CHAPTER 9: ANALYSIS OF HOUSEHOLD SURVEY DATA

Andrés Gutiérrez[1], Pedro Luis do Nascimento Silva[2]

2024-09-24

[1] Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

[2] SCIENCE, pedronsilva@gmail.com

Contents

List of Figures

List of Tables

Abstract

Analyzing complex household survey data requires knowing and properly applying the foundations of the design-based inference. The researcher will be faced to a small database that contains specific information that will allow her to make conclusions over the whole population.

The purpose of any analysis on this kind of datasets is not referred to make conclusions on the sample itself – which in most of the cases is a small subgroup of the population - but to the domains of interest and the whole population. Having that into account, the first step in any analysis plan should be devoted to defining the sampling design based on the selection mechanisms used to draw the final sample and the findings on the field related to nonresponse and lack of coverage.

The chapter covers three main topics of analysis: descriptive statistics; comparisons and association; and modeling of survey data. On the one hand, we introduce simple descriptive statistics, such as totals, frequencies, means and proportions, quantiles and some graphics; on the other, we delve deeper on complex relationships between the variables of the survey. All these analyses rely on the representativity principle of the design-based inference. This way, the reader will find a strong focus, not only on point estimates, but also on uncertainty measures. The chapter also presents a short discussion on the different approaches that can be used to estimate variances; the best way to visualize the estimates; and NSO practical experiences.

Introduction

A key concern of every agency that produces statistical information is with the *correct* use of the data that it produces. This is even reflected in the United Nations *Fundamental Principles of Official Statistics*, namely:

- **Principle 3.** To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.
- **Principle 4.** The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

Here we emphasize a particular aspect, aiming to empower users when analysing household survey data. The computer revolution, with the resulting ease of access computers, created favorable conditions for the increased use of statistical data, including those resulting from household sample surveys. Sometimes this data is used for purely descriptive purposes. Other times, however, its use is made for analytical purposes, involving the testing of hypothesis or the construction of models, when the objective is to draw conclusions that are also applicable to populations other than the one from which the sample was extracted. In such cases, standard statistical software may provide biased or misleading results. This chapter's purpose is to present the relevant models, methods and software to enable users to account for the complex survey design frequently used to conduct household sample surveys when analysing the resulting data.

What makes such data special for those who intend to use them for analytical purposes? The answer is that they are obtained through complex sample surveys of finite populations that often involve: *stratification*, *clustering* of units of analysis, *unequal probabilities of selection*, and *weighting adjustments* to compensate for non-response and/or improve precision.

Standard data analysis methods and software typically ignore these aspects, and may produce biased estimates of both the target parameters and the variances of these estimates. In this chapter we analyze the impact of simplifications made when using standard data analysis methods and software, and present the necessary adjustments to these procedures in order to appropriately incorporate the aspects highlighted here into the analysis.

In section 9.1, a short discussion on the fundamental principles of the design-based inference is presented, emphasizing that conclusions taken from probability sample surveys should be based on a pair: the point estimate and it associated margin of error (or any related measure). In section 9.2, we begin the journey with simple descriptive statistics: means, ratios, proportions and other typical descriptive parameters. Section 9.3 is devoted to more complex parameters that allow comparisons of the phenomenon of interest between subgroups for continuous and discrete variables. In this section we present standard tests to compare means and measure the degree of association between variables, and also consider the problem of correlation and association. Section 9.4 focuses on modelling survey outcomes. We first involve the reader in an discussion on the role of weighting when estimating regression coefficients. Then, we introduce some proper approaches to estimate complex parameters in linear and logistic regression models. Finally, section 9.5 presents a summary of ideas and tools for survey data visualization showing the best practices for creating graphics and maps in a context where uncertainty measures of estimates are important.

Most of the sections in the chapter present national experiences that will guide the reader on how NSOs are dealing with the different stages of the analysis of household survey data.

