CHAPTER 9: ANALYSIS OF HOUSEHOLD SURVEY DATA

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

2024-12-02

[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 *design-based inference*. The researcher will be faced to a small dataset containing information that will allow her to make conclusions over the whole population.

The purpose of any analysis on this kind is not to make conclusions about the sample itself – which in most of the cases is a small subset of the population - but to the the whole population and possibly domains or population subgroups of interest. Having that into account, the first step in any analysis plan should be devoted to understanding and specifying the sampling design used to draw the sample and the findings on the field related to nonresponse and lack of coverage.

This chapter covers three main topics of analysis: descriptive statistics; comparisons and associations; and modeling of survey data. First, we introduce simple descriptive statistics, such as totals, means, frequencies and proportions, quantiles and some graphics; next, we delve deeper on relationships between the survey variables. 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 some NSO practical experiences.

Introduction

A key concern for every agency that produces statistical information is ensuring the *correct* use of the data it provides. This concern is enshrined in the United Nations *Fundamental Principles of Official Statistics*, particularly in the following principles:

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

This chapter emphasizes empowering users to analyze household survey data accurately. The advent of the computer revolution, coupled with greater access to computational tools, has led to increased use of statistical data, including household survey data. 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. When using standard statistical software for such analyses, results can be biased or misleading if the complex sampling design is not properly accounted for.

The primary focus of this chapter is to present relevant models, methods, and software that enable users to incorporate complex designs into their analyses effectively. What makes household survey data special or challenging for those who intend to use them for analytical purposes is because they are collected through complex sampling methods that often involve:

- **Stratification**: Dividing the population into comprehensive distinct subgroups before sampling;
- **Clustering**: Grouping units and sampling groups rather than units to simplify data collection;
- **Unequal probabilities of selection**: Giving units different probabilities of being selected;

- **Weighting adjustments**: Correcting for non-response and/or improving precision.

Standard data analysis methods and software ignore these features, leading to biased estimates of both the target parameters and their associated variances. Here we analyze the impact of simplifications made when using standard data analysis methods and software, and present the necessary adjustments to these methods in order to appropriately incorporate the aspects highlighted above into the analysis.

Section 9.1 provides guidance on the preparation of a plan for the analysis of the household survey data. Such plans are important both to guide survey development and subsequently any secondary survey data analysis. In section 9.2, 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 9.3, we begin the journey with simple descriptive statistics: means, ratios, proportions and other typical descriptive parameters. Section 9.4 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 9.5 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 9.6 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.

The primary purpose of this chapter is to present and explain the fundamental concepts of the design-based paradigm in household surveys and demonstrate how to analyze complex household survey data. Throughout the chapter, national experiences are highlighted to illustrate how National Statistics Offices (NSOs) manage different stages of household survey data analysis. These practical examples provide a useful guide for applying the concepts and methods discussed in real-world contexts.

By the end of the chapter, readers will be equipped with the knowledge and tools needed to analyze household survey data effectively while accounting for the complexities of survey designs adopted in such surveys.

Chapter 1

Planning Analysis

Planning the analysis stage of a survey is an essential part of the overall survey planning process. Following the **Generic Statistical Business Process Model (GSBPM)** (https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51), this step corresponds to the subprocess labeled as *2.1 - Design Outputs*. At this stage, it is important to distinguish between two groups of analysts: primary data producers and secondary data users.

Proper planning and understanding of the survey design are crucial for both primary data producers and secondary data users. For primary data producers, creating a comprehensive tabular plan ensures alignment with survey objectives. For secondary users, clear research questions and attention to survey metadata enable accurate and meaningful analyses.

1.1 Primary data producers

They are responsible for planning and executing the survey to collect the intended data. For them, planning the analysis typically involves preparing a *tabular plan* — a document specifying the core set of tables to be produced once the survey data becomes available. This plan helps to ensure that the survey results align with the stated needs and objectives of the survey (see chapter 2).

Preparing a *tabular plan* generally requires defining three sets of specifications:

1. **Filter Conditions**: These may be used to define subgroups of the population for which specific tables will be produced. For example, in a survey where questions regarding occupation are asked only from individuals aged 15 or older, a filter condition might be 'if age > 14'. Such a condition means that only those in the relevant age group would be included in tables for the occupation related variables (status, type, income, etc.).

2. **Classification or Domain Variables**: These are variables used to subdivide the population into meaningful groups for analysis. For example, geographic areas (e.g., states or provinces), age groups, or sex might define rows in a table. These variables are often chosen to meet reporting requirements, such as providing estimates by province in national household surveys. A typical list of domain defining variables would include: geographic levels (provinces, etc.); type of region (urban x rural); sex; age groups; education; race / etchnicity; etc.

3. **Response or Survey Variables**: These are the variables being analyzed to understand how they vary across the defined domains. For instance, continuous survey variables (like income) might be used to create columns in a table, summarizing means, medians, or other statistics. These variables are all others that we do not use as classification ones. Categorical survey variables (like employment status) will generate a column for each category where the corresponding cross-classified frequencies will appear, with one row for each of the classes of the domain defining variables.

An important consideration when defining domains has to do with sample design, namely when defining strata and sample sizes. For most national household sample surveys, providing breakdown by province or state is required, and therefore stratification by provinces will be essential. Also, if precision requirements must be satisfied at the province level, then sample sizes that meet such requirements must be computed by province and then summed to obtain the country level sample size.

For domains defined by characteristics that are unavailable from the sampling frame, say age groups, for the case of household surveys that use area sampling, sample size calculations must take into account what the required total sample size must be such that estimates for the rarest group meet precision specifications. As an example, suppose that estimates are required by age groups such as young adults (18 to 29), adults (30 to 49), ageing adults (50 to 59) and elderly (60 and over). Assuming that the population distribution by these age groups is such that the ageing adults is the rarest group with 12.5% of the total population, and if a minimum sample size of 500 individuals in this group is required, then the total sample must be at least 500 / 0.125 = 4,000. That is, in order for the full sample to provide an expected sample of about 500 ageing adults we must sample at least 4,000 individuals for the survey.

1.2 Secondary Data Users

Secondary data users are all those who will access and analyze the survey data after it has been released, typically with access to only the public datasets and documentation provided by the data producers. Their first task is to define clear research questions and locate the relevant survey data and documentation (metadata). Good survey metadata

must describe the sampling design and estimation methods used (including details about stratification, clustering, and survey weights), both for descriptive parameters and for the corresponding measures of precision.

For example, based on the aforementioned, a well-defined research question posed by some secondary data user, can be such as **"Do rural households face digital exclusion compared to urban households?"**. Notice that it directs the analysis towards estimating relevant parameters and precision measures.

The question might involve testing whether the proportion of rural households with internet access is significantly lower than that of urban households. This clarity allows for more direct and accurate analyses. The null hypothesis behind this question is that $H_0)P_{Rural} = P_{Urban}$, which we wish to test against the alternative $H_A)P_{Rural} < P_{Urban}$, where P denotes the proportion of households having internet access in the domain identified.

A related but different set of hypothesis could be $H_0)C_{Rural} = C_{Urban}$ versus $H_A)C_{Rural} > C_{Urban}$ where C denotes the average cost of connection for internet access in the specified domain. Listing the research questions in this way would enable the subsequent analysis to progress more directly towards the estimation of the relevant parameters from the survey data, and corresponding precision measures, both of which are required to compute test statistics that would provide the evidence required to answer them.

But in order for such estimation to take place, the secondary data user will first have to find out details about how the sampling and estimation were done for the particular survey at hand. As we will discuss in subsequent sections of this chapter, it is essential to account for the survey weights when computing point estimates of both descriptive or model parameters, and to account for structural components of the sampling design and estimation process (stratification, clustering, unequal inclusion probabilities, non-response adjustment and calibration of survey weights, if any) when estimating variances or other measures of precision of the point estimates.

Users that disregard such aspects of the sampling design do so at their peril, and may end up producing biased estimates that will lead to incorrect inferences and decisions. The recommended practice is for data producers to provide sufficient detail about such sampling aspects as part of the metadata released with survey microdata, in order to enable secondary data users to consider these aspects when conducting their analyses of interest.

1.2.1 Quality Control for Secondary Analysts

A standard quality control step for secondary data analysts would consist of the following steps:

1. **Load the Data and Metadata**: Ensure that the survey microdata is properly linked to metadata describing the sampling design and estimation processes;

2. **Replicate Published Estimates**: Recreate some of the estimates provided by the survey producers, including measures of precision, to confirm that the data and design have been correctly interpreted and loaded;

3. **Compare Scenarios**: When the analyst is capable of replicating published estimates, she can then proceed towards the required new analysis for which no previous results are available. Repeat this analysis under two scenarios: ignoring and accounting for the sampling design. Compare the results and assess the impact of incorporating the sample design;

4. **Finalize the Analysis**: Use only the results that account for the sampling design in the final interpretation, ensuring that the design effects are appropriately incorporated.

Chapter 2

Accounting for the sampling design

When analysing household survey data, ignoring the sampling design undermines the representativeness, accuracy, and credibility of survey-based findings, which can lead to incorrect decisions. This is why accounting for the sampling design is essential when analyzing household survey data to ensure valid and unbiased estimates. As seen in the previous chapters, regular household surveys has two major characteristics:

- They use *complex sampling designs* (e.g., stratification, clustering, and unequal probabilities of selection) to represent the population efficiently. Ignoring the design can lead to biased population-level inferences.
- They define *sampling weights* for each sampling unit (primary, and remaining ones) to properly represent the population.

To illustrate this fact, we provide a simple example. Suppose a country has two regions: Region A with 100 people, and Region B with 900 people. Wealthy people lives in Region A, with an average income of \$10,000, less wealthy people live in Region B, with average income of \$2,000. The true population mean income is \$2,800, because:

$$\theta = \frac{(100 \times 10,000) + (900 \times 2,000)}{100 + 900} = 2,800.$$

Suppose a survey is conducted where 50 people are sampled from each Region. After data collection, it was found that the sample mean for Region A was \$10,000, while the sample mean for Region B was \$2,000. If the sampling design is ignored, all units in the sample will receive equal weights, regardless of their population sizes. This way, the mean income is biasedly estimated by:

$$\widehat{\theta} = \frac{(50 \times 10,000) + (50 \times 2,000)}{100} = 6,000.$$

When the sampling design is considered, weights are applied proportional to the population sizes of each neighborhood. This way, units in Region A would receive a weight of $\frac{100}{50} = 2$, while units in Region B would receive a weight of $\frac{900}{50} = 18$. In this escnario, the mean income is unbiasedly estimated by:

$$\hat{\theta} = \frac{(2 \times 50 \times 10,000) + (18 \times 50 \times 2,000)}{(2 \times 50) + (18 \times 50)} = 2,800.$$

Ignoring the sampling design causes that Region A (smaller but wealthier) dominates the estimate, even though it represents only 10% of the country. This creates a bias, making the average income seem higher than it actually is for the whole population. By considering the sampling design and using weights, and strata, the estimate correctly reflects the true contribution of each Region, avoiding bias.

In real applications, to account for the sampling design, we must ensure that primary sampling units (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.

This section discusses how survey data from a sample can be used to draw conclusions about an entire population using a design-based approach. This method assumes that the sample is selected through a well-defined probability sampling process, which ensures that every unit in the population has a known, non-zero chance of being included. Sampling weights, which reflect how much each sampled unit represents in the population, are essential tools in this approach. These weights allow analysts to make estimates that account for the sampling process and produce results that are representative of the population. A well-described survey design facilitates statistical analysis, supports effective data interpretation, and enables meaningful insights into complex phenomena. Not accounting for it may lead to biased estimates and misleading conclusions.

2.1 Parameters and estimators

Two common goals when analyzing survey data are to estimate the value of a characteristic for the whole population, such as total income, and to estimate the average value of that characteristic, such as the average income per person. These are referred to as the population total and the population mean, respectively. The design-based approach incorporates the probability sampling design in the inferential process.

Under probability sampling, every unit in the population has a known chance of being included in the sample. The *sample inclusion probabilities* are then used to calculate *basic sampling weights*. When the design is properly incorporated, estimates for population totals and means are unbiased. This means that, on average, the estimates will

equal the true population values if the survey were to be repeated many times under the same conditions.

The estimators used to make inferences about the parameters of the population use the sampling weights to create weighted sums of the survey data, which serve as estimates for the population values. If the weights are appropriately applied, the resulting estimates are consistent with the true values in the population.

As stated in the previous chapter, in some situations, the original sampling weights need adjustments to improve the accuracy of survey estimates. Adjustments may account for survey non-response, where weights of responding units are corrected to account for units that were selected for the survey but did not participate. These adjusted weights help to minimize biases in the estimates and make the results more reliable. If calibration was performed, the weights are modified to ensure that the weighted sums aligns more closely with known characteristics of the population, such as age or sex distributions.

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 $s = \{k_1, \ldots, k_n\} \subset U$ 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.

