

Padilla Terán, Alberto Manuel

**Working Paper**

## Variance estimator in complex surveys using linear regression with expansion factor as independent variable

Working Papers, No. 2017-07

**Provided in Cooperation with:**

Bank of Mexico, Mexico City

*Suggested Citation:* Padilla Terán, Alberto Manuel (2017) : Variance estimator in complex surveys using linear regression with expansion factor as independent variable, Working Papers, No. 2017-07, Banco de México, Ciudad de México

This Version is available at:

<https://hdl.handle.net/10419/174460>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

Banco de México  
Documentos de Investigación

Banco de México  
Working Papers

N° 2017-07

Variance Estimator in Complex Surveys using Linear  
Regression with Expansion Factor as Independent  
Variable

Alberto Manuel Padilla Terán  
Banco de México

June 2017

La serie de Documentos de Investigación del Banco de México divulga resultados preliminares de trabajos de investigación económica realizados en el Banco de México con la finalidad de propiciar el intercambio y debate de ideas. El contenido de los Documentos de Investigación, así como las conclusiones que de ellos se derivan, son responsabilidad exclusiva de los autores y no reflejan necesariamente las del Banco de México.

The Working Papers series of Banco de México disseminates preliminary results of economic research conducted at Banco de México in order to promote the exchange and debate of ideas. The views and conclusions presented in the Working Papers are exclusively the responsibility of the authors and do not necessarily reflect those of Banco de México.

# Variance Estimator in Complex Surveys using Linear Regression with Expansion Factor as Independent Variable\*

Alberto Manuel Padilla Terán<sup>†</sup>  
Banco de México

**Abstract:** In probability sampling, variance estimation of an estimated mean or total requires developing a mathematical expression that depends on the design used to extract a sample. These formulae can be difficult to build and sometimes involve computation of joint inclusion probabilities of selection, which can be hard to obtain. For some sampling designs it is not possible to obtain an unbiased estimator of the variance. These designs include the selection of one element or one large primary sampling unit within some strata, or systematic selection of units or primary sampling units within strata. The problem of variance estimation may also arise from an analytical perspective, while estimating means or totals in unplanned domains, it is possible to arrive at only one unit or cluster within some strata. In this article, we propose a linear regression variance estimator which is very simple to compute and gives a solution to the aforementioned problems. Some examples using different designs are given.

**Keywords:** Linear regression, variance estimation, expansion factor, unplanned domains, collapsed strata.

**JEL Classification:** C80, C83.

**Resumen:** En el muestreo probabilístico, para la estimación de varianza de una media se requieren fórmulas que dependen del diseño empleado para la extracción de la muestra. Estas fórmulas pueden ser difíciles de construir y, en ocasiones, involucran el cálculo de las probabilidades de selección conjuntas, el cual puede complicarse mucho. Para algunos diseños muestrales no es posible obtener un estimador insesgado de la varianza. Estos diseños incluyen la selección de un elemento o un conglomerado grande dentro de algunos estratos o el uso del muestreo sistemático de unidades o conglomerados dentro de estratos. El problema de estimación de varianza también puede surgir desde un punto de vista analítico al estimar medias o totales en dominios no planeados, ya que se puede terminar con una unidad o conglomerado dentro de algunos estratos. En este artículo se propone un estimador de varianza usando regresión lineal el cual es fácil de calcular y proporciona una solución a las situaciones arriba mencionadas. Se proporcionan algunos ejemplos empleando diferentes diseños.

**Palabras Clave:** Regresión lineal, estimación de varianza, factor de expansión, dominios no planeados, estratos colapsados.

---

\*The author would like to thank seminar participants at Banco de México and an anonymous referee for helpful comments.

<sup>†</sup> Dirección General de Investigación Económica. Email: [ampadilla@banxico.org.mx](mailto:ampadilla@banxico.org.mx).

# 1. Introduction

Variance estimation is of primary importance to inference in survey sampling. Once an estimator of a mean or a total of a finite population is computed, it is important to measure the quality of this estimator, which can be achieved by variance estimation. Actually, large probability samples can be subjected to different methods, such as stratification, clustering and probability proportional to some measure of size, to select the sample units (see for example the design of the National Survey of Victimization and Perception of Public Safety, ENVIPE 2014). In these designs, variance may be quite difficult to estimate due to clustering and/or probability proportional to size. In unplanned domains in particular, variance estimation can be difficult to compute, or the variance estimators do not exist. This is the case when only one unit or primary sampling unit in some strata or clusters is available (see Breidt et al., 2014, and the references cited therein for problems of variance estimation with one unit in stratification). Most of the extant literature in this field pertains to variance estimation that is specific to a type of survey design or a particular context. It is not practical to cite them all, but one can have a good idea of the type of situations with respect to variance estimation in Wolter (1985), chapter 11 of Särndal et al. (1992), and the Methodological Note of the ENVIPE 2014.

In practice, some variance estimators are really difficult to compute and require extensive knowledge of the survey design and the ability to handle the data using specialized software. To overcome these limitations, we propose a variance estimator of the Horvitz-Thompson estimator of the mean or total (Horvitz & Thompson, 1952). As it is easy to compute, this estimator is applicable to a wide variety of complex designs, with the exception of self-weighted designs, as explained in the final part of Section 3.2. An important feature of the proposed variance estimator is that it only requires the values of the variable of interest and the expansion factors at the element level (see Section 2.1). It can be implemented using

software that can compute least squares estimates from simple linear regression, or it can even be calculated using a spreadsheet with basic functions, such as variance and covariance.

The article is organized as follows. In Section 2, we introduce the notation and give a brief summary of the key points related to estimation under the so-called design-based approach. Section 3 begins with some linear regression results that are used to construct the proposed variance estimator. Next, we build a decomposition of the Horvitz-Thompson estimators of the total and the mean, which is necessary to derive the variance estimator using linear regression with the expansion factor as the independent variable. In Section 4, we apply the proposed method to three different sets of data and designs, namely a stratified sample design, an example of two-stage cluster design with unequal sizes, and an application to data from the ENVIPE 2014.

## **2. Some estimation methods employed in survey sampling**

### **2.1 Notation**

Let  $U$  denote a finite population of  $N$  elements labeled  $k = 1, \dots, N$ ,  $1 < N$ . It is customary to represent the finite population by its label  $k$  as:  $U = \{1, 2, \dots, k, \dots, N\}$ . The variable under study will be represented by  $y$ , while  $y_k$  will denote the value of  $y$  for the  $k$ -th population element,  $k \in U$ . In this case, the  $y_k$  are real numbers. It is important to note that some survey designs may include several stages or levels of aggregation of the elements in the population. For example, in a household survey, it is common to collect information on the individuals within the household, as well as that related to the household as a whole. In this case, it is necessary to distinguish the information at an individual (element) level from that at the household level.

The sample will be denoted by  $s$ , a subset of  $U$  of size  $I < n < N$  and will be represented by a column vector  $I = (I_1, \dots, I_k, \dots, I_N)' \in \{0,1\}^N$ . In this case,  $I_k$  is a sample membership indicator distributed as a Bernoulli random variable (see Chapter 2 of Särndal et al., 1992), and it is equal to 1 if the  $k$ th element is in the sample and 0 otherwise. It is worth mentioning that this indicator variable is the random element in finite population sampling and  $y_k$  has a numerical value. Thus, the density function induced by the design is discrete. This approach is also known in pertinent literature as design-based sampling.