The purpose of this chapter is defining and explaining basic concepts of the design-based paradigm in household surveys to analyze complex household survey data. In section 9.1, a short discussion on the fundamental principles of the design-based inference is presented, emphasizing that conclusions taken from this kind of surveys should be based on a pair: the point estimate and it associated margin of error (or any related measure). In section 9.2, we begin the journey with simple descriptive statistics: means, ratios, proportions and other parameters are part of this discussion. Section 9.3 is devoted to more complex parameters that allow comparisons of the phenomenon of interest between subgroups for continuous and discrete variables. In this section we present standard tests to compare means and measure the degree of association between variables. This section also deals with the problem of correlation and association. Section 9.4 focuses on modelling survey outcomes. We first involve the reader in an discussion on the role of weighting when estimating regression coefficients. Then, we introduce some proper approaches to estimate complex parameters in linear and logistic regression models. Finally, section 9.5 presents a summary on survey data visualization showing the best practices for creating graphics and maps in a context where uncertainty measures of estimates are important.

Most of the sections in the chapter present national experiences that will guide the reader on how currently NSOs are dealing with the different stages of the analysis of household survey data.

Chapter 1

The golden pair: sample design and estimator

Accounting for the sampling design is crucial for analyzing complex survey data. We must ensure that PSU, strata, and weights are available in the survey dataset to enable adequate analysis. Alternatively, when such information is not available, the dataset should at least contain replicate weights, or the analyst should have clear guidance on how to compute both point and variance estimates.

A well-described survey design facilitates statistical analysis, supports effective data interpretation, and enables meaningful insights into complex phenomena. Missing or incorrect design information may lead to biased estimates and misleading conclusions.

1.1 Parameters and estimators

Under a design-based approach this section presents the basic principles of inductive inference and how, using the *sampling weights* (from chapter VIII), one can get consistent estimators for population parameters of interest. We adopt the notation introduced in chapter VIII for presenting the expressions required here.

The *population total* $Y = \sum_U y_k$ and *mean* $\overline{Y} = \frac{Y}{N}$ of a survey variable y can be estimated by weighted estimators given by $\widehat{Y}_{HT} = \sum_s d_k\, y_k$ and $\overline{y}_H = \frac{\widehat{Y}_{HT}}{\widehat{N}_{HT}} = \frac{\sum_s d_k\ y_k}{\sum_s d_k}$, respectively. When the survey weights are calibrated and/or non-response adjusted, the above expressions may still be used, but with the calibrated or non-response adjusted weights, w_k say, replacing the design weights d_k, for all $k \in s$.

Here $s = \{k_1, \dots, k_n\} \subset U$ denotes the set of units in a sample selected from the population U using a *probability sampling design* $p(s)$, that ensures strictly positive first order inclusion probabilities $\pi_k = Pr(k \in s)$, $\forall\, k \in U$. These inclusion probabilities are assumed known $\forall\, k \in s$, at least to the data producers.

Under the design-based framework and assuming full response, $\widehat{Y}_{HT}$ is unbiased for Y and its sampling variance is given by

$$V_p\left(\widehat{Y}_{HT}\right) = \sum_{k \in U} \sum_{j \in U} \left(\frac{d_k d_j}{d_{kj}} - 1 \right) y_k y_j$$

where $d_{kj} = 1/\pi_{kj}$ and $\pi_{kj} = Pr(k, j \in s)$, $\forall\, k, j \in U$. This result assumes that the sampling design $p(s)$ is such that $\pi_{kj} > 0 \;\forall\, k, j \in U$.

Under full response, this variance can be estimated unbiasedly by

$$\widehat{V_p}\left(\widehat{Y}_{HT}\right) = \sum_{k \in s} \sum_{j \in s} \left(d_k d_j - d_{kj} \right) y_k y_j$$

While the above formula for variance estimation is general and covers the vast majority of sample designs used in the practice of household sample surveys, it is not used in practice because the second order inclusion probabilities π_{kj} (and corresponding pairwise weights d_{kj}) are generally unknown to survey data analysts. In fact, even data producers do not compute such pairwise weights, since there are more efficient methods for variance estimation that do not require having such weights.

