I \left\{ \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \check{t}_i \check{t}_j + \sum_{s_I} (1-\pi_{Ii}) \frac{V_i}{\pi_{Ii}^2} + \sum_{s_I} \frac{V_i}{\pi_{Ii}} \right\} \\ &= \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j + \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}} = V(\hat{t}). \end{aligned}$$

□

For $r = 2$, we obtain a two-stage sampling and for $r = 3$ a three-stage. From the general result, we can derive formulas for the variance and variance estimator in the particular case of two-stage sampling.

5.1 Two-stage sampling

5.1.1 Description

In this case, the second stage will consist in selecting for every $i \in s_I$, a sample s_i of elements from U_i, $s_i \subset U_i$ according to a design $p_i(.|s_I)$. Let $\pi_{k|i}$, $\pi_{k,l|i}$ be the first and second order inclusion probabilities with respect to the second stage sampling, $p_i(\cdot|s_I)$.
Using the invariance and independence properties, one can deduce the final inclusion probabilities π_k and π_{kl},

$$\pi_k = \pi_{Ii}\pi_{k|i} \quad \text{if} \quad k \in \mathcal{U}_i$$

$$\pi_{kl} = \begin{cases} \pi_{Ii}\pi_{k|i} & \text{if} \quad k = l \in \mathcal{U}_i \\ \pi_{Ii}\pi_{kl|i} & \text{if} \quad k \neq l \in \mathcal{U}_i \\ \pi_{Iij}\pi_{k|i}\pi_{l|j} & \text{if} \quad k \in \mathcal{U}_i \quad \text{and} \quad l \in \mathcal{U}_j, i \neq j \end{cases}$$

The resulting sample is

$$s = \bigcup_{i \in s_I} s_i.$$

We suppose that we have invariance and independence with respect to the second stage of selection. For every $i \in s_I$ let n_i be the size of s_i, then the size of s is:

$$n_s = \sum_{i \in s_I} n_i.$$

We construct the unbiased estimator $\hat{t}_{i\pi}$ of t_i with respect to the second stage; it results then the expression for $\hat{t}_{i\pi}$:

$$\hat{t}_{i\pi} = \sum_{s_i} \frac{y_k}{\pi_{k|i}}$$

From the general result for multi-stage sampling, we have the π-estimator for the total of the population, $t_y = \sum_\mathcal{U} y_k$ of the form:

$$\hat{t}_\pi = \sum_{i \in s_I} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} = \sum_{i \in s_I} \frac{1}{\pi_{Ii}} \left(\sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} \right)$$

with the variance :

$$V(\hat{t}_\pi) = V_{PSU} + V_{SSU}$$

where $V_{PSU} = \sum_{\mathcal{U}_\mathcal{I}} \sum_{\mathcal{U}_\mathcal{I}} \Delta_{Iij} \check{t}_i \check{t}_j$ is the variance due to the first stage sampling and $V_{SSU} = \sum_{\mathcal{U}_\mathcal{I}} \frac{V_i}{\pi_{Ii}}$ is the variance due to the second stage sampling. We have $V_i = V(\hat{t}_{i\pi}|s_I) = \sum_{\mathcal{U}_i} \sum_{\mathcal{U}_i} \Delta_{kl|i} \check{y}_{k|i} \check{y}_{l|i}$.
An unbiased estimator for V_i is :

$$\hat{V}_i = \sum_{s_I} \sum_{s_I} \check{\Delta}_{kl|i} \check{y}_{k|i} \check{y}_{l|i}.$$

Then, an unbiased estimator for $V(\hat{t}_\pi)$ is :

$$\hat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} \frac{\hat{t}_{j\pi}}{\pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}.$$

Remark 8 *: For particular conditions, we can obtain designs already studied.*

1. *for $s_I = \mathcal{U}_I$ we obtain stratified sampling (described in section 3) and we find the formula for the variance estimator;*

2. *for $s_i = \mathcal{U}_i$ for all i, then we have cluster sampling which we describe in next section.*

5.1.2 Two-stage sampling with SI designs at each stage

Let us denote by SI,SI the two-stage sampling with SI designs at both stages. At first stage, we select an SI sample s_I of size n_I from the population $\mathcal{U}_I$ of PSUs. For each $i \in s_I$, we draw an SI sample s_i of size n_i from U_i. The final sample is $s = \cup_{i \in s_I} s_i$.
The first and second order inclusion probabilities are

$$\pi_{Ii} = \frac{n_I}{N_I}, \quad \pi_{Iij} = \frac{n_I(n_I-1)}{N_I(N_I-1)} \quad \text{for the 1st stage}$$

$$\pi_{k|i} = \frac{n_i}{N_i}, \quad \pi_{kl|i} = \frac{n_i(n_i-1)}{N_i(N_i-1)}, k, l \in \mathcal{U}_i \quad \text{for the 2nd stage}$$

The π-estimators for the PSU totals $t_i = \sum_{\mathcal{U}_i} y_k$ (with respect to the second stage) are

$$\hat{t}_{i\pi} = \sum_{s_i} \frac{y_k}{\pi_{k|i}} = \sum_{s_i} N_i \frac{y_k}{n_i} = N_i \overline{y}_{s_i}$$

for all $i \in s_I$ and the π-estimator for the total $t_y = \sum_{\mathcal{U}} y_k$ is

$$\hat{t}_\pi = \sum_{s_I} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} = \frac{N_I}{n_I} \sum_{s_I} \hat{t}_{i\pi}.$$

The variance is

$$V(\hat{t}_\pi) = N_I^2 \frac{1-f_I}{n_I} S^2_{t\mathcal{U}_I} + \frac{N_I}{n_I} \sum_{\mathcal{U}_I} N_i^2 \frac{1-f_i}{n_i} S^2_{y\mathcal{U}_i}$$

where $f_I = n_I/N_I$, $f_i = n_i/N_i$, $S^2_{t\mathcal{U}_I} = \sum_{\mathcal{U}_I} (t_i - \overline{t}_{\mathcal{U}_I})^2/(N_I - 1)$ is the variance in $\mathcal{U}_I$ of the total t_i and $S^2_{y\mathcal{U}_i} = \sum_{U_i} (y_k - \overline{y}_{\mathcal{U}_i})^2/(N_i - 1)$ is the variance in U_i of y_k.

5.1.3 Two-stage sampling with R and SAS

5.2 Cluster sampling

5.2.1 Description

It is the simplest case of multistage sampling since we deal with only one-stage. In cluster sampling, the population is divided into N_I PSU which are called now *clusters*. Next, a sample s_I of clusters is selected as described in the first step of multi-stage sampling according to a sampling design $p_I(\cdot)$. Every individual from the selected cluster is observed.
The final sample is

$$s = \bigcup_{i \in s_I} U_i$$

of size $n_s = \sum_{s_I} N_i$.

The clusters may be confounded with strata since both are disjointed subpopulations of $\mathcal{U}$. Nevertheless, the way clusters and strata are build is not the same and we can say the same thing about the selection procedure. Whereas stratification generally increases precision compared to SI sampling, cluster sampling generally decreases it. Members of the same cluster tend to be more familiar that elements selected at random from the whole population (Lohr, 1999):

1. members of the household tend to have similar political views;
2. fish in the same lake tend to have the similar concentration of mercury.

By sampling everyone in the cluster, we repeat the same information instead of obtaining new one and the estimations are less precise than obtained with a SI sampling. Cluster sampling is used in practice because it is usually much cheaper and more convenient to sample in clusters than randomly in the population.

The first and second inclusion probabilities are

$$\pi_k = \pi_{Ii} \quad \text{if} \quad k \in U_i$$

$$\pi_{kl} \quad = \quad \begin{cases} \pi_{Ii} & \text{if} \quad k, l \in \mathcal{U}_i \\ \pi_{Iij} & \text{if} \quad k \in \mathcal{U}_i, \quad l \in \mathcal{U}_j, \quad i \neq j \end{cases}$$

The population total can be expressed as

$$t_y = \sum_{\mathcal{U}} y_k = \sum_{\mathcal{U}_I} t_i$$

where $t_i = \sum_{\mathcal{U}_i} y_k$.

Result 12 *Let us consider the cluster sampling.*

1. *The π-estimator of t_y is $\hat{t}_\pi = \sum_{s_I} \check{t}_i = \sum_{s_I} \dfrac{t_i}{\pi_{Ii}}$.*
2. *The variance of $\hat{t}_\pi$ is given by $V(\hat{t}_\pi) = \sum_{\mathcal{U}_I} \sum_{\mathcal{U}_I} \Delta_{Iij} \check{t}_i \check{t}_j$.*
3. *The variance estimator is $\hat{V}(\hat{t}_\pi) = \sum_{s_I} \sum_{s_I} \check{\Delta}_{Iij} \check{t}_i \check{t}_j$.*

If the first-stage sampling design is of fixed size, the Yates-Grundy variance formulas can be used for $\hat{t}_\pi$,

$$V(\hat{t}_\pi) = -\frac{1}{2}\sum_{\mathcal{U}_I}\sum_{\mathcal{U}_I}\Delta_{Iij}(\check{t}_i - \check{t}_j)^2$$

with the unbiased variance estimator

$$\hat{V}(\hat{t}_\pi) = -\frac{1}{2}\sum_{s_I}\sum_{s_I}\check{\Delta}_{Iij}(\check{t}_i - \check{t}_j)^2$$

Remark 9 *The systematic sampling is a particular case of cluster sampling obtained when the sample size from the first stage is one, $n_I = 1$ and N_I clusters corresponds to the $\boldsymbol{a}$ possible systematic samples.*

5.2.2 Simple random cluster sampling (SIC)

We apply the result 12 when the sample of clusters s_I of size n_I is selected according to a SI design among the N_I clusters. The inclusion probabilities are given by $\pi_{Ii} = n_I/N_I$ and

$$\hat{t}_\pi = \sum_{s_I} N_I \frac{t_i}{n_I} = N_I \bar{t}_{s_I}.$$

The variance of the π-estimator is given by

$$V(\hat{t}_\pi) = N_I^2 \frac{1-f_I}{n_I} S^2_{t\mathcal{U}_I}$$

where $f_I = n_I/N_I$ and $S^2_{t\mathcal{U}_I} = \sum_{\mathcal{U}_I}(t_i - \bar{t}_{\mathcal{U}_I})^2/(N_I - 1)$ is the variance of t_i with $\bar{t}_{\mathcal{U}_I} = \sum_{\mathcal{U}_I} t_i/N_I$. The variance estimator is given by

$$\hat{V}(\hat{t}_\pi) = N_I^2 \frac{1-f_I}{n_I} S^2_{ts_I}$$

where $S^2_{ts_I} = \sum_{s_I}(t_i - \bar{t}_{s_I})^2/(n_I - 1)$.