An important part of survey analysis is understanding the level of uncertainty in the estimates. Since we are working with a sample and not the entire population, there will always be some variability in the results. This variability is measured using the *sampling variance* of the estimate, which indicates how much the estimate might differ from the true population value in repeated samples. While there are theoretical formulas to calculate this variance, these can be complex and rely on information that is not always available to analysts. Under the design-based framework and assuming full response, $\widehat{Y}_{HT}$ is unbiased for Y and its sampling variance can be estimated unbiasedly by

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

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 $\pi_{kj} > 0\ \forall\, k, j \in U$.

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.

For this reason, simpler and more efficient methods are often used in practice, allowing analysts to quantify the uncertainty without requiring overly detailed information about the sampling design.

2.2 Approaches to Variance Estimation

When working with household surveys, the sample is usually only a small subset of the entire population. Because of this, it is important to provide not only the main estimates of interest, such as totals or averages, but also the level of uncertainty in these estimates. This uncertainty is often expressed as measures of variance or confidence intervals, which help us understand how much the estimates might differ if the survey was repeated.

Understanding and estimating uncertainty is a critical part of analyzing household survey data; by using proper methods, analysts can measure the reliability of their estimates. There are several methods to estimate the uncertainty in survey results: * One common approach is based on approximate formulas like *Taylor linearization*, which simplifies complex relationships between variables into linear ones;

- The *ultimate cluster* method, is used in surveys that collect data through multi-stage sampling, where groups of people (clusters) are selected at different stages, exactly as household survey data is collected;
- Estimating equations (Binder, 1983a), which comprises a unifying idea of sampling theory, provides a flexible framework for estimating totals, means, ratios, and other parameters as well as their corresponding variances.

2.2.1 Estimating equations

With the help of modern software, these methods can be implemented efficiently, ensuring accurate and meaningful analysis of survey data. Many population parameters can be written/obtained as solutions for *population estimating equations*. Variance estimation for these sample-based methods follows a consistent framework. Although the details can be technical, the key idea is that the same principles used to estimate totals can be applied to estimate variances. This generality makes the method simple and versatile, allowing it to be implemented in widely used software like the `R survey` package and the `Stata svy` functions. These tools automate much of the process, making

it accessible for analysts to estimate both population parameters and their associated uncertainties.

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. These equations provide a general way to define and calculate many population parameters, such as totals, means, and ratios. The concept is straightforward: population parameters can be defined as solutions to specific equations that involve all the units in the population. This approach is flexible and can be adapted to calculate many different types of parameters.

- 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 $\theta_U = \sum_{i \in U} y_i \ = \ Y$.

- For ratios of population totals, taking $z_i(\theta) = y_i - \theta x_i$, the corresponding population estimating equation is given by $\sum_{i \in U}(y_i - \theta x_i) = 0$. Solving for θ gives the *population ratio* $\theta_U = \sum_{i \in U} y_i / \sum_{i \in U} x_i \ = \ R$.

- Similarly, for population means, take $z_i(\theta) = y_i - \theta$.

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.

A consistent estimator for the variance of estimators obtained as solutions of sample estimating equations can be obtained as:

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

Where $\widehat{J}(\hat{\theta}) = \sum_{k \in s} d_k \, [\partial z_k(\theta)/\partial\theta]_{\theta=\hat{\theta}}$.

This approach implies that 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.

2.2.2 Ultimate Cluster Method

The *Ultimate Cluster* method is a straightforward and powerful approach for estimating the uncertainty (variance) of totals in surveys that use multi-stage cluster sampling designs. This method, proposed by Hansen et al. (1953), simplifies the complex nature of multi-stage designs by focusing only on the variation between the largest groups, known as Primary Sampling Units (PSUs). It assumes that the PSUs were selected randomly and independently, even if they were not actually chosen this way in the sampling process. This assumption allows for a simpler analysis while still providing reliable variance estimates.

The method considers only the variation between information available at the level of PSUs, and assumes 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 (households and individuals). The requirements for the application of this method are:

- One has unbiased estimators of totals for the variable of interest for each sampled PSU;
- Data are available for at least two sampled PSUs in each stratum (if the sample is stratified in the first stage);
- The survey dataset contains all the information regarding PSUs, strata and weights.

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

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

where U_{1h} and s_{1h} are the population and sample sets of PSUs in stratum h, $d_{hi} = 1/\pi_{hi}$, $\widehat{Y}_h = \sum_{i \in s_{1h}} d_{hi} \widehat{Y}_{hi}$ for $h = 1, \dots, H$. (See for example, (Shah et al., 1993), p. 4).

Although the method was originally proposed for estimation of variances of estimated totals, it can also be applied in combination with estimating equations approach to

obtain variance estimates for estimators of other population quantities that can be obtained as solutions to sample estimating equations. This makes the method versatile and useful for a wide range of applications in survey analysis.

One key assumption of the method is that, within strata, the PSUs were selected independently and with replacement. In reality, most surveys select PSUs without replacement, which is a more efficient design. However, the variance estimates produced by the *Ultimate Cluster* method are generally close enough to be useful, even under these conditions. This practical simplicity is why the method is widely used in survey analysis.

The *Ultimate Cluster* method is particularly attractive because of its simplicity. Survey practitioners often prefer it over more complex approaches that account for all stages of the sampling design. Although these detailed methods may provide slightly more accurate variance estimates, they are significantly harder to implement and require more detailed information about the sampling process. In contrast, this method offers a reasonable approximation that works well for most practical purposes, especially for estimating totals or averages. A discussion about Quality of this approximation and alternatives can be found in (Särndal et al., 1992), p. 153.

2.2.3 Bootstrap Method

When the user does not have access to information of PSUs or strata in the database, the *Ultimate Cluster* method cannot be used, and some other methods should be considered. Among them, we have replication-based methods; in particular, the *bootstrap* method, which comprises a powerful and flexible approach for estimating variances in surveys, particularly when dealing with complex survey designs that involve multiple stages or stratification. Originally proposed by Efron (1979), the version commonly used for household surveys is called the Rao-Wu-Yue Rescaling Bootstrap (Rao et al., 1992). This method is well-suited for stratified multi-stage sampling designs and has become a widely used tool for variance estimation with complex survey data.

The *bootstrap* method relies on creating many new "replicated" datasets, which are slightly different versions of the original sample. These replicated datasets mimic the process of repeatedly drawing samples from the population. By analyzing the variation in results across these datasets, we can estimate how much uncertainty there is in our estimates from the original sample.

1. First, we create a new sample for each stratum by randomly selecting primary sampling units (PSUs) from the original sample, allowing PSUs to be selected more than once (with replacement). Each selected PSU is included in the new dataset along with all its associated data. The size of this random sample with replacement is of $m_h - 1$ PSUs in each of the H design strata.

2. This process of creating new samples is repeated many times, usually hundreds or thousands, to produce multiple "replicated" datasets. That is, repeat Step 1 R times, and denote by $m_{hi}(r)$ the number of times that the PSU i of stratum h was selected for the sample in replicate r.

3. For each replicate, *bootstrap* weights are calculated for each unit. These weights account for how often each PSU appears in the replicate and ensure that the replicated datasets remain representative of the population. The *bootstrap* weight of unit k within PSU i of stratum h is $w_{hik}(r) = w_{hik} \times \frac{m_h}{m_h - 1} \times m_{hi}(r)$.

4. The parameter of interest, such as a total or mean, is estimated for each replicated dataset using the *bootstrap* weights. That is, for each replica r, calculate an estimate $\widehat{\theta}_{(r)}$ of the target parameter *theta* using the *bootstrap* weights $w_{hik}(r)$ instead of the original weights w_{hik}.

5. Finally, the variability of the results across all replicated datasets is used to estimate the variance. The idea is that the variation in these replicate estimates reflects the uncertainty in the original estimate. This estimate of the variance takes the following form:

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

where $\tilde{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{(r)}$ is the average of the replica estimates.

Whenever the original sampling weights w_{hik} receive non-response adjustments or are calibrated, the corresponding non-response adjustments and/or calibration of the basic weights must be repeated for each replica, so that the variance estimates adequately reflect the effects of the calibration and non-response adjustments on the uncertainty of the point estimates. This ensures that the variance estimates accurately reflect the additional uncertainty introduced by these adjustments.

The bootstrap method has several advantages. It works well for complex survey designs and can handle a wide range of parameters, including those that are difficult to estimate using traditional methods, such as medians or other nonlinear statistics. It also provides a way to estimate variances when other methods are not available or practical to use. The method is particularly helpful for survey analysts who may not have access to specialized software for calculating variances.

Many modern statistical software tools, including the `survey` package in R, support bootstrap replication and variance estimation, making it accessible to a wide range of users. While the bootstrap method is computationally intensive, requiring many replicated datasets to be created and analyzed, it is highly effective. It provides robust variance estimates even for complex parameters and remains one of the most flexible tools for analyzing survey data.

2.3 Using software to generate valid inferences

The design and analysis of information from household surveys must include extensive use of existing computational tools. This section reviews in detail the computational approaches of the statistical software used for each of the statistical processes required to publish official figures with high levels of accuracy and reliability. Specifically, for the following processes:

- Sample selection according to the defined sampling design;
- Generation of sampling weights for each individual and household;
- Modeling of nonresponse and statistical imputation;
- Calibration of sampling weights and adjustments for nonresponse;
- Estimation of sampling errors for each indicator of interest in the statistical production tables;
- Analysis of multivariate relationships between survey variables.

Nations (2005, Section 7.8) highlights the importance of including the structure of complex survey designs in the inference process for estimating official statistics from household surveys. It warns, with an empirical example, that failing to do so may result in biased estimates and underestimated sampling errors. Below are some key features that statistical software packages incorporate when managing data from complex survey designs, such as those found in household surveys. A more detailed review, including syntax and computational code, can be found in Heeringa et al. (2017a, Appendix A).

In general, these computational tools are designed to enhance the efficiency of variance approximation methods for complex samples, as well as replication techniques to estimate design-based variances (Westat, 2007). Some of these software packages are free to use, although most are licensed products requiring paid licenses. These products, in addition to providing descriptive statistics (such as means, totals, proportions, percentiles, and ratios), allow for fitting linear and logistic regression models. All resulting statistics are based on the survey design.

R

R is a free software increasingly used in social research, as it is likely to host the latest scientific findings implemented in this software (R Core Team, 2024). Being open-source, researchers can upload their own collections of computational functions to the official repository (`CRAN`) and make them available to the community. The `samplesize4surveys` package (Rojas, 2020) determines the sample size for individuals and households in repeated, panel, and rotational household surveys. The `sampling` (Tillé and Matei, 2016) and `TeachingSampling` (Gutiérrez, 2015) packages enable the selection of probabilistic samples from sampling frames under a wide variety of designs and algorithms. The `survey` package (Lumley, 2016), once the survey design is predefined using the `svydesign()` function, allows for analyzing household survey data and obtaining appropriate standard error estimates.

STATA

The `svy` environment provides tools for appropriate inference of official statistics from household surveys (STATA, 2017). The `svyset` command specifies variables identifying survey design features, such as sampling weights, clusters, and strata. The `svydescribe` command produces tables describing strata and sampling units at a given survey stage. Once survey design definitions are loaded, any model can be estimated, and the resulting statistics will be survey-design-based. The `svy` environment also supports predictive commands.

SPSS

The `complex samples` module in `SPSS` (IBM, 2017) supports the selection of complex samples through user-defined sampling schemes. Next, an analysis plan must be created by assigning design variables, estimation methods, and sample unit sizes. Once the sampling plan is defined, the module enables the estimation of counts, descriptive statistics, and crosstabulations. It is also possible to estimate ratios and regression coefficients in linear models, along with corresponding hypothesis test statistics. Finally, the module allows for estimating nonlinear models, such as logistic regressions, ordinal regressions, or Cox regressions.

SAS

This statistical software includes a procedure for selecting probabilistic samples called `SURVEYSELECT`, which integrates common selection methods such as simple random sampling, systematic sampling, probability proportional to size sampling, and stratified allocation tools. To analyze data from complex samples, specific procedures have been programmed (SAS, 2010):

- `SURVEYMEANS`: Estimates totals, means, proportions, and percentiles, along with their respective standard errors, confidence intervals, and hypothesis tests;
- `SURVEYFREQ`: Estimates descriptive statistics (e.g., totals and proportions) in one- and two-way tables, provides sampling error estimates, and analyzes goodness-of-fit, independence, risks, and odds ratios;
- `SURVEYREG` and `SURVEYLOGISTIC`: Fit linear and logistic regression models, respectively, estimating regression coefficients with associated errors and providing an exhaustive analysis of model properties;
- `SURVEYPHREG`: Fits survival models using pseudo-maximum likelihood techniques.

Chapter 3

Descriptive parameters

As stated before, household surveys play a critical role in tracking progress toward global objectives, such as the **Sustainable Development Goals (SDGs)**. For this purpose, descriptive analyses often include a range of specialized indicators designed to monitor outcomes like access to education, health services, and economic opportunities. These indicators are derived from the survey data and are essential for policymakers and organizations aiming to achieve sustainable development targets.

Descriptive parameters are the most commonly analyzed outputs from household survey data. These analyses focus on summarizing key characteristics of the population by estimating values for a variety of survey variables. The goal is to provide clear and meaningful insights into the population using data collected from a representative sample.