1.2 Uncertainty in household surveys

As the sample is typically a small subset of the population, it is important to obtain not only point estimates for the parameters of interest, but also the corresponding uncertainty measures and/or confidence intervals. In this subsection we present some approaches for variance estimation: approximate formulas from *Taylor linearization* and/or the *ultimate cluster* approach for variances under multi-stage cluster sampling. We also introduce replication methods and generalized variance functions, which are essential when PSU or strata are missing from the sample dataset.

A unifying idea of sampling theory is that of estimating equations - Binder (1983). Many population parameters can be written/obtained as solutions for *population estimating equations*. A generic population estimating equation is given by $ _{i U} z_i ( ) = 0 $, where $z_i(\bullet)$ is an *estimating function* evaluated for unit i and θ is a population parameter of interest.

For the case of the population total, take $z_i(\theta) = y_i - \theta/N$. The corresponding population estimation equation is given by $\sum_{i \in U}(y_i - \theta/N) = 0$, and solving for θ gives the population total $\theta_U = \sum_{i \in U} y_i = Y$. Similarly, take $z_i(\theta) = y_i - \theta$ for the population mean. As a final example, consider the ratio of population totals. Taking $z_i(\theta) = y_i - \theta x_i$, the corresponding population estimation equation is given by

$\sum_{i\in U}(y_i - \theta x_i) = 0$. Solving for θ gives the *population ratio* $\theta_U = \sum_{i\in U} y_i / \sum_{i\in U} x_i = R$.

The idea of defining population parameters as solutions to population estimating equations allows defining a general method for obtaining corresponding sample estimators. It is a matter of using the *sample estimating equations* $\sum_{k\in s} d_k \, z_k(\theta) = 0$. Under *probability sampling*, full response and with $d_k = 1/\pi_k$, the sample sum in the left hand side is unbiased towards the population sum in the corresponding population estimating equation. Solving the sample estimating equation yields consistent estimators for the corresponding population parameters.

The case of the population mean yields the sample estimating equation $\sum_{k\in s} d_k(y_k - \theta) = 0$, and by solving on θ, we recover the Hàjek estimator $\widehat{\theta} = \sum_{k\in s} d_k \, y_k \, / \sum_{k\in s} d_k = \overline{y}_H$. In the case of the population ratio, solving $\sum_{k\in s} d_k(y_k - \theta x_k) = 0$ on θ, yields the well-known estimator $\widehat{\theta} = \sum_{k\in s} d_k \, y_k \, / \sum_{k\in s} d_k \, x_k = \widehat{R}$.

The variance of estimators obtained as solutions of sample estimating equations can be obtained as:

$$V_p(\widehat{\theta}) \doteq \left[J(\theta_U)\right]^{-1} V_p \left[\sum_{k\in s} d_k \, z_k(\theta_U)\right] \left[J(\theta_U)\right]^{-1}$$

where $J(\theta_U) = \sum_{k\in U} \left[\partial z_k(\theta)/\partial\theta\right]_{\theta=\theta_U}$, and θ_U is the solution of the corresponding population estimating equation.

A consistent estimator of this variance is given by:

$$\widehat{V}_p(\widehat{\theta}) = \left[\widehat{J}(\widehat{\theta})\right]^{-1} \widehat{V}_p \left[\sum_{k\in s} d_k \, z_k(\widehat{\theta})\right] \left[\widehat{J}(\widehat{\theta})\right]^{-1}$$

where $\widehat{J}(\widehat{\theta}) = \sum_{k\in s} d_i \, \left[\partial z_k(\theta)/\partial\theta\right]_{\theta=\widehat{\theta}}$.