5.2.3 Efficiency of SIC

We introduce the *homogeneity coefficient*

$$\delta = 1 - \frac{S^2_{yW}}{S^2_{yU}} \tag{5.1}$$

where

$$S^2_{yW} = \frac{1}{N - N_I}\sum_{\mathcal{U}_I}\sum_{\mathcal{U}_i}(y_k - \overline{y}_{\mathcal{U}_i})^2$$

is the pooled within-cluster variance which can be written as

$$S^2_{yW} = \frac{\sum_{\mathcal{U}_I}(N_i - 1)S^2_{yU_i}}{\sum_{\mathcal{U}_I}(N_i - 1)}$$

where $S^2_{yU_i} = \sum_{\mathcal{U}_i}(y_k - \overline{y}_{\mathcal{U}_i})^2/(N_i - 1)$.

Result 13 *The homogeneity coefficient δ satisfies*

$$-\frac{N_I - 1}{N - N_I} \leq \delta \leq 1.$$

Proof From (5.1), we have $\delta \leq 1$. We use the ANOVA decomposition

$$SST = SSW + SSB \quad \text{or}$$

$$(N-1)S^2_{y\mathcal{U}} = (N - N_I)S^2_{yW} + \sum_{\mathcal{U}_I} N_i(\overline{y}_{\mathcal{U}_i} - \overline{y}_{\mathcal{U}})^2$$

which gives that $\dfrac{S^2_{yW}}{S^2_{y\mathcal{U}}} \leq \dfrac{N-1}{N-N_I}$. The lower bound on δ follows from the fact that $\delta = 1 - \dfrac{S^2_{yW}}{S^2_{y\mathcal{U}}}$.
□

Remark 10 *1. A small δ-value means $\frac{S^2_{yW}}{S^2_{yU}}$ proxy to unity or equivalently, elements in the same cluster dissimilar with respect to the study variable. We have in this case a low degree of homogeneity within clusters.*

2. At large δ-value means a high degree of homogeneity within clusters.

3. $\delta = 1$ means $S^2_{yW} = 0$.

4. $\delta = 0$ means $S^2_{yW} = S^2_{yU}$.

5.2.4 Comparison between SIC sampling and SI sampling

5.2.5 Cluster sampling with R and SAS

Chapter 6

Taylor linearization

In the previous section, a π-estimator was presented for the population total, with special attention on the variance estimator and its expression for different sampling designs.

But, in practice, in most of the cases, we are confronted with surveys that involve not only one, but several variables of interest and more generally, the unknown quantity depends on these variables through a general function. An example is the ratio of two unknown population totals

$$\mathcal{R} = \frac{\sum_{\mathcal{U}} y_k}{\sum_{\mathcal{U}} z_k} = \frac{t_y}{t_z}$$

where y and z are two variables of study. First, we will study the linear case, then the general case will be considered by using Taylor series expansions. The use of Taylor series implies the introduction of supplementary conditions on the population and on the sampling design. They will permit the development in Taylor series and also the convergence of it. The subject was treated by Wolter (1985), Särndal *et al.* (1992). We will give a short review of the main results.

6.1 π estimators for the linear case

Suppose that there are J variables of study $\mathcal{Y}_1, \ldots, \mathcal{Y}_j, \ldots, \mathcal{Y}_J$, and let y_{jk} be the value of $\mathcal{Y}_j$ for the k-th element of the population, for all $j = 1, \ldots, J$ and all $k = 1, \ldots, N$. Let $t_j = \sum_{\mathcal{U}} y_{jk}$ be the total of the $\mathcal{Y}_j$ variable for all $j = 1, \ldots, J$. The objective is to estimate these quantities, namely the components of the following vector:

$$\mathbf{t} = (t_1, \ldots, t_j, \ldots, t_J)'$$

For each variable of interest the theory of the π estimator, presented in the first section, can be applied. A sample s is drawn from $\mathcal{U}$, according to a sampling design $p(s)$, with the probabilities of inclusion of first and second order, π_k and π_{kl} and for all $k \in s$ we observe the value of the vector:

$$\mathbf{y_k} = (y_{1k}, \ldots, y_{jk}, \ldots, y_{Jk})'$$

and each t_j total is estimated by the π-estimator $\hat{t}_{j\pi} = \sum_s \breve{y}_{jk}$ so that the π-estimator of $\mathbf{t}$ is

$$\hat{\mathbf{t}}_\pi = (\hat{t}_{1\pi}, \ldots, \hat{t}_{j\pi}, \ldots, \hat{t}_{J\pi})'$$

We have the following results :

Result 14 *(Särndal et al. 1992): The variance-covariance matrix of $\hat{t}_\pi$ has the following expression:*

$$V(\hat{\mathbf{t}}_\pi) = E\left((\hat{\mathbf{t}}_\pi - \mathbf{t})(\hat{\mathbf{t}}_\pi - \mathbf{t})'\right)$$

and is a symmetric matrix with the j-th diagonal element given by the variance of $\hat{t}_{j\pi}$ and the elements jj' given by the covariance of $\hat{t}_{j\pi}$ and $\hat{t}_{j'\pi}$:

$$V(\hat{t}_{j\pi}) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \check{y}_{jk} \check{y}_{jl}$$

and

$$C(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \check{y}_{jk} \check{y}_{j'l}$$

The matrix $V(\hat{\mathbf{t}}_\pi)$ is estimated with no bias by the matrix $\hat{V}(\hat{\mathbf{t}}_\pi)$ such that the jth diagonal element is

$$\hat{V}(\hat{t}_{j\pi}) = \sum_{s} \sum_{s} \check{\Delta}_{kl} \check{y}_{jk} \check{y}_{jl}$$

and the jj' element is :

$$\hat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum_{s} \sum_{s} \check{\Delta}_{kl} \check{y}_{jk} \check{y}_{j'l}$$

We consider now a more general situation, namely when the requested estimator for a parameter population θ can be written as follows:

$$\theta = f(t_1, \ldots, t_J)$$

and f is a linear function. Then, applying the above result, an estimator for θ is given by:

$$\hat{\theta} = f(\hat{t}_1, \ldots, \hat{t}_J).$$

This is derived from the fact that if f is a linear function then θ can be written as:

$$\theta = a_0 + \sum_{j=1}^{J} a_j t_j = f(t_1, \ldots, t_J).$$

Consequently $\hat{\theta} = a_0 + \sum_{j=1}^{J} a_j \hat{t}_{j\pi} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi})$ where $\hat{t}_{j\pi} = \sum_s \frac{y_{jk}}{\pi_k}$ is the π-estimator for the total t_j. We can apply now the conclusions from the previous result in order to obtain the expressions for the variance-covariance matrix of θ and also for an estimator of this matrix.

Result 15 *For a parameter of the population, having the form :*

$$\theta = a_0 + \sum_{j=1}^{J} a_j t_j = f(t_1, \ldots, t_J)$$

an estimator is given by :

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi})$$

with the variance-covariance matrix :

$$V(\hat{\theta}) = \sum_{j=1}^{J} \sum_{j'=1}^{J} a_j a_j' C(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$$

$$\text{and } \hat{V}(\hat{\theta}) = \sum_{j=1}^{J}\sum_{j'=1}^{J} a_j a'_j \widehat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$$

where $C(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$ *and* $\widehat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$ *are given in the result from above.*

As a conclusion, if a parameter of interest is expressed by a linear function of the totals of q variables of study, then the expressions for V and $\hat{V}$ can be deduced in a simple way.

6.2 The general case

The case when f is not a linear function will be reduced to the former one. The main idea is to approximate f around the true value by a linear function, for which we know how to derive formulas for the variance and for an estimator of the variance, according to the case one. Under general conditions, we will show how the variance of the estimator can be approximated by the variance of a linear estimator. The approximation will be made by a first-order Taylor series expansion and the method is called *Taylor linearization*. We start with a consistent estimator $\hat{t}$ for t and a function f which satisfies several conditions, the expansion in Taylor series of f around the point $t = (t_1, \ldots, t_J)$ will give:

1. an approximate expression for the design variance of $\hat{\theta} = f(\hat{t})$;
2. a suitable estimator of the variance of $\hat{\theta}$.

But for a finite population $\mathcal{U}$ we can not define the consistency and asymptotic unbiasedness of an estimator. For achieving these conditions and also for allowing us to develop f in a Taylor series with a remainder of low order, we need supplementary conditions and mathematical results concerning the behaviour of a finite population and of the probabilities π_k, π_{kl} when n and N increase to infinite simultaneously. In order to obtain an infinite population, we will consider the initial population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ with the corresponding π_k.
We give here Särndal's way (1980) to obtain an infinite population. Isaki & Fuller (1982), Särndal *et al.* 1992 propose different ones. Särndal (1980) reproduces this population t-1 times. For all t, a sample s is selected from each $\mathcal{U}$ according to $p(s)$, with the same π_k, for all t. The resulting population will have $N_t = Nt$ elements from which we select a sample $s_{(t)}$ consisting of $n_{s_{(t)}} = \sum_{\gamma=1}^{t} n_{s_\gamma} = nt$ elements. Next, we allow $t \to \infty$ and it results that $N_t \to \infty$ and $n_{s(t)} \to \infty$ but n and N remains constant. This framework allows us to define the properties of consistency and asymptotic unbiasedness for an estimator. Now we can derive the results for Taylor linearization. Let us give two results which will be used in the next.

