Advanced Methods in Survey Sampling

T. Merkouris

Athens University of Economics and Business

Graduate Programm in Statistics

Contents

- Estimation theory in survey sampling
  - Estimation of (finite) population parameters
  - Estimation of subpopulation parameters
- Using auxiliary information in estimation
  - Method of generalized regression
  - Calibration.
- Estimation of variance in complex surveys
- Treatment of non-sampling errors
  - no-response
  - imputation.

Populations, subpopulations, variables

A finite population U of size N:

set of discrete units

$$U = \{1, \ldots, N\},$$

e.g., persons, households, businesses, schools, farms etc.

In a finite population U

- subpopulations are defined as subsets $U_d \subset U$
- there is no intrinsic randomness

Populations, subpopulations, variables

A study characteristic of the population defines a variable (study/target variable) y, with value y_i for unit $i \in U$.

The variable y is **non-random** --- no probability distribution function is defined. The values of y_i, $i \in U$, are unknown but fixed.

In sample surveys there is customarily a very large number of variables etc (continuous and categorical),
e.g., income, labour status, education level etc.

Parameters of finite populations

A *parameter* θ of a population is defined as a function of the values of a variable y

$$\theta = \theta(y_1, \ldots, y_N)$$

or many variables, e.g., y and z

$$\theta = \theta(y_1, \ldots, y_N, z_1, \ldots, z_N)$$

Parameters of finite populations

The main parameters are:

A population total

$$Y = \sum_{i=1}^{N} y_i, \quad \acute{\eta} \quad \sum_{U} y_i.$$

e.g., total number of employed persons, total production of wine.

(Totals are not defined in non-finite populations.)

$$Y = \sum_{i=1}^{N} y_i = N, \quad \text{when } y_i = 1, i \in U$$

Parameters of finite populations

A mean

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

e.g, mean personal income.

A ratio of totals

$$R = \frac{Y}{Z} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} z_i}$$

e.g., total produce of wine to total vineyard area.

A proportion

$$P = \frac{N_d}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad y_i = \begin{cases} 1, & i \in U_d \\ 0, & i \notin U_d \end{cases}$$

Parameters of finite populations

Variance of y

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right]$$

The variance S^2 is used in the sampling design and in the computation of the stanard error of estimators.

Standard deviation

$$S = \sqrt{S^2}$$

coefficient of variation, (cv)

$$cv = \frac{S}{\bar{Y}}$$

Random/probability sample

A random sample is a subset of the population $s \subset U$, of size n, which is selected so that

- The set $S = \{s_1, \ldots, s_M\}$ of all possible distinct samples s is well defined
- A known probability of selection $p(s)$ is assigned to each sample $s \in S$
- Each unit $i \in U$ has a non-zero probability of inclusion in the sample s
- Each sample s is selected with probability $p(s)$, with specified mechanism of randomness which is called *random sampling technique*.

The function $p(s)$ defines a probability distribution in S

$$p(s) \geq 0, \quad \sum_{s \in S} p(s) = 1$$

The probability $p(s)$ specifies the probability of selecting each unit $i \in U$, and the statistical properties of the estimators of parameters.

Probabilities of inclusion in the sample

The probability $P(i \in s)$ of inclusion of unit i in a sample s is denoted by π_i and given by

$$\pi_i = \sum_{s \ni i} p(s), \quad i = 1, \ldots, N$$

Example 1:

Consider the population $U = \{1, 2, 3, 4\}$ and a sample s of size $n = 2$.

All possible samples of this size are

$$s_1 = \{1, 2\}, s_2 = \{1, 3\}, s_3 = \{1, 4\}, s_4 = \{2, 3\}, s_5 = \{2, 4\}, s_6 = \{3, 4\}$$

Then

$$\begin{aligned} \pi_1 &= p(s_1) + P(s_2) + p(s_3) \\ \pi_2 &= p(s_1) + P(s_4) + p(s_5) \\ \pi_3 &= p(s_2) + P(s_4) + p(s_6) \\ \pi_4 &= p(s_3) + P(s_5) + p(s_6) \end{aligned}$$

If $p(s_1) = p(s_2) = p(s_3) = p(s_4) = p(s_5) = p(s_6) = 1/6$, then

$\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1/2$

Probabilities of inclusion in the sample

The probability $P(i \in s)$ of inclusion of unit i in a sample s is denoted by π_i and given by

$$\pi_i = \sum_{s \ni i} p(s), \quad i = 1, \ldots, N$$

Example 1:

Consider the population $U = \{1, 2, 3, 4\}$ and a sample s of size $n = 2$.

All possible samples of this size are

$$s_1 = \{1, 2\}, s_2 = \{1, 3\}, s_3 = \{1, 4\}, s_4 = \{2, 3\}, s_5 = \{2, 4\}, s_6 = \{3, 4\}$$

Then

$$\begin{aligned} \pi_1 &= p(s_1) + P(s_2) + p(s_3) \\ \pi_2 &= p(s_1) + P(s_4) + p(s_5) \\ \pi_3 &= p(s_2) + P(s_4) + p(s_6) \\ \pi_4 &= p(s_3) + P(s_5) + p(s_6) \end{aligned}$$

If $p(s_1) = 1/3, p(s_2) = 1/6, p(s_6) = 1/2, p(s_3) = p(s_4) = p(s_5) = 0$, then

$\pi_1 = 1/2, \pi_2 = 1/3, \pi_3 = 2/3, \pi_4 = 1/2$

Probabilities of inclusion in the sample

The procedure of random sampling requires $\pi_i > 0$ for each $i \in U$.

The probability $P(i \in s, j \in s)$ of joint inclusion of units i and j is a sample s is denoted by π_{ij} and given by

$$\pi_{ij} = \sum_{s \ni i,j} p(s), \quad i, j = 1, \ldots, N$$

In example 1, we have $\pi_{13} = p(s_2) = 1/6$

Probabilities of inclusion in the sample

The inclusion of a unit i in a random sample s is expressed by the random *indicator* variable

$$I_i(s) = \begin{cases} 1, & i \in s \\ 0, & i \notin s \end{cases}$$

For the samples of example 1, $I_3(s_4) = 1 \;\; I_3(s_5) = 0$

The inclusion of two units i and j in the same random sample s is expressed by the product $I_i(s)I_j(s)$.

The variable $I_i(s)$ is the only **random** variable, defined for each unit $i \in U$.

Properties:

$$E(I_i(s)) = P(I_i(s) = 1) = \pi_i$$

$$V(I_i(s)) = \pi_i(1 - \pi_i), \quad C(I_i(s), I_j(s)) = \pi_{ij} - \pi_i\pi_j$$

Probabilities of inclusion in the sample

In sample surveys of finite populations the population units may have **unequal** probabilities π_i.

This is due to random sampling that uses knowledge of the structure of the population to reduce the statistical error in the parameter estimation.

Unequal probabilities of selection imply a distribution in the sample that is different from the distribution in the population.

Example 2:
Consider simple random sampling from a population U, with $N = 50000$ $n = 3000$, and with the same probability $\pi_i = n/N = 3/50$, $i \in U$
The distribution (histogram) of a variable y in the sample is similar to its distribution in the population (see 1st and 2nd graph).

Probabilities of inclusion in the sample

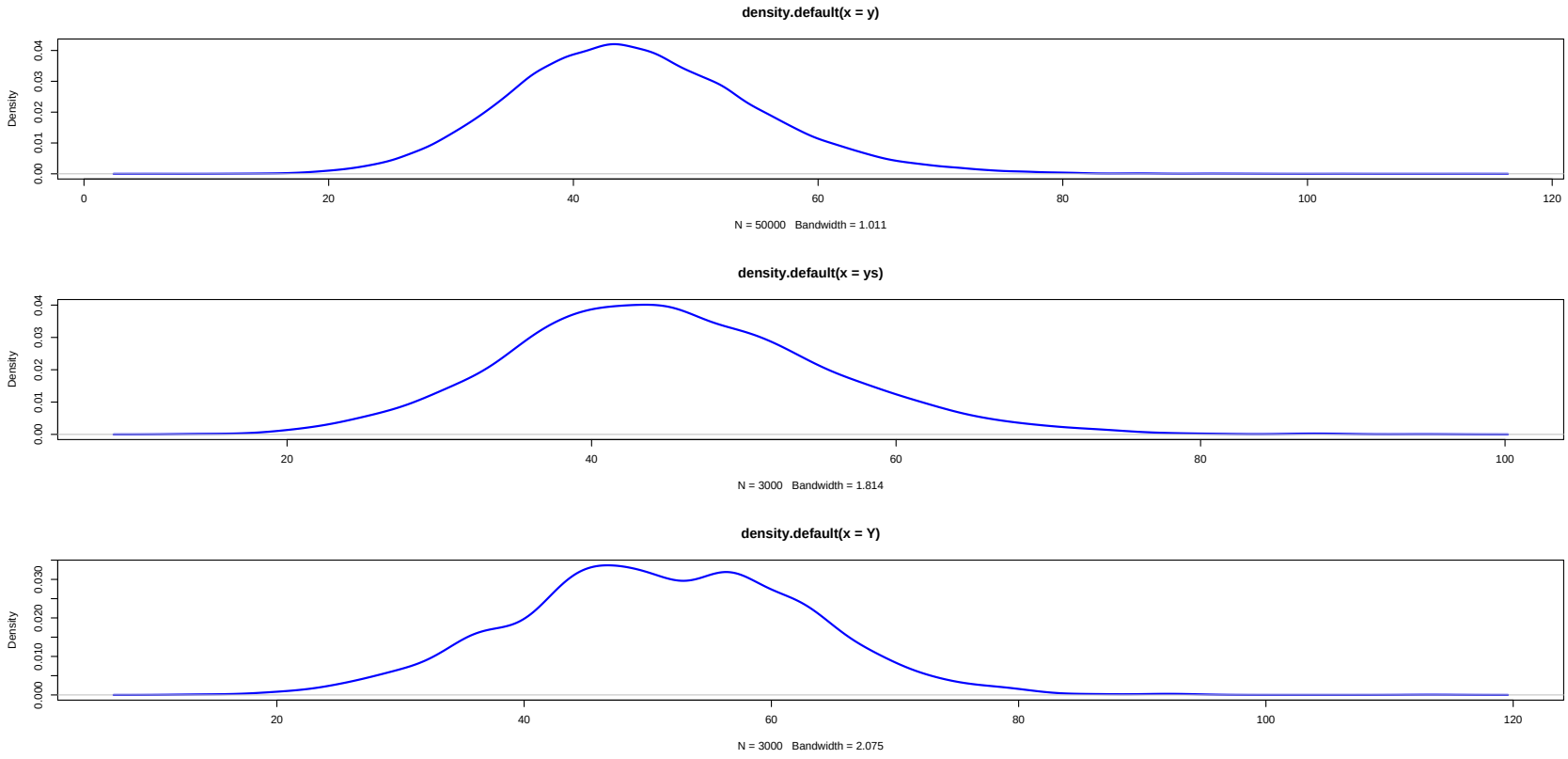
In sample surveys of finite populations the population units may have **unequal** probabilities π_i.