The sample covariance between two sets of observations will be used in this article and it is computed using the following expression (Eq. 1), where subscript  $s$  refers to sample information:

$$\text{cov}_s(x, y) = \frac{\sum_{i=1}^n (x_i - \hat{x})(y_i - \hat{y})}{n}, \quad (1)$$

$$\text{with } \hat{x} = \frac{\sum_{i=1}^n x_i}{n} \text{ and } \hat{y} = \frac{\sum_{i=1}^n y_i}{n}.$$

Equation (1) can also be written as:

$$\text{cov}_s(x, y) = \frac{\sum_{i=1}^n x_i y_i - n\hat{x}\hat{y}}{n}. \quad (2)$$

From the above expression, it is straightforward to obtain:

$$\sum_{i=1}^n x_i y_i = n \text{cov}_s(x, y) + n\hat{x}\hat{y}. \quad (3)$$

This expression will be used in the development of the variance estimator using linear regression.

Another quantity that will be utilized frequently in this work is the estimator of the coefficient of variation, defined as:

$$cv_s(y) = \frac{\hat{\sigma}_y}{\hat{y}}, \quad (4)$$

$$\text{with } \hat{\sigma}_y^2 = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}.$$

Recall that the Pearson correlation coefficient, also known as correlation coefficient, is computed as:

$$\rho_s(x, y) = \frac{\text{cov}_s(x, y)}{\hat{\sigma}_x \hat{\sigma}_y}, \quad (5)$$

and  $\hat{\sigma}_x, \hat{\sigma}_y > 0$ .

Note that the covariance exists and is zero if all  $x$  and/or  $y$  are equal, so  $\hat{\sigma}_x = 0$  and/or  $\hat{\sigma}_y = 0$ , but the correlation in (5) is not defined if one of both variances are zero.

## 2.2 Estimation under the design-based approach

The objective is to estimate a function  $t$  that depends on the  $y_k$ ,  $t = t(y_1, \dots, y_k, \dots, y_N)$ . For example, a total is written as  $y_U = \sum_{k=1}^N y_k$ . Since we are interested in estimating a total using the design-based approach, it is customary to use the Horvitz-Thompson estimator (Horvitz & Thompson, 1952). This estimator provides the following expression for a total:

$$\hat{y}_U = \sum_{k=1}^N I_k y_k / \pi_k = \sum_{k=1}^n y_k / \pi_k, \text{ with } \pi_k > 0, \text{ where } \pi_k = P(I_k = 1) \text{ is the first-order inclusion}$$

probability. The expansion factor is the inverse of the probability of selection  $\pi_i$  and will be denoted by  $F_i = 1/\pi_i$ , with  $\pi_i > 0$  and  $i \in \{1, \dots, n\}$ . Särndal et al. (1992) show that expansion factors satisfy

$$\sum_{i=1}^n F_i = N. \quad (6)$$

In practice, expansion factors are not just the inverse of the probabilities of selection, but rather also include nonresponse adjustments and calibrations to known totals of

subpopulations, and thus continue to satisfy Equation (6). The estimation method proposed in this article can be applied to any set of expansion factors as long as they satisfy Equation (6) and are not equal. Although  $F_i$  is represented with capital letter, it is not a random variable once a sample is drawn. We utilize this notation in order to maintain alignment with its use in some methodological notes, such as those issued by INEGI.

For variance computation and estimation under the design-based approach, it is also necessary to determine the second-order inclusion probabilities,  $\pi_{kl} = P(I_k I_l = 1)$ .

The variance of a Horvitz-Thompson estimator is, provided that  $\pi_k > 0$  and  $\pi_l > 0$ ,

$$v(\hat{y}_U) = \sum \sum_U c(I_k, I_l) \hat{y}_k \hat{y}_l = \sum \sum_U (\pi_{kl} - \pi_k \pi_l) \hat{y}_k \hat{y}_l. \quad (7)$$

In this expression,  $\hat{y}_k = y_k / \pi_k$ . An unbiased estimator of this variance is, provided that

$\pi_{kl} > 0$ :

$$\hat{v}(\hat{y}_U) = \sum \sum_s \hat{c}(I_k, I_l) \hat{y}_k \hat{y}_l = \sum \sum_s \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \quad (8)$$

In these expressions,  $c(I_k, I_l)$  and  $\hat{c}(I_k, I_l)$  denote the population and estimated covariances, respectively, between the sample indicator variables. In Chapter 2 of Särndal et al.'s (1992) book, it is shown that Equation (8) is an unbiased estimator of (7) and one can find the expressions for  $c(I_k, I_l)$  and  $\hat{c}(I_k, I_l)$ . Expressions (7) and (8) must be specifically determined for a particular design to obtain an explicit formula. It is important to mention that it is not always possible to build a closed formula for a variance estimator using Eq. (8), because the second-order inclusion probabilities are sometimes difficult to compute. Some designs, such as systematic sampling with equal first-order inclusion probabilities of selection have some  $\pi_{kl} = 0$ , in which case Eq. (8) is not defined. In this case, alternative variance estimators must be used (see Wolter, 1985).



Estimation in finite populations can also be made under a different approach, known in the literature as model-based design. Under this approach, it is assumed that the finite population is drawn from an infinite population (superpopulation), as discussed by Valliant et al. (2000).

### 3. Proposed variance estimator based on linear regression

The proposed variance estimator requires knowledge of factor expansions and the variable of interest at element level. Once the required data is obtained from a survey, the point estimator of a total or mean is the usual Horvitz-Thompson estimator (Horvitz & Thompson, 1952). Some authors, including Särndal et al. (1992), refer to it as expansion estimator. Recall that the Horvitz-Thompson estimator of the total is computed as:

$$\hat{y}_{HT} = \sum_{i=1}^n F_i y_i, \quad (9)$$

and the estimator of the mean is calculated using:

$$\hat{y}_{HT} = \frac{\sum_{i=1}^n F_i y_i}{\sum_{i=1}^n F_i}. \quad (10)$$

#### 3.1 Linear regression results

We will use results from linear regression. Hence, consider a variable  $y_i, i \in \{1, \dots, n\}$  which is a realization from a random variable  $Y$ , whose first and second moments are finite. Now consider the following model:

$$y_i = \beta_0 + \beta_1 F_i + \varepsilon_i. \quad (11)$$

In this equation,  $F_i, i \in \{1, \dots, n\}$  corresponds to the expansion factor and  $\varepsilon_i$  is the  $i$ -th error term. The expansion factors will be considered fixed, so that all analyses performed will be

conditioned on them. This is not unrealistic assumption because, in practice, once the expansion factors are computed, they are not subsequently modified.

The assumptions for model given by Eq. (11) are, as discussed by Dutta (1982) and Ott (1984): (i) the random error term  $\varepsilon_i$  has zero expectation; (ii)  $\varepsilon_1, \dots, \varepsilon_i, \dots, \varepsilon_n$  are independent of each other; (iii) the covariance between  $\varepsilon_i$  and  $F_i$  is zero; (iv) the variance  $\sigma_\varepsilon^2$  is constant for all settings of  $F$ ; (v) and  $\varepsilon$ , for a given setting of the independent variable  $F$ , is normally distributed with mean 0 and variance  $\sigma_\varepsilon^2$ .

In regression analysis (see for example Dutta, 1982), simple linear regression model estimators are given by:

$$\hat{\beta}_1 = \frac{\text{cov}_s(y, F)}{\hat{\sigma}_F^2} \quad (12)$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{F}. \quad (13)$$

where

$$\text{cov}_s(y, F) = \frac{\sum_{i=1}^n (y_i - \bar{y})(F_i - \bar{F})}{n} \quad (14)$$

and

$$\hat{\sigma}_F^2 = \frac{\sum_{i=1}^n (F_i - \bar{F})^2}{n}, \quad (15)$$

with  $\hat{\sigma}_F^2 > 0$ .