This approach implies that by one is able to estimate many population parameters and corresponding variances using essentially well known methods for estimating totals. Its simplicity and generality have enabled the development of software such as the `R survey` package, the `Stata svy` functions and others.

1.3 Ultimate Cluster Method

The central idea of the *Ultimate Cluster* method for variance estimation for estimators of totals in multi-stage cluster sampling designs, proposed by (**?**), is to consider only the variation between information available in the level of PSUs, and assume that these would have been selected with replacement from the PSU population. This idea is simple, but quite powerful, because it allows to accommodate a variety of sampling designs, involving stratification and selection with unequal probabilities (with or without replacement) of both PSUs as well as lower level sampling units. The requirements for the application of this method are that one has unbiased estimators of totals for the variable of interest for each sampled PSU, and that data are available for at least two sampled PSUs in each stratum (if the sample is stratified in the first stage).

Although the method was originally proposed for estimation of variances of estimated totals, it can also be applied in combination with Taylor linearization to obtain variance estimates for estimators of other population quantities that can be obtained as solutions to sample estimating equations.

Consider a multi-stage sampling design, in which n_h PSUs are selected in stratum h, $h = 1, \dots, H$. Let π_{hi} be the inclusion probability of PSU i stratum h, and by $\widehat{Y}_{hi}$ an unbiased estimator of the total Y_{hi} of the survey variable y for the i-th PSU in stratum h, $h = 1, \dots, H$. Hence an unbiased estimator of the population total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$ is given by $\widehat{Y}_{UC} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} d_{hi} \widehat{Y}_{hi}$, and the *ultimate cluster* estimator of the corresponding variance is given by:

$$\widehat{V}_{UC}\left(\widehat{Y}_{UC}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(d_{hi} \widehat{Y}_{hi} - \frac{\widehat{Y}_h}{n_h} \right)^2$$

where $d_{hi} = 1/\pi_{hi}$, $\widehat{Y}_h = \sum_{i=1}^{n_h} d_{hi} \widehat{Y}_{hi}$ for $h = 1, \dots, H$. (See for example, (**?**), p. 4).

Although often the selection of primary units can have Primary Cluster estimator presented here may provide a reasonable approximation of the corresponding variance of randomization. This is because sampling plans without replacement are generally more efficient than plans with replacement of equal size. Such an approximation is widely used by sampling practitioners to estimate variances of usual descriptive quantities such as totals and medium (with due adaptation) due to their simplicity, compared to the much greater complexity involved with the employment of variance estimators that attempt to incorporate all steps of plans sampling in several stages. A discussion about Quality of this approximation and alternatives can be found in (**?**), p. 153.

In some cases, sample replication methods (*bootstrap*, *jackknife*) can also be used to estimate variances, as we will see later.

1.4 Using software to generate valid inferences

In this part, we advocate to using specialized statistical software to generate efficient estimation processes. Those packages support complex survey data analysis by specifying the survey design using appropriate commands or functions.

Chapter 2

Descriptive parameters

When analyzing complex survey data, several descriptive parameters are meaningful and important. For example: poverty and unemployment rates are simple parameters that allow decision-making for governments; also, income distribution can be used to monitor inequality along time.

2.1 Frequencies

2.1.1 Point Estimation

The accurate estimation of absolute sizes and proportions in household surveys is essential for obtaining representative data that reflects the demographic and socioeconomic reality of a population. These figures serve as the basis for public policy decision-making, resource allocation, and the design of social programs.

The ability to understand the distribution of specific categories, such as poverty status, employment status, education level, among others, provides valuable information to address inequalities and promote equitable development.