This is due to random sampling that uses knowledge of the structure of the population to reduce the statistical error in the parameter estimation.

Unequal probabilities of selection imply a distribution in the sample that is different from the distribution in the population.

Example 2:
Consider simple random sample from five strata defined by the values of y in ascending order, with stratum sizes
$N1 = 20000, N2 = 12000, N3 = 10000, N4 = 5000, N5 = 3000$, sample sizes $n_1 = \cdots = n_5 = n/5 = 600$ and probabilities of selection $n_i/N_i, i = 1, \ldots, 5$
The distribution of the variable y in the sample is different from its distribution in the population (see 3rd graph).

Probabilities of inclusion in the sample



Probabilities of inclusion in the sample

In sample surveys of finite populations the population units may have **unequal** inclusion probabilities π_i.

This is due to random sampling that uses knowledge of the structure of the population to reduce the statistical error in the parameter estimation.

Unequal probabilities of selection imply a distribution in the sample that is different from the distribution in the population.

The representativeness of the sample is restored with the use of the sampling weights (see next slide).

Sampling/design weights (αναγωγικοί συντελεστές)

The sampling weight of unit $i \in U$ is defined as

$$w_i = \frac{1}{\pi_i} I_i(s), \quad i \in U$$

$$E(w_i) = \frac{1}{\pi_i} E(I_i(s)) = \frac{1}{\pi_i} \pi_i = 1$$

The weight of a unit that has not been selected in the sample is by definition equal to zero.

The weight of a selected unit is the inverse of its selection probability.

Sampling/design weights (αναγωγικοί συντελεστές)

The interpretation of w_i:

The weight w_i of the sample unit i is interpreted as the number of population units (the unit i included) that are "represented" by the sample unit i.
Thus, the sample unit i represents itself and $w_i - 1$ non-selected population units, and all the sample units together represent the whole population.

Suppose that in example 1 the sample $s_3 = \{1, 4\}$ is selected with simple random sampling, so that $w_1 = 2$ and $w_4 = 2$. Then the sample s_3 produces the pseudo-population $\{1, 1, 4, 4\}$.

Sampling/design weights (αναγωγικοί συντελεστές)

Consider a population of twelve units, with the respective values of a variable y

$$U = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}, y_{12}\}.$$

Supoose that a simple random sample of size $n = 2$ from U is $s = \{y_3, y_8\}$. In this case we have $\pi = n/N = 2/12$ for all units of U, and $w = 6$ for both units of s.

In view of the interpretation of the weights, regarding the representativeness of the sample, it is possible to expand the sample creating the pseudo-population

$$\{y_3, y_3, y_3, y_3, y_3 y_3, y_8, y_8, y_8, y_8, y_8, y_8\},$$

where each of the the units 3 and 8 of the sample represents 6 units of the the population, with the same value of the variable y.

Sampling/design weights (αναγωγικοί συντελεστές)

Suppose now that a sample of size 4 from U is $s = \{y_2, y_5, y_8, y_{11}\}$. Here we have $\pi = n/N = 4/12$, and $w = 3$ for all four units of the sample.

Te pseudo-population that can be created based on this sample and the weights of the four sample units are

$$\{y_2, y_2, y_2, y_5, y_5, y_5, y_8, y_8, y_8, y_{11}, y_{11}, y_{11}\}$$

Obviously, a larger sample, with larger probability of inclusion (and smaller weight) has better representativeness.

Sampling/design weights (αναγωγικοί συντελεστές)

The "weighted" sample values $w_i y_i$ of a variable y restore the disproportionality of the sample, relative to the population, which is due to the unequal probabilities of selection of the sampling units.

The sampling weights are used in the inductive inference, i.e., using the sample to draw conclusions about the population.

Statistical theory for finite populations

Statistics for finite populations is primarily **descriptive**, i.e., it focuses on the estimation of population parameters.

The estimation of population parameters is based on a sample s of size n selected randomly with probability $p(s)$.

The uncertainty about the estimation is due to the fact that only a part of the population is surveyed.
Whereas the characteristics of the population remain fixed, their estimation depends on the chosen sample.

Statistical theory for finite populations

The estimator of a parameter $\theta = \theta(y_1, \ldots, y_N)$ is function of the random sample, and is denoted by

$$\hat{\theta} = \hat{\theta}(s) = \hat{\theta}(y_1, \ldots, y_n)$$

The estimator $\hat{\theta}(s)$ is random variable. The only random element is the set s that defines which units comprise the sample. The difference in the values of $\hat{\theta}(s)$ from sample to sample is the sample variance of $\hat{\theta}(s)$.

The inference for finite populations is based in the concept of the assumed repetition of the random sampling, with sampling design $p(s)$, which results in the selection of different samples.

According to this principle of assumed repetition of sampling, the expected value $E(\hat{\theta})$ of $\hat{\theta}(s)$ is given by

$$E(\hat{\theta}) = \sum_{s \in S} p(s)\hat{\theta}(s)$$

This is a weighted average (σταθμικός μέσος) of the possible values $\hat{\theta}(s)$ of $\hat{\theta}$, with the probabilities $p(s)$ as weights.

When $p(s)$ is constant for all samples $s \in S$, of the same size n, then

$$E(\hat{\theta}) = \frac{1}{M} \sum_{s \in S} \hat{\theta}(s),$$

where M is the total number of samples of size n.

The estimator $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$, i.e., if it is "on average" equal to the estimated parameter θ.

A measure of sampling variance of $\hat{\theta}$, denoted by $V(\hat{\theta})$, is given by

$$V(\hat{\theta}) = \sum_{s \in S} p(s) \left[\hat{\theta}(s) - E(\hat{\theta}(s)) \right]^2 .$$

The standard error (τυπικό σφάλμα) of $\hat{\theta}$ is defined as $\sqrt{V(\hat{\theta})}$.

A commonly used index of reliability of an unbiased estimator is its relative standard error, known as coefficient of variation (συντελεστής μεταβλητότητας), defined as $CV(\hat{\theta}) = \sqrt{V(\hat{\theta})}/\theta$ and expressed as percentage.

The estimation of the variance $V(\hat{\theta})$, denote by $\hat{V}(\hat{\theta})$, is computed using the survey data. An estimate of the coefficient of variation is $\sqrt{\hat{V}(\hat{\theta})}/\hat{\theta}$.

Parameter estimation

The probability function $p(s)$ is of theoretical interest, as a mathematical tool used in the foundation of the probability theory of survey sampling and in the derivation of the statistical properties of the estimators, but is usually quite complicated in its use.
In practice it is much easier to derive the expected value and the variance of estimators knowing only the probabilities π_i and π_{ij}.

Since the main parameters associated with a variable y are functions of the population total $Y = \sum_i y_i$, the methodology of estimation deals primarily with this basic parameter.

Parameter estimation

For a sample $s = \{y_1, \ldots, y_n\}$, the estimator Horvitz-Thompson of Y is defined as the linear combination (weighted sum)

$$\hat{Y} = \sum_{i=1}^{N} w_i y_i = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i$$

The estimator $\hat{Y}$ is the sum of the weighted values $w_i y_i$ of the variable y.

In the special case of total $Y = N$,

$$\hat{N} = \sum_{i=1}^{N} w_i = \sum_{i=1}^{n} \frac{1}{\pi_i}$$

Example: In simple random sampling, $\pi_i = n/N$, so that $w_i = 1/\pi_i = N/n$ and $\hat{N} = \sum_{i=1}^{n} \frac{1}{\pi_i} = N$

Parameter estimation

The estimator $\hat{Y}$ is unbiased, i.e., $E(\hat{Y}) = Y$.

$$E(\hat{Y}) = \sum_{i=1}^{N} E(w_i) y_i = \sum_{i=1}^{N} y_i = Y$$

The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

An unbiased estimator of $V(\hat{Y})$ computed using the sample $s = \{y_1, \ldots, y_n\}$ is given by

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

Parameter estimation

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

Because of the double sum this formula is difficult to use in practice. For specific sampling designs $p(s)$ a simplification of the formula for fast computations is possible.

Also, for some sampling designs $p(s)$ it is very difficult to compute the probabilities π_{ij}. Then, methods of approximate estimation of $V(\hat{Y})$ are used; they will be discussed later.

In simple random sampling, $\pi_i = n/N$, $\pi_{ij} = n(n-1)/N(N-1)$, and the formula for $\hat{V}(\hat{Y})$ takes the simple form

$$\hat{V}(\hat{Y}) = \frac{N(N-n)}{n} \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2, \qquad \text{where } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Parameter estimation

The estimator of the population mean $\bar{Y}$ is given by

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{1}{N}\sum_{i=1}^{N} w_i y_i = \frac{1}{N}\sum_{i=1}^{n} \frac{1}{\pi_i} y_i$$

An alternative estimator of $\bar{Y}$ that is used when is uknown is given by

$$\tilde{\bar{Y}} = \frac{\hat{Y}}{\hat{N}} = \frac{\sum_1^N w_i y_i}{\sum_1^N w_i} = \frac{\sum_1^N (1/\pi_i) y_i}{\sum_1^N 1/\pi_i}$$

In some sampling designs the estimators $\hat{\bar{Y}}$ and $\tilde{\bar{Y}}$ are identical. Even when N is known and the two estimators differ, the estimator $\tilde{\bar{Y}}$ is preferred because it usually has smaller variance.

Parameter estimation

The non-linear estimator $\tilde{\tilde{\bar{Y}}}$ is **approximately** (for large samples) unbiased. The **approximate** variance of $\tilde{\tilde{\bar{Y}}}$ is given by

$$\begin{aligned} V(\tilde{\tilde{\bar{Y}}}) &= \frac{1}{N^2} V\left[\sum_{i=1}^{n} w_i(y_i - \bar{Y})\right] \\ &= \frac{1}{N^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)(y_i - \bar{Y})(y_j - \bar{Y}). \end{aligned}$$

An estimator of $V(\tilde{\tilde{\bar{Y}}})$ is given by

$$\begin{aligned} \hat{V}(\tilde{\tilde{\bar{Y}}}) &= \frac{1}{\hat{N}^2} \hat{V}\left[\sum_{i=1}^{n} w_i(y_i - \tilde{\tilde{\bar{Y}}})\right] \\ &= \frac{1}{\hat{N}^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)(y_i - \tilde{\tilde{\bar{Y}}})(y_j - \tilde{\tilde{\bar{Y}}}) \end{aligned}$$

Parameter estimation

The estimator of a population proportion P is given by

$$\tilde{P} = \frac{\hat{N}_d}{\hat{N}} = \frac{\sum_1^{N_d} w_i}{\sum_1^N w_i} = \frac{\sum_1^{n_d} 1/\pi_i}{\sum_1^n 1/\pi_i},$$

where n_d is the size of the subset of s_d of s that corresponds to the subpopulation U_d.