Similarly, the covariance between the intercept and the slope is expressed as:

$$\text{cov}_s(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{F}\sigma_\varepsilon^2}{\sum_{i=1}^n (F_i - \bar{F})^2}. \quad (16)$$

The variances for the slope and intercept are given by:

$$\hat{v}(\hat{\beta}_0) = \hat{\sigma}_\varepsilon^2 \frac{\sum_{i=1}^n F_i^2}{n \sum_{i=1}^n (F_i - \bar{F})^2} \quad (17)$$

and

$$\hat{v}(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (F_i - \bar{F})^2}. \quad (18)$$

In Eq. (16) to (18),

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} \quad (19)$$

and the predicted value is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 F_i. \quad (20)$$

### 3.2 A decomposition of the Horvitz-Thompson estimator

Let us return for a moment to the sample covariance. If we use expression (3) with  $y_i$  and  $F_i$ , we obtain

$$\sum_{i=1}^n F_i y_i = n \text{cov}_s(F, y) + n\bar{F}\hat{y}. \quad (21)$$

The left-hand side of Equation (21) is equal to (9) and corresponds to the Horvitz-Thompson estimator of the total,

$$\hat{y}_{HT} = n\bar{F}\hat{y} + n \text{cov}_s(F, y). \quad (22)$$

By dividing both sides of (21) by  $n\bar{F}$  we obtain:

$$\frac{\sum_{i=1}^n F_i y_i}{\sum_{i=1}^n F_i} = \frac{\text{cov}_s(F, y)}{\bar{F}} + \hat{y}, \quad (23)$$

where the left-hand side is equal to (10), thus

$$\hat{y}_{HT} = \hat{y} + \frac{\text{cov}_s(F, y)}{\bar{F}} \quad (24)$$

is the Horvitz-Thompson estimator of the mean.

Equation (24) can also be expressed as:

$$\hat{y}_{HT} = \hat{y} [1 + \rho_s(F, y) cv(F) cv(y)] \quad (25)$$

The main results of these expressions are summarized below.

**Proposition 1.** Let  $(y_i, F_i)$ ,  $i \in \{1, \dots, n\}$  be an array containing the variable of interest  $y$ , from which the Horvitz-Thompson estimator of the mean is built and  $F_i$  the expansion factor associated to element  $i$  in a sample of size  $n$ . Compute  $\hat{y}$ ,  $\bar{F}$ ,  $cv(F)$ ,  $cv(y)$ ,  $\text{cov}_s(F, y)$  and  $\rho_s(F, y)$ . The Horvitz-Thompson estimator of the mean can be expressed as

$$\hat{y}_{HT} = \hat{y} + \frac{\text{cov}_s(F, y)}{\bar{F}} \quad \text{or} \quad \hat{y}_{HT} = \hat{y} [1 + \rho_s(F, y) cv(F) cv(y)].$$

The estimator of the total is obtained by multiplying these equations by  $\sum_{i=1}^n F_i$ .

Proof of Equation (25) is given in Annex 2.

Equations (22) to (24) are important because they exhibit a decomposition of the Horvitz-Thompson estimators of a total and mean. The estimator of a total can be expressed as a sum of a product of equal-weighted averages from  $F$  and  $y$  and an effect of a linear association between these two variables via the covariance. The estimator of the mean is therefore the sum of the sample mean and an effect of linear association between the interest variable and the expansion factor divided by the mean of the expansion factors. Note that expression given in Eq. (25) enables us to assess the effect of the expansion factor, the variable of interest  $y$

and a measure of linear dependence between them. To the best of the author's knowledge, the decompositions expressed in (22) to (25) have not been previously published. Equation (24) can be seen as an additive decomposition of the Horvitz-Thompson estimator into the sample mean and the effect of the expansion factors.

Since we are interested in variance estimation of the mean, we will restrict our focus on expression (24). Thus, the total is computed by multiplying the estimator of the mean by  $\sum_{i=1}^n F_i$ . It is important to point out that the proposed variance estimator is not defined when the expansion factors  $F_i$  have the same value. This is the case, for example, in stratified sampling with proportional allocation and in general in the so-called self-weighted samples. This type of sampling occurs in sampling designs with equal sampling fractions in strata or clusters.

### 3.3 Proposed variance estimator using linear regression

In this section, we present the main results obtained in this work, namely two expressions for the variance estimators using simple linear regression with the expansion factor as the independent variable.

In the first step, an expression for the Horvitz-Thompson estimator is developed in terms of the regression coefficients and the expansion factors.

From Eq. (12) and (13) we have:

$$\text{cov}_s(F, y) = \hat{\beta}_1 \hat{\sigma}_F^2 \quad (26)$$

and

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{F}. \quad (27)$$

From Equation (24) we obtain:

$$\hat{y} = \hat{y}_{HT} - \frac{\text{cov}_s(F, y)}{\bar{F}}. \quad (28)$$

In the next step, the covariance from Eq. (26) is substituted into (28) and  $\hat{y}$  from Eq. (27) into the left-hand side of (28):

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{F} = \hat{y}_{HT} - \frac{\hat{\beta}_1 \hat{\sigma}_F^2}{\bar{F}}.$$

From this equation and using expression (4), we obtain:

$$\hat{y}_{HT} = \hat{\beta}_0 + \hat{\beta}_1 \bar{F} [1 + cv^2(F)]. \quad (29)$$

It is important to mention that Equation (29) is an identity, and it is not an approximation because it yields the same value as the Horvitz-Thompson estimator of the mean,

$$\hat{y}_{HT} = \frac{\sum_{i=1}^n F_i y_i}{\sum_{i=1}^n F_i}. \text{ Consequently, } \hat{y}_{HT} \text{ computed using the last formula and } \hat{y}_{HT} \text{ calculated}$$

using Eq. (29) produce estimators of equal value. Equation (29) relates the Horvitz-Thompson estimator of the mean to the coefficients of a simple linear regression model and the expansion factors. This result is summarized below.

**Proposition 2.** Let  $(y_i, F_i)$ ,  $i \in \{1, \dots, n\}$  be an array containing the variable of interest  $y$ , utilized in building the Horvitz-Thompson estimator of the mean and  $F_i$  the expansion factor associated with element  $i$  in a sample of size  $n$ . Compute  $cv(F)$  and fit the model  $y_i = \beta_0 + \beta_1 F_i + \varepsilon_i$  to  $(y_i, F_i)$ , where  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are obtained from the outcome of the regression. The Horvitz-Thompson estimator of the mean is thus given by  $\hat{y}_{HT} = \hat{\beta}_0 + \hat{\beta}_1 \bar{F} [1 + cv^2(F)]$ .

Finally, if we apply the variance used in the theory of linear regression to Eq. (29), conditioned on the expansion factors  $F_i$ ,  $i \in \{1, \dots, n\}$  and the expressions provided in Section 3.1, we obtain:

$$\hat{v}_{LS}(\hat{y}_{HT} | F) = \hat{v}(\hat{\beta}_0) cv^2(F). \quad (30)$$

Proof of Equation (30) is given in Annex 2.

This result shows that the proposed variance estimator is affected by the variation between values of the dependent variable through  $\hat{\beta}_0$  (recall from Equation (13) that  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{F}$ ) and a measure of relative variation expansion factors determined from the means of the coefficient of variation.

This is the main result of the present work and is summarized below.