The most basic and frequently estimated parameters include **frequencies**, **proportions**, **means**, and **totals**. Means and totals provide average and cumulative values, respectively, which are useful for understanding population-level behaviors and trends. Frequencies can show the number of households/people in a specific category (e.g. number of poor people), while proportions can represent the share of households/people meeting a particular condition (e.g. poverty rate).

In recent years, the scope of descriptive analysis has expanded beyond these basic parameters. Analysts now estimate more complex metrics, such as **quantiles** of numeric variables, which help describe the distribution of values (e.g., the median income of households). Other widely used metrics include measures of **poverty** and **inequality** indicators, which are crucial for understanding economic disparities and informing policy decisions (e.g. FGT and Gini indices) - see Jacob et al. (2024).

3.1 Frequencies and proportions

One of the most fundamental tasks in household survey analysis is estimating the size of subpopulations, namely the number of people or households in specific categories, as well as the proportions they represent within the population. These estimates are crucial because they provide a snapshot of the demographic and socioeconomic profile of a population. Policymakers and planners use this information to make decisions about resource allocation, public policy design, and the development of social programs.

For example, understanding how many people live below the poverty line, how many are unemployed, or how many have completed a certain level of education provides valuable insights. These insights help address inequalities, support the design of targeted interventions, and promote equitable development across communities. The ability to understand the distribution across categories provides valuable information to address inequalities and promote equitable development.

To estimate the size of a population or subpopulation, analysts focus on categorical variables, which divide the population into distinct groups. For example, categories could represent different poverty levels, employment statuses, or education levels. The size of a population refers to the total number of individuals or households in the survey data who fall into a specific category. Population size estimates are calculated by combining the information collected from survey samples with *sampling weights*. These weights indicate how many people or households each surveyed unit represents in the broader population. A sampling estimator of a population size is given by the following expression:

$$\widehat{N} = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}$$

where s_{hi} is the sample of households or individuals in PSU i of stratum h; s_{1h} is the sample of PSUs within stratum h; and w_{hik} is the weight (expansion factor) of unit k within PSU i in stratum h.

Subpopulation size estimates work similarly but focus on a subset of the population defined by a specific characteristic. For example, if we want to estimate the number of people in a particular category, we would identify the relevant group in the survey data and sum up their weights. This approach allows analysts to estimate not only the total population size but also the size of specific groups of interest. This way, a binary variable should be defined, $I(y_{hik} = d)$. It will take the value one if unit k from PSU i in stratum h belongs to category d in the discrete variable y. A sampling estimator for this parameter is given by the following expression:

$$\widehat{N}_d = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\, I(y_{hik} = d)$$

Proportions describe the relative size of specific groups within the population. For instance, the proportion of households living below the poverty line is a critical measure for understanding socioeconomic disparities. To estimate a proportion, analysts calculate the weighted average of the binary variable. This approach ensures that the estimate accurately reflects the population distribution. As mentioned by Heeringa et al. (2017b), by recoding the original response categories into simple indicator variables y with possible values of 1 and 0 (e.g., 1=Yes, 0=No), the estimator for a proportion is defined as follows:

$$\hat{p}_d = \frac{\widehat{N}_d}{\widehat{N}} = \frac{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\, I(y_{hik} = d)}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}}$$

As this defines a nonlinear estimator, we can apply Taylor linearization to obtain the approximate variance of the above estimator by defining the corresponding estimating function as $z_{hik} = I(y_{hik} = d) - \hat{p}_d$. Many statistical packages provide proportion estimates and standard errors on a percentage scale.

When the target proportions are close to 0 or 1, special methods are used to ensure confidence intervals remain meaningful; notice that the limits of the traditional symmetric normal confidence intervals may fall outside the permissible range for proportions. This would have no interpretation due to the nature of the parameter. To address this issue, alternative confidence interval estimates, as proposed by Rust et al. (2007) and Dean and Pagano (2015) are available. One alternative based on using the logit transformation of the estimated proportion is:

$$CI\left(\hat{p}_d\ ;\ 1-\alpha\right) = \frac{exp\left[ln\left(\frac{\hat{p}_d}{1-\hat{p}_d}\right) \pm \frac{t_{1-\alpha/2,\,df} \times se(\hat{p}_d)}{\hat{p}_d(1-\hat{p}_d)}\right]}{1 + exp\left[ln\left(\frac{\hat{p}_d}{1-\hat{p}_d}\right) \pm \frac{t_{1-\alpha/2,\,df} \times se(\hat{p}_d)}{\hat{p}_d(1-\hat{p}_d)}\right]}$$

3.2 Totals, means and ratios

In household surveys, analyzing numerical data often involves estimating key descriptive measures such as means, totals, and ratios. These measures summarize important characteristics of the population and provide valuable insights for decision-making. The estimation process can be applied to the entire population or specific subgroups, depending on the research objectives. As mentioned by Heeringa et al. (2017a), the estimation of population totals or averages for a variable of interest, along with the estimation of corresponding variances, has played a crucial role in the development of probability sampling theory. Estimators of population means, proportions and ratios are all dependent on estimating component population totals, as we show in the sequence.

The estimation of population totals is a fundamental task in survey analysis. A total represents the sum of a specific variable (e.g., total income or total expenditure) across the entire population. For example, if the goal is to estimate the total income of all households in a country, we combine data from the sample using weights that account for the survey design and ensure representativeness. For single numeric survey variables, the simplest estimates are for totals and means. Ratios are often used to obtain summaries that relate two numeric variables. Estimates for such parameters can be obtained either for the entire population or disaggregated by domains of interest, depending on the research needs.

Once the sampling design is defined, which was done in the previous section, the estimation process for the parameters of interest is 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 $i = 1, 2, ..., m_h$, the estimator for the population total can be written as:

$$\widehat{Y} = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\, y_{hik}$$

Under full response, the Ultimate Cluster variance estimator for $\widehat{Y}$ was provided in section 9.2. Modern statistical tools, such as the `survey` package in R, make it straightforward to calculate totals and their associated uncertainties.

The confidence interval of level $1 - \alpha$ for the population total Y is given by:

$$\widehat{Y} \pm z_{1-\alpha/2} \times \sqrt{\widehat{V}_{UC}\left(\widehat{Y}\right)}$$

with $z_{1-\alpha/2}$ denoting the quantile of the Gaussian distribution leaving an area of $\alpha/2$ to its right.

Population means, or averages, are also very important and provide an understanding of the central tendency of a variable. For instance, the average income of households can indicate the general economic well-being of a population. A mean is calculated as the total of a variable divided by the population size. Since estimating a mean involves both totals and population sizes, the accuracy of a mean estimate depends on the accurate estimation of both components. Specialized techniques, such as resampling methods or Taylor linearization, are used to estimate the uncertainty associated with means. The estimation of the population means is a very important task in household surveys. An estimator of the population mean can be written as the ratio of two estimated population totals, as follows:

$$\widehat{\overline{Y}} = \frac{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\, y_{hik}}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}} = \frac{\widehat{Y}}{\widehat{N}}.$$

Since $\widehat{\overline{Y}}$ is a nonlinear estimator, there is no closed-form formula for exact the variance of this estimator. For this reason, either resampling methods or Taylor series approximations must be used. The latter may be achieved remembering that for the survey mean the sampling estimating equation requires defining $\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}(y_{hik} - \theta) = 0$, therefore we can apply the variance estimator given in section 9.2 with $z_{hik} = y_{hik} - \widehat{\overline{Y}}$.

Ratios provide insights into the relationship between two variables. For example, the ratio of household expenditures to income can reveal patterns in spending behavior. A ratio is calculated by dividing one total by another, such as total expenditures by total income. The accuracy of a ratio depends on the precise estimation of both totals. Ratios are particularly useful for creating indicators that help compare groups or track progress over time. As another example, ODS indicator N.17.6.1 is defined as the ratio of the number of broadband subscriptions per 100 inhabitants in a country or region.

Since a ratio is the quotient of two estimators of 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:

$$\widehat{R} = \frac{\widehat{Y}}{\widehat{X}} = \frac{\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}\ y_{hik}}{\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}\ x_{hik}}$$

For variance estimation, all you need to do is specify the estimating function as $z_{hik} = y_{hik} - \widehat{R}\, x_{hik}$, when y and x are the numerator and denominator variables, respectively, and apply the variace estimator given in section 9.2.

3.3 Correlations

Correlation analysis is a useful method for understanding the relationship between two numeric variables in survey data. For example, you might be interested in knowing whether household income and expenditure are related, and if so, how strongly. The Pearson correlation coefficient is commonly used to measure this relationship as it quantifies the strength and direction of a linear relationship between two variables. Its value ranges from -1 to 1:

- A **positive value** indicates that as one variable increases, the other also tends to increase;
- A **negative value** indicates that as one variable increases, the other tends to decrease;

- A value close to **zero** suggests little to no linear relationship between the variables.

When analyzing survey data, the correlation is estimated using the survey weights. These weights ensure that the estimated correlation reflects the relationships in the entire population, not just the sample. Weighted correlations adjust for the complex survey design, accounting for stratification, clustering, and unequal probabilities of selection. To compute the correlation coefficient, we look at how the variables vary together (their covariance) and normalize this by their individual variations. This normalized measure ensures the correlation is unaffected by the units of measurement of the variables, making it easier to interpret.

The Pearson correlation coefficient between two numeric survey variables, say x and y, can be estimated using

$$\hat{\rho}_{xy} = \frac{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \left(y_{hik} - \widehat{\overline{Y}}\right) \left(x_{hik} - \widehat{\overline{X}}\right)}{\sqrt{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \left(y_{hik} - \widehat{\overline{Y}}\right)^2} \sqrt{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \left(x_{hik} - \widehat{\overline{X}}\right)^2}}$$

Modern statistical software, such as R, provides functions to calculate weighted Pearson correlation coefficients directly. Tools like the `survey` package ensure that the correlations are estimated correctly, accounting for the survey design. This allows analysts to obtain accurate and meaningful measures of association.

3.4 Percentiles and inequality measures

Percentiles and quantiles are useful tools for analyzing the distribution of data beyond just the average. These measures divide data into segments to show how values are spread. For example, the 10th percentile indicates the value below which 10% of the data falls, while the median (50th percentile) divides the data into two equal halves. These measures help describe not only central tendencies but also the spread and variation within a dataset. For instance, identifying the top 10% of income earners might guide tax policy, while finding the bottom 15% could inform subsidy programs. The estimation of percentiles relies on the cumulative distribution function (CDF), which represents the proportion of the population with values less than or equal to a given number. Once the CDF is calculated using survey data and weights, percentiles and quantiles can be derived. The CDF for a variable y in a finite population of size N is defined as follows:

$$F(t) = \frac{1}{N} \sum_{h=1}^{H} \sum_{i \in U_{1h}} \sum_{k \in U_{hi}} I(y_{hik} \leq t)$$

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

$$\widehat{F}(t) = \frac{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} I(y_{hik} \leq t)}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}}$$

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 1/2. Thus, the estimated median is the smallest value where the estimated CDF is greater than or equal to 1/2. Following Heeringa et al. (2017b), 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:

$$\widehat{F}(y_{(j)}) \leq q \leq \widehat{F}(y_{(j+1)})$$

Hence, the estimator of the q-th quantile $y_{(q)}$ is given by:

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

Quantiles are inherently nonlinear measures, making their variance estimation more complex. Kovar et al. (1988) present results from a simulation study where they recommend using the *Balanced Repeated Replication* (BRR) technique.

Economic inequality is a critical area of focus for governments and international organizations. The **Gini coefficient** is a widely used measure to quantify inequality in income or wealth distributions. It is derived by comparing the income distribution of a target population to a perfectly equal distribution. In household surveys, it is calculated using weights that account for the survey design, ensuring representativeness. A normalized version of these weights is often used to simplify the calculations. The Gini coefficient ranges from 0 to 1, where 0 indicates perfect equality (everyone has the same income) and values closer to 1 indicate greater inequality. The Gini coefficient is critical for tracking changes in income distribution over time and comparing inequality levels across regions or countries.

Following the estimating equation proposed by Binder and Kovacevic (1995), the estimator for the Gini coefficient is given by:

$$\widehat{G} = \frac{2 \times \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w^*_{hik} \widehat{F}_{hik}\, y_{hik} - 1}{\widehat{\overline{Y}}}$$

where w^*_{hik} is a normalized sampling weight, defined as

$$w^*_{hik} = \frac{w_{hik}}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}}$$

and $\widehat{F}_{hik}$ represents the estimated CDF for individual k in cluster i of stratum h. Osier (2009) and Langel and Tillé (2013) provide important computational details for estimating the variance of this complex estimator.

3.5 NSO – Practical example

In this subsection a NSO will share how they do disseminate its results on basic descriptive statistics, how they publish the resulting tables and how do they deal with the suppression of estimates that do not reach expected quality.

Chapter 4

Associations between variables

4.1 Motivation and concepts

Household surveys often collect data on categorical variables, such as employment status, educational attainment, or access to services. Understanding whether two categorical variables are related, or *associated*, is an important aspect of survey analysis. For example, are employment status and access to the internet connected in a meaningful way? Categorical variables are those that divide the population into distinct groups or categories. For example:

- **Employment status** might have categories like "employed," "unemployed," and "not in the labor force";
- **Educational attainment** might include categories such as "primary," "secondary," and "tertiary."