Notice that $\tilde{P} = \tilde{\tilde{Y}}$ if we define $y_i = 1$ when $i \in U_d$ and $y_i = 0$ when $i \notin U_d$. Therefore, the arguments on approximate unbiasedness and variance of $\tilde{\tilde{Y}}$ apply to $\tilde{P}$.

In simple random sampling, the estimator is $\hat{P} = \hat{N}_d/N$, and

$$\hat{V}(\hat{P}) = \frac{N-n}{N(n-1)}\hat{P}(1-\hat{P})$$

Parameter estimation

The estimator of a population ratio $R = Y/Z$ is given by

$$\hat{R} = \hat{Y}/\hat{Z}$$

The non-linear estimator $\hat{R}$ is approximately (for large samples) unbiassed, with approximate variance

$$\begin{aligned} V(\hat{R}) &= \frac{1}{Z^2} V\left[\sum_{i=1}^{n} w_i(y_i - Rz_i)\right] \\ &= \frac{1}{Z^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)(y_i - Rz_i)(y_j - Rz_j). \end{aligned}$$

Parameter estimation

An estimator of $V(\hat{R})$ is given by

$$\begin{aligned}\hat{V}(\hat{R}) &= \frac{1}{\hat{Z}^2}\hat{V}\left[\sum_{i=1}^{n} w_i(y_i - \hat{R}z_i)\right] \\ &= \frac{1}{\hat{Z}^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\pi_{ij}}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)(y_i - \hat{R}z_i)(y_j - \hat{R}z_j).\end{aligned}$$

In simple random sampling the formula $\hat{V}(\hat{R})$ takes the simple form

$$\hat{V}(\hat{R}) = \frac{N}{\hat{Z}^2}\frac{N-n}{n}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{R}z_i)^2.$$

Parameter estimation

The estimation of the population variance of a variable y

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

is based on the equivalent expression of S^2 as function of totals:

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right].$$

Then

$$\hat{S}^2 = \frac{1}{\hat{N}-1}\left[\sum_{i=1}^{n} w_i y_i^2 - \frac{1}{\hat{N}}\left(\sum_{i=1}^{n} w_i y_i\right)^2\right] = \frac{1}{\hat{N}-1}\sum_{i=1}^{n} w_i (y_i - \tilde{\bar{Y}})^2,$$

where $\tilde{\bar{Y}} = \hat{Y}/\hat{N}$.

Parameter estimation

In the same way we can estimate the covariance of two variables y and z

$$\begin{aligned} S_{yz} &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})(z_i-\bar{Z}) \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i z_i - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} z_i\right)\right]. \end{aligned}$$

Then

$$\begin{aligned} \hat{S}_{yz} &= \frac{1}{\hat{N}-1}\left[\sum_{i=1}^{n} w_i y_i z_i - \frac{1}{\hat{N}}\left(\sum_{i=1}^{n} w_i y_i\right)\left(\sum_{i=1}^{n} w_i z_i\right)\right] \\ &= \frac{1}{\hat{N}-1}\sum_{i=1}^{n} w_i(y_i-\tilde{\bar{Y}})(z_i-\tilde{\bar{Z}}). \end{aligned}$$

Estimation for sub-populations

In large scale surveys, we may also be interested in estimation of parameters for various sub-populations (domains), which are **not** strata.

Let $U_d \subset U$ be a sub-population of size (number of units) N_d.

Suppose that we want to estimate the total $Y_d = \sum_{U_d} y_i$ of the sub-population for a variable y. We define a new variable y_d so that

$$y_{di} = \begin{cases} y_i, & i \in U_d \\ 0, & i \notin U_d \end{cases}$$

Estimation for sub-populations

Then Y_d is the total (for the whole U) for the new variable y_d, that is

$$Y_d = \sum_{U_d} y_i = \sum_{U} y_{di}.$$

Now we can use the general theory for the estimation of the total Y_d. Let then s be a sample from U and consider $s_d = s \cap U_d$, (i.e., the subset of s that belongs to the sub-population U_d).

The estimator of Y_d is

$$\hat{Y}_d = \sum_{s_d} w_i y_i = \sum_{s} w_i y_{di} \quad (= \sum_{s} \frac{1}{\pi_i} y_{di}).$$

Estimation for sub-populations

According to the general theory of estimation

$$E(\hat{Y}_d) = \sum_U E(w_i) y_{di} = Y_d$$

$$V(\hat{Y}_d) = \sum_U \sum_U \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_{di} y_{dj}$$

$$\hat{V}(\hat{Y}_d) = \sum_s \sum_s \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_{di} y_{dj}$$

Estimation for sub-populations

The HT estimator of N_d is

$$\hat{N}_d = \sum_{s_d} w_i \quad (= \sum_{s_d} \frac{1}{\pi_i})$$

The HT estimator of the mean $\bar{Y}_d = Y_d / N_d$ of the sub-population U_d is

$$\hat{\bar{Y}} = \hat{Y}_d / \hat{N}_d$$

The HT estimator P_d is

$$\hat{P}_d = \frac{\hat{N}_d}{N}$$

If N is unknown, then

$$\hat{P}_d = \frac{\hat{N}_d}{\hat{N}} = \frac{\sum_{s_d} w_i}{\sum_s w_i} = \frac{\sum_{s_d} 1/\pi_i}{\sum_s 1/\pi_i}.$$

Estimation for sub-populations

The size n_{s_d} of the sample s_d is random, and can be written as

$$n_{s_d} = \sum_{U} I(i \in s_d) = \sum_{U_d} I(i \in s)$$

and then $E(n_{s_d}) = \sum_{U_d} \pi_i$.

Example:
In the case of simple random sampling, when $\pi_i = n/N$, it follows that $E(n_{s_d}) = nN_d/N$, that is, the expected size of the sample s_d is proportional to the size of U_d.

Using auxiliary variables in estimation

In survey sampling it is very useful to define **auxiliary** variables, continuous or gategorical.

Examples: geographic place, gender, age,
land area, business type, etc.

The values of an auxiliary variable may be known before sampling, e.g., geographic place, or become known only for the the sampled units, e.g., the age of persons.

The auxiliary variabls are used:

- in the design of sampling
- in the definition of sub-populations
- in the improvement of the estimators of population parameters.

Using auxiliary variables in estimation

The estimator Horvitz-Thompson $\hat{Y}$ uses the sample values of y.

Suppose that we have information on an auxiliary vector variable $\mathbf{x}$ with p components, i.e., $\mathbf{x} = \left(x_1, \ldots, x_p\right)'$
either with the values $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)', i \in U$ or with the total $\mathbf{X} = \sum_U \mathbf{x}_i$.

This information (for the entire population) can be used for the improvement of the estimation of $\hat{Y}$ if y correlates with $\mathbf{x}$.

Using auxiliary variables in estimation

Suppose that the relationship of y with $\mathbf{x}$ is such that for each $i \in U$ the value y_i is approximated ("predicted") by the linear combination $\mathbf{x}_i^{'}\mathbf{B} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, i.e., $y_i \approx \mathbf{x}_i^{'}\mathbf{B}$.

A suitable vector coefficient $\mathbf{B}$ is determined with method of least squares, which minimizes the distance function (sum of squares) $\sum_U (y_i - \mathbf{x}_i^{'}\mathbf{B})^2$. This $\mathbf{B}$ is given by

$$\mathbf{B} = \left(\sum_U \frac{\mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_U \frac{\mathbf{x}_i y_i}{q_i},$$

where q_i are known constants (usually $q_i = 1$).

For univariate x (and $q_i = 1$) , the univariate B has the form

$$B = \frac{\sum_U x_i y_i}{\sum_U x_i^2}.$$

Using auxiliary variables in estimation

Suppose that the relationship of y with $\mathbf{x}$ is such that for each $i \in U$ the value y_i is approximated ("predicted") by the linear combination $\mathbf{x}_i^{'}\mathbf{B} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, i.e., $y_i \approx \mathbf{x}_i^{'}\mathbf{B}$.

A suitable vector coefficient $\mathbf{B}$ is determined with method of least squares, which minimizes the distance function (sum of squares) $\sum_U (y_i - \mathbf{x}_i^{'}\mathbf{B})^2$. This $\mathbf{B}$ is given by

$$\mathbf{B} = \left(\sum_U \frac{\mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_U \frac{\mathbf{x}_i y_i}{q_i},$$

where q_i are known constants (usually $q_i = 1$).

For univariate x (and $q_i = 1$) , the univariate B has the form

$$B = \frac{\sum_U x_i y_i}{\sum_U x_i^2}.$$

With this $\mathbf{B}$ a relationship of linear regression í is defined between y and $\mathbf{x}$, is the regression coefficient, and the differences $e_i = y_i - \mathbf{x}_i^{'}$ are the residuals of the linear regression.

Using auxiliary variables in estimation

The coefficient

$$\mathbf{B} = \left(\sum_{U} \frac{\mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_{U} \frac{\mathbf{x}_i y_i}{q_i},$$

is a population parameter. Estimation of $\mathbf{B}$ using a sample s is given by

$$\hat{\mathbf{B}} = \left(\sum_{s} \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_{s} \frac{w_i \mathbf{x}_i y_i}{q_i},$$

For the sample values of y the residuals are $\hat{e}_i = y_i - \mathbf{x}_i^{'} \hat{\mathbf{B}}$.

Asymptotically (i.e., for large samples) $\hat{\mathbf{B}}$ is approximatelly equal to $\mathbf{B}$.

The regression estimator

The **regression estimator** of the total Y is defined

$$\begin{aligned}\hat{Y}^{GR} &= \hat{Y} + \hat{\mathbf{B}}'(\mathbf{X} - \hat{\mathbf{X}}) \\ &= \sum_s w_i y_i + \sum_s \frac{w_i y_i \mathbf{x}_i'}{q_i} \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i)\end{aligned}$$

In expanded form

$$\hat{Y}^{GR} = \hat{Y} + \hat{\beta}_1(X_1 - \hat{X}_1) + \cdots + \hat{\beta}_p(X_p - \hat{X}_p)$$

The regression estimator

Properties of $\hat{Y}^{GR}$:

Asymptotically, $\hat{Y}^{GR}$ is given approximately by

$$\hat{Y}^{GR} \approx \hat{Y} + \mathbf{B}^{'}(\mathbf{X} - \hat{\mathbf{X}}).$$

It follows that $(\hat{Y}^{GR}) \approx Y$, i.e., $\hat{Y}^{GR}$ is approximately unbiased estimator of Y.