**Proposition 3.** Let  $(y_i, F_i)$ ,  $i \in \{1, \dots, n\}$  be an array containing the variable of interest  $y$ , from which the Horvitz-Thompson estimator of the mean is built and  $F_i$  denote the expansion factor associated with element  $i$  in a sample of size  $n$ . Compute  $cv(F)$  and fit the model  $y_i = \beta_0 + \beta_1 F_i + \varepsilon_i$  to  $(y_i, F_i)$ , where  $\hat{v}(\hat{\beta}_0)$  is obtained through regression analysis. Then, the variance estimator for the Horvitz-Thompson estimator of the mean is given by  $\hat{v}_{LS}(\hat{y}_{HT} | F) = \hat{v}(\hat{\beta}_0) cv^2(F)$ .

Another expression for  $\hat{v}_{LS}(\hat{y}_{HT} | F)$ , which is useful for analyzing the effects on the variance from its different components, is given by:

$$\hat{v}_{LS}(\hat{y}_{HT} | F) = \frac{\hat{\sigma}_y^2}{n-2} [1 - \rho_s^2(y, F)] [1 + cv^2(F)]. \quad (31)$$

Proof of Equation (31) is given in Annex 2.

Note that the variance estimator of Equation (31) can be obtained without computing the regression estimators and the corresponding standard errors.

## 4. Examples

**4.1 Example 1:** Consider the population of building heights given in Table 5.4, pages 119-121, from Scheaffer et al. (1987). This population contains heights of 355 buildings (expressed in feet) located in 10 US cities. Some information about this population is shown in Table 1 below.

Table 1. Building heights in feet

Stratum (h) City	Total of buildings $N_h$	Relative size $W_h=N_h/N$	Height (feet) Average $y_h$	Minimum	Maximum	Variance	Coefficient of variation
Atlanta	18	0.051	454.5	374	723	14,635.0	27%
Chicago	43	0.121	620.7	460	1,454	44,247.9	34%
Dallas	38	0.107	484.2	327	939	19,922.4	29%
Detroit	26	0.073	415.3	325	720	8,132.7	22%
Houston	53	0.149	512.8	328	1,002	29,300.6	33%
Los Angeles	23	0.065	491.1	347	858	22,998.1	31%
New York	66	0.186	644.4	505	1,472	41,313.8	32%
Philadelphia	28	0.079	418.2	340	548	3,764.4	15%
Pittsburgh	22	0.062	467.2	330	841	20,014.8	30%
San Francisco	38	0.107	489.5	355	853	11,113.3	22%
Total	355		523.0				

Source: own construction based on pages 119-121 from Scheaffer et al. (1987).

The tabulated information pertains to tall buildings in the selected cities, as can be seen from column “Maximum”, and by comparing their heights with the average for each stratum. Variances within strata have non-similar values, ranging from 3,764.4 in Philadelphia to 44,247.9 in Chicago. This variability affects some variance estimator types, including those obtained using simple random sampling within strata. Thus, if simple random sampling within each stratum is used, due to some large values in some cities, it is highly probable that the estimator will have a poor coverage.

In this example, the original data pertaining to building heights will be used to simulate a variable in the population correlated with it and from which a sample will be drawn. The values were generated by simulations using part of a method developed by Padilla (2015).



Specifically, if  $u_i$  and  $v_i$  are realizations from two distribution functions with finite first two moments, then  $a_i = u_i + bv_i$  (where  $b$  is a constant) is also a realization of the sum of the two distributions, provided that the second moment is finite for both distributions. By applying this method, it is possible to control the correlation between  $a_i$  and  $u_i$  to some extent, given the realizations from  $u_i$  and  $v_i$  (for more details, see Padilla, 2015).

Since we are aware of the presence of high values within strata, a new stratum will be built, to which the 10% tallest buildings from each city will be allocated. This corresponds to 40 buildings, resulting in 315 buildings located in the 10 analyzed cities. The new simulated variable was generated using 315 random numbers from a continuous uniform distribution with lower and upper values, 325 and 850, respectively. As a result,  $u_i$  and  $v_i$  had the same uniform distribution and  $b = 0.2$ . The new variable for the new stratum corresponding to the 40 tallest buildings was also generated by drawing 40 random numbers from a continuous uniform distribution with 491 and 1472 as the lower and upper values, respectively. As in the previous case,  $u_i$  and  $v_i$  had the same uniform distribution and  $b = 0.2$ .

Table 2. Building heights (expressed in feet) within a new stratum  
New simulated variable, correlation of 0.5 with original building heights

Stratum (h)	Total of buildings	Relative size	New variable Height (feet)	Minimum	Maximum	Variance	Coefficient of variation
City	$N_h$	$W_h=N_h/N$	Average $y_h$				
Atlanta	16	0.045	600.2	448	919	20,706.7	24%
Chicago	38	0.107	731.4	462	937	27,150.8	23%
Dallas	34	0.096	665.4	417	931	24,339.7	23%
Detroit	23	0.065	673.8	421	928	25,399.3	24%
Houston	47	0.132	685.5	420	987	25,403.8	23%
Los Angeles	20	0.056	698.3	454	901	21,118.7	21%
New York	59	0.166	716.8	467	964	20,498.6	20%
Philadelphia	25	0.070	733.6	447	905	17,401.6	18%
Pittsburgh	19	0.054	663.8	409	914	23,212.6	23%
San Francisco	34	0.096	651.7	411	939	20,674.6	22%
Tallest buildings	40	0.113	1,133.4	641	1,581	83,712.5	26%
Total	355		739.3			49,713.1	

Source: own construction based on pages 119-121 from Scheaffer et al. (1987).

Since this table contains information pertaining to the new variable, it is not possible to compare the averages or variances; however, the coefficients of variation are comparable. In Table 1, this quantity varies from 15% to 34%, while in Table 2 this range is reduced to 18% and 26%. In fact, only the coefficient of variation corresponding to the new stratum “The tallest buildings” has the highest value, 26%, while the remaining quantities cluster around 23%.

Table 3. Sample size, almost proportional allocation

Stratum (h)	Total of buildings	Relative size	Sample size	Expansion factor
City	$N_h$	$W_h=N_h/N$	$n_h$	$F_{hi}$
Atlanta	16	0.045	3	5.333
Chicago	38	0.107	7	5.429
Dallas	34	0.096	6	5.667
Detroit	23	0.065	4	5.750
Houston	47	0.132	8	5.875
Los Angeles	20	0.056	4	5.000
New York	59	0.166	10	5.900
Philadelphia	25	0.070	5	5.000
Pittsburgh	19	0.054	4	4.750
San Francisco	34	0.096	6	5.667
Highest buildings	40	0.113	7	5.714
Total	355		64	

Source: own construction based on pages 119-121 from Scheaffer et al. (1987).

The sample was allocated proportionally to the relative sizes  $W_h$ ,  $n_h = nW_h$ , and its size was rounded up to the nearest integer. For this reason, in Table 3, the expression “almost proportional allocation” was used.

Table 4. Main results from the estimation of the new variable “height”

Quantity	Value
average $y_{hi}$	710.17
average $F_{hi}$	5.55
$\sigma(F)$	0.36
$cv(F)$	6.4%
$v_{LS}(\beta_0)$	172,736.9
$sd_{LS}(\beta_0)$	415.6
$\hat{v}_{LS}(\hat{y}_{HT}   F)$	711.3
$sd_{LS}(\hat{y}_{HT}   F)$	26.7
$cv(\hat{y}_{HT})$	3.8%

Expression (31) yields the following values for each term:

$$\frac{\hat{\sigma}_y^2}{n-2} = 712.782, 1 - \rho_s^2(y, F) = 0.994 \text{ and } 1 + cv^2(F) = 1.004.$$

The product of the three components is 711.3, as shown in Table 4 for  $\hat{v}_{LS}(\hat{y}_{HT} | F)$ .