This section introduces methods to describe and infer associations between pairs of categorical variables. When analyzing associations between two categorical variables, we are interested in whether the distribution of one variable depends on the categories of the other. To assess the relationship between two categorical variables, analysts examine how often different combinations of categories occur. For example, they might count how many individuals fall into each pairing of employment status and educational attainment. These counts are then used to calculate proportions, which describe the relative frequency of each pairing within the population.

Analyzing associations between categorical variables is useful in many contexts, such as:

- **Policy Development**: Understanding the relationship between education and employment helps design effective workforce policies;

- **Program Evaluation**: Assessing whether access to healthcare varies by income level can inform targeted interventions;
- **Social Research**: Studying connections between demographic factors and access to services provides insights into societal trends.

In practice, this analysis often starts with a **contingency table**, a grid that shows the counts or proportions of units in each combination of categories for the two variables. For example, one axis of the table might list employment statuses, while the other lists levels of educational attainment.

We start by defining some notation. Let x and y denote two categorical variables, having R and C classes respectively. In order to formulate hypothesis tests for the independence between x and y, we need to consider a *superpopulation model*. We assume that the pairs (x_{hik}, y_{hik}) correspond to observations from identically distributed random vectors $(X; Y)$, that have joint distribution specified by

$$P_{rc} = Pr\left(X = r \ ; \ Y = c\right) \quad \text{for } r = 1, ..., R \text{ and } c = 1, ..., C$$

with $\sum_r \sum_c P_{rc} = 1$.

If a census could be carried out to collect data on x and y from every unit in the population, we could calculate the population counts of units having classes (r, c) for (x, y) given by:

$$N_{rc} = \sum_{h=1}^{H} \sum_{i \in U_{1h}} \sum_{k \in U_{hi}} I\left(x_{hik} = r \ ; \ y_{hik} = c\right)$$

and the corresponding population proportions as $p_{rc} = N_{rc}/N_{++}$, where $N_{++} = \sum_r \sum_c N_{rc}$ denotes the total number of units in the population.

Under the superpopulation model, the population proportions p_{rc} could be used to estimate (or approximate) the unknown probabilities P_{rc}. Since in most instances we will have samples, not censuses, the population proportions p_{rc} must be estimated using weighted estimators provided in the previous sections.

4.2 Cross-tabulations

Cross-tabulations, also known as contingency tables, are a fundamental tool in survey analysis. They organize data into a table format, showing the frequency distribution of two or more categorical variables. By summarizing relationships between these variables, cross-tabulations help researchers identify patterns and associations that might otherwise go unnoticed. This type of analysis is widely used in research and policy

decision-making, as it provides a straightforward way to explore how different variables interact. For example, a contingency table might examine how employment status varies by educational attainment, or how access to the internet differs between urban and rural households.

Tests of independence can be used to assess whether the cross-classified variables are related or independent. This type of analysis is important in many research and decision-making settings. In the specialized literature, cross-tabulations are also referred to as *contingency tables*. Here a table is a two-dimensional array with rows indexed by $r = 1, ..., R$ and columns indexed by $c = 1, ..., C$. Such tables are widely used in household survey analysis as they summarize the relationship between categorical variables in terms of frequency counts.

A contingency table aims to succinctly represent the association between different categorical variables. It is a grid with rows and columns that represent the categories of two variables. Each cell in the table contains the frequency or proportion of observations that fall into the corresponding combination of categories. The rows might represent categories of a domain defining variable such as "education level" (primary, secondary, tertiary). The columns might represent categories of another variable, such as "employment status" (employed, unemployed, not in the labor force). The table can also include **marginal totals**, which summarize the data for each row or column, and a **grand total**, representing the overall population.

In household surveys, frequencies in contingency tables are calculated using survey weights. These weights ensure that the estimates accurately reflect the entire population, accounting for the sampling design. For each cell, the weighted frequency represents the estimated number of individuals in the population with the corresponding combination of categories. For instance: we consider the case of a two-way contingency table. For most household sample surveys, a typical tabular output comprises the weighted frequencies that estimate the population frequencies, as follows:

		y		
x	1	...	C	row marg.
1	$\widehat{N}_{11}$	...	$\widehat{N}_{1C}$	$\widehat{N}_{1+}$
...	...	$\widehat{N}_{rc}$	...	...
R	$\widehat{N}_{R1}$	...	$\widehat{N}_{RC}$	$\widehat{N}_{R+}$
col. marg.	$\widehat{N}_{+1}$	...	$\widehat{N}_{+C}$	$\widehat{N}$

where the estimated frequency in cell (r, c) is obtained as

$$\widehat{N}_{rc} = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \, I\left(x_{hik} = r \; ; \; y_{hik} = c\right)$$

and $\widehat{N}_{r+} = \sum_c \widehat{N}_{rc}$, $\widehat{N}_{+c} = \sum_r \widehat{N}_{rc}$ and $\widehat{N}_{++} = \sum_r \sum_c \widehat{N}_{rc}$.

Weighted frequencies can also be converted into **proportions**, which indicate the relative size of each group compared to the total population. Proportions are particularly useful when comparing groups of different sizes or when focusing on the relative distribution of categories. The estimated proportions from these weighted sample frequencies are obtained as follows:

$$\hat{p}_{rc} = \frac{\widehat{N}_{rc}}{\widehat{N}_{++}}$$

$\hat{p}_{r+} = \widehat{N}_{r+}/\widehat{N}_{++}$, and $\hat{p}_{+c} = \widehat{N}_{+c}/\widehat{N}_{++}$.

Two-way tables can also display the estimates of population relative frequencies, as shown below:

		y		
x	1	...	C	row marg.
1	$\hat{p}_{11}$	...	$\hat{p}_{1C}$	$\hat{p}_{1+}$
...	...	$\hat{p}_{rc}$	...	...
R	$\hat{p}_{R1}$	...	$\hat{p}_{RC}$	$\hat{p}_{R+}$
col. marg.	$\hat{p}_{+1}$	...	$\hat{p}_{+C}$	1

While tables are a clear way to present data, visualizations such as stacked bar charts or heatmaps can enhance understanding by highlighting patterns and differences between categories. These visuals complement contingency tables, making it easier to communicate findings to a broad audience. More on this will be elaborated in section 9.7.

4.3 Testing for independence

In household surveys, it is often important to determine whether two categorical variables are associated or independent (i.e., whether the distribution of one variable is unaffected by the categories of the other). For example, is there a relationship between "educational level" and "employment status"? To answer such questions, *independence tests* are used. These tests compare the observed data with what would be expected if the two variables were completely unrelated.

To perform these tests, analysts rely on models that assume the data comes from a larger, hypothetical population (a *superpopulation*). The observed data from the survey is treated as a sample from this superpopulation, and the analysis aims to draw conclusions about the larger population. The starting point for testing independence

is the **null hypothesis**, which assumes that the two variables are independent. This means the likelihood of being in any combination of categories is simply the product of their marginal probabilities.

To test this hypothesis, observed frequencies (or proportions) in a contingency table are compared with the expected frequencies under the null hypothesis. If the observed and expected values differ significantly, the null hypothesis of independence is rejected, suggesting an association between the variables. This way, it is possible to perform independence tests to verify whether x and y are associated. Following Heeringa et al. (2017b), the null hypothesis that x and y are independent is defined as:

$$H_0) \quad P^0_{rc} = P_{r+} \times P_{+c} \quad \forall \; r = 1, \dots, R \text{ and } c = 1, \dots, C.$$

Hence, to test the independence hypothesis we compare the estimated proportions $\hat{p}_{rc}$ with the estimated expected population proportions under the null P^0_{rc}. If there is a large difference between them, then the independence hypothesis would not be supported by the data.

Testing for independence in survey data requires adjustments to account for the sampling design, which often includes stratification, clustering, and unequal probabilities of selection. The **Rao-Scott adjustment** modifies traditional chi-square tests to incorporate these design effects. The test statistic is adjusted for the survey design using a measure called the **generalized design effect (GDEFF)**, which accounts for the complexity of the sampling design. It follows a chi-square distribution under the null hypothesis. Therefore, the following Pearson Rao-Scott adjusted test statistic X^2_{RS} (Rao and Scott, 1984) is defined by:

$$X^2_{RS} = \frac{n_{++}}{GDEFF} \sum_r \sum_c \frac{\left(\hat{p}_{rc} - \widehat{P}^0_{rc}\right)^2}{\widehat{P}^0_{rc}}$$

where $\widehat{P}^0_{rc} = \hat{p}_{r+} \times \hat{p}_{+c}$ estimates the cell frequencies under the null hypothesis and $GDEFF$ is an estimate of the generalized design effect (Heeringa et al. (2017b) p. 177). Under the null hypothesis of independence, the large sample distribution of X^2_{RS} is $\chi^2_{[(R-1)(C-1)]}$.

When the sample size or degrees of freedom is small, adjustments to the X^2_{RS} test statistic can improve accuracy. These adjustments use an **F-distribution** instead of the chi-square distribution, making the tests more robust for smaller datasets. As mentioned by Heeringa et al. (2017b), it was Fay (1979), along with Fellegi (1980), who began proposing corrections to Pearson's chi-square statistic based on a generalized design effect. Rao and Scott (1984) later expanded the theory of generalized design effect corrections for these statistical tests, as did Thomas and Rao (1987).

The Rao-Scott adjustment requires the calculation of generalized design effects, which are analytically more complex than Fellegi's approach. Nevertheless, Rao-Scott adjusted statistics are now the standard for analyzing categorical survey data in software systems such as R, Stata and SAS.

4.4 Tests for group comparisons

Comparing group means is a common goal in household survey analysis. For example, researchers might ask: "Is there a significant difference in average income between male and female headed households?" To answer such questions, statistical tests are used, adapted to account for the complexities of survey data, such as stratification, clustering, and unequal selection probabilities. This section explains the methods for testing differences in means, adjusted for survey design, with examples to illustrate their application.

To determine whether the means of two groups are significantly different we will introduce t-test and contrasts adjusted for the sampling design.

4.4.1 Hypothesis Test for the Difference of Means

A hypothesis test is a statistical procedure used to evaluate evidence in favor of or against a statement or assumption about a population. In this process, a null hypothesis (H_0) is proposed, representing the initial statement that needs to be tested, and an alternative hypothesis (H_1), which is the statement opposing the null hypothesis. These statements may be based on some belief or past experience and will be tested using the evidence gathered from the survey data. If it is suspected that the parameter θ is equal to a particular value θ_0, the possible combinations of hypotheses that can be tested are:

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases} \quad \begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta > \theta_0 \end{cases} \quad \begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta < \theta_0 \end{cases}$$

One of the two hypotheses will be considered true only if the statistical evidence, which is obtained from the sample, supports it. The decision about which hypothesis is true is based on the statistical evidence gathered from the data. This process is called **hypothesis testing**.

In many cases, important parameters of interest, such as differences in means or weighted sums of means, can be expressed as a linear combination of various descriptive statistics. These combinations are often used in constructing economic indices or comparing population means. The variance of these combinations is important for understanding the precision of the estimate. That is, parameters can be expressed as a

linear combination of measures of interest. The most common cases are differences in means, weighted sums of means used to construct economic indices, etc. Thus, consider a function that is a linear combination of J descriptive statistics, as shown below:

$$f(\theta_1, ..., \theta_J) = \sum_{j=1}^{J} a_j \theta_j$$

where the a_j are *known* constants. An estimator of this function is given by:

$$\hat{f}\left(\hat{\theta}_1, ..., \hat{\theta}_J\right) = \sum_{j=1}^{J} a_j \hat{\theta}_j$$

And its variance is calculated as follows:

$$Var\left(\sum_{j=1}^{J} a_j \hat{\theta}_j\right) = \sum_{j=1}^{J} a_j^2 \, Var\left(\hat{\theta}_j\right) + 2 \times \sum_{j=1}^{J-1} \sum_{k>j}^{J} a_j a_k \, Cov\left(\hat{\theta}_j, \hat{\theta}_k\right)$$

As seen in the variance expression for the estimator, it requires the variances of the individual estimators, as well as the covariances of pairs of estimators.

In the context of comparing means between two populations, there are several potential hypotheses that can be tested. On the one hand, the null hypothesis may state that the means of two populations are equal. On the other, the alternative hypothesis could suggest that the means are different, or that one is greater than or less than the other.

Of particular interest is analyzing the difference in population means. In order to formulate the hypothesis tests for this case, we need to consider a *superpopulation model*. We assume that y_{hik} correspond to observations from identically distributed random variables Y having means $\mu_{y,j}$ if unit k belongs to domain j, with $j = 1, 2$. Then we can define the difference in population means between domains 1 and 2 as $\mu_{y,1} - \mu_{y,2}$. As an example, consider that $\mu_{y,1}$ is the average household income for households with male heads of household, and $\mu_{y,2}$ is the average household income for households with female heads.