Alternatively,

$$\begin{aligned}
\hat{Y}^{GR} &\approx \mathbf{B}^{'}\mathbf{X} + (\hat{Y} - \mathbf{B}^{'}\hat{\mathbf{X}}) \\
&= \mathbf{B}^{'}\mathbf{X} + \sum_{s} w_i(y_i - \mathbf{x}_i^{'}\mathbf{B}) \\
&= \mathbf{B}^{'}\mathbf{X} + \sum_{s} w_i e_i
\end{aligned}$$

The regression estimator

It follows that asymptotically $V(\hat{Y}^{GR}) \approx V(\sum_s w_i e_i)$

Remark: $\sum_s w_i e_i$ is the estimator of the total $\sum_U e_i$.
Hence

$$V(\hat{Y}^{GR}) \approx \sum_U \sum_U \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) e_i e_j$$

An asymptotically unbiased estimator of $V(\hat{Y}^{GR})$ is

$$\hat{V}(\hat{Y}^{GR}) \approx \sum_s \sum_s \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \hat{e}_i \hat{e}_j$$

The estimator $\hat{Y}^{GR}$ has asymptotically smaller variance than the estimator $\hat{Y}$ if the residuals $e_i = y_i - \mathbf{x}_i' \mathbf{B}$ have smaller variance than the values y_i.

Special cases of the regression estimator

The **ratio estimator** estimator:

Consider a univariate x, such that approximately $y_i \approx Bx_i$, and let $q_i = x_i$.

Then $B = Y/X$ (why?), $\hat{B} = \hat{Y}/\hat{X}$, and from the general formula of $\hat{Y}^{GR}$ follows the estimator

$$\hat{Y}^R = \hat{Y}\frac{X}{\hat{X}} \quad (= X\hat{B})$$

The variance of $\hat{Y}^R$ follows from the general formula of $V(\hat{Y}^{GR})$.

The estimator $\hat{Y}^R$ has smaller variance than the estimator $\hat{Y}$ when the values y_i are scattered in small distance from straight line passing through zero (small residuals $e_i = y_i - x_i B$).

Special cases of the regression estimator

When $x_i = 1$, so that $y_i \approx B$, $X = N$ and $B = Y/N$, then

$$\hat{Y}^R = \hat{Y}\frac{N}{\hat{N}} \quad (= N\hat{B}).$$

This variant of the ratio estimator can be used in sampling designs in which $\hat{N} \neq N$.

Generalizations of the estimator $\hat{Y}^R$ are defined when different linear regressions $y_i \approx Bx_i$ are defined for different sub-populations that comprise the population U (see "poststratification" below).

Special cases of the regression estimator

Poststratified estimator:

Poststratification is the division of the sample in subsets (poststrata) which correspond to specific sub-populations.
Poststratification is done after the collection of the data, when the sampling units are identified as members of these sub-populations.

Example: In a survey of people, poststratification by specific age groups is possible if the age is one of the auxiliary information collected by the sample.

Special cases of the regression estimator

Let us consider poststrata $U_1, \ldots, U_G$ with respective known sizes $N_1, \ldots, N_G$ and samples $s_1, \ldots, s_G$.
Assume then different linear approximations by poststratum, $y_i \approx B_g$ for $i \in U_g$, so that $B_g = Y_g/N_g$ and $\hat{B}_g = \hat{Y}_g/\hat{N}_g$.

The poststratified estimator is defined as

$$\hat{Y}^{PS} = \sum_{g=1}^{G} \hat{Y}_g \frac{N_g}{\hat{N}_g} \qquad \left(= \sum_{g=1}^{G} N_g \hat{\bar{Y}}_g\right).$$

The variance of $\hat{Y}^{PS}$ follows from the general formula of $V\left(\hat{Y}^{GR}\right)$.

The estimator $\hat{Y}^{PS}$ is more efficient than $\hat{Y}^{R} = \hat{Y}\frac{N}{\hat{N}}$ when B_g differ significantly.

The asymptotic properties of $\hat{Y}^{PS}$ require adequately large sizes n_g or small number of poststrata.

Important properties of $\hat{Y}^{GR}$

The estimator $\hat{Y}^{GR}$ can be written alternatively as

$$\begin{aligned}
\hat{Y}^{GR} &= \sum_s w_i y_i + \sum_s \frac{w_i y_i \mathbf{x}_i'}{q_i} \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i) \\
&= \sum_s w_i \left[1 + \frac{\mathbf{x}_i'}{q_i} \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i) \right] y_i \\
&= \sum_s c_i y_i
\end{aligned}$$

where $c_i = w_i g_i$, with $g_i = 1 + \frac{\mathbf{x}_i'}{q_i} \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i)$

The estimator has $\hat{Y}^{GR}$ a linear form, with respect to y_i, like the HT estimator $\hat{Y} = \sum_s w_i y_i$. The weights c_i are independent of y.

Important properties of $\hat{Y}^{GR}$

Substituting y_i with $\mathbf{x}_i$ it follows from the formula

$$\hat{Y}^{GR} = \sum_s w_i y_i + \sum_s \frac{w_i y_i \mathbf{x}_i^{'}}{q_i} \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i)$$

that the regression estimator of the total $\mathbf{X}$ is $\hat{\mathbf{X}}^{GR} = \mathbf{X}$.

In this case, the weights c_i are "`calibrated" to the known population total $\mathbf{X}$, that is $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$.

Calibration

Calibration is a procedure of adjusting the sampling weights in the linear form $\sum_s w_i y_i$ so that the new weights c_i (the calibrated weights) satisfy the calibration constraint $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$, i.e., the estimation procedure reproduces exactly known population totals $\mathbf{X}$.

The same weights c_i produce the calibration estimator

$$\hat{Y}^C = \sum_s c_i y_i$$

of the total Y of any variable y.

Calibration

It follows easily that

$$\hat{Y}^C = \hat{Y} + \sum_s (c_i - w_i) y_i.$$

The estimator $\hat{Y}^C$ will be approximately unbiased if $E\left[\sum_s (c_i - w_i) y_i\right] \approx 0$, i.e., if the differences $c_i - w_i$ are small.

The proper weights c_i can be determined through the minimization of the distance function $\sum_s q_i (c_i - w_i)^2 / w_i$ under the constraint $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$. Usually $q_i = 1$.
The minimization gives

$$c_i = w_i g_i, \text{ where } g_i = 1 + \frac{\mathbf{x}_i'}{q_i} \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i),$$

i.e.,

$$\hat{Y}^C = \hat{Y}^{GR}$$

Calibration

The basic objective of calibration is the consistency of specific estimates with the corresponding population totals which are already known from other sources (e.g., administrative, census etc).

The calibration produces an estimator of linear form, identical to the regression estimator, without using any assumption of linear relationship (regression) of y with $\mathbf{x}$.

The adjustment factors g_i $c_i = w_i g_i$ depend on the observations $\mathbf{x}_i$, but are independent of y. They can viewed as a measure of difference between sample and population. It holds that $g_i \to 1$ when $n \to N$.

Since $\hat{Y}^C = \hat{Y}^{GR}$, the estimator $\hat{Y}^C$ can also be written as $\hat{Y}^C = \hat{Y} + \hat{\mathbf{B}}'(\mathbf{X} - \hat{\mathbf{X}})$.

Calibration

Special case 1:

Consider the categorical variable $\mathbf{x}$ with p categories which correspond to a partition of the population into p population groups $U_1, \ldots, U_p$. Suppose that the sizes of these groups, $N_1, \ldots, N_p$ are known.

The value of the variable $\mathbf{x}$ for unit $i \in U$ is defined as

$$\mathbf{x}_i = (\delta_{i1}, \ldots, \delta_{ip},)', \qquad \text{where } \delta_{ij} = \begin{cases} 1, & i \in U_j \\ 0, & i \notin U_j \end{cases}$$

and indicates which group the unit i belongs to.

Then the total of $\mathbf{x}$ is

$$\sum_{i \in U} \mathbf{x}_i = (N_1, \ldots, N_p)'.$$

Calibration

Let s be a random sample from U, and $s_j = s \cap U_j$ be set of the sample units that belong to gategory (group) j. Then

$$\sum_{i \in s} w_i \mathbf{x}_i = (\hat{N}_1, \ldots, \hat{N}_p)', \qquad (\hat{N}_j = \sum_{i \in s_j} w_j)$$

Calibration that satisfies the constraint

$$\sum_{i \in s} c_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i, \qquad \text{i.e., } (\hat{N}_1^c, \ldots, \hat{N}_p^c)' = (N_1, \ldots, N_p)')$$

gives $g_i = N_j / \hat{N}_j$ if $i \in U_j$ (so that $c_i = w_i N_j / \hat{N}_j$) and the calibration estimator of the total Y, of any variable y, is given by

$$\hat{Y}^C = \sum_{i \in s} c_i y_i = \sum_{j=1}^{p} \sum_{i \in s_j} c_i y_i = \sum_{j=1}^{p} \frac{N_j}{\hat{N}_j} \sum_{i \in s_j} w_i y_i = \sum_{j=1}^{p} \hat{Y}_j \frac{N_j}{\hat{N}_j},$$

where $\hat{Y}_j$ is the estimator of the total of y for category j.

Calibration

Remarks:

- In this special case of calibration, the estimator $\hat{Y}^C$ is the same as the poststratified estimator!
- The estimator $\hat{Y}^C$ has simple form and its construction is simple.
- Since $\hat{Y}^C = \hat{Y}^{GR}$, the estimator $\hat{Y}^C$ can be written as $\hat{Y}^C = \hat{Y} + \hat{\mathbf{B}}'(\mathbf{N} - \hat{\mathbf{N}})$, where $\mathbf{N} = (N_1 \ldots, N_p)'$, $\hat{\mathbf{N}} = (\hat{N}_1 \ldots, \hat{N}_p)'$.

  Analytically: $\hat{Y}^C = \hat{Y} + \hat{B}_1(N_1 - \hat{N}_1) + \cdots + \hat{B}_p(N_1 - \hat{N}_p)$
- Usual cases of such calibration: the p categories of the variable $\mathbf{x}$ are age groups (in surveys of persons), or types of businesses (in business surveys).

Special case 2:

Consider the categorical variable $\mathbf{x}$ with $p + q$ categories which correspond to two different partitions of a population into p groups $U_{11}, \ldots, U_{1p}$ and q groups $U_{21}, \ldots, U_{2q}$. Suppose that the sizes of the groups $N_{11}, \ldots, N_{1p}$ and $N_{21}, \ldots, N_{2q}$, respectively, are known.

Example: Partition of population of persons by age groups and geographic regions.