Result 16 *(Wolter 1985): Let* $f : \xi_J \to R$ *be a real valued function defined on a q dimensional Euclidian space, continuous, differentiable with continuous partial derivatives of order two in an open sphere containing* $\mathbf{a} = (a_1, \ldots, a_J)'$. *For* $\mathbf{X}_n = (X_{1n}, \ldots, X_{Jn})'$ *satisfying:*

$$\mathbf{X}_n = \mathbf{a} + O_p(r_n) \text{ where } r_n \to 0,$$

we have:

$$f(\mathbf{X}_n) = f(\mathbf{a}) + \sum_{j=1}^{J} (X_{jn} - a_j) \frac{\partial f}{\partial x_j}(\mathbf{a}) + O_p(r_n^2).$$

As mentioned, we want to estimate a parameter expressed by a nonlinear function of totals of J variables of interest:

$$\theta = f(t_1, \ldots, t_J) \text{ where } t_j = \sum_{\mathcal{U}} y_{kj}, j = 1, \ldots, J.$$

with $f : \xi_J \to R$ continuous, differentiable with continuous partial derivatives.

We define an estimator for θ by substitution. It is obtained by replacing each total with the corresponding π estimator:

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi}).$$

We can take as vectors $\mathbf{a}$ and $\mathbf{X}_n$ from **Result 16** the vectors $\mathbf{t}$ and $\mathbf{t}_\pi$ as defined at the beginning of this section,

$$\mathbf{t} = (t_1, \ldots, t_J)' \text{ and}$$
$$\hat{\mathbf{t}}_\pi = (\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi}).$$

We make the following assumptionswhich have the following properties :

1. $N^{-1}(\hat{\mathbf{t}}_\pi - \mathbf{t}) \xrightarrow{P} 0$ and
2. $N^{-1}(\hat{\mathbf{t}}_\pi - \mathbf{t}) = O_p(n^{-\frac{1}{2}})$
3. $n^{\frac{1}{2}} N^{-1}(\hat{\mathbf{t}}_\pi - \mathbf{t}) \xrightarrow{\mathcal{L}} N(0, \Sigma)$.

For $N^{-1}\mathbf{t}$, $N^{-1}\hat{\mathbf{t}}_\pi$ given above and $r_n = n^{-\frac{1}{2}}$ the conditions from **Result 16** are satisfied. We suppose also that it exists $\alpha > 0$ such that $f(N^{-1}t) = N^{-\alpha} f(t)$.
We give below the first-order Taylor expansion of $f(N^{-1}\hat{\mathbf{t}}_\pi)$ around $f(N^{-1}\hat{t}_\pi)$.

$$\begin{aligned} N^{-\alpha} f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi}) &= N^{-\alpha} f(t_1, \ldots, t_J) \\ &+ N^{-\alpha} \sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j) \left. \frac{\partial f(v_1, \ldots, v_J)}{\partial v_j} \right|_{(v_1,\ldots,v_J)=(t_1,\ldots,t_J)} \\ &+ O_p\left(\frac{1}{n}\right) \end{aligned}$$

and if we note with $\alpha_j = \left. \frac{\partial f(v_1,\ldots,v_J)}{\partial v_j} \right|_{(v_1,\ldots,v_J)=(t_1,\ldots,t_J)}$ for $j = 1, \ldots, J$ we obtain the equivalent formula :

$$N^{-\alpha}\hat{\theta} = N^{-\alpha}\theta + N^{-\alpha} \sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j)\alpha_j + O_p\left(\frac{1}{n}\right)$$

and

$$N^{-\alpha}\hat{\theta} = N^{-\alpha}\theta + O_p(n^{-\frac{1}{2}}).$$

In particular $\hat{\theta}$ can be approximated by $N^{-\alpha}(\hat{\theta} - \theta) \simeq N^{-\alpha} \sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j)\alpha_j$. By assumption 3, we obtain that $N^{-\alpha}(\hat{\theta} - \theta)$ is asymptotically normal with mean zero and variance equal to the variance of $\sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j)\alpha_j$.

$$V(\sum_{j=1}^{J} (\hat{t}_{j\pi} - t_j)\alpha_j) = \boldsymbol{\alpha}\boldsymbol{\Sigma}\boldsymbol{\alpha}'$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ is the vector of the partial derivatives of f calculated in $\mathbf{t} = (t_1, \ldots, t_q)$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{t}_\pi - \mathbf{t}$,

$$V(\hat{\theta}) \simeq \boldsymbol{\alpha}\boldsymbol{\Sigma}\boldsymbol{\alpha}' = V\left(\sum_{i=1}^{J} \alpha_j(\hat{t}_{j\pi} - t_j)\right)$$

To obtain a variance estimator, we will substitute sample-based estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$. Suppose that the estimator $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\Sigma}}$ are available; then an estimator for MSE $(\hat{\theta})$ is given by:

$$\widehat{V}(\hat{\theta}) = \hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\alpha}}'.$$

As it can be observed, the expressions for approximative variance and variance estimator are complicated because they involve the calculation of the variance-covariance matrix $\boldsymbol{\Sigma}$; for the variance estimation, we need to calculate an estimator for each element of $\hat{\boldsymbol{\Sigma}}$. Woodruf (1971) gives a simple method for which the variance estimation is simplified. This method is a generalization of the Keyfitz's (1957) method for obtaining the variances for specific types of estimates derived from specific sample designs. It consists in reordering the components of the sum $\sum_{j=1}^{J} \alpha_j \hat{t}_{j\pi}$. We have:

$$\begin{aligned}\sum_{j=1}^{J} \alpha_j \hat{t}_{j\pi} &= \sum_{j=1}^{J} \alpha_j \left(\sum_s \frac{y_{jk}}{\pi_k}\right) = \sum_s \frac{1}{\pi_k}\left(\sum_{j=1}^{J} \alpha_j y_{jk}\right) \\ &= \sum_s \frac{u_k}{\pi_k} = \sum_s \breve{u}_k.\end{aligned}$$

where $u_k = \sum_{j=1}^{J} \alpha_j y_{jk}$. As it can be observed, $\sum_{j=1}^{J} \alpha_j \hat{t}_{j\pi}$ can be written equivalently as the π-estimator for the total of the new introduced quantities u_k. Because the u_k for all $k \in \mathcal{U}$ depend on α_j which on their turn, depend on the $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$ which are unknown we obtain that u_k are unknown and so they can not be used. The quantities α_j are estimated by $\hat{\alpha}_j$ the corresponding π-estimator

$$\hat{\alpha}_j = \left.\frac{\partial f(v_1, \ldots, v_J)}{\partial v_j}\right|_{(v_1,\ldots,v_J)=(\hat{t}_{1\pi},\ldots,\hat{t}_{J\pi})}$$

and then the u_k are estimated by:

$$\hat{u}_k = \sum_{j=1}^{J} \hat{\alpha}_j y_{kj}.$$

It results then $V(\hat{\theta}) \simeq V(\sum_s \breve{u}_k)$. Now we can give the result obtained by Woodruff (1971).

Result 17 *(Woodruff 1971)*
1. An approximatively unbiased estimator for the population parameter

$$\theta = f(t_1, \ldots, t_J)$$

is given by the substitution estimator:

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{J\pi})$$

where $\hat{t}_{j\pi}$ *is the corresponding* π *estimator of* t_j.
2. Using the Taylor linearization, the approximative variance is :

$$V(\hat{\theta}) \simeq \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$$

where $u_k = \sum_{j=1}^{J} \alpha_j y_{kj}$ *and* $\alpha_j = \left.\frac{\partial f(v_1,\ldots,v_J)}{\partial v_j}\right|_{(v_1,\ldots,v_J)=(t_1,\ldots,t_J)}$ *and*
3. The estimated variance has the expression:

$$\widehat{V}(\hat{\theta}) = \sum_{s} \sum_{s} \breve{\Delta}_{kl} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

where $\hat{u}_k = \sum_{j=1}^{J} \hat{\alpha}_j y_{kj}$ *and* $\hat{\alpha}_j$ *is the* π *estimator for* α_j.

Supposing that the general conditions given by Isaki & Fuller (1982) are fulfilled, $\hat{u}_k$ is a consistent estimator for u_k and as a result $\hat{V}(\hat{\theta})$ is consistent for $V(\hat{\theta})$. For a design of fixed size, we have the alternative formulas for the approximative variance and variance estimation:

$$V(\hat{\theta}) \simeq -\frac{1}{2} \sum \sum_{\mathcal{U}} \Delta_{kl} \left(\breve{u}_k - \breve{u}_l\right)^2$$

$$\hat{V}(\hat{\theta}) = -\frac{1}{2} \sum \sum_{s} \breve{\Delta}_{kl} \left(\breve{u}_k - \breve{u}_l\right)^2$$

In large sample, we estimate the approximative variance of $\hat{\theta}$ given in Woodruff's result and the found value can be considered as an estimator of the true variance.

From the above result, we can summarize and give the steps when the Taylor technique is applied, namely:

- For the population parameter $\theta = f(t_1, \ldots, t_J)$, expressed as a function of the J totals, we derive the substitution estimator $\hat{\theta}$ which is approximately unbiased for θ. The variance and variance estimator of $\hat{\theta}$ must be calculated.
- The linearized variable, u_k, is derived for all $k \in \mathcal{U}$; these quantities are calculated in the population, u_k being expressed as functions of the J totals $t_1, \ldots, t_J$.
- The unknown quantities u_k are estimated by $\hat{u}_k$.
- The parameter is approximated by $\hat{\theta} \simeq \theta + (\sum_s \breve{u}_k - \sum_{\mathcal{U}} u_k)$.
- According to the result of Woodruf we have

$$V(\hat{\theta}) \simeq V\left(\sum_{s} \frac{u_k}{\pi_k}\right) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}.$$

- A Horvitz-Thompson variance estimate, based on the unknown u_k can be obtained:

$$\hat{V}(\hat{\theta}) = \sum_{s} \sum_{s} \breve{\Delta}_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}.$$

- The estimated variance based on the sample estimators $\hat{u}_k$ has the expression:

$$\hat{V}(\hat{\theta}) = \sum_{s} \sum_{s} \breve{\Delta}_{kl} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

6.3 Examples

In the following, we consider several of non-linear parameter and apply in the Taylor technique.

Estimation of a ratio

Let us consider the ratio between two unknown population totals $t_y = \sum_{\mathcal{U}} y_k$ and $t_x = \sum_{\mathcal{U}} x_k$:

$$R = \frac{t_y}{t_x} = \frac{\sum_{\mathcal{U}} y_k}{\sum_{\mathcal{U}} x_k} = f(t_y, t_x)$$

where $f(v_1, v_2) = \frac{v_1}{v_2}$; we want to estimate R.

1. The substitution estimator of R is $\hat{R} = \dfrac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}$ where $\hat{t}_{y\pi}, \hat{t}_{x\pi}$ are the π estimators for t_y, t_x.