This difference in means can be consistently estimated by:

$$\widehat{\overline{Y}}_1 - \widehat{\overline{Y}}_2$$

where $\widehat{\overline{Y}}_j$ is the sample estimator of $\mu_{y,j}$ $(j = 1, 2)$.

Considering the parameter of interest in this section, the hypotheses to test are typically:

- Null hypothesis: There is no difference between the means.

- Alternative hypothesis: There is a difference, which could be in either direction (greater or less).

To test one of these hypothesis, the following test statistic is used, which follows a t-student distribution with df degrees of freedom, calculated as the difference between the number of PSUs m in the sample and the number of strata H.

$$t = \frac{\widehat{\overline{Y}}_1 - \widehat{\overline{Y}}_2}{se\left(\widehat{\overline{Y}}_1 - \widehat{\overline{Y}}_2\right)} \sim t_{[df]}$$

Where:

$$\widehat{se}\left(\widehat{\overline{Y}}_1 - \widehat{\overline{Y}}_2\right) = \sqrt{\widehat{Var}\left(\widehat{\overline{Y}}_1\right) + \widehat{Var}\left(\widehat{\overline{Y}}_2\right) - 2\,\widehat{Cov}\left(\widehat{\overline{Y}}_1\,;\widehat{\overline{Y}}_2\right)}$$

If a confidence interval is needed for the difference in means, it is constructed using the estimated difference and the standard error, along with the appropriate critical value from the t-distribution. This interval provides a range of plausible values for the true difference in means, offering a more complete understanding of the data. It would be constructed as follows:

$$\widehat{\overline{Y}}_1 - \widehat{\overline{Y}}_2 \;\pm\; t_{[df]}\;\widehat{se}\left(\widehat{\overline{Y}}_1 - \widehat{\overline{Y}}_2\right)$$

4.5 NSO – Practical example

In this part an NSO will share its experiences on dealing with statistical comparisons among groups and how do they present the results in tables.

Chapter 5

Regression: modelling survey data

Regression modeling is a powerful tool for analyzing relationships between variables in survey data. It allows researchers to estimate how one or more independent variables (predictors) influence a dependent variable (outcome). For instance, consider a researcher who is modeling household income (dependent variable) as a function of education level and employment status (independent variables) using household survey data. Such data are often obtained from surveys that adopted complex sampling designs that include stratification, clustering, and unequal probabilities of selection. These features must be accounted for to ensure valid inferences when fitting regression models, since ignoring the sampling design can lead to:

- **Biased Estimates**: Unequal probabilities of selection mean some household/individuals represent more of the population than others. Without weights, the model may disproportionately reflect oversampled groups.
- **Underestimated Standard Errors**: Clustering and stratification affect the variability of estimates. Ignoring these features can result in overly narrow confidence intervals and inflated significance levels.

In this section, we explore how survey weights and sampling design features are incorporated into regression model specification and fitting. We also discuss a parsimonious solution to the challenges posed by weighting. Modeling survey data requires careful consideration of the sampling design to ensure valid inferences. Incorporating survey weights and adjusting for clustering and stratification allows researchers to produce accurate, representative, and reliable results.

5.1 To weight or not to weight?

When performing regression analysis on survey data, a key question arises: should we include survey weights in the estimation of regression parameters and their associated

standard errors? This question has sparked debate among researchers - Skinner et al. (1989), Pfeffermann (2011) - as there are trade-offs between accounting for the complex design features and simplifying the model for ease of interpretation and efficiency.

On the one hand, when including sampling features we are making sure to achieve **Population Representativeness**, because survey weights ensure that the regression model reflects the true population distribution, correcting for oversampling or undersampling of certain groups. Also we will obtain **accurate variance estimates**, because we are adjusting for stratification, clustering, and unequal selection probabilities, providing valid standard errors and confidence intervals.

On the other hand, including sampling design features may yield to an **increasing of variance**, and can inflate the variance of parameter estimates, particularly if the weights vary widely. Also, extreme or highly variable weights can lead to unstable estimates, where certain observations disproportionately influence the model, making it unstable. For explanatory or analytical purposes (e.g., understanding relationships between variables), unweighted models can sometimes provide more efficient and stable estimates.

To answer the question: when to weight?, we can distinguish two scenarios:

- **Descriptive Inference**: Always weight. The primary goal is to reflect the population, and survey weights are essential for accuracy.
- **Analytical Inference**: Consider unweighted or weight-adjusted models. If the goal is to explore relationships or test hypotheses, weighting may not always be necessary, particularly if the model structure includes key design variables (e.g., strata or clusters).

5.2 Some inferential approaches to modelling data

When working with survey data, one of the key challenges is understanding and addressing the variability inherent in the data. This variability comes from two primary sources. The first source is the **sampling design**, which refers to the way the data was collected. The second source of variability arises from the **model itself**, which is used to analyze the data and make inferences about the population.

To combine these two sources of variability into a coherent framework, advanced inferential methods are required. These methods aim to respect the structure of the survey design while also accounting for the assumptions and uncertainties within the model. Two main approaches used for this purpose are **pseudo-likelihood** and **combined inference** (XXXXX citation neeeded).

The **pseudo-likelihood approach** modifies the traditional likelihood methods used in statistical modeling to account for the complexities of the survey design. In simpler

terms, it adjusts the standard modeling techniques to ensure that they properly incorporate the way the sample was drawn. This adjustment is crucial because ignoring the sampling design can lead to biased estimates and incorrect conclusions about the population.

On the other hand, **combined inference** seeks to integrate the information from the survey design and the model in a unified way. This approach ensures that the uncertainties from both sources —sampling and model— are reflected in the final results. By blending these components, combined inference provides a more comprehensive view of the variability and helps produce more reliable estimates.

5.3 Linear models

A regression model seeks to explain how changes in one or more independent (explanatory) variables affect a dependent (response) variable. In its simplest form, linear regression examines the relationship between a single independent variable and a dependent variable. The dependent variable is the outcome of interest, while the independent variable represents factors that may influence it. The model also includes an error term, which captures unexplained variability in the data.

5.3.1 Basic Definitions

As noted by Heeringa et al. (2017a), the first authors to empirically discuss the impact of complex sampling designs on regression model inferences were Kish and Frankel (1974), who highlighted the challenges posed by complex sampling designs. Later, Fuller (1975) developed a variance estimator for regression model parameters based on Taylor linearization with unequal weighting of observations under stratified and two-stage sampling designs.

As is well known, the use of regression model theory requires certain statistical assumptions to be met, which can sometimes be challenging to verify in practice. In this regard, Shah et al. (1977) discuss some aspects related to the violation of these assumptions and provide appropriate methods for making inferences about the estimated parameters of linear regression models using survey data.

Similarly, Binder (1983b) obtained the sampling distributions of estimators for regression parameters in finite populations and related variance estimators in the context of complex samples. Skinner et al. (1989) studied the properties of variance estimators for regression coefficients under complex sample designs. Later, Fuller (2002) provided a summary of estimation methods for regression models containing information related to complex samples. Finally, Pfeffermann (2011) discussed various approaches to fitting linear regression models to complex survey data, presenting empirical support for the use of the "*q-weighted*" method, which is recommended in this document.

A simple linear regression model is defined as $y = \beta_0 + \beta_1 x + \varepsilon$, where y represents the dependent variable, x is the independent variable, and β_0 and β_1 are the model parameters. The variable ε is known as the *random error* of the model.

For more complex situations, multiple linear regression models allow for the inclusion of several independent variables. This approach helps to account for the simultaneous effects of multiple factors on the outcome. In these models, each independent variable is associated with a coefficient, which indicates the strength and direction of its relationship with the dependent variable. Notice that a positive coefficient suggests that as the corresponding independent variable increases, the dependent variable also increases. Generalizing the previous model, multiple linear regression models are defined by allowing the dependent variable to interact with two or more variables, as presented below:

$$y = x\beta + \varepsilon = \sum_{j=0}^{p} \beta_j x_j + \varepsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Another way to write the multiple regression model is:

$$y_k = x_k \beta + \varepsilon_k$$

Where, $x_k = \left(1 \ , x_{1k} \ , \ldots \ , x_{pk}\right)$ and $\beta = \left(\beta_0 \ , \ , \beta_1 \ , \ldots \ , \beta_p\right)^T$.

The subscript k refers to the sample element or respondent in the dataset. Regression models are built on several assumptions about the data. Heeringa et al. (2017a) present some considerations for regression models, which are described below:

- $E\left(\varepsilon_k \mid x_k\right) = 0$. That is, the average error for any given value of the independent variable is assumed to be zero, meaning that the model does not systematically over- or under-predict outcomes.
- $Var\left(\varepsilon_k \mid x_k\right) = \sigma^2_{y|x}$. That is, the variability of the errors should be constant across all levels of the independent variables, a property known as homoscedasticity (homogeneity of variance).
- $\varepsilon_k \mid x_k \sim N\left(0, \sigma^2_{y|x}\right)$ (normality of errors), meaning that the residuals conditioned on the covariates follow a normal distribution. This property also extends to the response variable y_k.
- $cov\left(\varepsilon_k, \varepsilon_l \mid x_k, x_l\right) = 0 \ \forall k \neq l$. That is, the errors for different observations should be independent, meaning that the outcome for one observation does not influence another. This way, the residuals in different observed units should not be correlated with the values given by their predictor variables.

When these assumptions are met, regression models can provide accurate and unbiased estimates of relationships between variables. The predicted values from the model represent the expected outcomes based on the observed values of the independent variables. This makes regression a useful tool for understanding patterns in data and making informed predictions.

Once the linear regression model and its assumptions are defined, it can be deduced that the best unbiased linear estimator is defined as the expected value of the dependent variable conditioned on the independent variables x, as:

$$E\left(y \mid x\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

5.3.2 Estimation of Parameters

When working with survey data, the goal of regression analysis is to estimate the relationships between variables *in the population*. If a complete census were available, the calculation of regression parameters would be straightforward because we would have all the data. However, in practice, we work with survey samples, which introduce additional complexities, particularly when these samples are drawn using complex designs.

When estimating the parameters of a linear regression model considering that the observed information comes from surveys with complex samples, the standard approach to estimating regression coefficients and their standard errors is altered. The main reason for this change is that data collected through a complex survey generally does not have an identical distribution, and the assumption of independence cannot be maintained since the sample design is constructed with dependencies (as most complex designs include stratification, clustering, unequal selection probabilities, etc.).

In this context, when fitting regression models with such datasets, using conventional estimators derived from traditional methods (such as maximum likelihood, for example) will induce bias because these methods assume the sample data are independently and identically distributed and come from a specific probability distribution (binomial, Poisson, exponential, normal, etc.).

For illustrative purposes, the estimation of the parameter β_1 and its variance for a simple linear regression will be shown. The extension to multiple regression parameter estimation is algebraically complex and beyond the scope of this book. Below is the estimation of the slope and its variance in a simple linear regression model:

$$\widehat{\beta_1} = \frac{\sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \left(y_{hik} - \widehat{\overline{Y}}\right) \left(x_{hik} - \widehat{\overline{X}}\right)}{\sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \left(x_{hik} - \widehat{\overline{X}}\right)^2}$$

Understanding the uncertainty in the estimated parameters is essential for making valid inferences. In regression analysis with complex samples, variance estimation involves calculating measures that reflect the variability introduced by the sampling design. These calculations often rely on weighted sums and include adjustments for dependencies in the data.

For multiple regression, the variance of each coefficient is computed while considering its relationship with other coefficients. This results in a variance-covariance matrix, which summarizes the variability and relationships between all the estimated coefficients. The matrix provides a comprehensive view of the precision of the estimates and is a key tool for interpreting regression results. As a generalization, according to Kish and Frankel (1974), the variance estimation of coefficients in a multiple linear regression model requires weighted totals for the squares and cross-products of all combinations of y and $x = \{1, x_1, \ldots, x_p\}$.

5.4 Working with weights

When analyzing data from complex surveys, a critical question arises: how should survey weights be used in regression models? Heeringa et al. (2017a) addresses the problem of how to correctly weight regression models and whether expansion factors should be used to estimate regression coefficients when working with complex survey data. In this context, two main approaches exist for incorporating weights into regression models when working with complex survey data:

First, the **design-based approach** focuses on making inferences about the entire population. Here, survey weights are essential to ensure unbiased estimates of the regression coefficients. These weights account for the survey design, including unequal probabilities of selection. However, this approach has a limitation: it does not protect against model misspecification. If the model does not correctly describe the relationships in the population, the estimates will still be unbiased for the specified model but not necessarily meaningful for the population. If the researcher fits a poorly specified model, unbiased estimates of the regression parameters would be obtained in a model that does not correctly describe the relationships in the finite population.

In contrast, the **model-based approach** argues that weights are unnecessary if the model is correctly specified for the sample, and the sampling is non-informative. This approach assumes that the relationships between the variables are well-represented by the model, regardless of the sampling design. Using weights in this case can increase the variability of the estimates, leading to unnecessarily larger standard errors.

The choice between these two approaches depends on the context and the sensitivity of the inferences to the inclusion of weights. A practical recommendation is to fit regression models both with and without weights using statistical software and compare the results. If including weights leads to significant changes in the regression coefficients

or conclusions, it indicates that either the model may not be correctly specified or the sampling was informative, and therefore weighted estimates should be preferred. On the other hand, if weights only increase the standard errors without altering the coefficients meaningfully, the model is likely well-specified, and weights may not be necessary.