The value of the variable $\mathbf{x}$ for unit $i \in U$ is defined as

$$\mathbf{x}_i = \left(\delta_{i11}, \ldots, \delta_{i1p}, \delta_{i21}, \ldots, \delta_{i2q}\right)',$$

where $\delta_{i1j} = \begin{cases} 1, & i \in U_{1j} \\ 0, & i \notin U_{1j} \end{cases}$ $\delta_{i2k} = \begin{cases} 1, & i \in U_{2k} \\ 0, & i \notin U_{2k} \end{cases}$

Calibration

The total of $\mathbf{x}$ is

$$\sum_{i \in U} \mathbf{x}_i = (N_{11}, \ldots, N_{1p}, N_{21}, \ldots, N_{2q})'.$$

Let s be a random sample from U, and $s_{1j} = s \cap U_{1j}$, $s_{2k} = s \cap U_{2k}$ are the partitions of the sample that correspond to the partitions of the population. Then

$$\sum_{i \in s} w_i \mathbf{x}_i = (\hat{N}_{11}, \ldots, \hat{N}_{1p}, \hat{N}_{21}, \ldots, \hat{N}_{2q})', \quad (\hat{N}_{1j} = \sum_{i \in s_{1j}} w_i), \quad (\hat{N}_{2k} = \sum_{i \in s_{2k}} w_i)$$

Calibration that satisfies the constraints

$$\sum_{i \in s} c_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$$

equates the sizes of the $p + q$ groups with their estimates obtained from sample s.

Calibration

In this special case, the calibrated weights c_i do not have a simple form, but $\hat{Y}^C$ can be written as

$$\hat{Y}^C = \hat{Y} + \hat{\mathbf{B}}_1'(\mathbf{N}_1 - \hat{\mathbf{N}}_1) + \hat{\mathbf{B}}_2'(\mathbf{N}_2 - \hat{\mathbf{N}}_2),$$

where $\mathbf{N}_1 = (N_{11} \ldots, N_{1p})'$, $\hat{\mathbf{N}}_1 = (\hat{N}_{11} \ldots, \hat{N}_{1p})'$ and $\mathbf{N}_2 = (N_{21} \ldots, N_{2q})'$, $\hat{\mathbf{N}}_2 = (\hat{N}_{21} \ldots, \hat{N}_{2q})'$.

Calibration

Crossing of the two partitions produces $p \times q$ groups U_{jk}, $j = 1, \ldots, p$, $k = 1, \ldots, q$ with associated sizes N_{jk}. This is equivalent to single partition of U into $p \times q$ groups with respect to the two characteristics concurrently, e.g., with classification of each unit of a population of persons by geographic region and age group.

If te sizes N_{jk} are known, the calibration that equates $\hat{N}_{jk}$ to N_{jk} ($p \times q$ equations) reduces to the first special case, which produces the simple poststratified estimator.

Calibration

Remarks:

In large scale surveys, calibration is carried out for multiple partition of the population, e.g., by gender, by age groups and by geographic regions. The calibration procedure is then an extension of the procedure for *case* 2.

Calibration for multiple partition with a large number of groups implies large number of calibration constraints (equations of estimates with totals). This may have undesirable consequences:

- some negative clibration factors g_i, and hence negative calibration weights $c_i = w_i g_i$.
- the sample in the different groups is not adequate for the asymptotic properties of $\hat{Y}^C$.

Calibration in estimation for sub-populations

Suppose that calibration has been carried out using some vector auxiliary variable $\mathbf{x}$, so that $\sum_s c_i \mathbf{x}_i = \sum_U \mathbf{x}_i$.
Then the calibrated weights c_i can be used to estimate the total Y (or any parameter associated with any variable y) and for any U_d.

The estimation procedure is the same as in the production of the estimate of $\hat{Y}_d$, but using the weights c_i instead of w_i, that is

$$\hat{Y}_d^C = \sum_s c_i y_{di}.$$

Calibration in estimation for sub-populations

Using the analytic expression

$$c_i = w_i g_i = w_i \left[1 + \frac{\mathbf{x}_i^{'}}{q_i} \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} (\mathbf{X} - \sum_s w_i \mathbf{x}_i) \right]$$

the estimator $\hat{Y}_d^C$ can take the alternative form of a regression estimator

$$\hat{Y}_d^C = \hat{Y}_d + \hat{\mathbf{B}}_d^{'}(\mathbf{X} - \hat{\mathbf{X}}),$$

where $\hat{\mathbf{B}}_d = \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_s \frac{w_i \mathbf{x}_i y_{di}}{q_i} = \left(\sum_s \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_{s_d} \frac{w_i \mathbf{x}_i y_i}{q_i}.$

Calibration in estimation for sub-populations

Remarks:

The calibration has not been done specifically for U_d, i.e., $\sum_{s_d} c_i \mathbf{x}_i \neq \sum_{U_d} \mathbf{x}_i$.

In regression terms this means that the auxiliary variable $\mathbf{x}$ has been used for the improvement of the estimation at the level of the population U and not at the level of the sub-population U_d.

The result of this is that the improvement of the estimation of Y_d is small or negligible (the smaller the U_d the smaller the improvement).

Calibration in estimation for sub-populations

If we want calibration at the level of U_d, i.e., $\sum_{s_d} c_i \mathbf{x}_i = \sum_{U_d} \mathbf{x}_i$, or we want to improve te estimation of Y_d, the the procedure of constructing $\hat{Y}_d^C$ is restricted to s_d, and then

$$\hat{Y}_d^C = \hat{Y}_d + \hat{\mathbf{B}}_d^{'}(\mathbf{X}_d - \hat{\mathbf{X}}_d),$$

where $\hat{\mathbf{B}}_d = \left(\sum_{s_d} \frac{w_i \mathbf{x}_i \mathbf{x}_i^{'}}{q_i} \right)^{-1} \sum_{s_d} \frac{w_i \mathbf{x}_i y_i}{q_i}$.

Calibration at the level of U_d requires that the totals $\mathbf{X}_d$ be available, which may not be true (especially for very small U_d).

Also, for small U_d or/and for numerous auxiliary variables, the sample may not be large enough for the asymptotic properties of $\hat{Y}_d^C$ to hold.

Variance estimation in complex surveys

We saw in previous chapters that for any sampling design the variance of the estimator $\hat{Y}$ is given by

$$V(\hat{Y}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

If $\pi_{ij} > 0$ for all units $i, j \in U$, an unbiased estimator of $V(\hat{Y})$ that is computed using the sample $s = \{y_1, \ldots, y_n\}$ is given by

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

Variance estimation in complex surveys

As already shown, the variance of the important non linear functions of totals $\tilde{\bar{Y}} = \hat{Y}/\hat{N}$, $\hat{P} = \hat{N}_d/\hat{N}$, and $\hat{R} = \hat{Y}/\hat{Z}$ are calculated approximately (for large samples) using suitable variants of the basic formula of the variance of the estimator of a total.

Variance estimation in complex surveys

In practice, the general use of the basic formula

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j$$

is problematic for the following reasons:

- Because of the double sum the formula is computationally difficult in large samples.
- For many sampling designs it is very difficult (or impossible) to compute the probabilities π_{ij}.
- The formula cannot be used in the case of non-smooth functions of totals (e.g., median, quartiles).

In such problematic cases other, approximate, methods of estimation of the variance are used.

A simple approximate method of estimation of $V(\hat{Y})$:

An approximate estimator of $V(\hat{Y})$ is given by the formula

$$\tilde{V}(\hat{Y}) = \frac{1}{n(n-1)} \sum_s \left(\frac{y_i}{\pi_i/n} - \hat{Y} \right)^2$$

This simplified estimator is calculated as if sampling has been done with replacement (although in reality sampling has been done without replacement), thus circumventing the probabilities π_{ij} and the double sum.

In the case of simple random sampling we have $\tilde{V}(\hat{Y}) = \frac{N^2}{n(n-1)} \sum_s (y_i - \bar{y})^2$, where $\bar{y} = \sum_s y_i/n$. If the sampling ratio $f = n/N$ is very small, so that $1 - f \approx 1$, then $\tilde{V}(\hat{Y}) = \hat{V}(\hat{Y})$ (which is the variance given by the general formula).

This approximate method is applied also to stratified sampling, with H strata and stratum sample sizes n_h:

$$\tilde{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{s_h} \left(\frac{y_i}{\pi_i/n_h} - \hat{Y}_h \right)^2$$

The simplification of the calculations for $\tilde{V}(\hat{Y})$, compared to $\hat{V}(\hat{Y})$, is significant. However, $\tilde{V}(\hat{Y})$ is not an unbiased estimator of $V(\hat{Y})$.
In many cases, the bias is positive (overestimation), and then $\tilde{V}(\hat{Y})$ can be used as upper bound estimate of the variance $V(\hat{Y})$.

The general methodology

Let θ be a parameter, with estimator $\hat{\theta}$ calculated using a sample s.

Consider a number of (say K) proper subsets $s_1, \ldots, s_K$ of the sample s, and the different estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ that are calculated using these K different subsets .

An alternative estimator of θ which is based on the full sample s is the average

$$\hat{\theta}^{\star} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k$$

Resampling methods of variance estimation

Example 1:

$$\theta = Y, \qquad \hat{\theta} = \hat{Y} = \sum_{s} w_i y_i, \qquad \hat{\theta}_k = \hat{Y}_k = \sum_{s_k} w_i y_i, \quad k = 1, \ldots, K$$

$$\hat{\theta}^{\star} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k = \frac{1}{K} \sum_{k=1}^{K} \hat{Y}_k$$

Example 2:

$$\theta = \frac{Y}{Z}, \qquad \hat{\theta} = \frac{\hat{Y}}{\hat{Z}}, \qquad \hat{\theta}_k = \frac{\hat{Y}_k}{\hat{Z}_k}, \quad k = 1, \ldots, K$$

$$\hat{\theta}^{\star} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{Y}_k}{\hat{Z}_k}$$

Resampling methods of variance estimation

We consider two variance estimators of $\hat{\theta}^\star$:

$$\hat{V}_1 = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}^\star \right)^2$$

$$\hat{V}_2 = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta} \right)^2$$

It follows from the identity

$$\sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta} \right)^2 = \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}^\star \right)^2 + (\hat{\theta}^\star - \hat{\theta})^2$$

that $\hat{V}_2 \geq \hat{V}_1$.

Resampling methods of variance estimation

When the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are uncorrelated and have the same expected value, then $\hat{V}_1$ is unbiased estimator of $V(\hat{\theta}^\star)$.

Both $\hat{V}_1$ and $\hat{V}_2$ are used also for the estimation of $V(\hat{\theta})$, assuming that $V(\hat{\theta}^\star)$ and $V(\hat{\theta})$ are almost equal.

The approximate variance estimators in the rest of this chapter are of the form $\hat{V}_1$ and $\hat{V}_2$. But $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are usually correlated, and both $\hat{V}_1$ and $\hat{V}_2$ are then biased estimators of $V(\hat{\theta}^\star)$ and $V(\hat{\theta})$.