2. The linearized variable of R is :

$$\begin{aligned} u_k &= y_k \left.\frac{\partial f(v_1, v_2)}{\partial v_1}\right|_{(v_1,v_2)=(t_y,t_x)} + x_k \left.\frac{\partial f(v_1, v_2)}{\partial v_2}\right|_{(v_1,v_2)=(t_y,t_x)} \\ &= y_k \frac{1}{t_x} + x_k \left(\frac{-t_y}{t_x^2}\right) \\ &= \frac{1}{t_x}\left(y_k - R x_k\right). \end{aligned}$$

3. $\hat{R}$ is approximately estimated by

$$\hat{R} \simeq R + \left(\sum_s \frac{u_k}{\pi_k} - \sum_{\mathcal{U}} u_k\right) = R + \frac{1}{t_x}\sum_s \frac{y_k - R x_k}{\pi_k}.$$

because $\sum_{\mathcal{U}} u_k = 0$ in this case.

4. The approximated variance of $\hat{R}$ is:

$$V(\hat{R}) \simeq \left(\frac{1}{t_x}\right)^2 \sum_{\mathcal{U}}\sum_{\mathcal{U}} \Delta_{kl} \frac{y_k - R x_k}{\pi_k} \frac{y_l - R x_l}{\pi_l}.$$

5. The variance estimator is obtained by replacing u_k by $\hat{u}_k = \dfrac{1}{\hat{t}_{x\pi}}\left(y_k - \hat{R} x_k\right)$ in the expression of the estimated Horvitz-Thompson variance formula:

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_{x\pi}^2} \sum_s \sum_s \check{\Delta}_{kl} \frac{y_k - \hat{R} x_k}{\pi_k} \frac{y_l - \hat{R} x_l}{\pi_l}.$$

Estimation of a mean

Another example is the derivation of an estimator for the mean of the population.
There are two situations : N is known and not. In the first situation, an unbiased estimator for $\overline{y}_{\mathcal{U}} = \frac{1}{N}\sum_{\mathcal{U}} y_k$ is :

$$\widehat{\overline{y}}_{\mathcal{U}\pi} = \frac{1}{N}\sum_s \frac{y_k}{\pi_k}$$

with the variance and variance estimation, respectively :

$$V(\widehat{\overline{y}}_{\mathcal{U}\pi}) = \frac{1}{N^2}\sum_{\mathcal{U}}\sum_{\mathcal{U}}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

$$\widehat{V}(\widehat{\overline{y}}_{\mathcal{U}\pi}) = \frac{1}{N^2}\sum_{s}\sum_{s}\breve{\Delta}_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

derived using the general theory of the π estimator.

For N unknown, we need an estimator for N. The quantity N can be regarded as the total of the variable $x_k = 1$ for all $k \in U$, then $N = t_x$ and the substitution estimator is $\hat{N} = \sum_s \frac{1}{\pi_k}$. Then $\overline{y}_{\mathcal{U}}$ can be regarded as a ratio of two totals variable. Using the first example, an estimator for $\overline{y}_{\mathcal{U}}$ is :

$$\tilde{y}_s = \frac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}} = \frac{1}{\hat{N}}\sum_s \frac{y_k}{\pi_k}$$

with the approximate variance :

$$V(\tilde{y}_s) \simeq \frac{1}{N^2}\sum_{\mathcal{U}}\sum_{\mathcal{U}}\Delta_{kl}\frac{y_k - \overline{y}_{\mathcal{U}}}{\pi_k}\frac{y_l - \overline{y}_{\mathcal{U}}}{\pi_l}$$

and the variance estimator :

$$\hat{V}(\tilde{y}_s) = \frac{1}{\hat{N}^2}\sum_{s}\sum_{s}\hat{\Delta}_{kl}\frac{y_k - \tilde{y}_s}{\pi_k}\frac{y_l - \tilde{y}_s}{\pi_l}.$$

For SI and STSI designs the two estimators $\tilde{y}_s$ and $\widehat{\overline{y}}_{\mathcal{U}\pi}$ are identical. Generally, the estimator $\tilde{y}_s$ gives better results than $\widehat{\overline{y}}_{\mathcal{U}\pi}$ even if N is known. We give below several arguments in favor of $\tilde{y}_s$.

1. Compare the variances of $\tilde{y}_s$ and $\widehat{\overline{y}}_{\mathcal{U}\pi}$. The estimator $\tilde{y}_s$ is preferred when $y_k - \overline{y}_U$ are all small.

2. Another reason is that $\tilde{y}_s$ works better when we have designs of variable sizes such as Bernoulli or Poisson designs. More exactly, $\overline{y}_{\mathcal{U}}$ is variable due to variable sample size which is not the case for $\tilde{y}_s$.

3. A third argument in favor of $\tilde{y}_s$ arises in the case when π_k are poorly (or negatively) correlated with the y_k values. In this situation, $\tilde{y}_s$ enjoys a kind of adaptability or insensitivity to unlucky samples that is lacking in $\overline{y}_{\mathcal{U}}$ because its denominator remains fixed.

Ratio estimator

Suppose we have a study variable, y, and that we have the auxiliary variable x; x_k is the value of x for the k-th element of the population and x_k is known for all k in $\mathcal{U}$. The objective is to estimate $t_y = \sum_{\mathcal{U}} y_k$; t_y can be written :

$$t_y = t_x\frac{t_y}{t_x}$$

$t_x = \sum_{\mathcal{U}} x_k$ is a known quantity; then $R = \frac{t_y}{t_x}$ is the ratio of two totals and we can apply the results from the first example. It follows :

1. $\hat{t}_y = t_x \hat{R}$, where $\hat{R} = \dfrac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}$

2. $V(\hat{t}_y) \simeq \sum \sum_{\mathcal{U}} \Delta_{kl} \dfrac{y_k - Rx_k}{\pi_k} \dfrac{y_l - Rx_l}{\pi_l}$

3. $\hat{V}(\hat{t}_y) = \dfrac{t_x^2}{\hat{t}_{x\pi}^2} \sum_s \sum_s \check{\Delta}_{kl} \dfrac{y_k - \hat{R}x_k}{\pi_k} \dfrac{y_l - \hat{R}x_l}{\pi_l}$.

Estimation of the regression coefficient

Let us derive an approximately unbiased estimator for the coefficient of multiple regression. Let $\mathcal{X}_1, \ldots, \mathcal{X}_q$ be q auxiliary variables and $B_1, \ldots, B_q$ be the coefficients of regression of $\mathcal{Y}$ through $\mathcal{X}_1, \ldots, \mathcal{X}_q$. We intend to obtain an approximately unbiased estimator for the vector of regression coefficients $\mathbf{B} = (B_1, \ldots, B_q)'$ whose variance or approximative variance can be calculated. We denote by $\mathbf{x}_k$ the vector containing the values taken by the auxiliary variables for the k-th element in the population,

$$\mathbf{x_k} = (x_{1k}, \ldots, x_{qk})' \text{ for all } k \in \mathcal{U}$$

and $\mathbf{T} = \sum_{\mathcal{U}} \mathbf{x_k} \mathbf{x_k}'$

1. The vector of regression coefficients has the expression:

$$\mathbf{B} = (B_1, \ldots, B_q)' = (\sum_{\mathcal{U}} \mathbf{x_k} \mathbf{x_k}')^{-1} (\sum_{\mathcal{U}} \mathbf{x_k} y_k).$$

It can be observed that $\mathbf{B}$ can be written as a function of totals. Thus it is possible to apply the method of Taylor linearization. We have $\mathbf{B} = f(t_q, t_z) = \dfrac{t_q}{t_z}$ where:

$$t_q = \sum_{\mathcal{U}} q_k \text{ the total of the new variable } q_k = \mathbf{x_k} y_k, k \in \mathcal{U},$$

and

$$t_z = \sum_{\mathcal{U}} z_k \text{ the total of the new variable } z_k = \mathbf{x_k} \mathbf{x_k'}, k \in \mathcal{U}.$$

2. The substitution estimator for $\mathbf{B}$ is:

$$\hat{\mathbf{B}} = \left(\sum_s \frac{\mathbf{x_k} \mathbf{x_k'}}{\pi_k} \right)^{-1} \left(\sum_s \frac{\mathbf{x_k} y_k}{\pi_k} \right).$$

3. The linearized variable of $\mathbf{B}$ is:

$$\begin{aligned} u_k &= q_k \left. \frac{\partial f(v_1, v_2)}{\partial v_1} \right|_{(v_1, v_2) = (t_q, t_z)} + z_k \left. \frac{\partial f(v_1, v_2)}{\partial v_2} \right|_{(v_1, v_2) = (t_q, t_z)} \\ &= q_k \frac{1}{t_z} + z_k \left(\frac{-t_q}{t_z^2} \right) \\ &= \mathbf{T}^{-1} \mathbf{x_k} (y_k - \mathbf{x_k'} \mathbf{B}). \end{aligned}$$

4. We can obtain an approximative expression for $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} \simeq \mathbf{B} + \left(\sum_s \breve{u}_k - \sum_{\mathcal{U}} u_k \right) = \mathbf{B} + \mathbf{T}^{-1} \left(\sum_s \frac{\mathbf{x_k} y_k}{\pi_k} - \hat{\mathbf{T}}\mathbf{B} \right).$$

5. The approximated variance of $\hat{\mathbf{B}}$ has the expression:

$$V(\hat{\mathbf{B}}) \simeq \mathbf{T}^{-1}\mathbf{V}\mathbf{T}^{-1}$$

where $\mathbf{V} = (v_{jj'})_{i,j=1,\ldots,q}$, $v_{ij} = v_{ji}$ and $v_{jj'} = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} \frac{x_{jk}(y_k - \mathbf{x_k'B})}{\pi_k} \frac{x_{j'l}(y_l - \mathbf{x_l'B})}{\pi_l}$.

