13.2. Sampling error

The variance estimation is based on the assumption that the PSUs were selected according to a PPS with replacement scheme. As the clusters (one or more unified city blocks) are used as primary sampling units (PSUs) in the sample design, the variance procedure estimates the variance from the variation among the PSUs.

w_{hijk} (>0) stands for the survey weight (extrapolation factor) attached to the sample individual k ($k = 1$), as one individual is surveyed, in each sampling household) belonging to the sampling household of order j ($j = 1,\ldots,n_{hi}$) that belongs to the selected cluster of order i of the stratum h ($h = 1, \ldots, H$).

Estimation of survey characteristics

Let y_{hijk} be the value of variable y of the ultimate unit (individual) of the household of order j, belonging to the hi primary sampling unit (cluster). Moreover, Y stands for the total population, which is derived by adding the characteristic y of all ultimate units included in all strata h. The form of the estimator on the basis of the multistage-stage sample design is:

$$\hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{a_h} \sum_{j=1}^{n_{hi}} w_{hijk} y_{hijk}$$

Estimation of a ratio

Let x_{hijk} be the value of the characteristic x of the ultimate unit of the household of order j, belonging to the hi primary sampling unit (cluster). Moreover, X stands for the total population, which is derived by adding the characteristic x of all ultimate units included in all strata h. The form of the estimator $\hat{R}$ in the case of a multi-stage sample design is:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{a_h} \sum_{j=1}^{n_{hi}} w_{hijk} y_{hijk}}{\sum_{h=1}^{H} \sum_{i=1}^{a_h} \sum_{j=1}^{n_{hi}} w_{hijk}\, x_{hijk}}$$

Variance estimation

In order to calculate the variance of the estimated characteristics, the following steps should be followed:

a) For every selected primary sampling unit (cluster) i of the stratum h, we calculate the quantity T_{hi} using the following formula:

$$T_{hi} = a_h \sum_{j=1}^{n_{hi}} w_{hijk} y_{hijk}$$

Where:

a_h: Number of primary sampling units

w_{hijk} : Weight attached to the sample individual k that belong to the household of primary sampling unit (cluster) i of the stratum h

y_{hijk} : Value of variable y of the ultimate unit (individual) of the household of order j, belonging to the hi primary sampling unit (cluster)

b) Since T_{hi} has been calculated for every primary sampling unit (cluster) i $(i = 1,\dots,a_h)$ of the stratum h ,then $V(\hat{Y})$ is calculated as (Rao, 1988):

$$V(\hat{Y}) = \sum_{h=1}^{H} \frac{1}{a_h(a_h-1)} \left[\sum_{i=1}^{a_h} T_{hi}^2 - \frac{1}{a_h} \left(\sum_{i=1}^{a_h} T_{hi} \right)^2 \right]$$

For the estimation of the variance of a ratio $\hat{R} = \frac{\hat{Y}}{\hat{X}}$ additional steps should be followed, below:

a) For every selected primary sampling unit (cluster) i of the stratum h, we calculate the quantity F_{hi} using the following formula:

$$F_{hi} = a_h \sum_{j=1}^{n_{hi}} w_{hijk} x_{hijk}$$

b) Since $T_{hi}$and F_{hi} have been calculated for every primary sampling unit (cluster) i ($i = 1,2,\dots,a_h$) of the stratum h, then $V(\hat{X})$ is calculated as:

$$V(\hat{X}) = \sum_{h=1}^{H} \frac{1}{a_h(a_h-1)} \left[\sum_{i=1}^{a_h} F_{hi}^2 - \frac{1}{a_h} \left(\sum_{i=1}^{a_h} F_{hi} \right)^2 \right]$$

The variance of $\hat{R}$ can be calculated using the following formula:

$$V(\hat{R}) = \frac{V(\hat{Y}) + \hat{R}^2 V(\hat{X}) - 2\hat{R}Cov(\hat{X},\hat{Y})}{\hat{X}^2}$$

where:

$$Cov(\hat{X},\hat{Y}) = \sum_{h=1}^{H} \frac{1}{a_h(a_h-1)} \left[\sum_{i=1}^{a_h} T_{hi}F_{hi} - \frac{1}{a_h} \left(\sum_{i=1}^{a_h} T_{hi} \right) \left(\sum_{i=1}^{a_h} F_{hi} \right) \right]$$

The variance estimator $\hat{V}(\hat{\theta})$ has to be adjusted to take unit non-response into account. Different methods can be used: methods based on the assumption that respondents are missing at random or completely at random within e.g. strata or constructed response homogeneity groups, methods using the two-phase approach, etc.

Sampling error estimation method : Analytic method