The method of Random Groups

Let s be a sample from population U that is partitioned in K non-overlapping random groups (sub-samples) $s_1, \ldots, s_K$, so that $s = \cup_{k=1}^{K} s_k$. These groups are not (statistical) independent.

We assume that s is partitioned by a random mechanism so that each random group s_k has the same sampling design with that of the full sample s.

Let $\hat{\theta}_1, \ldots, \hat{\theta}_K$ be estimators of θ, where $\hat{\theta}_k$ is calculated only with the data of s_k, $k = 1 \ldots, K$.

Resampling methods of variance estimation

We consider two alternative estimators of θ: the estimator $\hat{\theta}_{RG}$, which is the average

$$\hat{\theta}_{RG} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k,$$

and the estimator $\hat{\theta}$ which is calculated with the full sample s, without its partition to the K groups.

We consider the alternative variance estimators:

$$\hat{V}_{RG1} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}_{RG} \right)^2$$

and

$$\hat{V}_{RG2} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta} \right)^2$$

Resampling methods of variance estimation

Both $\hat{V}_{RG1}$ and $\hat{V}_{RG2}$ are estimators of $V(\hat{\theta}_{RG})$ and of $V(\hat{\theta})$, but they are not unbiased estimators.

The bias of $\hat{V}_{RG1}$ as estimators of $V(\hat{\theta}_{RG})$ is given by the formula

$$E(\hat{V}_{RG1}) - V(\hat{\theta}_{RG}) = -\frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{k \neq l} C(\hat{\theta}_k, \hat{\theta}_l).$$

If all the pairs have the same covariance (say C), then
$E(\hat{V}_{RG1}) - V(\hat{\theta}_{RG}) = -C$.

It holds that $\hat{V}_{RG2} \geq \hat{V}_{RG1}$.

When the sample s is selected with stratified sampling, then the method of random groups is applied to each stratum separately.

Resampling methods of variance estimation

Remarks:

The method of random groups is computationally very simple.

In many cases, the partition of the sample into random groups with the same sampling design is not simple.

The estimation of the variance of θ is unstable, i.e., it has large variance, because of the small number of random groups used in practice.

In practice the use of this method is not as common as other, more advanced, methods which are based on the concept of resampling.

The method Jackknife

Let s be a random sample from a population U, and $\hat{\theta}$ be the estimator of a parameter θ.

The sample s is partitioned in K random groups $s_1, \ldots, s_K$ of equal size $m = n/K$. These groups are random (sub)samples from the full sample s.

We assume that the selection of the sub-samples $s_1, \ldots, s_K$ is done with simple random sampling, even when the full sample s has not been selected with simple random sampling.

Resampling methods of variance estimation

With each sub-sample s_k, k, $k = 1, \ldots, K$, we associate an estimator $\hat{\theta}_{(k)}$, of the same type as the estimator $\hat{\theta}$, but calculated using the data that are left after the omission of s_k, i.e., using the data in $s - s_k$.

Note 1: The different sets $s - s_k, \quad k = 1, \ldots, K$ overlap.

Note 2: In the calculation of $\hat{\theta}_{(k)}$, the weights w_i for all $i \in s - s_k$ are multiplied by $K/(K-1)$ to counterbalance the loss of s_k.

For $k = 1, \ldots, K$ we define the "pseudo-value"

$$\hat{\theta}_k = k\hat{\theta} - (K-1)\hat{\theta}_{(k)}$$

The Jackknife estimator of θ is defined as the average of the pseudo-values $\hat{\theta}_k$

$$\hat{\theta}_{JK} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$$

The Jackknife variance estimator of $\hat{\theta}_{JK}$ is

$$\begin{aligned} \hat{V}_{JK} &= \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}_{JK} \right)^2 \\ &= \frac{K-1}{K} \sum_{k=1}^{K} \left(\hat{\theta}_{(k)} - \hat{\bar{\theta}} \right)^2 \end{aligned}$$

where $\hat{\bar{\theta}} = \sum_{k=1}^{K} \hat{\theta}_{(k)}/K$. Here $\hat{V}_{JK}$ is used as estimator of $V(\hat{\theta})$, and $V(\hat{\theta}_{JK})$.

For good accuracy of the variance estimator $\hat{V}_{JK}$ a sufficient number of sub-samples is required (large K). The maximum possible number of sub-samples is obtained in the special case where $K = n, m = 1$.

Resampling methods of variance estimation

Example: Sample s of size n with simple random sampling.

$$\pi = n/N, \quad w_i = N/n, \quad \theta = Y, \quad \hat{\theta} = \hat{Y} = \sum_s w_i y_i = (N/n) \sum_s y_i.$$

Consider K sub-samples $s_1, \ldots, s_K$ of equal sizec $m = n/K$. Then $K/(K-1) = n/(n-m)$.

$$\hat{\theta}_{(k)} = \frac{K}{K-1} \frac{N}{n} \sum_{s-s_k} y_i = \frac{N}{n-m} \sum_{s-s_k} y_i = N\hat{\bar{Y}}_{s-s_k}$$

$$\hat{\theta}_k = K\hat{\theta} - (K-1)\hat{\theta}_{(k)} = \frac{N}{m} \sum_{s_k} y_i = N\hat{\bar{Y}}_{s_k}$$

$$\hat{\theta}_{JK} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k = \frac{N}{n} \sum_s y_i = N\hat{\bar{Y}}_s \quad (= \hat{\theta})$$

Resampling methods of variance estimation

$$\hat{V}_{JK} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_k - \hat{\theta}_{JK} \right)^2 = \frac{N^2}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\bar{Y}}_{s_k} - \hat{\bar{Y}}_s \right)^2$$

$$= \frac{N^2(K-1)}{K} \sum_{k=1}^{K} \left(\hat{\bar{Y}}_{s-s_k} - \hat{\bar{Y}}_s \right)^2$$

When $K = n$ (and $m = 1$), then

$$\hat{V}_{JK} = \frac{N^2}{n(n-1)} \sum_{s} \left(y_i - \hat{\bar{Y}}_s \right)^2 = \frac{1}{1-f} \hat{V}(\hat{Y}),$$

where $\hat{V}(\hat{Y})$ is unbiased estimator of $V(\hat{Y})$ according to the general formula. In this case, the approximate variance $\hat{V}_{JK}$ is larger than the variance $\hat{V}(\hat{Y})$, with the difference approaching zero when $f = n/N$ approaches zero.

Resampling methods of variance estimation

Continuing the example, consider now the ratio estimator $\hat{\theta} = \hat{R} = (\hat{Y}/\hat{X})X = (\sum_s y_i / \sum_s x_i)X$, for simple random sampling.

For $K = n$, $m = 1$ we have $\hat{\theta}_{(k)} = (\sum_{s-k} y_i / \sum_{s-k} x_i)X$, $k = 1, \ldots, n$, where $s - k$ denotes the sample s without the unit k.
The Jackknife estimator of $V(\hat{\theta})$ is

$$\hat{V}_{JK} = \frac{n-1}{n} \sum_{k=1}^{n} (\hat{\theta}_{(k)} - \hat{\bar{\theta}})^2,$$

where $\hat{\bar{\theta}} = \sum_{k=1}^{n} \hat{\theta}_{(k)}/n$.

For an application of the above example we consider the survey data (file cherry.csv in e-class) from a population of 2967 cherry trees. This survey, which is described in the book Lohr (2009), had as objective the estimation of the total timber volume (in cubic feet) for the population of the cherry trees.

Measurements of height, diameter and volume were done on a sample of 31 cherry trees which were selected with simple random sampling. The total diameter X of all cherry trees was known: $X = 41837$ feet. The very strong correlation between diameter and volume, $\rho = 0,96$, and the knowledge of X, justifies the use of the ratio estimator for the estimation of the total volume Y with auxiliary variable the diameter.

With common weight $N/n = 2967/31 = 95,71$ for all sample units, we calculate the estimates $\hat{Y} = 89517.26$, $\hat{X} = 39307.96$, and $\hat{R} = (\hat{Y}/\hat{X})X = 95272.16$. Also, we calculate the Jackknife variances $\hat{V}_{JK}(\hat{Y}) = 76729654$ and $\hat{V}_{JK}(\hat{R}) = 30765141$.

The relative difference of variances $(\hat{V}_{JK}(\hat{R}) - \hat{V}_{JK}(\hat{Y}))/\hat{V}_{JK}(\hat{Y}) = -0,599$ shows that the variance of the ratio estimator $\hat{R}$ is smaller than the variance of the estimator $\hat{Y}$ almost by 60%.

Jackknife for cluster sampling

Suppose that the sample s consists of K clusters, which form a random sample from a population of N_c clusters.

In this case the clusters are the random groups of the sample on which the Jackknife method is applied.

Then if $\hat{\theta}$ is the estimate calculated with the full sample s, the estimate $\hat{\theta}_{(k)}$ is the estimate calculated without cluster k.

In multistage sampling, the Jackknife method is applied to the first stage of sampling, with the K primary sampling units (PSU) forming the random groups of the sample, regardless of the number of secondary or tertiary units.
The estimate $\hat{\theta}_{(k)}$ is calculated with the data left after the omission of the PSU k.

Jackknife for stratified sampling

Consider stratified sampling with H strata, and suppose that the sample in stratum h $(h = 1, \ldots, H)$ is partitioned randomly into K_h groups (sub-samples).

Let $\hat{\theta}$ be the estimate of θ which is calculated with the full sample s. Then $\hat{\theta}_{(hk)}$ is the estimate of θ which is calculated with the data left in the sample after the omission of group k in stratum h.

The Jackknife estimator of $V(\hat{\theta})$ is

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{K_h - 1}{K_h} \sum_{k=1}^{K_h} \left(\hat{\theta}_{(hk)} - \hat{\theta}\right)^2$$

Jackknife for stratified multistage sampling

In stratified multistage sampling the Jackknife method is applied separately to each stratum in the first stage of sampling, with random groups the K_h selected PSU in stratum h $(h = 1, \ldots, H)$. The K_h PSU, denoted by s_{hk}, form the sample s_h of stratum h, i.e., $s_h = \bigcup_{k=1}^{K_h} s_{hk}$.

Let $\hat{\theta}_{(hk)}$ be the estimate of θ when the units of PSU k in stratum h (i.e., s_{hk}) is omitted. Then, in the calculation of $\hat{\theta}_{(hk)}$ the weights of the units in the rest of the PSU in stratum h are inflated by $K_h/(K_h - 1)$, and the weights of the units in the rest of the strata do not change.