6. The linearized variable u_k is estimated by $\hat{u}_k = \hat{\mathbf{T}}^{-1}\mathbf{x_k}(y_k - \mathbf{x_k'}\hat{\mathbf{B}})$, $\hat{\mathbf{T}} = \sum_s \frac{\mathbf{x_k x_k'}}{\pi_k}$ and the estimator for the variance of $\mathbf{B}$:

$$\hat{V}(\hat{\mathbf{B}}) = \hat{\mathbf{T}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{T}}^{-1}$$

where $\hat{\mathbf{V}} = \left(\hat{v}_{jj'}\right)$ is a $q \times q$ matrix with elements:

$$\hat{v}_{jj'} = \sum_s \sum_s \breve{\Delta}_{kl} \frac{\hat{u_k}}{\pi_k} \frac{\hat{u_l}}{\pi_l}.$$

Remark 11 *:The same result would have been obtained more easily if we had considered* $\mathbf{B}$ *as a ratio of the totals of the variables* q_k *and* z_k *and we had applied directly the linearized variable for a ratio derived at first example.*

Chapter 7

"Methods de redressement "

7.1 Model Approach

The previous sections treated *the parameter vector* $\mathbf{y} = (y_1, \ldots, y_N)$ as a non-random quantity, the only randomness being the sampling design $p(\cdot)$. In the present section, we will consider that $\mathbf{y} = (y_1, \ldots, y_N)$ is the outcome of a vector random variable $\mathbf{Y} = (Y_1, \ldots, Y_N)$ with distribution ξ. We call *superpopulation model* a specified set of conditions for the class of distributions of which ξ belongs to. The main aspect of the statistical analysis in the *superpopulation model* is thus that $\mathbf{y}$ is treated as the outcome of $\mathbf{Y}$ about which certain features are assumed known. The *superpopulation model* ξ can be regarded as a mathematical device used to make explicit the theoretical derivations.

Among the first having used the *superpopulation model*, we mention Cochran (1939, 1946), Deming and Stephan (1941), Madow and Madow (1944).

Although the use of a *superpopulation model* ξ is not accepted by all survey practitioners, there are situations when it is arguable that this approach will perform much better. We mention two such situations. The first one is the inclusion of the treatment of nonsampling errors in survey sampling (Särndal *et al.* 1992, ch. 14). Secondly, it is possible under a *superpopulation model* ξ to make comparison of variances of two p-unbiased strategies, fact which entails the resolution of some of the nonexistence problems in uniformly minimum variance p-unbiased estimation (Cassel *et al.* 1977).

In the case of a *superpopulation model* ξ, we have two kinds of randomness: one already existing, the sampling design $p(\cdot)$ and the new one introduced by the joint distribution ξ of $Y_1, \ldots, Y_N$. We need supplementary notations induced by ξ. Let $Q = Q(Y_1, \ldots, Y_N)$ be a function of $Y_1, \ldots, Y_N$ and we denote by $E_\xi(Q)$ the expectation of Q with respect to ξ defined as follows

$$E_\xi(Q) = \int Q d\xi.$$

In the same manner, we can define other statistical quantities as variance and covariance with respect to ξ. In the next, we will use the index p for all the quantities calculated with respect to the design p; for example, E_p is the p-expectation and ξ for the model.

In the new frame, we call *statistic* a function $T = T(\mathcal{D})$ where $\mathcal{D} = \{(k, Y_k) : k \in S\}$, S being a random variable with values in $\mathcal{S}$ the set of all possible sample s. So, the function T, for any given value s of S, depends on those Y_k for which $k \in s$. The statistic T for $S = s$ used for making inference about the population mean $\overline{Y} = N^{-1} \sum_{\mathcal{U}} Y_k$ is called *predictor* and T for $Y_k = y_k$ is called an *estimate* for $\overline{y} = N^{-1} \sum_{\mathcal{U}} y_k$.

Definition 2 *:*

1. *T is called p-unbiased for $\overline{Y}$ if and only if, for a given design p, $E_p(T(\mathbf{y})) = \overline{y}$ for all $\mathbf{y} = (y_1, \ldots, y_N)$.*
2. *T is called ξ-unbiased for $\overline{Y}$ if and only if, for a given design ξ, $E_\xi(T(\overline{Y}) - \overline{Y}) = 0$ for all $s \in \mathcal{S}$.*
3. *T is called $p\xi$-unbiased for $\overline{Y}$ if and only if, for given p and ξ $E_\xi E_p(T(\overline{Y}) - \overline{Y}) = 0$.*

For a *superpopulation model* approach, the choice of a strategy (p, T) will be dictated by the objective to minimise the ξ-expected p-MSE,

$$E_\xi \mathrm{MSE}_p(p, T) = E_\xi E_p(T - \overline{Y})^2.$$

Although the objective is the same, there are two different ways of obtaining the desired minimum called the *model-based approach* and the *design-based approach*. For the first one (Brewer 1963, Royall 1970, 1971), the sampling design is of minor importance. The objective is to choose T such that for every given sample s, T minimizes $E_\xi(T - \overline{Y})^2$. We will not develop here this approach but we mention its application in repeated surveys (Blight & Scott 1973, Scott & Smith 1974). For the second one (Cassel *et al.* 1976, Särndal 1980, Särndal *et al.* 1989), the support is on the sampling design p. We look for an estimate T of $\overline{y}$ such that T minimizes $E_p(T - \overline{y})^2$. We give a brief presentation of the main results in the case of the *design-based approach*.

Finally, we consider only the *noninformative* designs fact which allows us to interchange the order in calculating the expectations with respect to p and ξ.

Cassel, Särndal & Wretman (1976) introduce the p-unbiased *generalized difference estimator*

$$T_{GD} = \sum_s \frac{Y_k - e_k}{N\pi_k} + \sum_{\mathcal{U}} \frac{e_k}{N}$$

for an arbitrary vector $\mathbf{e} = (e_1, \ldots, e_N)$; T_{GD} is ξ-unbiased if $e_k = \mu_k$, for ξ defined as follows (Cassel, Särndal & Wretman 1976)

$$\xi : \begin{cases} E_\xi(Y_k) = \mu a_k + b_k = \mu_k; \\ E_\xi(Y_k - \mu_k)^2 = a_k^2 \sigma^2 = \sigma_k^2; \\ E_\xi\{(Y_k - \mu_k)(Y_l - \mu_l)\} = a_k a_l \rho \sigma^2 = \sigma_{kl} \text{ for } k \neq l; \end{cases}$$

where $a_k > 0, b_k$ for $k = 1, \ldots, N$ are known numbers with $\sum_{k=1}^N a_k = N$ and μ, σ^2 and ρ are unknown and $\frac{-1}{N-1} \leq \rho \leq 1$.

We have the following optimality result:

Result 18 *: (Cassel, Särndal and Wretman 1976)*
Under the above model, the optimal strategy (p, T) with p a fixed size design of size n and T a p-unbiased linear estimator of $\overline{Y}$ is $(p_0 : T_{GD0})$, where

$$T_{GD0} = \sum_s \frac{Y_k - b_k}{na_k} + \sum_{\mathcal{U}} \frac{b_k}{N}$$

and $p_0 = p_0(s)$ is the sampling design with the inclusion probabilities $\pi_k = fa_k$, $f = \frac{n}{N}$.

The estimator T_{GD0} is also ξ and $p\xi$-unbiased. The HT estimator, $\hat{t}_{y\pi}$ belongs to the class T_{GD} if $\mathbf{e} = 0$ or if $e_k = \pi_k$ for p a fixed size design. From the above result, we have in general $E_\xi V_p(p, \hat{t}_{y\pi}) \geq E_\xi V_p(p_0, T_{GD0})$ with equality for $p = p_0$ and $b_k \propto a_k$.

We consider in the next a particular model ξ (Särndal, 1980). Let us consider $\mathcal{X}_1, \ldots, \mathcal{X}_q$ auxiliary variables with $\mathbf{x}'_\mathbf{k} = (x_{k1}, \ldots, x_{kq})$ and that the variables $Y_1, \ldots, Y_N$ are independent. For the regression model:

$$E_\xi(Y_k) = \mathbf{x}'_\mathbf{k}\boldsymbol{\beta} = \mu_k$$
$$V_\xi(Y_k) = \sigma_k^2 = \sigma^2 v_k$$

where $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_q)$ and σ^2 are unknown, $v_k = v(\mathbf{x_k})$ is a known function for all k in U, T_{GD0} becomes:

$$\begin{aligned} T_{GR} &= \sum_s \frac{Y_k}{N\pi_k} + \sum_{j=1}^{q} \beta j \left(\frac{1}{N} \sum_U x_{kj} - \sum_s \frac{x_{kj}}{N\pi_k} \right) \\ &= N^{-1} \left[\hat{t}_{y\pi} + (t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi})'\boldsymbol{\beta} \right] \end{aligned}$$

with $\hat{t}_{y\pi} = \sum_s \frac{Y_k}{\pi_k}$ and $\hat{t}_{\mathbf{x}\pi} = \sum_s \frac{\mathbf{x}_k}{\pi_k}$. For the auxiliary information, we need only to know $t_\mathbf{x} = \sum_\mathcal{U} \mathbf{x_k}$. If we suppose that all β_j are known, then the above optimality result give that T_{GR} is optimal for $\pi_k \propto v_k$ for all k.

The estimator T_{GR} can be viewed as a correction of the ξ-model biased but p-unbiased estimator $N^{-1}\hat{t}_{y\pi}$,

$$T_{GR} = N^{-1}\hat{t}_{y\pi} - B_\xi(N^{-1}\hat{t}_{y\pi})$$

where $B_\xi(N^{-1}\hat{t}_{y\pi}) = N^{-1}(t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi})'\boldsymbol{\beta}$ is the ξ-bias of $N^{-1}\hat{t}_{y\pi}$ (Thompson 1997).

More realistic, β_j are unknown, for all $j = 1, \ldots, q$. Särndal (1980) proposes to estimate the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ by design-based $\hat{\boldsymbol{\beta}}_s$ as follows

$$\hat{\boldsymbol{\beta}}_\boldsymbol{s} = \boldsymbol{G}'_\boldsymbol{s}\boldsymbol{Y}_\boldsymbol{s} = \left(\boldsymbol{W}'_\boldsymbol{s}\boldsymbol{X}_\boldsymbol{s}\right)^{-\mathbf{1}} \boldsymbol{W}'_\boldsymbol{s}\boldsymbol{Y}_\boldsymbol{s}$$

where $\mathbf{W_s} = (w_{kj})_{k=\overline{i,n}, j=\overline{i,q}}$ and the w_{kj} may or may not depend on the known quantities $\mathbf{x}'_\mathbf{k}$ and v_k, $\mathbf{X}_s = (\mathbf{x}'_\mathbf{k})_{k\in s}$ and $\mathbf{Y}_s = (Y_k)_{k\in s}$. It results that the vector $(w_{k1}, \ldots, w_{kq})$ is the vector of weights applied to unit k. The resulting estimator is the so-called *generalized regression estimator*:

$$\hat{T}_{GR} = N^{-1} \left[\hat{t}_{y\pi} + (t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi})'\hat{\boldsymbol{\beta}}_s \right]. \tag{7.1}$$

This estimator can be obtained without implying the variance structure of the model and it has the following properties:

Result 19 *(Särndal 1980)*
1. The estimator $\hat{\boldsymbol{\beta}}_s$ is ξ-unbiased for $\boldsymbol{\beta}$:

$$E_\xi(\hat{\boldsymbol{\beta}}_s) = \boldsymbol{\beta}.$$

2. $\hat{T}_{GR}$ with $\boldsymbol{\beta}_s$ given by (7.1) is model unbiased under the ξ model.