To address the challenges of using raw sampling weights, several adjustments have been proposed to balance accuracy and efficiency. Some are listed below:

1. **Senate Sampling Weights**: This approach scales weights so that the sum of the weights equals the sample size rather than the population size. The goal is to retain representativeness while reducing the variability of weights. This adjustment is particularly useful in large samples where the raw weights are excessively variable. This approach preserves relative differences in weights.

$$w_k^{\text{Senate}} = w_k \times \frac{n}{\sum w_k}$$

2. **Normalized Weights**: these weights are used to rescale the raw weights to sum to 1. This adjustment ensures that the overall weight does not inflate variances unnecessarily. This approach is useful when comparing models with different subsets of data or when variance inflation is a concern.

$$w_k^{\text{Normalized}} = \frac{w_k}{\sum w_k}$$

3. **Pfeffermann Model Weights**: this approach incorporates weights into the likelihood function of the regression model, allowing the model to use weights adaptively. This method combines the benefits of weighting and model-based inference. This approach also adjusts for the variability introduced by weights while retaining design-based properties. It is ideal for models requiring both descriptive and analytical inference.

The third solution (called the *q-weighted approach*) was proposed by Pfeffermann (2011) who suggested a slightly different specification of the weights. This adjustment modifies the original weights to reflect the relationships in the data more accurately, reducing variability while still accounting for the complex survey design. It also balance the benefits of both design- and population-based approaches. The steps in this approach are as follows:

1. A regression model is fitted to the original survey weights, using the predictor variables in the primary regression model of interest.

2. Predicted survey weights are obtained for each case based on the predictor variables.

3. The original survey weights are divided by these predicted weight values, creating adjusted weights.

4. These adjusted weights are then used in the final regression model.

The q-weighted approach provides a middle ground. It retains the benefits of the design-based approach by incorporating survey weights but reduces the variance typically associated with their use. At the same time, it considers the relationships captured by the model, aligning more closely with the model-based perspective. This makes it particularly useful for situations where the choice between the two paradigms is unclear or when both perspectives have merit.

5.5 Model Diagnostics

When using statistical models with household survey data, it is essential to evaluate the quality of the models to ensure the validity of the conclusions. This involves a series of checks and analyses that focus on the assumptions and performance of the model. These checks help confirm whether the model adequately describes the data and whether the results can be trusted. Model diagnostics begin with evaluating whether the model fits the data well. This involves analyzing several aspects:

- **Model fit**: It is important to determine whether the model provides an adequate fit to the data, i.e. explains a good portion of the variability of the response;
- **Distribution of errors**: Examine whether the errors are normally distributed;
- **Error variance**: Check whether the errors have constant variance;
- **Error independence**: Verify that the errors can be assumed to be uncorrelated;
- **Influential data points**: Identify if any data points have an unusually large influence on the estimated regression model;
- **Outliers**: Detect points that do not follow the general trend of the data, known as outliers.

5.5.1 Coefficient of determination

The coefficient of determination, also known as the multiple correlation coefficient (R^2), is a common measure of goodness-of-fit in a regression model. This coefficient estimates the proportion of variance in the dependent variable explained by the model and ranges between 0 and 1. A value close to 1 indicates that the model explains a large proportion of that variability, while a value near 0 suggests the opposite. For surveys with complex sampling designs, the weighted estimator of R^2 is given by:

$$\widehat{R}^2_\omega = 1 - \frac{\widehat{SSE}_\omega}{\widehat{SST}_\omega}$$

Where $\widehat{SSE}_\omega$ is the weighted sum of squared errors given by

$$\widehat{SSE}_\omega = \sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} \left(y_{hik} - x_{hik}\widehat{\beta}\right)^2$$

and $\widehat{SST}_\omega$ is the total weighted sum of squares given by

$$\widehat{SST}_\omega = \sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} \left(y_{hik} - \widehat{\overline{Y}}\right)^2$$

This estimator adjusts the R^2 calculation to reflect the characteristics of the sampling design, such as stratification and unequal selection probabilities, ensuring that survey weights are considered when evaluating the goodness-of-fit of the model.

5.5.2 Standardized Residuals

Residuals are the differences between observed and predicted values. Analyzing residuals is critical for diagnosing whether the model violates key assumptions. For example:

- Residuals should show no specific pattern when plotted against predicted values or independent variables;
- If the residuals exhibit a pattern, this could indicate non-constant variance (heteroscedasticity) or a non-linear relationship.

Graphical analysis is often used to detect issues, with plots of residuals against predicted values serving as a common diagnostic tool. A careful study of the residuals should help the researcher conclude whether the fitting process has not violated the assumptions

or if, on the contrary, one or more assumptions are not met, requiring a review of the model specification or of the fitting procedure. To analyze the residuals, Pearson residuals (Heeringa et al., 2017a) are defined as follows:

$$r_{p_k} = (y_k - \hat{\mu}_k) \sqrt{\frac{w_k}{V(\hat{\mu}_k)}}$$

Where $\hat{\mu}_k$ is the estimated expected value of y_k under the fitted model, and w_k is the survey weight for unit k in the complex sample dataset. Finally, $V(\hat{\mu}_k)$ is the variance function of the outcome. These residuals are used to perform normality and constant variance analyses.

If the assumption of constant variance is not met, the estimators remain unbiased and consistent, but they are no longer efficient. That is, they are no longer the best in the sense that they no longer have the smallest variance among all unbiased estimators. One way to analyze the assumption of constant variance in the errors is through graphical analysis. This is done by plotting the model residuals against $\hat{y}$ or the model residuals against x_k. If these plots reveal any pattern other than a cloud of points with constant spread, it can be concluded that the error variance is not constant.

5.5.3 Influential Observations

Another set of techniques used for model analysis involves examining influential observations. Certain data points can have a disproportionately large impact on the model. These influential points may not necessarily be outliers but could still affect model parameters significantly. An observation is deemed influential if, when removed from the data set, it causes a significant change in the model fit. It is important to note that an influential point may or may not be an outlier. To detect influential observations, it is essential to clarify what type of influence is being sought. For instance, an observation may be influential for parameter estimation but not for error variance estimation. Common techniques for identifying influential observations include:

1. **Cook's Distance**: Measures the impact of removing a data point on the overall model fit.

2. $D_f Beta_{(i)}$ **Statistic**: This statistic assesses the effect of removing a data point on individual regression coefficients, and measures the change in the estimated regression coefficient vector when the observation is removed.

3. $D_f Fits_{(i)}$ **Statistic**: it evaluates the influence of a data point on the overall model fit, and measures the change in the model fit when a particular observation is removed.

5.6 Inference on Model Parameters

After confirming that the model fits well and satisfies its assumptions, the next step is to assess whether the independent variables significantly contribute to explaining the dependent variable. This is done by testing the significance of the regression coefficients. If a coefficient is statistically significant, it suggests that the associated variable has a meaningful relationship with the dependent variable.

Given the distributional properties of the regression coefficient estimators, a natural test statistic for evaluating the significance of these parameters is based on the t-distribution and is described as follows:

$$t = \frac{\hat{\beta}_j - \beta_j}{se\left(\hat{\beta}_j\right)} \sim t_{df-p}$$

where the degrees of freedom (df) for a household survey (complex samples) is given by the number of PSUs m minus the number of strata H and p is the number of predictor variables in the fitted model.

This test statistic evaluates the hypotheses $H_0 : \beta_j = 0$ versus the alternative $H_1 : \beta_j \neq 0$. Similarly, a confidence interval of $(1-\alpha) \times 100\%$ for β_j can be constructed, as follows:

$$\hat{\beta}_j \pm t_{1-\frac{\alpha}{2}, df}\, se\left(\hat{\beta}_j\right)$$

5.6.1 Estimation and Prediction

According to Neter et al. (1996), linear regression models are essentially used for two purposes. One is to explain the variable of interest in terms of predictors that may be found in surveys, administrative records, censuses, etc. Additionally, they are also used to predict values of the variable under study, either within the range of values collected in the sample or outside of this range. The first purpose has been addressed throughout this chapter, and the second is achieved as follows:

$$\widehat{E}(y_k \mid x_{obs,k}) = x_{obs,k}\hat{\beta}$$

The variance of the predicted value is estimated as follows:

$$\widehat{Var}\left(\widehat{E}\left(y_k \mid x_{obs,k}\right)\right) = x'_{obs,k}\, \widehat{Cov}\left(\hat{\beta}\right)\, x_{obs,k}$$

Chapter 6

Tables

Tables are a fundamental tool for disseminating statistics from household survey data. They serve to organize and present numerical results efficiently, minimizing the need for lengthy text descriptions. When well-designed, tables enhance clarity and make it easier for users and wider audiences to interpret survey results. It is therefore important to discuss some core principles and ideas to the preparation and production of tables with survey results.

Before we enter detailed discussions, it is important to distinguish three main types of tables that can be used for presenting the results of a survey:

- **Presentation Tables**: These tables are designed to highlight key findings and are often included in reports or presentations; they are concise and focus on results that support specific messages or conclusions;
- **Reference Tables**: These tables provide comprehensive details and are aimed at users who need in-depth information; they are typically larger, covering a wide range of variables and subgroups or domains;
- **Long Tables**: These tables are structured for use in databases or data systems; they contain raw or minimally processed data, organized for further analysis or integration with other datasets.

Regardless of the type of table, certain principles should guide their design to ensure they are effective and user-friendly. According to Miller (2004), two fundamental principles should always be considered:

- **Principle 1.** Make it easy for your reader to find and understand the numbers presented in your tables, using clear and concise labels for rows and columns, highlighting key results, and avoiding excessive detail that might overwhelm the reader;

- **Principle 2.** Draw the layout and labels of the table in a simple and direct way, helping to focus attention on the results you want to show, using logical and intuitive ordering of rows and columns, grouping related variables or categories together and minimizing clutter by avoiding unnecessary lines, colors, or decorations.

6.1 Presentation tables

The primary goal of *presentation tables* is to communicate key results clearly and effectively. They are designed to support the accompanying text by organizing data in a way that emphasizes significant patterns, trends, or stories revealed by the survey. These tables should help readers quickly grasp the main findings without being overwhelmed by excessive detail. These are generally small tables, used to highlight certain key results obtained from the survey, to be presented in press releases, executive summaries, scientific articles or reports, or on landing web pages that showcase survey outputs. They are not expected to provide all results on a topic, but rather to highlight key results that should draw the attention of a reader to some of the main stories the data have produced.

In presentation tables, the data should be presented concisely, and organized to support the text with the analysis of the corresponding data. They should be designed in such a way to help readers learn about the key results on the topic provided by the survey. Short, well-designed and formatted tables can provide a lot of information that readers can absorb quickly. This applies to tables published in any vehicle: reports, press releases, articles, electronic publications or websites. The example below illustrates the idea.

Presentation tables should have rows (and possibly columns) sorted in a way that helps the reader perceive patterns, such as high or low figures. Such tables will often sacrifice detail in exchange for readability and understanding. Numbers should be presented with no more than 3 or 4 digits altoghether. If they area population counts, use thousands. If the figures are percentages, use no more than a single decimal digit, or even present only percentages rounded to the nearest integer, if the precision of the estimates do not warrant providing decimals (e.g. margins of error larger than 1%).

6.1.1 Example of presentation table and corresponding text - include in a box.

Below is an example of a presentation table designed to highlight key findings on gender representation among different managerial levels and age groups. The accompanying text contextualizes the data, emphasizing the underrepresentation of women in managerial roles across all ages.

Among middle and senior managers, women are outnumbered at all ages. The under-representation of women was observed in all age groups. Relative to their share among non-managers, women were outnumbered among middle and senior managers. In all age groups, women accounted for about 4 in 10 middle managers and 3 in 10 senior managers.

Table 2 - Share of women (%) by age group and occupation

Age group	Non-managers	Middle managers	Senior managers
25 to 34 years	44.6	40.3	28.4
35 to 44 years	45.7	38.7	31.3
45 to 54 years	48.3	40.5	31.7

Note: The category "women" includes women, as well as some non-binary people. Source: Statistics Canada, Census of Population, 2021. https://www150.statcan.gc.ca/n1/pub/36-28-0001/2024010/article/00005-eng.htm

This table concisely presents the percentage of women in different occupational categories (non-managers, middle managers, and senior managers) across three age groups. It highlights how a well-designed presentation table can summarize key findings and complement textual analysis. By organizing data clearly and emphasizing critical patterns, such tables enhance the readability and impact of survey results.

6.2 Reference tables

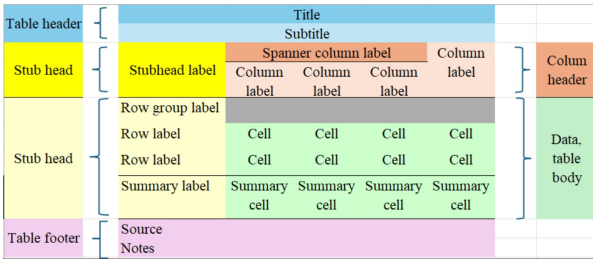
Reference tables are longer tables designed to present more comprehensive sets of results from statistical studies. These tables are typically included in reports to provide detailed information, but they should remain manageable in size. A good rule of thumb is to limit them to a maximum of 200 rows and 12 columns. Larger tables should be considered for dissemination as database-like tables, accessible via downloads or interactive websites.