We will compare the estimated variance  $\hat{v}_{LS}(\hat{y}_{HT} | F)$  with two population variances corresponding to simple random sampling without replacement, *srswor*, and stratified random sampling, *strs*. While this is not a fair comparison, because  $\hat{v}_{LS}(\hat{y}_{HT} | F)$  is affected by random error, it will serve as a suitable measure of the estimator’s adequacy.

Table 5. Computation of population variance under stratified random sampling

City	$N_h$	$W_h$	$n_h$	$F_h$	$f_h$	$S_h^2$	$V_{srswor}$	$W_h^2 v_{srswor}$
Atlanta	16	0.045	3	5.333	0.188	20,706.7	5,608.1	11.4
Chicago	38	0.107	7	5.429	0.184	27,150.8	3,164.2	36.3
Dallas	34	0.096	6	5.667	0.176	24,339.7	3,340.7	30.6
Detroit	23	0.065	4	5.750	0.174	25,399.3	5,245.5	22.0
Houston	47	0.132	8	5.875	0.170	25,403.8	2,635.0	46.2
Los Angeles	20	0.056	4	5.000	0.200	21,118.7	4,223.7	13.4
New York	59	0.166	10	5.900	0.169	20,498.6	1,702.4	47.0
Philadelphia	25	0.070	5	5.000	0.200	17,401.6	2,784.3	13.8
Pittsburgh	19	0.054	4	4.750	0.211	23,212.6	4,581.4	13.1
San Francisco	34	0.096	6	5.667	0.176	20,674.6	2,837.7	26.0
Tallest buildings	40	0.113	7	5.714	0.175	83,712.5	9,866.1	125.3
Total	355	1	64	5.462	0.183	49,713.1	634.6	385.1

Source: own construction based on pages 119-121 from Scheaffer et al. (1987).

In Table 5, the quantities required for computing the population variance under stratified random sampling are given. The values under column *Vsrswor* correspond to population variance values within strata obtained through simple random sampling without replacement,  $v(\hat{y}_h) = (1 - n_h/N_h)S_h^2/n_h$ , where  $S_h^2 = \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)^2 / (n_h - 1)$ . The symbol  $f_h = n_h/N_h$  denotes the sampling fraction. The value 634.6 corresponding to the last row labelled “Total” is the population value of the variance under *srswor*,  $v(\hat{y}_{HT}) = (1 - n/N)S^2/n$ . In this expression,  $S^2 = \sum_{i=1}^N (y_i - \bar{y})^2 / (n - 1)$ . The last column contains values  $W_h^2 v(\hat{y}_h)$ , which are part of the equation for the population variance of stratified random sampling without replacement for the Horvitz-Thompson estimator of the mean,  $v_{strs}(\hat{y}_{HT}) = \sum_{h=1}^H W_h^2 v(\hat{y}_h) = 385.1$ .

Table 6. Comparison of variance estimator under regression with *srswor* and *strs*

	srswor	strs
variance	1.12	1.85
standard deviation	1.06	1.36

Table 6 provides the comparison of the variance estimator using linear regression to simple random sampling and stratified sampling. The quantities pertain to the ratio of the variance estimator using regression to the corresponding population variance minus one. From the tabulated data, it is evident that the proposed variance estimator is similar to simple random sampling. While it is not as efficient as stratified sampling, this was expected because, setting aside the variance under equal allocation, stratified sampling is more efficient than simple random sampling.

Estimation with a single element in sample in one stratum.

Now suppose that, in the last population, we have only one element in the sample in the first stratum of Table 5 corresponding to Atlanta, while other data remain unchanged. The revised quantities in Table 5 are given in Table 7.

Table 7. Computation of population variance under stratified random sampling

City	$N_h$	$W_h$	$n_h$	$F_h$	$f_h$	$S^2_h$	$V_{srswor}$	$W^2_h v_{srswor}$
Atlanta	16	0.045	1	16.000	0.063	20,706.7	19,412.5	39.4
Chicago	38	0.107	7	5.429	0.184	27,150.8	3,164.2	36.3
Dallas	34	0.096	6	5.667	0.176	24,339.7	3,340.7	30.6
Detroit	23	0.065	4	5.750	0.174	25,399.3	5,245.5	22.0
Houston	47	0.132	8	5.875	0.170	25,403.8	2,635.0	46.2
Los Angeles	20	0.056	4	5.000	0.200	21,118.7	4,223.7	13.4
Nueva York	59	0.166	10	5.900	0.169	20,498.6	1,702.4	47.0
Philadelphia	25	0.070	5	5.000	0.200	17,401.6	2,784.3	13.8
Pittsburgh	19	0.054	4	4.750	0.211	23,212.6	4,581.4	13.1
San Francisco	34	0.096	6	5.667	0.176	20,674.6	2,837.7	26.0
Tallest buildings	40	0.113	7	5.714	0.175	83,712.5	9,866.1	125.3
Total	355	1	62	5.726	0.175	49,713.1	661.8	413.2

Source: own construction based on pages 119-121 from Scheaffer et al. (1987).

With this new data, the expansion factor in Atlanta is now 16, the total sample size is 62, the estimator of the mean is 708.33. In addition, the coefficient of variation of the expansion factors is 23.8% and the variance estimator under regression is 765.93; thus, the standard error is 27.68. This last quantity can be compared with  $sd_{LS}(\hat{y}_{HT} | F) = 26.7$  from Table 4. The difference of one unit between these two standard errors is insignificant considering the fact that the problem of estimating the variance with a unit in one stratum is not trivial, as discussed by Wolter (1985), Cochran (1986), Breidt et al. (2014) and Padilla (2016). In this case, the estimator of the design-based variance estimator,  $\hat{v}_{strs}(\hat{y}_{HT})$ , does not exist, because the variance within the stratum pertaining to Atlanta,  $s_h^2 = \sum_{hi=1}^{n_h} (y_{hi} - \hat{y}_h)^2 / (n_h - 1)$  cannot be computed with only one observation,  $n_h = 1$ .

**4.2 Example 2:** Consider the data and results related to the sample in Example 9.1, given in Table 9.1, pages 236 and 237 from Scheaffer et al. (1987). This is an example of estimation from a sample obtained by a two-stage clustered design. The first stage corresponds to 10 factories, while the second relates to the machines within each factory, using a sampling fraction of 20%. Both samples were extracted by simple random sampling without replacement. There are 90 factories across the United States, with varying numbers of machines, resulting in clusters of unequal sizes. The objective is to estimate the average idle time (in hours) due to machine malfunction. This was achieved by visual inspection of the repairs log for each machine included in the sample. Some information about this population is shown in Table A.2 in Annex 1. This table contains almost all the information given in Example 9.1, Scheaffer et al. (1987). Moreover, in the sample of 10 factories, the number of machines within factories varies from 40 to 65. Labels  $B_i$  and  $b_i$  refer to the number of machines in a particular factory and sample, respectively. *Average*  $y_i$  corresponds to the machine idle time due to malfunction and repairs in the  $i$ -th factory and  $s_i^2$  is the sample variance between the machine idle times in the  $i$ -th factory. The last column of Table A.2 contains the coefficient of variation for  $y$  within each factory. The values this quantity takes vary considerably across factories, ranging from 45% to 101%. All values from the sample used in Example 9.1 given by Scheaffer et al. (1987) are displayed in Table A.1 column (a) in Annex 1. The column labelled “factory” also corresponds to this example. Columns (b) and (c) in Table A.1 contain values generated by simulations using part of a method developed by Padilla (2015). The method is based on the original  $y$  and  $F$  series, to which a series of random numbers from a specified distribution multiplied by a constant is added. The random numbers presented in column (b) are derived from a normal distribution with zero mean and  $SD = 2$ . In column (c), a student t distribution with eight degrees of freedom was used.