2.1.1.1 Size Estimates

In this section, the processes for estimating categorical variables will be carried out. First, one of the most important parameters is the size of a population, which represents the cardinality of that set; in other words, the total number of individuals that comprise it. In terms of notation, the population size is estimated as follows:

$$\hat{N}_{\omega} = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}$$

Similarly, the size estimate in a subpopulation, defined by a dichotomous variable $I(y_i = d)$, which takes the value of one if individual i belongs to category d in the discrete variable, is given by the following expression:

$$\hat{N}_{\omega}^{d} = \sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} I(y_i = d)$$

2.2 Means, proportions, and ratios

After conducting the graphical analysis of trends in the continuous survey variables, it is necessary to obtain point estimates of the measured parameters. These estimates can be obtained either generally for the entire population or disaggregated by domains of interest, depending on the research needs. In the context of household surveys, point estimates refer to the estimation of totals, averages, ratios, means, etc. As mentioned by Heeringa et al. (2017), the estimation of totals or averages for a variable of interest in the population, along with the estimation of its variance, has played a crucial role in the development of probability sampling theory. These estimates allow for unbiased and accurate results, providing valuable insights into what is happening in the studied households and enabling informed public policy decision-making.

2.2.1 Total Estimates

Once the sampling design is defined, which was done in the previous section, the estimation processes for the parameters of interest are carried out. For the estimation of totals with complex sampling designs that include stratification ($h = 1, 2, ..., H$) and subsampling in PSUs (assumed to be within stratum h) indexed by $\alpha = 1, 2, ..., a_h$, the estimator for the total can be written as:

$$\hat{y}_{\omega} = \sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}$$

Where $n_{h\alpha}$ is the sample size of households or individuals in PSU α of stratum h; a_h is the sample size of PSUs within stratum h; H is the total number of strata in the sampling design. Finally, $y_{h\alpha i}$ and $\omega_{h\alpha i}$ correspond respectively to the observation of the variable of interest and the weight (expansion factor) of element i associated with PSU α within stratum h. The unbiased variance estimator for this total estimator $\hat{y}_{\omega}$ is:

$$\widehat{var}\left(\hat{y}_{\omega}\right) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \left[\sum_{\alpha=1}^{a_h} \left(\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i} \right)^2 - \frac{\left(\sum_{\alpha=1}^{a_h} \omega_{h\alpha i} y_{h\alpha i} \right)^2}{a_h} \right]$$

As can be seen, calculating the total estimate and its estimated variance is complex. However, these calculations can be performed in `R` using the `svytotal` function. The confidence interval is given by the following expression:

$$\hat{y}_\omega \pm 1.96 \times \sqrt{\widehat{var}\left(\hat{y}_\omega\right)}$$

2.2.2 Estimation of Averages

The estimation of the population mean or average is a very important parameter in household surveys. According to Gutiérrez (2016), an estimator of the population mean can be written as a nonlinear ratio of two estimated finite population totals, as follows:

$$\hat{\bar{y}}_\omega = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}} = \frac{\hat{y}_\omega}{\hat{N}_\omega}.$$

It should be noted that if y is a binary variable, the weighted mean estimates the population proportion. On the other hand, since $\hat{\bar{y}}_\omega$ is not a linear statistic, there is no closed-form formula for the variance of this estimator. For this reason, resampling methods or Taylor series expansions must be used. In this particular case, using Taylor series, the variance of the estimator is as follows:

$$var\left(\hat{\bar{y}}_\omega\right) \doteq \frac{var\left(\hat{y}_\omega\right) + \hat{\bar{y}}_\omega^2 \times var\left(\hat{N}_\omega\right) - 2 \times \hat{\bar{y}}_\omega \times cov\left(\hat{y}_\omega, \hat{N}_\omega\right)}{\hat{N}_\omega^2}.$$

As can be observed, calculating the variance estimation involves complex components to compute analytically, such as the covariance between the estimated total and the estimated population size.