Reference tables will typically take core classification, domain definition or *explanatory* variables to define the rows, and have the *outcome* classification or output variables define the columns. In both directions, sorting should typically be such that it is easier for the readers to locate the data that they are most interested in, either using alphabetic sorting or well known classifications. Such tables have in many cases been replaced by access to interactive databases that allow the interested user to obtain the tables they want from a website.

Tables (of all types) should be *self-sustaining*. The idea is that each table should have the necessary metadata, so that if copied from one location to another it still makes

sense. If you can get your tables to be *self-sustaining*, they will be easier to understand correctly, either in or out of the original context.

Anatomy of a table. Figure 9.1 presents the essential components of a table.



The following are the essential components of a well-designed table:

- The *title* (and optional subtitle) of a table is mandatory and must provide a clear and precise indication of the data that will be presented in the table. These elements, combined, must answer the questions about what, where and when regarding the data to be presented inside the table. Be concise and avoid using verbs.

- *Column header* elements should identify the data that is displayed in each column of the table. They must also provide much of the relevant metadata: unit of measurement, time period, geographical area, etc.

- *Row headers* and *stub* elements, provided as the first column in the table, should identify the data that is displayed in each row of the table.

- *Source* of the data must always be provided at the bottom of the table, and must indicate the organization responsible and the name of the survey or study that produced the results contained in the table. The omission of the citation of the source prevents the reader from seeking more information about the data presented, and should be avoided.

- *Notes* are optional, but they can be used to provide additional details about the data as needed to understand and use it correctly. Avoid using long texts, which if needed, would be better placed in a document that is then cited in the Notes section. If there is more than one Note, number sequentially, and use the numbers to indicate the corresponding calls inside the table. Make sure that the calls to *Notes* are sufficiently distinct from the actual figures / numbers inside te table to avoid confusion.

- *Data* is the most important piece of information that the user expects to get from the table. Therefore, it is essential to present them in a way that is easy to extract the relevant information. For some tables, depending on the message you want to convey, it may be easier to search for information by rows or columns. This

should be the most important consideration when deciding whether to present the table in portrait or landscape orientation. Dividing lines, dotted lines, shading, and even spacing can be helpful in guiding the reader to read the table in the 'right' direction.

When designing tables to present statistical data, ensuring clarity and consistency is crucial. Start by maintaining uniform spacing across columns to enhance readability, while avoiding unnecessary text or excessive width that can distract from the data. Time series data should always be organized in chronological order, preferably ascending for reference tables, to provide a clear and logical progression of information. Categorize data using standard classifications to facilitate understanding and comparison across different datasets.

The arrangement of rows and columns should follow a clear and logical order, with numerical data aligned to the right to ensure decimal points are neatly aligned. Decimal places should be limited to what is necessary for precision, and rounding should aim for 3-4 significant digits to simplify the data while preserving its integrity. Avoid blank cells, which can cause confusion; instead, use appropriate symbols to indicate missing or "not applicable" values, ensuring the table remains informative and complete.

Finally, these practices collectively improve the usability of the table, making it easier for readers to analyze and interpret the data. By adhering to these guidelines, you create a presentation that is both professional and accessible, promoting effective communication of statistical insights.

The recommendations provided here to reference tables should also apply to longer tables provided as databases, but these can have additional resources if they are embedded on websites. For example, there may be support for users to sort tables using the values in each column, which would be useful for large tables where the user may be looking for the higher (or lower) values in a given column.

6.3 Low-Quality Estimates

National Statistical Offices routinely produce descriptive statistics, such as totals, averages, proportions, and ratios, based on survey data. These statistics provide valuable insights into key characteristics of the population, such as income levels, employment rates, or access to education. To ensure this information reaches a wide audience, NSOs often use a variety of dissemination channels, including:

- **Public Reports**: Comprehensive reports summarizing key findings from household surveys;
- **Online Platforms**: Interactive data visualization tools and downloadable datasets on official websites;

- **Press Releases**: Brief summaries of major findings designed to capture public and media attention.

These dissemination efforts aim to make the data understandable and actionable for policymakers, researchers, and the general public. When publishing tables of results, NSOs strive for clarity and usability. Tables are typically organized to highlight trends, comparisons, and distributions of key variables. Common features of published tables include:

- **Aggregated Data**: Grouping data by domain defining variables like age, sex, region, or socioeconomic status;
- **Confidence Intervals**: Including measures of uncertainty to provide context for the estimates;
- **Metadata**: Offering detailed explanations of the data collection methods, definitions, and limitations.

By presenting data in a user-friendly format, NSOs ensure their publications are accessible to a diverse audience.

Not all estimates derived from survey data meet the necessary quality standards for publication. Estimates may be suppressed if they are based on small sample sizes, have high variance, or are otherwise unreliable. NSOs use established criteria to determine when suppression is necessary, ensuring that the released data maintains its credibility. To address this issue we can use the following approaches:

1. **Quality Thresholds**: NSOs set predefined thresholds for measures like the coefficient of variation (CV) or standard errors;
2. **Flagging and Suppression**: Estimates that fall below these thresholds are either flagged with warnings about their reliability or omitted entirely from published tables;
3. **Transparency**: NSOs provide clear documentation explaining why certain estimates are suppressed, maintaining transparency and trust.

Chapter 7

Data visualization

In this section we discuss how to present data and estimates resulting from household surveys using graphics. Effective graphs can reveal patterns, trends, and relationships in the data, making it easier to interpret findings and communicate them to diverse audiences. While standard plots can still be used to show distributions and associations from the raw (unweighted) sample data, these can be misleading for the corresponding population distributions and associations. Therefore it is recommended that modified plots that account for survey weights be used instead.

For example, a bar chart showing income distribution should incorporate weights to properly represent the income distribution for the entire population. Similarly, scatter plots exploring associations between variables should use weighted markers or density adjustments to ensure the relationships are accurately depicted. In addition, regarding the display of survey estimates, which are subject to sampling error, it is important to convey this message by presenting not only point estimates, but also standard errors or confidence intervals.

When presenting survey estimates, it is essential to recognize that these estimates are subject to sampling error. To effectively communicate this uncertainty, graphs should include measures such as standard errors or confidence intervals. For instance:

- Confidence intervals can be added to bar charts or line graphs to show the range of plausible values for an estimate;
- Error bars in scatter plots can illustrate the variability associated with specific data points.

Incorporating these elements into visualizations helps ensure that viewers understand the inherent uncertainty in the survey estimates, fostering more informed interpretations. When the survey units have different sampling weights, these should be taken into account when preparing graphs with their data. The main reason is that weights

can be interpreted as the number of population units that each sample unit represents. Hence, it is evident that unequal weights need to be considered in the elaboration of graphs based on such sample data.

When graphs are created *without* considering weights, the visual representation reflects the sample characteristics rather than the population. This discrepancy can distort distributions, proportions, or relationships between variables. Incorporating weights ensures that the graphs provide a more accurate representation of the population.

7.1 Bar charts

When the data of interest are categorical, their descriptive analysis will be done using contingency tables. Bar charts are commonly used to visualize categorical data. For survey data, descriptive analysis of categorical variables typically begins with contingency tables that summarize weighted counts or proportions. These tables can then be used to create bar charts, ensuring the results reflect population-level characteristics rather than just sample data. Ideally one should also aim to display error lines overlaying bars to indicate their respective confidence interval widths, thus conveying the uncertainty of the corresponding point estimates. Obtaining the weighted counts or proportions and their confidence intervals can be easily done using tools from several software packages, e.g. the `survey` package in R.

As an example, the bar chart in Figure 9.2 presents a comparison of the number of individuals (`Nd`) between rural and urban zones, with error lines indicating the confidence intervals for each estimate. According to the values in the table, the urban zone shows a slightly higher `Nd` value than the rural zone, with 78,164 individuals in the urban area compared to 72,102 in the rural area. This difference suggests a higher concentration of people in the urban zone.

The confidence intervals allow us to assess the precision of these estimates. In the rural zone, the confidence interval ranges from 66,039 to 78,165 individuals, while in the urban zone, the confidence range goes from 72,526 to 83,802 individuals. This overlap between the intervals indicates that, although the urban zone has a higher number of individuals, the difference is not pronounced enough to be statistically significant.

Furthermore, the standard deviation of `Nd` is 3,062 for the rural zone and 2,847 for the urban zone, reflecting similar variability in both zones. This suggests that the estimates are consistent in terms of relative uncertainty, without major differences in data dispersion between the zones.

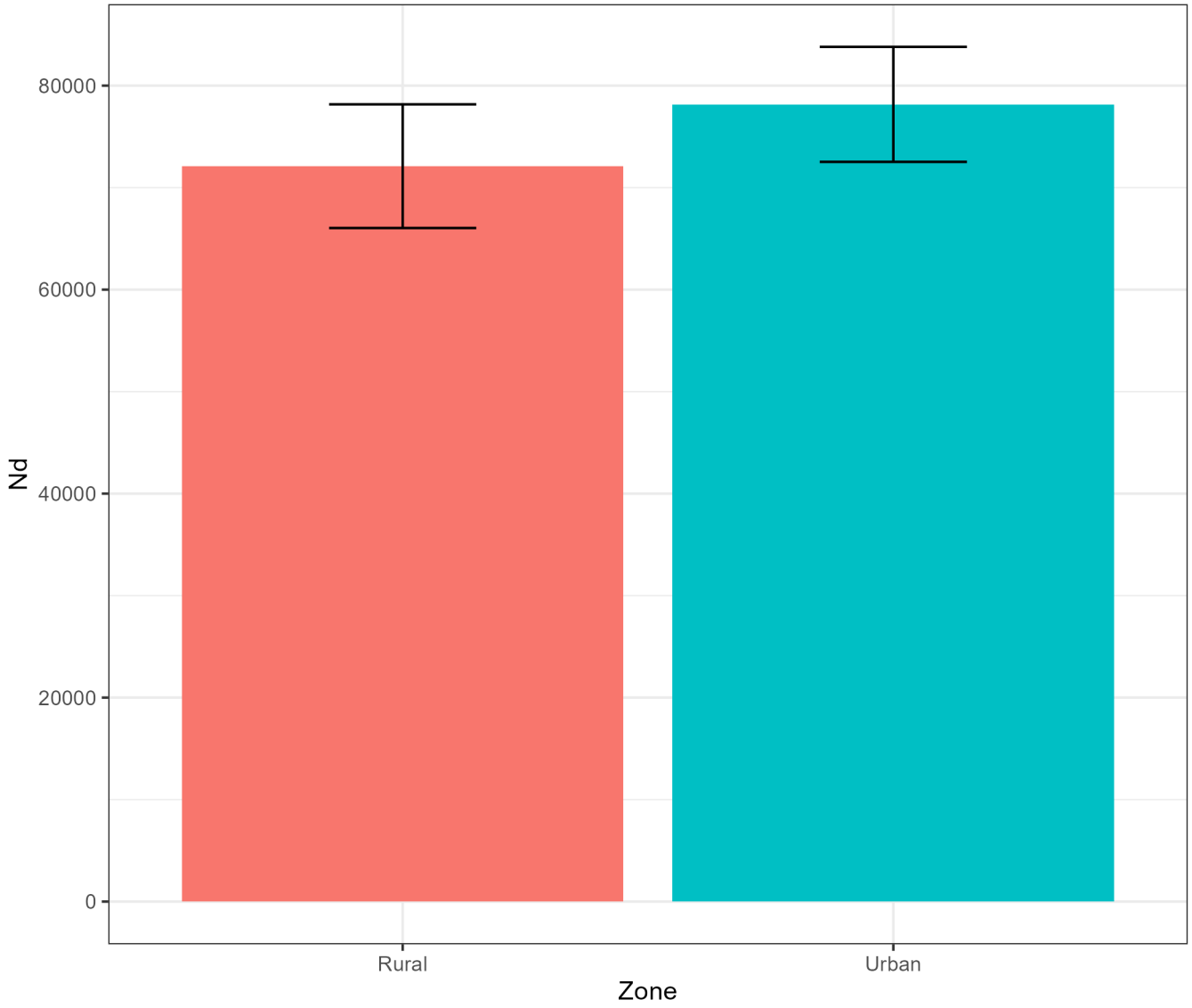


Figure 7.1: Distribution of Population by Area

Table 7.1: Population distribution by area

Zone	Number of Individuals (Nd)	Standard Error (Nd_se)	Lower Limit (Nd_low)	Upper Limit (Nd_upp)
Rural	72,102	3,062	66,039	78,165
Urban	78,164	2,847	72,526	83,802

7.2 Histograms

Histograms serve to present the distribution of a single numeric (continuous) survey variable or response. If one had a census, then the histogram is a powerful tool to describe the underlying distribution, even for very large datasets. When displaying sample data, however, the sampling weights must be taken into account when estimating frequencies or relative frequencies of population units having values in the specified histogram bins. Modern survey analysis tools can easily provide weighted histograms where the sampling weights are incorporated.