In Table 8, the variance estimations under regression,  $\hat{v}_{LS}(\hat{y}_{HT} | F)$ , are presented, based on two-stage unequal cluster sampling using simple random sampling in both stages for cases (a), (b) and (c), labelled  $\hat{v}_D(\hat{y})$ . Recall that case (a) corresponds to Example 9.1, Scheaffer et

al. (1987). While we reproduce the result of the variance estimation reported in Example 9.1,  $\hat{v}_D(\hat{y})=0.037$ , we do not include the expression for this variance. Those interested in computation of this variance should refer to Scheaffer et al. (1987). Closer inspection of data presented in Table 8 reveals that the variance regression estimator is greater than the variance under two-stage cluster sampling. Note that, in case (a), the correlation between variable  $y$  and the expansion factors is almost zero, which implies a poor fit of the regression model and a large value of the standard error of  $\hat{\beta}_0$ .

Table 8. Main results, comparison of variance estimator under regression with the variance from two-stage unequal cluster sampling

Quantity	(a)	(b)	(c)
$sd_{LS}(\hat{y}_{HT}   F) =$	0.313	0.179	0.154
$\hat{v}_{LS}(\hat{y}_{HT}   F) =$	0.098	0.032	0.024
$sd_D(\hat{y}) =$	0.193	0.312	0.296
$\hat{v}_D(\hat{y}) =$	0.037	0.097	0.088
$\rho_s(y, F) =$	0.023	0.100	-0.219
$sd_{LS}/sd_D =$	162%	57%	52%
$\hat{v}_{LS}/\hat{v}_D =$	264%	33%	27%

Source: own construction based on pages 236-237 from Scheaffer et al. (1987).

In cases (b) and (c), the variance estimator under regression is about one third of  $\hat{v}_D(\hat{y})$ . More specifically, the correlation between variable  $y$  and the expansion factors is 0.1 and -0.22, respectively, which implies a better fit of the regression model than was achieved in case (a), as well as a smaller value of the standard errors of  $\hat{\beta}_0$ . This can be confirmed via expressions (17), (19) and (A7) in Annex 2, where the variance of the residuals, part of the estimated variance of  $\hat{\beta}_0$ , decreases when  $\rho_s(y, F)$  increases and/or  $\sigma_y^2$  decreases. This behavior can be seen through the components of the variance estimator under regression, given by Equation (31). The value of each component is shown in Table 9, where it can be seen that for cases (b) and (c) the values of  $\sigma_y^2$  are about 1/3 of the corresponding  $\sigma_y^2$  obtained in case

(a). This is the main contributor to the improvement in variance estimation compared to the values of  $\hat{v}_D(\hat{y})$ .

Table 9. Components of the variance estimator under regression

Quantity	(a)	(b)	(c)
$\hat{\sigma}_y^2/(n-2)=$	0.098	0.032	0.025
$1 - \rho_s^2(y, F) =$	0.999	0.990	0.952
$1 + cv^2(F) =$	1.001	1.001	1.001
$\hat{v}_{LS}(\hat{y}_{HT}   F) =$	0.098	0.032	0.024

Estimation with a single primary sampling unit.

Now suppose that, in case (a), the sample from Factory 1 contains only one value that corresponds to Machine 5, i.e.,  $y_5 = 11$  (see table A1 in Annex 1), with the remaining data unchanged. While the number of factories (clusters) remains the same, as Factory 1 has only one observation in the sample, the total sample size reduces to 95. In this scenario, the estimator of the mean exists, but the design-based variance estimator,  $\hat{v}_D(\hat{y})$ , does not because the variance within factories cannot be computed for Factory 1 with only one observation. Nonetheless, the proposed variance estimator under regression can be computed using Equation (31), as shown below:

$$\frac{\hat{\sigma}_y^2}{n-2} = 0.11, 1 - \rho_s^2(y, F) = 0.999 \text{ and } 1 + cv^2(F) = 1.001, \text{ so } \hat{v}_{LS}(\hat{y}_{HT} | F) = 0.11.$$

This value can be compared to the corresponding entry in Table 9 under (a),  $\hat{v}_{LS}(\hat{y}_{HT} | F) = 0.098$ . The ratio of the variance estimators is just  $0.11/0.098 = 1.12$ . The increase in variance, due to loss of information, is not a matter of concern. In these situations, the proposed variance estimator under regression is an option to variance estimation without



using difficult to apply methods, as discussed by Wolter (1985), Cochran (1986) and Breidt et al. (2016). The Horvitz-Thompson estimators of the mean for case (a) with  $n = 104$  and  $n = 95$  are 4.598 and 4.581, respectively.

**4.3 Example 3:** We use data from the National Survey of Victimization and Perception of Public Safety (ENVIPE in Spanish) 2014, which is administered by the National Institute of Statistics and Geography (INEGI in Spanish). This is an annual survey conducted in the 32 states of México that includes both urban and rural households and applies to persons aged 18 and older. It aims to elucidate respondents' perceptions of public safety and institutional performance in relation to security. It also includes items on a variety of topics related to crime-affected households. It is important to note that the survey design is complex, since it includes stratification, clustering and the use of probability proportional to a measure of size to select the units in sample. Moreover, the inverse of the first-order probabilities of selection have been adjusted by non-response and are calibrated to known totals. This is comprised in the expansion factors released in public data from ENVIPE.

The sample data used in this example corresponds to the state of Aguascalientes. Its data set contains 1,858 entries. We are interested in estimating the proportion of households affected by crime in Aguascalientes and computing a variance estimator, along with the 90% confidence interval. This same estimation is computed for the rural area of Aguascalientes.

#### **4.3.1 Estimating the proportion of households affected by crime in Aguascalientes**

The sample corresponding to Aguascalientes contains 1,858 entries, the sum of the expansion factors is 325,126 households,  $\sum_{i=1}^{1858} F_i = 325,126$ , and  $\sum_{i=1}^n y_i F_i = 94,343$ . The sample mean of the expansion factors is 174.9 and the coefficient of variation is 14%. As the Horvitz-Thompson estimator of the mean is 0.29, it indicates that nearly 30% of households in Aguascalientes have been affected by crime.

The regression estimator yields  $\sqrt{\hat{v}(\hat{\beta}_0)} = 0.0753$  and the square root of the variance estimator using linear regression is  $\sqrt{\hat{v}_{LS}(\hat{y}_{HT} | F)} = 0.0106$ . Based on this information, 90% confidence level for the mean estimator is  $(0.273, 0.308)$ .

#### 4.3.2 Estimating the proportion of rural households affected by crime in Aguascalientes

The sample corresponding to the rural part of Aguascalientes contains 351 entries, and the sum of the expansion factors is 59,049 households,  $\sum_{i=1}^{351} F_i = 59,049$ , and  $\sum_{i=1}^n y_i F_i = 9,455$ . The sample mean of the expansion factors is 168.7 and the coefficient of variation is 8.5%. The Horvitz-Thompson estimator of the mean is 0.16, indicating that 16% of the rural households have been affected by crime. The regression estimator yields  $\sqrt{\hat{v}(\hat{\beta}_0)} = 0.2302$  and the square root of the variance estimator using linear regression is  $\sqrt{\hat{v}_{LS}(\hat{y}_{HT} | F)} = 0.0195$ . This information yields 90% confidence level for the mean estimator of  $(0.128, 0.192)$ .

It is important to mention that the computations for estimating the variance using linear regression in the examples discussed above were performed in a spreadsheet using the regression analysis tool and by extracting the value of the estimate of the standard error of the intercept. The confidence intervals in the last two examples were calculated under the assumption of normality of the Horvitz-Thompson estimator. In this case, the estimator is a proportion. However, due to the sample size, the low values of the coefficients of variation and the fact that the estimated proportions are far from zero—see the relative values of  $\hat{v}_{LS}(\hat{y}_{HT} | F)$  compared to the mean estimators—the assumption of normality seems reasonable.