2.2.3 Proportion Estimation

The estimation of a proportion for a binary response variable requires a direct extension of the ratio estimator shown in the previous chapter. As mentioned by Heeringa et al. (2017), by recoding the original response categories into a single indicator variable y_i with possible values of 1 and 0 (e.g., yes $= 1$, no $= 0$), the estimator for a proportion is defined as follows:

$$\hat{p}_\omega^d = \frac{\hat{N}_\omega^d}{\hat{N}_\omega} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}\ I(y_i = d)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}}$$

By applying Taylor linearization to the above estimator, its variance is given by the following expression:

$$var\left(\hat{p}_{\omega}^{d}\right) \doteq \frac{var\left(\hat{N}_{\omega}^{d}\right) + (\hat{p}_{\omega}^{d})^{2} var\left(\hat{N}_{\omega}\right) - 2\,\hat{p}_{\omega}^{d}\, cov\left(\hat{N}_{\omega}^{d}, \hat{N}_{\omega}\right)}{(\hat{N}_{\omega})^{2}}$$

It is common to observe that many statistical packages opt to generate proportion estimates and standard errors on a percentage scale.

As is well known in the specialized literature, when the estimated proportion of interest is close to zero or one, the limits of the traditional confidence interval, based on the sampling design, may fall outside the permissible range for proportions. This would have no interpretation due to the nature of the parameter. For this reason, to address this issue, alternative confidence interval estimates based on the sampling design can be used, as proposed by Rust et al. (2007) and Dean and Pagano (2015). Thus, the confidence interval using the $Logit\,(p)$ transformation is given by:

$$IC\left[logit\left(p^{d}\right)\right] = \left\{ln\left(\frac{\hat{p}_{\omega}^{d}}{1-\hat{p}_{\omega}^{d}}\right) \pm \frac{t_{1-\alpha/2,\,gl} \times se\left(\hat{p}_{\omega}^{d}\right)}{\hat{p}_{\omega}^{d}\left(1-\hat{p}_{\omega}^{d}\right)}\right\}$$

Therefore, the confidence interval for p^{d} would be:

$$IC\left(p^{d}\right) = \left\{\frac{exp\left[ln\left(\frac{\hat{p}_{\omega}^{d}}{1-\hat{p}_{\omega}^{d}}\right) \pm \frac{t_{1-\alpha/2,\,gl} \times se\left(\hat{p}_{\omega}^{d}\right)}{\hat{p}_{\omega}^{d}\left(1-\hat{p}_{\omega}^{d}\right)}\right]}{1 + exp\left[ln\left(\frac{\hat{p}_{\omega}^{d}}{1-\hat{p}_{\omega}^{d}}\right) \pm \frac{t_{1-\alpha/2,\,gl} \times se\left(\hat{p}_{\omega}^{d}\right)}{\hat{p}_{\omega}^{d}\left(1-\hat{p}_{\omega}^{d}\right)}\right]}\right\}$$

2.2.4 Relationship Between Variables

In many household survey analyses, it is not sufficient to examine individual variables in isolation. For instance, analyzing the average income of men and women in a country is informative, but comparing the income difference between men and women is crucial for addressing the gender pay gap. This section provides computational tools for estimating ratios and explores hypothesis testing for differences in means, including more complex contrasts.

2.2.4.1 Estimation of Ratios

A particular case of a non-linear function of totals is the population ratio. This is defined as the quotient of two population totals for continuous characteristics of interest. In household surveys, there are times when estimating such a parameter is necessary. For example, estimating the ratio of expenditures to income, the number of men per woman, or the number of pets per household in a specific country.

Since the ratio is the quotient of two totals, both the numerator and the denominator are unknown quantities and thus need to be estimated. The point estimator for a ratio in complex surveys is the quotient of the estimators for the totals, as defined by:

$$\hat{R}_\omega = \frac{\hat{y}_\omega}{\hat{x}_\omega} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} x_{h\alpha i}}$$

However, because the ratio estimator is a quotient of two estimators (i.e., a quotient of two random variables), calculating the variance of this estimator is not straightforward. To address this, Taylor linearization methods are applied as demonstrated by Gutiérrez (2016).