Analytically, if w_i denotes generally the weight of unit i, in any stratum and PSU, then the adjusted weights for the calculation of $\hat{\theta}_{(hk)}$ are defined as follows:

$$w_{(hk)i} = \begin{cases} w_i, & \text{if, } i \notin s_h \\ \frac{K_h}{K_h - 1} w_i, & \text{if, } i \in s_h - s_{hk} \\ 0, & \text{if , } i \in s_{hk} \end{cases}$$

Then the Jackknife estimator of $V(\hat{\theta})$ is

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{K_h - 1}{K_h} \sum_{k=1}^{K_h} \left(\hat{\theta}_{(hk)} - \hat{\theta}\right)^2$$

Example:

Estimation of total, where $\hat{\theta} = \hat{Y} = \sum_s w_i y_i$.
The replicate $\hat{Y}_{(hk)}$ is

$$\hat{Y}_{(hk)} = \sum_s w_{(hk)i} y_i = \sum_{s-s_h} w_i y_i + \sum_{s_h - s_{hk}} \frac{K_h}{K_h - 1} w_i y_i$$

$\hat{V}_{JK}(\hat{Y}) = ?$

Resampling methods of variance estimation

The resampling procedure, involving the omission of a (PSU) s_{hk} for the calculation of the estimate $\hat{\theta}_{(hk)}$, consecutively for all K_h PSU of each stratum h, is independent of the variable and parameter of interest.

In practice, the ommision of s_{hk} from the full sample s is done implicitly with the adjustment of $w_i = 0$ for each unit $i \in s_{hk}$, (and with the proper adjustment of the weights of the rest of the units of s as described above). In this way, as many sets of adjusted weights for all units of s are created as are in total the PSU in s.

These sets of weights are used for the calculation of the estimate $\hat{\theta}_{(hk)}$ of any parameter $\theta_{(hk)}$ for any variable, in the same way that the full sample estimate $\hat{\theta}$ is calculated, and may form additional columns in the data file.

Resampling methods of variance estimation

Remarks:

The Jackknife method is useful for the estimation of the variance of the estimator of any parameter. However, it is not satisfactory when the estimated parameter is not a smooth function of totals (e.g., median and quartiles).

When there is a large number of strata, with many PSU per stratum, the method requires many calculations.

The method can be applied also to the calibration estimator $\hat{\theta}^C$. To this end, in each repetition calibration is carried out again with the reduced sample for the production of the calibration estimator $\hat{\theta}^C_{(k)}$, which is calculated with the calibrated weights.

The method Bootstrap

The method Bootstrap uses resampling for the selection of a number of sub-samples (replicates), with replacement, from the full sumple.

These sub-samples are used to calculate replicate estimates of θ, and thereby estimate the variance $V(\hat{\theta})$.

This procedure is described for stratified multistage sampling.

Resampling methods of variance estimation

Let K_h be the number of PSU in the sample of stratum h. In each repetition of the procedure of selecting sub-samples, $K_h - 1$ PSU are selected with simple random sample with replacement.

This is done independently for each stratum $h = 1, \ldots, H$, and thus a bootstrap replicate is created, consisting of $\sum_{h=1}^{H}(K_h - 1)$ PSU.

This procedure is repeated R times, producing R bootstrap replicates.

Let $m_{hk}(r)$ be the number of times that PSU k of stratum h is selected in replicate r $(r = 1, \ldots, R)$. Note: $0 \leq m_{hk}(r) \leq K_h - 1$.

Resampling methods of variance estimation

In replicate r the sampling weights are adjusted as follows:

$$w_i(r) = w_i \frac{K_h}{K_h - 1} m_{hk}(r), \text{ for unit } i \text{ in PSU } k \text{ of stratum } h.$$

For each r let $\hat{\theta}_{(r)}$ be the replicate estimate of θ, calculated in the same way as the estimate $\hat{\theta}$ but using the weights $w_i(r)$ instead of the initial weights w_i.

Then, the bootstrap estimator of the variance $V(\hat{\theta})$ is

$$\hat{V}_B(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\theta}_{(r)} - \hat{\theta} \right)^2.$$

Resampling methods of variance estimation

As in the method of Jackknife, in the method of Bootstrap the repeated procedure of forming sub-samples (replicates) is independent of any variable and parameter.

In each repetition r the weights of all units are adjusted as described above (for the non-selected PSU the weights are zero), and thus a new set of weights is created for the full sample to be used for the calculation of the estimate $\hat{\theta}_{(r)}$ of any parameter θ for any variable, in the same way that the estimate $\hat{\theta}$ is calculated.

Resampling methods of variance estimation

Remarks:

The number R is arbitrary, but usually it is $R = 1000$ or $R = 500$ or smaller.

The method of Bootstrap gives good estimate of variance for both smooth and non-smooth functions of totals (e.g., quartiles).

The method of Bootstrap usually requires fewer calculations than the method of Jackknife.

When the procedure of estimation includes calibration, then in addition to the calibration of the weights of the full sample to calculate the estimate $\hat{\theta}^C$, the adjusted weights $w_i(r)$ must be calibrated in each relicate r to calculate the replicate estimate $\hat{\theta}^C_{(r)}$.

Unit nonresponse

Almost invariably in surveys, some sampling units do not respond completely, in the sense that none of the required information is collected from them.

Causes for non-response

The main causes of non-response include:

- inability to communicate with the sampling units
- absence
- inability to respond (e.g., language, illiteracy)
- illness
- inaccessible units
- refusal

Consequences of non-response

Possible bias, because of violation of the basic principle of randomness of the sample.
The non-responding units may be systematically different from the responding, so that the responding part of the sample not to be representative of the population.
The responding part of the sample is representative of the part of the population which would be responded to the survey, which is rarely the same with the entire survey population.

The size of bias depends on the relationship of the respondents' characteristics with the survey variables, and increases with the non-response rate.
The stronger the relationship of the value y_i of a variable y for unit i and the probability of non-response of the unit, the larger the bias of estimates related with this variable.

Nonresponse

For example, suppose that in a survey of personal income the persons of high income have higher probability of non-response than persons of low income.
The result will be that for variables related positively with income, persons with high values of these variables will not be represented adequately in the sample.
Anyway, regardless of this relationship, in the case of estimation of totals it is obvious the bias (underestimation) resulting from the loss of sampling units.

It should be noted that in any case of non-response the size of the bias cannot be estimated.

Nonresponse

The variance of the estimators is also affected by non-response. The loss of information due to non-response (information from fewer sampling units) results in an increase of the variance of an estimator if the variance of the associated variable in the responding part of the sample remains the same as the variance in the full sample (or is larger). However, this is not more likely to happen.

In the above example of higher non-response rate for persons of high income, the variance of income in the respondents is smaller than that in the full sample. Therefore, there will be bias in the estimation (underestimation) of the variance of estimators related to income.

Measure of response

Let n_α $(n_\alpha < n)$ be the size of the subset of the sample for which there is response. A measure of response is given by the response rate

$$p_a = \frac{n_\alpha}{n}.$$

This measure, usually expressed as *percentage,* indicates the degree of success in eliciting response from the units of the selected sample.

Nonresponse

An alternative measure of response is given by the weighted response rate

$$\tilde{p}_\alpha = \frac{\sum_{i=1}^{n_\alpha} w_i}{\sum_{i=1}^{n} w_i} = \frac{\hat{N}_\alpha}{\hat{N}},$$

where $w_i = 1/\pi_i$ is the weight of the responded unit i, and $\hat{N}_\alpha$ is an estimate of the number of population units which would respond given their selection in the sample.

To $\tilde{p}_\alpha$ is interpreted as an estimate of the average probability of response from the members of the population.

The measures p_a and $\tilde{p}_\alpha$ may differ substantially. They are, however, equivalent when the weights of all units are equal.

These two measures do not give the size of bias resulting from non-response. Low non-response may cause large bias if the relationship of non-response with survey variables is strong.

Handling the problem of non-response

1. Prevention of non-response

Prevention of the problem at the stage of designing the survey. Factors related to possible non-response include:

- Subject (content) of the survey
- design of the questionnaire
- selection, training and supervision of interviewers
- method of data collection
- time and conditions in conducting the survey
- frequency of data collection in the case of repeated survey.

2. Reduction of the non-response

Steps towards a reductionof non-response during the survey include:

- Call-backs, follow-ups
  Repeated efforts for communication
  Different days and times (in face to face or telephone interviews)
  Different methods of data collection (e.g., telephone follow-up in mail survey)
- Survey of non-respondents
  A random sample of non-responded units is selected, and a special effort is made to collect data from all its units.
  If this process is successful, it may achieve unbiasdeness with suitable methodology. However, this is a time consuming process, and because of that it is rarely implemented.

3. Adjustment of the weights of the responded units

The objective of such an adjustment is to increase the weights of the responded units so that these units represent the non-responded units as well.
Observe first that whereas the estimate of the population size N based on the full sample

$$\hat{N} = \sum_{i=1}^{n} w_i,$$

is unbiased, and for some sampling designs is exactly equal to N, the estimate based on the reduced sample becomes

$$\hat{N}_\alpha = \sum_{i=1}^{n_\alpha} w_i,$$

which is an underestimation of N because of the loss of $n - n_\alpha$ sample units.

This observation suggests the way of adjusting the weights for restoring the unbiasedness of the estimator of N.
Specifically, the weights of all the responded units are multiplied by the common factor $\hat{N}/\hat{N}_\alpha$, which is the inverse of the weighted response rate $\tilde{p}_\alpha$.
Thus, the adjusted weight of each responded unit i is

$$\tilde{w}_i = w_i \frac{1}{\tilde{p}_\alpha} = w_i \frac{\hat{N}}{\hat{N}_\alpha} = w_i \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n_\alpha} w_i}.$$

Then, the estimator resulting from the use of the adjusted weights is

$$\tilde{N} = \sum_{i=1}^{n_\alpha} \tilde{w}_i = \frac{1}{\tilde{p}_\alpha} \sum_{i=1}^{n_\alpha} w_i = \frac{1}{\tilde{p}_\alpha} \hat{N}_\alpha = \hat{N},$$

that is, the unbiased estimator that would be obtained from the full sample.

The interpretation of $\tilde{p}_\alpha$ as estimated average probability of response of the population units, leads to the interesting view of the adjusted weight of each selected and responded unit as the inverse of the probability of selection and response of this unit, i.e., $\tilde{w}_i = 1/\pi_i \tilde{p}_\alpha$.

The adjusted weights are used in the estimation of any other parameter, but they do not restore the unbiasedness of the estimation. So, the estimator of the total Y

$$\tilde{Y} = \sum_{i=1}^{n_\alpha} \tilde{w}_i y_i = \frac{1}{\tilde{p}_\alpha} \sum_{i=1}^{n_\alpha} w_i y_i$$

is not unbiased, but its bias may be lower than the one that would result without the weight adjustment.

Nonresponse

The estimated average probability of response $\tilde{p}_\alpha$ does not depend on the survey variables, and is the same for all the responded units of the sample.