## 5. Conclusions

An estimator for the variance of the Horvitz-Thompson estimator of the mean using simple linear regression with the expansion factor as an independent variable was proposed. The examples presented in this article confirm that the variance computation is straightforward and only requires least squares estimation, which is available in commercial and free software and, as we mentioned, in a spreadsheet. From an analytical point of view, we also derived an expression for decomposing the proposed variance estimator into the element variance of the variable of interest, the correlation between this variable and the expansion factors, and the coefficient of variation of the expansion factors. The proposed variance estimator is not intended to substitute the traditional design-based variance estimators when they can be calculated. In cases when some strata contain only one element or cluster, which can arise in domain estimation, or when the variance is estimated with a sample extracted using probability proportional to some measure of size, the proposed variance estimator is a viable option. We did not conduct analysis to validate the adequacy of the regression model because this model was built to reproduce the Horvitz-Thompson estimator of the mean or total. Certainly, this is a topic for future research, together with the inclusion of different explanatory variables.

## References

Breidt, F. J., Opsomer, J.D. & Sánchez-Borrego, I., Nonparametric variance estimation under fine stratification: an alternative to collapsed strata. *Journal of the American Statistical Association*. Vol. 111, Iss. 514, 2016.

Cochran, W., *Técnicas de Muestreo*, Ed. CECSA, México, 1986.

Dutta, M., *Métodos Económicos*, South-Western Publishing Co., Cincinnati, Ohio, 1982.

Horvitz, D.G. & Thompson, D.J., A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, Vol. 47, No. 260, 1952, pp. 663-685.

INEGI, Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública 2014 ENVIFE, Síntesis Metodológica y bases de datos.

Ott, L., *An Introduction to Statistical Methods and Data Analysis*, 2nd ed. Duxbury Press, Boston, Mass., 1984.

Padilla, A.M. *Simulación de realizaciones de dos variables aleatorias con correlación dada*. Tópicos en Probabilidad y Estadística, Libro electrónico, Facultad de Ciencias Físico Matemáticas, BUAP, 2015. ISBN: 978-607-487-909-4.

Padilla, A. *Estimación de razón-remuestreo en muestreo estratificado*. Documentos de Investigación, No. 2016-02, Banco de México.

Särndal, C.E., Swensson, B. & Wretman, J.H., *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

Scheaffer, R.L., Mendenhall, W. & Ott, L., *Elementos de Muestreo*. Grupo Editorial Iberoamérica, S.A. de C.V., México, 1987.

Valliant, R., Dorfman, A. and Royall, R. *Finite Population Sampling and Inference: a prediction approach*. John Wiley and Sons, New York, 2000.

Wolter, K. M., *Introduction to variance estimation*. Springer-Verlag, New York, 1985.

# Annex 1

Table A1. Data used in Example 2

Factory and column (a) are sourced from Scheaffer et al. (1987); columns (b) and (c) are generated by the author as explained in Example 2.

Machine	Factory	(a) $y_i$	(b) $y_i$	(c) $y_i$
1	1	5.0	6.4	3.6
2	1	7.0	5.2	4.5
3	1	9.0	8.3	7.1
4	1	0.0	4.9	4.0
5	1	11.0	3.4	6.2
6	1	2.0	2.4	4.4
7	1	8.0	3.0	5.6
8	1	4.0	3.3	4.7
9	1	3.0	3.4	4.1
10	1	5.0	1.9	5.4
11	2	4.0	5.3	4.5
12	2	3.0	4.1	3.1
13	2	7.0	2.0	3.8
14	2	2.0	4.3	4.3
15	2	11.0	8.7	5.9
16	2	0.0	6.7	2.7
17	2	1.0	3.1	2.1
18	2	9.0	4.6	6.0
19	2	4.0	0.0	3.8
20	2	3.0	4.5	8.4
21	2	2.0	5.2	4.1
22	2	1.0	2.2	0.0
23	2	5.0	5.4	4.5
24	3	5.0	2.5	4.9
25	3	6.0	3.4	4.6
26	3	4.0	6.0	6.0
27	3	11.0	6.3	7.7
28	3	12.0	1.8	7.0
29	3	0.0	5.0	4.0
30	3	1.0	3.0	2.7
31	3	8.0	6.1	2.5
32	3	4.0	4.3	4.8
33	4	6.0	3.0	6.4
34	4	4.0	2.2	7.0
35	4	0.0	6.8	3.4
36	4	1.0	4.7	8.1
37	4	0.0	4.2	4.9
38	4	9.0	4.7	5.9
39	4	8.0	4.9	7.9
40	4	4.0	3.1	7.4
41	4	6.0	4.8	6.0
42	4	10.0	5.7	6.1
43	5	11.0	7.3	4.4
44	5	4.0	1.9	1.8
45	5	3.0	3.1	5.1
46	5	1.0	5.0	2.9
47	5	0.0	4.6	3.1
48	5	2.0	3.7	2.6
49	5	8.0	9.3	5.8
50	5	6.0	4.1	6.4
51	5	5.0	4.4	1.5
52	5	3.0	6.3	2.9
53	6	12.0	6.7	5.9
54	6	11.0	5.2	7.0
55	6	3.0	3.6	3.7
56	6	4.0	5.9	3.0
57	6	2.0	3.7	2.8
58	6	0.0	7.7	3.2
59	6	0.0	2.6	2.9
60	6	1.0	2.6	3.7
61	6	4.0	8.0	3.6
62	6	3.0	2.0	3.2
63	6	2.0	2.7	3.3
64	6	4.0	5.8	4.7
65	7	3.0	5.2	6.3
66	7	7.0	6.8	4.5
67	7	6.0	1.8	2.7
68	7	7.0	6.7	3.5
69	7	8.0	5.9	5.0
70	7	4.0	3.7	3.3
71	7	3.0	5.8	3.5
72	7	2.0	3.6	4.7
73	8	3.0	3.5	4.9
74	8	6.0	8.5	4.2
75	8	4.0	4.0	5.4
76	8	3.0	2.8	6.6
77	8	2.0	2.8	6.2
78	8	2.0	4.9	3.7
79	8	8.0	6.0	6.5
80	8	4.0	3.4	4.3
81	8	0.0	3.3	5.6
82	8	4.0	4.6	4.7
83	8	5.0	5.1	3.9
84	8	6.0	5.2	4.6
85	8	3.0	5.3	3.6
86	9	6.0	3.5	6.1
87	9	4.0	3.0	3.8
88	9	7.0	3.7	4.3
89	9	3.0	2.6	3.7
90	9	9.0	3.6	7.2
91	9	1.0	4.6	2.8
92	9	4.0	4.6	6.2
93	9	5.0	5.4	4.3
94	10	6.0	6.4	3.4
95	10	7.0	10.0	5.6
96	10	5.0	4.7	3.6
97	10	10.0	6.3	6.8
98	10	11.0	6.2	6.7
99	10	2.0	3.2	4.3
100	10	1.0	3.0	3.0
101	10	4.0	6.0	4.9
102	10	0.0	7.2	2.1
103	10	5.0	4.6	3.5
104	10	4.0	3.2	3.9

## Example 2: Selected information about factories

Table A.2 Original data table 9.1, page 236, Scheaffer et al. (1987)

Factory	$B_i$	$b_i$	average $y_i$	$s_i^2$	$cv(y)$
1	50	10	5.40	11.38	62%
2	65	13	4.00	10.67	82%
3	45	9	5.67	16.75	72%
4	48	10	4.80	13.29	76%
5	52	10	4.30	11.12	78%
6	58	12	3.83	14.88	101%
7	42	8	5.00	5.14	45%
8	66	13	3.85	4.31	54%
9	40	8	4.88	6.13	51%
10	56	11	5.00	11.80	69%
	522	104			

Table A.3 Simulated  $y_i$  as explained in example 2, using  $N(0,2)$ .