Thus we can use the model ξ to obtain ξ-unbiased estimators for $\boldsymbol{\beta}_s$ but the basic properties are not dependent on whether the model ξ holds or not. It implies that we have a model-assisted and not a model-dependent estimator.

If T_{GR} was p-unbiased, $\hat{T}_{GR}$ loses this property because $\hat{\boldsymbol{\beta}}_s$ as a nonlinear function is not p-unbiased for $\boldsymbol{\beta}$. Särndal (1980) states that an exact design-unbiasedness can entail loss in efficiency and he proposes the asymptotic design-unbiasedness (ADU) and consistency of $\hat{T}_{GR}$ as minimum requirements. Robinson & Särndal (1983) gives conditions for which $\hat{T}_{GR}$ is ADU and consistent for $\overline{Y}$. At the same time, they give an approximation for the design mean squared of order $O(n^{-1})$. These conditions do not require the superpopulation model to be true, but optimality is reached if the model is true. Robinson & Särndal (1983) state that in the case of perfect correctness of the model, the mean square error is minimized for $p(s)$ with the inclusion probabilities $\pi_k \propto \sigma_k$.

Remark 12 *There is a particular case when $\hat{T}_{GR}$ is exactly p-unbiased, namely for a model ξ for which $E_\xi(Y_k) = \beta x_k$ and a sampling design p with $\pi_k \propto x_k$. In this case, $\hat{T}_{GR} = N^{-1}\hat{t}_{y\pi}$.*

We return to the expression of $\hat{\boldsymbol{\beta}}_s = \boldsymbol{G}'_s\boldsymbol{Y}_s = (\boldsymbol{W}'_s\boldsymbol{X}_s)^{-1}\boldsymbol{W}'_s\boldsymbol{Y}_s$. Särndal (1980) discusses two particular choices for the matrix $\boldsymbol{W}_s$: π-inverse weighting and best linear unbiased weighting.

1. The π-inverse weighting is obtained for weights $\boldsymbol{W}_s$ such that for some vector $\mathbf{c} = (c_1, \ldots, c_q)'$ we have

   $$\mathbf{1}'_s\boldsymbol{\Pi}_s^{-1} = \mathbf{c}'\boldsymbol{W}'_s$$

   for $\boldsymbol{\Pi}_s = \mathrm{diag}(\pi_k)_{k\in s}$ and $\mathbf{1}_s$ the column vector composed of n ones. In this situation the estimator $\hat{T}_{GR}$ given by (7.1) has the appealing form

   $$\hat{T}_{GR} = N^{-1}t'_{\mathbf{x}}\hat{\boldsymbol{\beta}}_s$$

   since

   $$\hat{t}_{y\pi} - \hat{t}'_{\mathbf{x}\pi}\hat{\boldsymbol{\beta}}_s \;=\; \mathbf{1}'_s\boldsymbol{\Pi}_s^{-1}(\boldsymbol{Y}_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}}_s) = \mathbf{c}'\boldsymbol{W}'_s(\boldsymbol{Y}_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}}_s) = 0$$

   with $\hat{\boldsymbol{\beta}}_s = (\boldsymbol{W}'_s\boldsymbol{X}_s)^{-1}\boldsymbol{W}'_s\boldsymbol{Y}_s$.
   Särndal gives two examples of $\boldsymbol{W_s}$ satisfying the relation (7.2).

   - The first one is

     $$\boldsymbol{W_s} = \boldsymbol{\Pi_s^{-1}X_s}$$

     and the model ξ includes the intercept. In this case, $\hat{\boldsymbol{\beta}}_{s,1} = \boldsymbol{(X'_s\Pi_s^{-1}X_s)^{-1}X'_s\Pi_s^{-1}Y_s}$.

- Another possibility is

$$\boldsymbol{W}_s = \boldsymbol{\Pi}_s^{-1}\boldsymbol{V}_s^{-1}\boldsymbol{X}_s$$

and the model variance $\mathbf{V}_s = \text{diag}(v_k)_{k\in s}$ satisfies $v_k = \mathbf{c}'\mathbf{x_k}$. In this case

$$\hat{\boldsymbol{\beta}}_{GREG} = (\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s'\boldsymbol{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\boldsymbol{Y}_s.$$

With this expression for $\hat{\boldsymbol{\beta}}_{GREG}$, $\hat{T}_{GR}$ becomes $\hat{T}_{GREG}$ discussed by Särndal, Swensson and Wretman (1989). Särndal, Swensson and Wretman (1992) obtain $\hat{T}_{GREG}$ through a different way and they analyze it for a general model variance v_k.

2. The best linear unbiased weighting is obtained for $\boldsymbol{W}_s = \boldsymbol{V}_s^{-1}\boldsymbol{X}_s$ which gives for $\boldsymbol{\beta}$ the estimator

$$\hat{\boldsymbol{\beta}}_{BLU} = \boldsymbol{(X_s'V_s^{-1}X_s)^{-1}X_s'V_s^{-1}Y_s}.$$

When inserted into (7.1) this gives the special case of $\hat{T}_{GR}$ to be denoted $\hat{T}_{BLU}$. This weighting merits attention because it is, for many statisticians, the natural way to estimate under the model ξ.

Montanari (1987) derives the expression for the coefficient of regression β who minimizes the p-variance of

$$T_{GR} = N^{-1}\left[\hat{t}_{y\pi} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi})'\boldsymbol{\beta}\right]$$

as $\boldsymbol{\beta}_{opt} = \left[\text{Var}\left(\sum_U \frac{\mathbf{x_k}}{\pi_k}\right)\right]^{-1} \text{Cov}\left(\sum_U \frac{\mathbf{x_k}}{\pi_k}, \sum_U \frac{Y_k}{\pi_k}\right)$. Unfortunately, the expression of $\boldsymbol{\beta}_{opt}$ depends on unknown quantities making so impossible its derivation.

At this point of discussion a question arises: which of the two estimator presented here, $\hat{T}_{GREG}$ and $\hat{T}_{BLU}$, is preferable? $\hat{T}_{BLU}$ being obtained for the best linear unbiased estimator for $\boldsymbol{\beta}$ is preferred by many statisticians while survey practitioners, such as Särndal *et al.* (1992, pp 519), argue for $\hat{T}_{GREG}$. Their argument is that $\hat{T}_{GREG}$ is obtained for the sample-weighted estimator $\hat{\boldsymbol{\beta}}_{GREG}$ which is more robust than the unweighted $\hat{\boldsymbol{\beta}}_{BLU}$, namely $\hat{\boldsymbol{\beta}}_{GREG}$ remains design-consistent even if the model is wrong. Besides, $\hat{\boldsymbol{\beta}}_{GREG}$ is the solution of the optimal sample estimating function (Godambe & Thompson 1986, Godambe 1995):

$$\sum_s \frac{1}{\sigma_k^2\pi_k}\mathbf{x_k}(Y_k - \mathbf{x_k'}\boldsymbol{\beta}) \quad \text{for}$$

$$\sigma_k^2 = \sigma^2 v_k.$$

The regression coefficient $\hat{\boldsymbol{\beta}}_{GREG}$ can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GREG} &= \boldsymbol{(X_s'V_s^{-1}\Pi_s^{-1}X_s)^{-1}X_s'V_s^{-1}\Pi_s^{-1}Y_s} \\ &= \left(\sum_s \frac{\mathbf{x_k}\mathbf{x_k'}}{\sigma_k^2\pi_k}\right)^{-1}\sum_s \frac{\mathbf{x_k}Y_k}{\sigma_k^2\pi_k}\end{aligned}$$

The same formula for $\hat{\boldsymbol{\beta}}_{GREG}$ would have been obtained if we had used the substitution method described in chapter concerning Taylor linearization method for estimating the coefficient of regression

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X_N'V_N^{-1}X_N)^{-1}X_N'V_N^{-1}Y_N} \tag{7.2}$$

(Särndal *et al.* 1992, pp 193). The estimator $\hat{\boldsymbol{\beta}}_{GREG}$ is not p-unbiased for $\boldsymbol{\beta}$ but is approximately p-unbiased and ξ-unbiased, $E_\xi(\hat{\boldsymbol{\beta}}_{GREG}) = \boldsymbol{\beta}$. Under the conditions of consistency given by Fuller & Isaki (1982), Fuller (2002) $\hat{\boldsymbol{\beta}}_{GREG}$ is consistent for $\boldsymbol{\beta}$. More precisely, we have $\hat{\boldsymbol{\beta}}_{GREG} = \boldsymbol{\beta} + O_p(n^{-1/2})$.
Särndal, Swensson and Wretman (1989) give equivalent expressions for $\hat{T}_{GREG}$. Let's introduce the following notations:

- the predicted value for the k-th element is denoted by $\hat{Y}_k = \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_s$,
- $e_{ks} = Y_k - \hat{Y}_k$ is the k-th regression residual,
- $E_k = Y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}$ is the population fit residual, with $\hat{\boldsymbol{\beta}}$ given by (7.2) and the solution of the normal equation,
$$\sum_{\mathcal{U}} \frac{1}{\sigma_k^2}\mathbf{x_k}(Y_k - \mathbf{x}_\mathbf{k}'\boldsymbol{\beta}) = 0$$
- $g_{ks} = 1 + \left(t_x - \hat{t}_{\mathbf{x}\pi}\right)' \hat{T}^{-1}\frac{\mathbf{x_k}}{\sigma_k^2}$ where $\hat{T} = \sum_s \frac{\mathbf{x_k}\mathbf{x_k}'}{\sigma_k^2\pi_k}$ and we suppose that $\hat{T}^{-1}$ exists. The quantities g_{ks} are known in the literature as the g-weights and they were introduced by Särndal *et al.* (1989).