If the true probabilities of response are the same for all the population units, and the response of each unit is independent from the response of the other units, then the non-respondents are as if the have been selected randomly from the sample, and the respondents constitute a representative sample.
The assumption of such response mechanism is implicit when the non-response is not taken into consideration.

In this unrealistic case, the adjusted weights produce an unbiased estimator of total $\tilde{Y} = \sum_{i=1}^{n_\alpha} \tilde{w}_i y_i$ for any variable y.

Nonresponse

More realistic is the assumption that for a partition of the sample to different classes according to some characteristic(s), the probability of response is (almost) the same for the units that belong to the same class.

For such a class an estimate of this probability is the weighted rate of response of the units of the class, which can be used for the adjustment of the weights of all responded units of the class. In other words a different $\tilde{p}_\alpha$ is used for each class.

Let K be the number of such classes, of size N_k, $(k = 1, \ldots, K)$, and let $n_{k\alpha}$ be the number of respondents in class k. Then

$$\tilde{Y} = \sum_{k=1}^{K} \tilde{Y}_k = \sum_{k=1}^{K} \frac{1}{\tilde{p}_{k\alpha}} \sum_{i=1}^{n_{k\alpha}} w_i y_i = \sum_{k=1}^{K} \frac{\hat{N}_k}{\hat{N}_{k\alpha}} \sum_{i=1}^{n_{k\alpha}} w_i y_i.$$

The weights of the respondents in class k are increased uniformly by $1/\tilde{p}_{k\alpha} = \hat{N}_k/\hat{N}_{k\alpha}$, so that the non-respondents of this class are also represented in the sample.

Nonresponse

Example: Assume that for each unit of a sample of persons the age is known, and that the sample is partitioned into age classes as shown in the table.

	Age					
	15-24	25-34	35-44	45-64	65+	óc
n_k	202	220	180	195	203	1000
$n_{k\alpha}$	124	187	162	187	203	863
$\sum_{i=1}^{n_k} w_i$	30322	33013	27046	29272	30451	150104
$\sum_{i=1}^{n_{k\alpha}} w_i$	18693	28143	24371	28138	30451	129796
$\tilde{p}_{k\alpha}$	0,6165	0,8525	0,9011	0,9613	1,000	
$1/\tilde{p}_{k\alpha}$	1,622	1,173	1,110	1,040	1,000	

The weight of each respondent of age between 15 and 24 is multiplied by 1,622, and likewise for the weights of the respondents in the other age classes. In the class 65+ there was full response, and thus the weights did not change. Note that $1/\tilde{p}_\alpha = \hat{N}/\hat{N}_\alpha = 1{,}156$.

Nonresponse

In practice, for substantial reduction of the bias the classes should be specified so that the units in each class to be as similar as possible, with respect to main variables, and the weighted rates of response in different classes to be as dissimilar as possible.

Common characteristics of specification of the classes are geographic variables, and other variables that exist in the survey frame. For example, in a business survey, variables such as type of of business and size of business.

Nonresponse

Also, characteristics specified by *paradata*, that is, data produced by the sampling process. These may be data recorded by the interviewers, such as day and time of call for the interview, approach tactic, outcome of call, etc.
Other paradata are observations for each sampled household, such as dwelling type, security system, indication of presence of children, observations for the neighborhood, etc.

Often, to facilitate the process, the strata of the sample are used as classes.

Nonresponse

The adjustment of the weights for nonresponse may cause increase of the variance of estimates when the classes of adjustment are numerous and contain few sampling units.

Then, the probabilities of response are not estimated accurately, which results in an increase of the variance of estimates. Also, regardless of the number of classes, some of them may require large adjustment factors, the result of which is an increase of the variance of the estimates.

Methods of a different adjustment of the weight of each unit --- using an estimate of the probability of its response --- which are based on the use of auxiliary variables and on model assumptions, exist in the bibliography but they are rarely used in practice.

The estimator

$$\tilde{Y} = \sum_{k=1}^{K} \frac{\hat{N}_k}{\hat{N}_{k\alpha}} \sum_{i=1}^{n_{k\alpha}} w_i y_i$$

is of the same form as the poststratified estimator. The difference is that in poststratification the sizes of the poststrata N_j are known, whereas in the adjustment of the weights of the respondents by class the sizes of the classes N_k are unknown and estimated by $\hat{N}_k$.

Remark: In the adjustment of the weights of the respondents by class the adjustment factors $\hat{N}_k/\hat{N}_{k\alpha}$ are always greater than one, whereas in poststratification the adjustment factors may be any positive number (why?)

Nonresponse

Poststratification is a form of adjustment of the weights for non-response, with the weights of each poststratum g multiplied by $N_g/\hat{N}_{ga}$, where N_g is the size of the poststratum g and $\hat{N}_{ga}$ is the estimate of N_g which is based on the respondents of this poststratum, so that the sum of the adjusted weights of the poststratum to be equal to N_g (calibration).

In a survey we may have an adjustment of the weights by class for non-response correction, and poststratification with poststrata that are different from the classes or are partly overlapping with them. As in the case of the classes, the units of the same poststratum should have almost the same probability of response, for the poststratification to result in bias reduction.

Item nonresponse

For some sample units there may be partial response, in the sense that these units do not respond to some questions. This happens when the respondent refuses or is unable to respond to some questions, or when some wrong answers could not be corrected in the process of input editing.

The process of substituting non available or wrong data with suitable data for the creation of a complete data file is known as **imputation**.

ή ό (Nonresponse)

There exist several metods of imputation. Good methods of imputation can preserve known relationships among variables and reduce the bias due to partial non-response.

Imputation is used only for substitution of non available or wrong data, not for total non-response.

The imputation was in the past a manual process, but now days more often automated systems of imputation are used.

The artificial small survey data set (source: book of Lohr (2009)) shown in in the following table will be used to explain different imputation methods.
The number "1" in the last two columns indicates that the respondent answered yes in the question.

Nonresponse

Person	Age	Sex	Years of Education	Crime Victim	Violent Crime Victim
1	47	M	16	0	0
2	45	F	?	1	1
3	19	M	11	0	0
4	21	F	?	1	1
5	24	M	12	1	1
6	41	F	?	0	0
7	36	M	20	1	?
8	50	M	12	0	0
9	53	F	13	0	?
10	17	M	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	M	16	1	0
15	44	M	14	0	0
16	45	M	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

Methods of Imputation

These methods are divided to those that use data only from the responded and other auxiliary data, and to those that use data from other respondents.

Deductive imputation
This method determines the missing value with certainty using logical constraints and other data from the same unit, e.g., a missing term from a sum. It is the ideal but the less frequent type of imputation.
In the example of the table, for person 9 there is missing value in the last question. However, the response in the second last question implies logically that the missing value should be 0.

Historic imputation

This method is more useful in longitudinal surveys, especially for variables that are stable over time. It uses values reported by the same unit in a previous measurement.

In cases where the response in a previous measurement can determine with certainty the current response, this method is a special case of deductive imputation.

Mean imputation

With this method the imputed missing value for some variable is the mean of the values of the responding sample units. For the same variable, the mean value is used for each unit for which imputation is required.

This method is used only for quantitative variables, and often as last resort. It preserves totals and means, but distorts distributions and relationships among variables.
Also, it causes artificial concentration of values around the mean, implying artificial reduction of the variance of the values of the variable for which we do imputation.

In the example, for persons 2, 4 and 6 the value for years of education is missing. For all these three persons the imputed value is the mean of the 17 responses in the same question: 12,70. After imputation the mean of all 20 persons is the same with the mean of the 17 respondents.

Cell mean imputation

The sample is partitioned into cells (classes) so that the units of the same class are similar. In each such class we do imputation using the mean value of the respondents of the class. The distortion of the distribution of the values of the variable and the artificial reduction of the variance are less severe than with the simple mean imputation.

For the data of the example, the sample is partitioned in four classes by age and sex.

		Age	
		≤ 34	≥ 35
		Persons	Persons
Sex		3,5,10,14	1,7,8,15,16
	F	Persons	Persons
		4,12,13,19,20	2,6,9,11,17,18

Nonresponse

For persons 2 and 6, the missing value for the years of education is imputed with the mean of the four women of age equal or higher than 35 years who responded to the question: 12,25. For person 4, the imputed value is the mean of the four women of age equal or lower than 34 years who responded to the question: 11,25.

After imputation, in each class the mean value is the same with the mean of the respondents.

Hot-deck imputation

This is a class of methods that create a more authentic variability of the imputed values than the mean imputation method. These methods give always feasible values because the imputed values belong to respondents of the same survey.

Random hot-deck imputation

With this method the missing value is imputed with the value of a "donor", who is selected randomly from the respondents. This is the simplest type of hot-deck imputation.

Cell random hot-deck imputation

This is a variant of the previous method, in which suitable classes of sample units are created, as in the cell mean imputation. For a unit of a class, the donor is randomly selected from the respondents of the same class.

In the example, for person 10 the answers to the last two questions are missing. In the class of this person, persons 3, 5 and 14 have responded to both questions, and so one of these three persons is randomly selected as donor.

Sequential hot deck imputation
With this method the missing value is imputed with the corresponding value from the last preceding responded unit of the same class in the data file. The advantage of this non-random procedure is the easy sequential processing of the file. The disadvantage is that it often makes multiple use use of the same donor.

In the example, for person 19 the answer to the second last question is missing. Person 13 was the last of the same class that responded, and so the value 1 was used in the imputation.

Nearest-Neighbor Hot-Deck Imputation
The missing value of a variable is imputed with the value of a respondent who is the "nearest", according to some distance function for the values of this variable, defined by known auxiliary information.

For example, if age and sex are used to define distance, then the respondent of the same sex and the nearest age is selected as donor.

In our example, the missing values for person 10 are imputed with the values of person 3, who is of the same sex and of the nearest age with person 10.

Regression Imputation

This method employs regression of the variable for which we need imputation on a set of variables for which there is response from all units. The regression equation is used then for "prediction" of the missing values.

In the example, we have only 18 responses for the variable "crime victim" (perhaps too few for fitting a model to the data), but a logistic regression with regressor the age gives the following model for the predicted probability $\hat{p}$ for a person to be "crime victim",

$$\log \frac{\hat{p}}{1-\hat{p}} = 2.5643 - 0.0896 \times age.$$

The predicted probability for a person of age 17 to be "crime victim" is $\hat{p} = 0.74$. Since this probability is higher than 0.5, the missing value for person 10 is imputed with the value 1.

Cold-deck imputation

With this method, the imputed values are from previous survey or from historic data.

Suggested bibliography

Lohr, S.L. (2009). *Sampling: Design and Analysis*. Second Edition, Brooks/Cole. Cengage Learning.

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

Lumley, T. (2010). *Complex Surveys. A Guide to Analysis Using R*. Wiley.