Factory	$B_i$	$b_i$	average $y_i$	$s_i^2$	$cv(y)$
1	50	10	4.22	3.95	47%
2	65	13	4.32	4.83	51%
3	45	9	4.27	2.86	40%
4	48	10	4.42	1.81	30%
5	52	10	4.98	4.67	43%
6	58	12	4.70	4.49	45%
7	42	8	4.93	3.09	36%
8	66	13	4.58	2.47	34%
9	40	8	3.90	0.86	24%
10	56	11	5.52	4.30	38%
	522	104			

Table A.4 Simulated  $y_i$  as explained in example 2, using  $st(8)$ .

Factory	$B_i$	$b_i$	average $y_i$	$s_i^2$	$cv(y)$
1	50	10	4.97	1.21	22%
2	65	13	4.10	4.11	49%
3	45	9	4.92	3.22	36%
4	48	10	6.31	2.06	23%
5	52	10	3.65	2.86	46%
6	58	12	3.92	1.73	33%
7	42	8	4.17	1.34	28%
8	66	13	4.94	1.09	21%
9	40	8	4.81	2.32	32%
10	56	11	4.35	2.23	34%
	522	104			

## Annex 2

Proof of Equation (25):  $\hat{y}_{HT} = \hat{y}[1 + \rho_s(F, y)cv(F)cv(y)]$ .

From Equation (5), we have  $cov_s(y, F) = \rho_s(y, F)\sigma_y\sigma_F$ . Substitution of this expression in

Equation (24),  $\hat{y}_{HT} = \hat{y} + \frac{cov_s(F, y)}{\bar{F}}$ , yields  $\hat{y}_{HT} = \hat{y} + \frac{\rho_s(y, F)\sigma_y\sigma_F}{\bar{F}}$ . This expression can

be written as:  $\hat{y}_{HT} = \hat{y}\left[1 + \frac{\rho_s(y, F)\sigma_y\sigma_F}{\hat{y}\bar{F}}\right]$ .

Using  $cv(F) = \frac{\sigma_F}{\bar{F}}$  and  $cv(y) = \frac{\sigma_y}{\hat{y}}$  in the last expression for  $\hat{y}_{HT}$  we obtain the result.

Proof of Equation (30):  $\hat{v}_{LS}(\hat{y}_{HT} | F) = \hat{v}(\hat{\beta}_0)cv^2(F)$

From Equation (16) and (18), we have:

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\bar{F}v(\hat{\beta}_1) \quad (A1)$$

$$v(\hat{\beta}_0) = \bar{F}^2(1 + cv^2(F))v(\hat{\beta}_1) \quad (A2)$$

After applying the variance on Equation (29),  $\hat{y}_{HT} = \hat{\beta}_0 + \hat{\beta}_1\bar{F}[1 + cv^2(F)]$  and using (A1),

we obtain:

$$\begin{aligned} v_{LS}(\hat{y}_{HT} | F) &= v(\hat{\beta}_0) + \bar{F}^2(1 + cv^2(F))^2 v(\hat{\beta}_1) - 2\bar{F}v(\hat{\beta}_1)\bar{F}(1 + cv^2(F)) \\ &= v(\hat{\beta}_0) + \bar{F}^2(1 + cv^2(F))^2 v(\hat{\beta}_1) - 2\bar{F}^2(1 + cv^2(F))v(\hat{\beta}_1) \\ &= v(\hat{\beta}_0) + v(\hat{\beta}_1)\bar{F}^2(1 + cv^2(F))[1 + cv^2(F) - 2] \\ &= v(\hat{\beta}_0) + v(\hat{\beta}_1)\bar{F}^2(1 + cv^2(F))[cv^2(F) - 1] \end{aligned} \quad (A3)$$

We substitute  $v(\hat{\beta}_1) = \frac{v(\hat{\beta}_0)}{\bar{F}^2(1+cv^2(F))}$  from equation (A2) into (A3), resulting in:

$$\begin{aligned} v_{LS}(\hat{y}_{HT} | F) &= v(\hat{\beta}_0) + \frac{v(\hat{\beta}_0)\bar{F}^2(1+cv^2(F))[cv^2(F)-1]}{\bar{F}^2(1+cv^2(F))} \\ &= v(\hat{\beta}_0)[1+cv^2(F)-1] \\ &= v(\hat{\beta}_0)cv^2(F). \end{aligned}$$

Proof of Equation (31):  $\hat{v}_{LS}(\hat{y}_{HT} | F) = \frac{\hat{\sigma}_y^2}{n-2} [1 - \rho_s^2(y, F)] [1 + cv^2(F)].$

We will use the expression for the variance between elements given by:  $\hat{\sigma}_y = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}$

. From this equation, it is straightforward to obtain:

$$\sum_{i=1}^n y_i^2 = n\hat{\sigma}_y^2 + n\hat{y}^2 = n(\hat{\sigma}_y^2 + \hat{y}^2) \quad (\text{A4})$$

Recall from Equation (17) that  $\hat{v}(\hat{\beta}_0) = \hat{\sigma}_\varepsilon^2 \frac{\sum_{i=1}^n F_i^2}{n \sum_{i=1}^n (F_i - \bar{F})^2}$ . If we substitute (A4) into this

expression, using  $F$  instead of  $y$ , this yields:

$$\begin{aligned} \hat{v}(\hat{\beta}_0) &= \frac{\hat{\sigma}_\varepsilon^2}{n} \frac{n(\sigma_F^2 + \bar{F}^2)}{n\sigma_F^2} \\ &= \frac{\hat{\sigma}_\varepsilon^2}{n} (1 + cv^2(F)) \end{aligned} \quad (\text{A5})$$

Now, Eq. (19),  $\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$ , provides the estimator of the residuals. Using Equation

(20),  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 F_i$ , we have:



$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 F_i)^2}{n-2}.$$

Substituting Equation (13),  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{F}$ , into the last expression results in:

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &= \frac{\sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{F} - \hat{\beta}_1 F_i)^2}{n-2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_{i=1}^n (F_i - \bar{F})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (y_i - \bar{y})(F_i - \bar{F})}{n-2} \\ &= \frac{n}{n-2} (\hat{\sigma}_y^2 + \hat{\beta}_1^2 \sigma_F^2 - 2\hat{\beta}_1 \text{cov}(y, F)). \end{aligned} \quad (\text{A6})$$

From Equation (12),  $\hat{\beta}_1 = \frac{\text{cov}_s(y, F)}{\sigma_F^2}$ ; thus,  $\hat{\beta}_1 = \frac{\rho_s(y, F) \hat{\sigma}_y \sigma_F}{\sigma_F^2}$  and by substituting it in (A6)

we obtain:

$$\begin{aligned} &= \frac{n}{n-2} (\hat{\sigma}_y^2 + \rho_s^2(y, F) - 2\rho_s^2(y, F) \hat{\sigma}_y^2) \\ &= \frac{n}{n-2} (1 - 2\rho_s^2(y, F)) \hat{\sigma}_y^2 \end{aligned} \quad (\text{A7})$$

Now, we replace expression (A7) in (A5) and obtain the final result:

$$\hat{v}(\hat{\beta}_0) = \frac{\hat{\sigma}_y^2}{n-2} (1 - 2\rho_s^2(y, F)) (1 + \text{cv}^2(F)).$$