Result 20 *The generalized regression estimator $\hat{T}_{GREG}$ can be written in the following equivalent expressions*

$$\begin{aligned}
\hat{T}_{GR} &= N^{-1}\left[\hat{t}_{y\pi} + (t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi})'\hat{\boldsymbol{\beta}}_{GREG}\right] && (7.3)\\
&= \sum_s g_{ks}\frac{y_k}{\pi_k} && (7.4)\\
&= \sum_U \hat{Y}_k + \sum_s \frac{e_{ks}}{\pi_k} && (7.5)\\
&= \sum_U \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}} + \sum_s g_{ks}\frac{E_k}{\pi_k} && (7.6)
\end{aligned}$$

In the particular case of a model ξ with $v_k = c'\mathbf{x_k}$, for all $k \in \mathcal{U}$, Särndal *et al.* (1992) prove that the regression residuals e_{ks} have the property

$$\sum_s \frac{e_{ks}}{\pi_k} = 0$$

and Thompson (1997) proves that under the same model, the population fit residualsE_k, for $k \in U$ satisfy

$$\sum_U E_k = 0.$$

Then, the regression estimator $\hat{T}_{GREG}$ becomes

$$\hat{T}_{GREG} = N^{-1}\sum_{\mathcal{U}} \hat{Y}_k = N^{-1}\sum_{\mathcal{U}} \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_s = N^{-1}\sum_s g_{ks}\frac{Y_k}{\pi_k};$$

where $g_{ks} = (t_x)'\hat{T}^{-1}\frac{\mathbf{x_k}}{\sigma_k^2}$.

Result 21 *(Särndal et al. 1992) The generalized regression estimator $\hat{T}_{GREG}$ is approximated by:*

1.

$$\hat{T}_{GREG0} = N^{-1}\left\{\hat{t}_{y\pi} + \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi}\right)'\hat{\boldsymbol{\beta}}\right\} = N^{-1}\left\{\sum_{\mathcal{U}} \mathbf{x}'_{\mathbf{k}}\hat{\boldsymbol{\beta}} + \sum_{s}\check{E}_k\right\}$$

2. The approximate variance is given by :

$$V(\hat{T}_{GREG}) \simeq N^{-2}\sum_{\mathcal{U}}\sum_{\mathcal{U}}\Delta_{kl}\check{E}_k\check{E}_l;$$

3. the variance estimator is given by:

$$\hat{V}_g(\hat{T}_{GREG}) = N^{-2}\sum_{s}\sum_{s}\check{\Delta}_{kl}\left(g_{ks}\check{e}_{ks}\right)\left(g_{ls}\check{e}_{ls}\right).$$

From the expression of the approximate variance of $\hat{T}_{GREG}$, we can give another variance estimator, replacing E_k by its sample-based counterpart e_{ks}:

$$\hat{V}_1 \simeq \sum\sum_{s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{e_{ks}}{\pi_k}\frac{e_{ls}}{\pi_l}.$$

When s is of fixed size, the Yates-Grundy variance estimator $\hat{V}_1$ is:

$$\hat{V}_{YG} = -\frac{1}{2}\sum\sum_{s}\frac{\Delta_{kl}}{\pi_{kl}}\left[\frac{e_{ks}}{\pi_k} - \frac{e_{ks}}{\pi_k}\right]^2.$$

For any given model, the steps involved in deriving the regression estimator are summarized as follows:

1. Derive $\hat{\boldsymbol{\beta}}_s$, $\hat{Y}_k$ and e_{ks} to find $\hat{T}_{GREG}$.
2. Identify the g_{ks} needed to obtain the variance estimator $\hat{V}(\hat{T}_{GREG})$.
3. Find $\hat{\boldsymbol{\beta}}$ and E_k which are required for the approximate variance.

We consider below two particular models and we derive the regression estimator for a general sampling design. Then, we discuss the properties of the estimator for specified sampling design.

7.1.1 The common ratio model and the ratio estimator

A regression model assuming that y_k/x_k is constant is called *a common ratio model* or simply *ratio model* :

$$E_\xi(Y_k) = \beta x_k$$

$$V_\xi(Y_k) = \sigma_k^2 = \sigma^2 x_k$$

where the parameters β and σ are unknown. This regression is a straight line through the origin.

Result 22 *The regression estimator for the total generated by the ratio model is the ratio estimator*

$$\hat{t}_{GREG} = (\sum_U x_k)\frac{\sum_s y_k/\pi_k}{\sum_s x_k/\pi_k} = (\sum_U x_k)\hat{B}. \tag{7.7}$$

The approximative variance is obtained from result (21) by setting

$$E_k = y_k - Bx_k$$

where $B = \sum_U y_k / \sum_U x_k$ *and the variance estimator is obtained by setting*

$$e_{ks} = y_k - \hat{B}x_k$$

and for all $k \in s$,

$$g_{ks} = \frac{\sum_U x_k}{\sum_s \frac{x_k}{\pi_k}}.$$

Proof □

The ratio estimator under SI sampling

Under the SI design, the estimator given by (7.7) becomes

$$\hat{t}_{ratio} = (\sum_U x_k)\frac{\sum_s y_k}{\sum_s x_k} = N\overline{x}_U\frac{\overline{y}_s}{\overline{x}_s}.$$

We use result (22). We have

$$\hat{B} = \frac{\sum_s y_k/\pi_k}{\sum_s x_k/\pi_k} = \frac{\overline{y}_s}{\overline{x}_s}$$

and the approximative variance of $\hat{t}_{ratio}$ is the variance of the Horvitz-Thompson estimator

$$E_k = y_k - Bx_k.$$

We have

$$AV(\hat{t}_{ratio}) = N^2\frac{1-f}{n}S^2_{EU} \quad \text{with}$$

$$\begin{aligned} S^2_{EU} &= \frac{1}{N-1}\sum_U E_k^2 = \frac{1}{N-1}\sum_U (y_k - Bx_k)^2 \\ &= S^2_{yU} + B^2S^2_{xU} - 2BS_{xyU} \end{aligned}$$

since $\sum_U E_k/N = \sum_U (y_k - Bx_k)/N = 0$. In order to calculate the variance estimator, we derive

$$g_{ks} = \frac{\sum_U x_k}{\sum_s x_k/\pi_k} = \frac{\overline{x}_U}{\overline{x}_s}$$

and

$$e_{ks} = y_k - \hat{B}x_k.$$

The variance estimator is

$$\begin{aligned}\hat{V}(\hat{t}_{ratio}) &= \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_{ks}e_{ks}}{\pi_k} \frac{g_{ls}e_{ls}}{\pi_l} = \left(\frac{\overline{x}_U}{\overline{x}_s}\right)^2 N^2 \frac{1-f}{n} S_{es}^2 \\ &= \left(\frac{\overline{x}_U}{\overline{x}_s}\right)^2 N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_s (y_k - \hat{B}x_k)^2\end{aligned}$$

where

$$\begin{aligned}S_{es}^2 &= \frac{1}{n-1}\sum_s (e_{ks} - \overline{e}_s)^2 = \frac{1}{n-1}\sum_s e_{ks}^2 \\ &= S_{ys}^2 + \hat{B}^2 S_{xs}^2 - 2\hat{B}S_{xys}\end{aligned}$$

since $\overline{e}_s = 0$.

7.1.2 The common mean model

In many populations where a strong linear relationship exists between the study variable $\mathcal{Y}$ and a single auxiliary variable $\mathcal{X}$, the population regression line will intercept the $\mathcal{Y}$ axis at some distance from the origin. A model with an intercept will give a better regression estimator than the common ratio model discussed in the above section.
The *simple regression model* states that for $k \in U$,

$$E_\xi(Y_k) = \beta_1 + \beta_2 x_k \tag{7.8}$$
$$V_\xi(Y_k) = \sigma^2 \tag{7.9}$$

where β_1, β_2 and σ are unknown parameters. We have $\boldsymbol{x}_k' = (1, x_k)$, $\boldsymbol{\Sigma} = diag(\sigma^2)$ and the parameters β_1 and β_2 are estimated under the model ξ by

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (\boldsymbol{X}_N'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_N$$

where

$$\boldsymbol{X}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_N \end{pmatrix}$$

We have

$$(\boldsymbol{X}_N'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_N)^{-1} = \frac{\sigma^2}{N\sum_U (x_k - \overline{x}_U)^2} \begin{pmatrix} \sum_U x_k^2 & -t_x \\ -t_x & N \end{pmatrix}$$

and

$$\boldsymbol{X}_N'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_N = \frac{1}{\sigma^2} \begin{pmatrix} t_y \\ \sum_U x_k y_k \end{pmatrix}$$

which give

$$\hat{\beta}_1 = \overline{y}_U - \hat{\beta}_2 \overline{x}_U$$
$$\hat{\beta}_2 = \frac{\sum_U (x_k - \overline{x}_U)(y_k - \overline{y}_U)}{\sum_U (x_k - \overline{x}_U)^2}$$

These parameters are estimated using the sampling design p by

$$\begin{pmatrix} \hat{\beta}_{1s} \\ \hat{\beta}_{2s} \end{pmatrix} = \begin{pmatrix} \tilde{y}_s - \hat{\beta}_{2s}\tilde{x}_s \\ \frac{\sum_s (x_k - \tilde{x}_s)(y_k - \tilde{y}_s)/\pi_k}{\sum_s (x_k - \tilde{x}_s)^2/\pi_k} \end{pmatrix}$$

for $\tilde{y}_s = \frac{\sum_s y_k/\pi_k}{\hat{N}}$, $\tilde{x}_s = \frac{\sum_s x_k/\pi_k}{\hat{N}}$ and $\hat{N} = \sum_s \frac{1}{\pi_k}$.