2.3 Percentiles and inequality measures

In household surveys, it is always necessary to estimate dispersion measures of the studied variables. For example, to understand how disparate incomes are in a given country, which helps inform public policy decisions. Therefore, studying these parameters is crucial. Below is the estimator for the standard deviation:

$$s_\omega\left(y\right) = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}\left(y_{h\alpha i} - \hat{\bar{y}}_\omega\right)^2}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} - 1}$$

Non-central location measures (percentiles) are calculated to determine characteristic points of the data distribution beyond central values. Key non-central location measures include the median, quartiles, and percentiles. In most household surveys, not only are totals, means, and proportions estimated; for some indicators, it is necessary to estimate other parameters, such as medians and percentiles.

The median is a measure of central tendency that, unlike the mean, is not easily influenced by outliers, and is thus considered a robust measure. The median is the value that divides the population into two equal parts, implying that half of the population's observations fall above the median and the other half fall below.

On the other hand, the estimation of income percentiles in a given country can define the onset of public policy. For example, a tax could be imposed on individuals in the top 10% of the income distribution, or transport subsidies could be provided to those in the bottom 15% of the income distribution.

Quantile estimation is based on results related to weighted total estimators, using an estimation of the cumulative distribution function (CDF) of the population. Specifically, the CDF for a variable y in a finite population of size N is defined as follows:

$$F\left(x\right)=\frac{\sum_{i=1}^{N}I\left(y_{i}\leq x\right)}{N}$$

Where $I\left(y_{i}\leq x\right)$ is an indicator variable that takes the value 1 if y_{i} is less than or equal to a specific value x, and 0 otherwise. An estimator of the CDF in a complex sampling design is given by:

$$\hat{F}_{\omega}\left(x\right)=\frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_{h}}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}I\left(y_{i}\leq x\right)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_{h}}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}}$$

Once the CDF is estimated using the survey design weights, the q-th quantile of a variable y is the smallest value of y such that the CDF is greater than or equal to q. As is well known, the median is the value where the CDF is greater than or equal to 0.5. Thus, the estimated median is the value where the estimated CDF is greater than or equal to 0.5. Following the recommendations of Heeringa et al. (2017), to estimate quantiles, one first considers the order statistics denoted as $y_{(1)},\ldots,y_{(n)}$ and finds the value of j $(j=1,\ldots,n)$ such that:

$$\hat{F}_{\omega}\left(y_{j}\right)\leq q\leq\hat{F}_{\omega}\left(y_{j+1}\right)$$

Thus, the estimation of the q-th quantile $y_{(q)}$ in a complex sampling design is given by:

$$\hat{y}_{(q)}=y_{j}+\frac{q-\hat{F}_{\omega}\left(y_{j}\right)}{\hat{F}_{\omega}\left(y_{j+1}\right)-\hat{F}_{\omega}\left(y_{j}\right)}\left(y_{j+1}-y_{j}\right)$$

For the variance estimation and confidence intervals of quantiles, Kovar et al. (1988) present results from a simulation study where they recommend using the *Balanced Repeated Replication* (BRR) technique. The previously mentioned estimators and procedures for estimating percentiles and their variances are implemented in `R`. Specifically, the median estimation can be done using the function `survey_median`.

References

Bibliography

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3):279–292.

Dean, N. and Pagano, M. (2015). Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys. *Journal of Survey Statistics and Methodology*, 3(4):484–503.

Gutiérrez, H. A. (2016). *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Ediciones de la U, segunda edición edition. Google-Books-ID: UlVmE5pkRwIC.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2017). *Applied survey data analysis*. Chapman and Hall CRC statistics in the social and behavioral sciences series. CRC Press.

Kovar, J. G., Rao, J. N. K., and Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16(Suppl.):25–45.

Rust, K. F., Hsu, V., and Westat (2007). Confidence intervals for statistics for categorical variables from complex samples.