Result 23 *The regression estimator of the total is*

$$\hat{t}_{yr} = N\left[\tilde{y}_s + \hat{\beta}_{2s}(\overline{x}_U - \tilde{x}_s)\right].$$

The approximative variance is obtained for

$$E_k = y_k - \overline{y}_U - \hat{\beta}_2.$$

The variance estimator is obtained for

$$e_{ks} = y_k - \tilde{y}_s - \hat{\beta}_{2s}(x_k - \tilde{x}_s) \quad \textit{and}$$

$$g_{ks} = \frac{N}{\hat{N}}\left[1 + (x_k - \tilde{x}_s)\frac{\overline{x}_U - \tilde{x}_s}{\tilde{S}^2_{xs}}\right].$$

7.2 Calibration technique

Until now, no use of auxiliary information was made. Or, in sample survey auxiliary information on the finite population is often used to improve the precision of estimators of the population total. Several approaches are conceivable. We have *the calibration approach* described below, which does not rely explicitly on a model with the more recent *the model-calibration approach*, or *the model-assisted approach* described in the next section and when the inference is based upon a model of superpopulation taking into account at the same time the sampling design. We present in the next the principles of *the calibration approach* developed by Deville & Särndal (1992) and Deville, Särndal & Sautory (1993) and the implementation of this method in Calmar.

The objective is the estimation of the population total of the variable of interest $\mathcal{Y}$ denoted $t_y = \sum_{\mathcal{U}} y_k$ in the presence of univariate or multivariate auxiliary information for which the only request is that we know its population total, namely we do not need to know the value taken by an auxiliary variable for all the units in the population.

Let $\mathcal{X}_1, \ldots, \mathcal{X}_q$ be q auxiliary variables and for $k \in \mathcal{U}$ and let $\mathbf{x_k} = (x_{k1}, \ldots, x_{kq})'$ be the q-vector with the values of the auxiliary variables for the k-th element in the population. We suppose that the total $t_{\mathbf{x}} = \sum_U \mathbf{x_k}$ is known; the vector $(\mathbf{x_k}, y_k)'$ is observed for all $k \in s$. Let $\hat{t}_{y\pi} = \sum_s \frac{y_k}{\pi_k}$ be the π-estimator of t_y and we note with $d_k = \frac{1}{\pi_k}$ the π-weight corresponding to y_k, for all k in s.

The calibration technique consists in finding a new set of weights $\{w_k\}_{k\in s}$ which satisfies the conditions :

1. $\{w_k\}_{k\in s}$ are as close as possible to $\{d_k\}_{k\in s}$ in the sense of a distance between w_k and d_k.

2. $\{w_k\}_{k\in s}$ satisfy the calibration equations:

$$\sum_s w_k \mathbf{x_k} = t_\mathbf{x}$$

which means that the new weights must estimate well the auxiliary information.

The calibrated estimator is denoted by $\hat{t}_{yw}$, so that it is related with the w_k weights. We consider a function distance $G_k(w,d)$ such that :

1. for every fixed $d > 0, G_k(w,d) > 0$, differentiable with respect to w , strictly convex, defined on an interval $D_k(d)$ such that $d \in D_k(d)$;
2. $G_k(d,d) = 0$;
3. $g_k(w,d) = \dfrac{\partial G_k(w,d)}{\partial w}$ is continuous and the function that transforms the interval $D_k(d)$ in $Im_k(d)$ is one-to-one fashion.

The request that w_k would be as close as possible to d_k is equivalent to minimize the average distance $E_p\{\sum_s G_k(w_k,d_k)\}$. Deville & Särndal (1992) apply the Lagrange multipliers method which leads to the following weights, called *calibration weights*:

$$w_k = d_k F_k(\mathbf{x_k'}\boldsymbol{\lambda})$$

where $F_k(0) = 1, q_k = F_k'(0) > 0$, $d_k F_k$ is the reciprocal mapping of $g_k(\cdot, d_k)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_j, \ldots, \lambda_q)'$ is the vector of Lagrange multipliers. We determine $\boldsymbol{\lambda}$ from the calibration equations :

$$t_\mathbf{x} = \sum_s w_k \mathbf{x_k} = \sum_s d_k F_k(\mathbf{x_k'}\boldsymbol{\lambda})\mathbf{x_k}.$$

Deville & Särndal (1992) suppose conditions which ensure that the above equation has a unique solution belonging to $C = \cap_{k\in\mathcal{U}}\{\boldsymbol{\lambda} : \mathbf{x_k'}\boldsymbol{\lambda} \in \mathrm{Im}_k(d_k)\}$ with a probability tending to one. With $\boldsymbol{\lambda}$ determined, we can write the calibration estimator for t_y :

$$\hat{t}_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x_k'}\boldsymbol{\lambda}) y_k.$$

Several remarks on the derivation of the calibration estimator must be made:

1. The vector $\boldsymbol{\lambda}$ is determined solving the calibration system:

$$\phi_s(\boldsymbol{\lambda}) = t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi} \text{ where}$$

$$\phi_s(\boldsymbol{\lambda}) = \sum_s d_k\{F_k(\mathbf{x_k}'\boldsymbol{\lambda}) - 1\}\mathbf{x_k}.$$

Deville & Särndal (1992) propose Newton's algorithm for obtaining $\boldsymbol{\lambda}$, assuring that this method converge quickly.

2. Different choices of the distance function lead to different estimators. The most important case is $F_k(u) = 1 + q_k u$ when we obtain the generalized regression estimator:

$$\hat{t}_{yreg} = \sum_s w_k y_k = \hat{t}_{y\pi} + (t_\mathbf{x} - \hat{t}_{\mathbf{x}\pi})'\hat{\boldsymbol{\beta}}_s$$

where $\hat{t}_{\mathbf{x}\pi} = \sum_s d_k \mathbf{x_k}$ (respectively $\hat{t}_{y\pi}$) is the π-estimator for the total of $\mathbf{x_k}$ (respectively of $\mathcal{Y}$) and

$$\hat{\boldsymbol{\beta}}_s = \left(\sum_s d_k q_k \mathbf{x_k} \mathbf{x_k}' \right)^{-1} \sum_s d_k q_k \mathbf{x_k} y_k.$$

Focusing on the weights, the calibration technique leads to the same generalized regression estimator obtained by Särndal (1980) when a regression of $\mathcal{Y}$ through the $\mathcal{X}_1, \ldots, \mathcal{X}_q$ was considered. So, we have two different approach, the calibration technique and the regression, which leads to the same estimator.

3. With the assumption that a solution $\boldsymbol{\lambda}$ exists, the different choices for F_k can lead to negative weights, which are not desired for an estimation of the variance. Deville & Särndal (1992) modify properly the function F_k such that the resulting weights are positive.

Suppose now that $n, N \to \infty$ and :

- (C1) $\lim N^{-1} t_{\mathbf{x}} < \infty$;
- (C2) $N^{-1}(\hat{t}_{\mathbf{x}\pi} - t_{\mathbf{x}}) \to 0$ in design probability ;
- (C3) $n^{\frac{1}{2}} N^{-1}(\hat{t}_{\mathbf{x}\pi} - t_{\mathbf{x}})$ converges in distribution to $N(0, A)$;
- (C4) $\max ||\mathbf{x_k}|| = M < \infty$;
- (C5) $\max F_k''(0) = M' < \infty$.

We have the following result:

Result 24 *(Deville and Särndal 1992): Under the supposed conditions, $\hat{t}_{yw}$ has the following properties :*

1. *$\hat{t}_{yw}$ is design-consistent and at least asymptotically design-unbiased (ADU)*

$$N^{-1}(\hat{t}_{yw} - \hat{t}_{y\pi}) = O_p(n^{-\frac{1}{2}})$$

   if there exists a solution $\boldsymbol{\lambda}$ of the calibration equations.

2. *$\hat{t}_{yw}$ is asymptotically equivalent to the regression estimator $\hat{t}_{yreg}$ for any F_k that satisfies the condition (C5) from above.*

$$N^{-1}(\hat{t}_{yw} - \hat{t}_{yreg}) = O_p(n^{-1})$$

   and consequently $V(\hat{t}_{yw}) \simeq V(\hat{t}_{yreg})$.

From the second point, it results that the choice of F_k is not of great importance for the derivation of the variance of $\hat{t}_{yw}$ because all the estimators are asymptotically equivalent with the regression estimator. This result has important consequences on the derivation of the variance and variance estimation of $\hat{t}_{yw}$. We have :

$$V(\hat{t}_{yw}) \simeq V(\hat{t}_{reg}) = \sum_{\mathcal{U}} \sum_{\mathcal{U}} \Delta_{kl} (d_k E_k)(d_l E_l)$$

where $E_k = y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = (\sum_{\mathcal{U}} q_k \mathbf{x}_\mathbf{k}\mathbf{x}_\mathbf{k}')^{-1} \sum_{\mathcal{U}} q_k \mathbf{x}_\mathbf{k} y_k$. To estimate the variance, we need an estimator for $\hat{\boldsymbol{\beta}}$, which is given by :

$$\hat{\boldsymbol{\beta}}_{ws} = \left(\sum_s w_k q_k \mathbf{x}_\mathbf{k}\mathbf{x}_\mathbf{k}'\right)^{-1} \left(\sum_s w_k q_k \mathbf{x}_\mathbf{k} y_k\right)$$

and then

$$\hat{V}(\hat{t}_{ws}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_k)(w_l e_l)$$

where $e_k = y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}_{ws}$.

In conclusion, we can summarize the following:

1. For a sample s and for chosen F_k, we solve the calibration equation for obtaining $\boldsymbol{\lambda}$.
2. When $\boldsymbol{\lambda}$ is determined, we derive the calibration estimator:

$$\hat{t}_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x}_\mathbf{k}'\boldsymbol{\lambda}) y_k.$$

3. The variance estimate is equal to the variance estimate for the regression estimator, with the residuals e_k of $\mathcal{Y}$ on the calibrated variables:

$$\hat{V}(\hat{t}_{ws}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_k)(w_l e_l)$$

7.2.1 Calibration method with CALMAR

The emphasis in this section is on improving estimates in the presence of auxiliary information by using regression models. In this case, the only requirement about the auxiliary information is that the population total must be known. When the value of an auxiliary variable for each unit in the population is known, more complex models may be used. A model of regression as studied above will improve our estimate if it reduces its variance. This is achieved if the population fit residuals $E_k = y_k - \mathbf{x}_\mathbf{k}'\hat{\boldsymbol{\beta}}$ are small, namely that it exists a strong linear relationship between the variable of interest and the auxiliary variable. On the contrary case, the variance could be large. This justifies the use of more general models, as the nonparametric ones. We study in more details this situation in the following.