Histograms are often seen as precursors to density function estimates. A density estimate can be thought of as a histogram with a large number of bins, providing a smoother view of the data distribution. The `survey` package in R includes functionality for plotting smoothed density estimates that account for sampling weights, offering a more detailed representation of the population.

A common example of visualization in this type of analysis is the use of histograms to represent the distribution of variables such as income. These charts allow us to observe the distribution of the variable of interest in the expanded population and to understand its shape, dispersion, and general trends.

It is also common to perform graphical analyses broken down by subgroups, such as geographic areas (urban and rural) or thematic characteristics like sex (male and female). This approach helps identify key differences among specific population subgroups, for instance, by examining income distribution in men and women over the age of 18. Such breakdowns help visualize and communicate potential gaps between subgroups of interest.

In this way, charts help to communicate results in an accessible manner, offering a clear and straightforward visual representation for audiences who may not be familiar with the technical details of estimation methods.

In 7.2 the horizontal axis (x) represents income levels, spanning from 0 to over 4000 monetary units, while the vertical axis (y) indicates frequency, meaning the number of individuals within each income range.

The distribution shows that most of the population is concentrated at lower income levels, with a particularly high frequency near 0. As income levels rise, frequency declines

sharply, indicating a right-skewed (positively skewed) distribution with a smaller proportion of people at higher income levels. The light gray bars visually emphasize this concentration at lower incomes, highlighting a significant disparity in the population's income distribution.

As an example, Figure 7.3 presents two histograms illustrating the distribution of income and expenditure by sex. In the histogram on the left, titled "Income Histograms by Sex," we observe the income distribution, where blue bars represent men and pink bars represent women. The majority of the population, both male and female, is concentrated in the lower income levels, showing a right-skewed distribution. In the lower income levels, there are more men than women, while at higher income levels, the difference is less pronounced.

In the histogram on the right, titled "Expenditure Histograms by Sex," the distribution of expenditure is shown, also broken down by sex. Similar to income, most of the population of both sexes is concentrated in the lower expenditure levels, with a right-skewed trend. There is also a higher proportion of men in the lower expenditure levels, while at higher levels, the representation between sexes is more balanced. These histograms exemplify the similarity in the income and expenditure distributions between men and women, although men appear to be slightly more represented in the lower levels of both variables.

Histograms, especially when weighted for survey design, are invaluable for exploring and presenting the distribution of continuous variables. Subgroup analyses further enhance their utility, enabling the identification of disparities and trends across different population segments. Combined with smoothed density estimates, histograms provide a comprehensive and accurate view of the population's numeric variables.

7.3 Scatter Plots

Scatter plots are the tool of choice to explore relationships between two continuous variables, potentially revealing patterns or trends in the data. These plots face the two challenges discussed above. First one needs to try and convey in the plot that the different sample observations carry different weights. For small to moderate sample sizes this can be done by plotting circles or dots of varying sizes where the symbol size represents the corresponding observation sampling weight. Plots like these can be obtained using standard bubble plot tools or the scatter plot available in the `survey` package in R. As stated by Lumley (2010), when dealing with large datasets, displaying all the data points in a scatter plot can be overwhelming and cluttered. Several strategies can help address this issue:

1. **Subsampling**: Select a smaller, manageable subsample from the full dataset. The subsample should be selected with probabilities proportional to the sampling

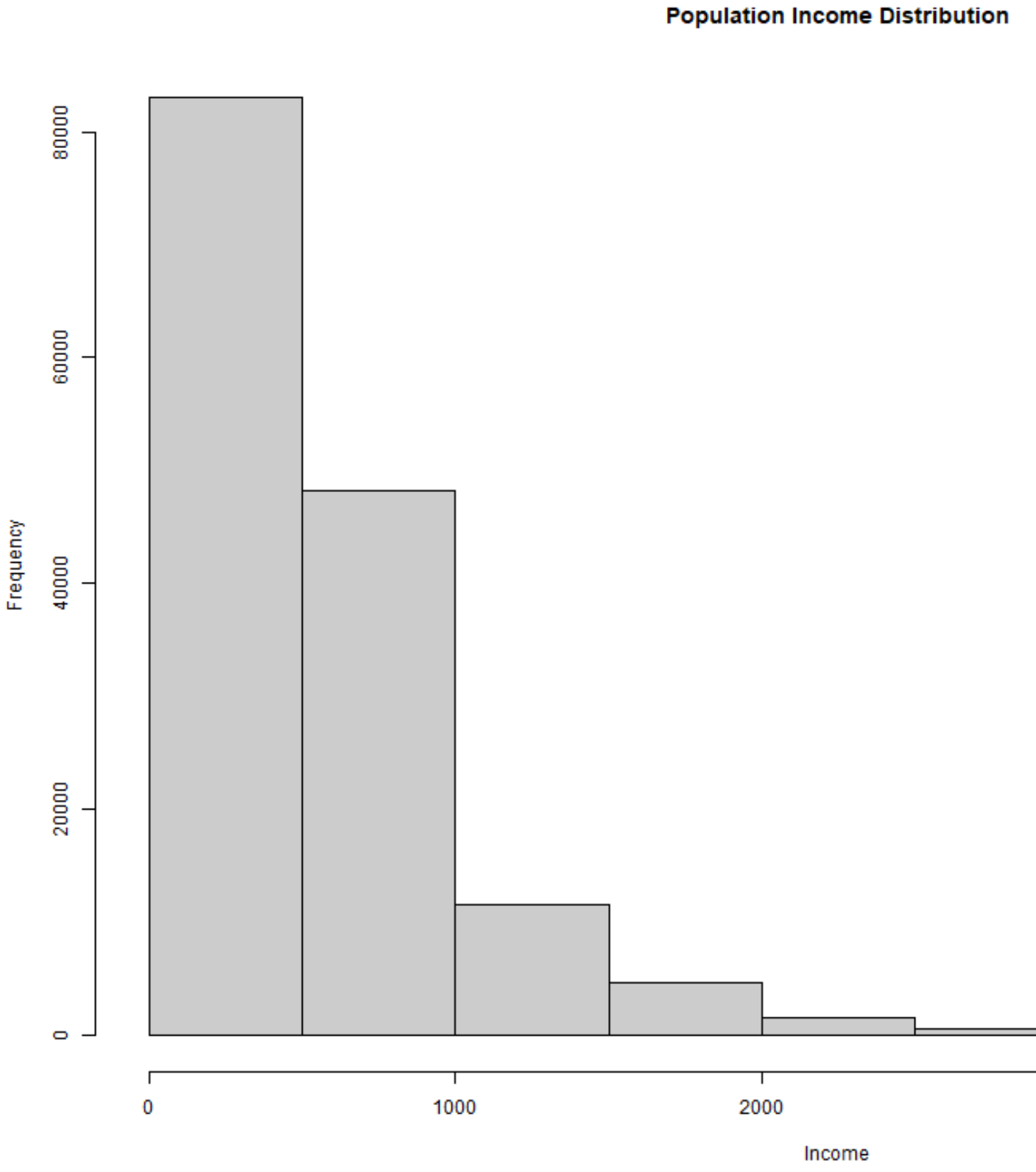


Figure 7.2: Distribution of Population Income

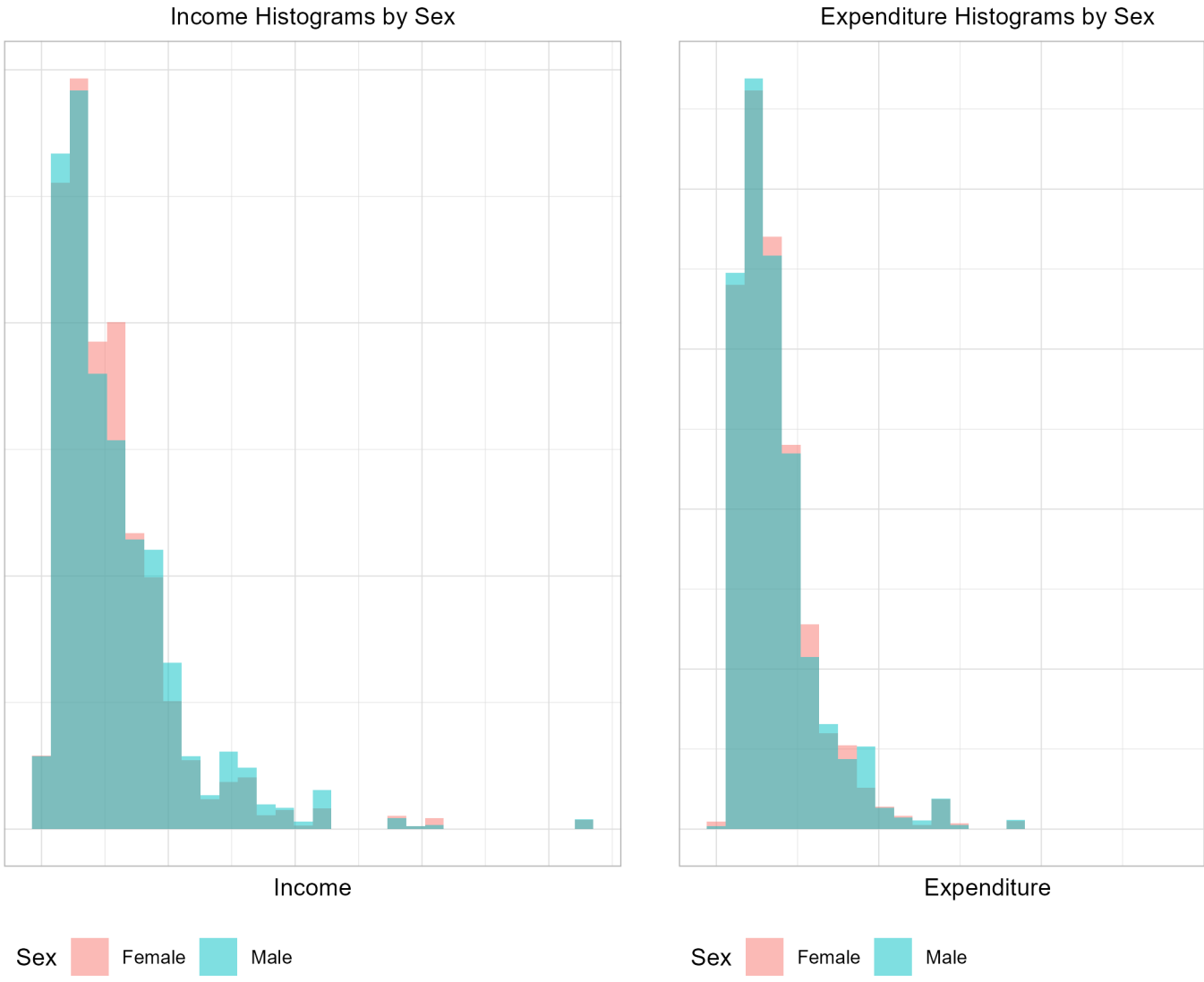


Figure 7.3: Histograms of Income and Expenditure by Sex

weights, ensuring that it behaves approximately like a simple random sample from the population. The resulting scatter plot maintains representativeness while being easier to interpret. The subsample obtained in this way behaves approximately as a simple random sample from the survey population.

2. **Hexagonal Binned Scatter Plots**: Divide the plot area into a grid of hexagons. Instead of plotting individual points, represent each hexagon with shading or size based on the total sampling weights of the points within that hexagon. This approach condenses the data into a clear and interpretable visualization. With complex household survey data, the number of points in a hexagonal bin should be replaced by the sum of the weights for points in the bin.

3. **Smoothed Scatter Plots**: Avoid plotting individual points altogether and instead estimate and display trends. For example, calculate specific quantiles (e.g., quartiles) of the y-axis variable conditioned on the x-axis variable and smooth these values across the range of the x-axis. This approach highlights trends while minimizing visual clutter.

The figure 7.4, illustrates the weighted relationship between income and expenditure in a population. In this plot, the size of the points represents the weight assigned to each observation. A high concentration of points is observed at lower income and expenditure levels, suggesting that most of the population has low income and low expenditure.

Although there is an upward trend, indicating that income and expenditure tend to increase together, the dispersion of points reveals that higher expenditure is not always associated with proportionally higher income. Some larger points, corresponding to observations with greater weight, are distributed across different levels of income and expenditure without concentrating in a single area. Additionally, a few isolated points at high expenditure levels may represent outliers with considerably higher-than-average expenditure. Overall, this plot suggests a positive relationship between income and expenditure, accompanied by significant variability and some exceptional cases.

Scatter plots are a versatile and effective way to explore relationships between variables in survey data. By incorporating sampling weights and adopting strategies to manage large datasets, they can provide clear, meaningful insights into population-level patterns. Whether using weighted points, hexagonal binning, or smoothing techniques, scatter plots remain a cornerstone of data visualization for continuous variables.

7.4 NSO – Practical example

In this subsection we will include the experience of a NSO on displaying information through graphics.

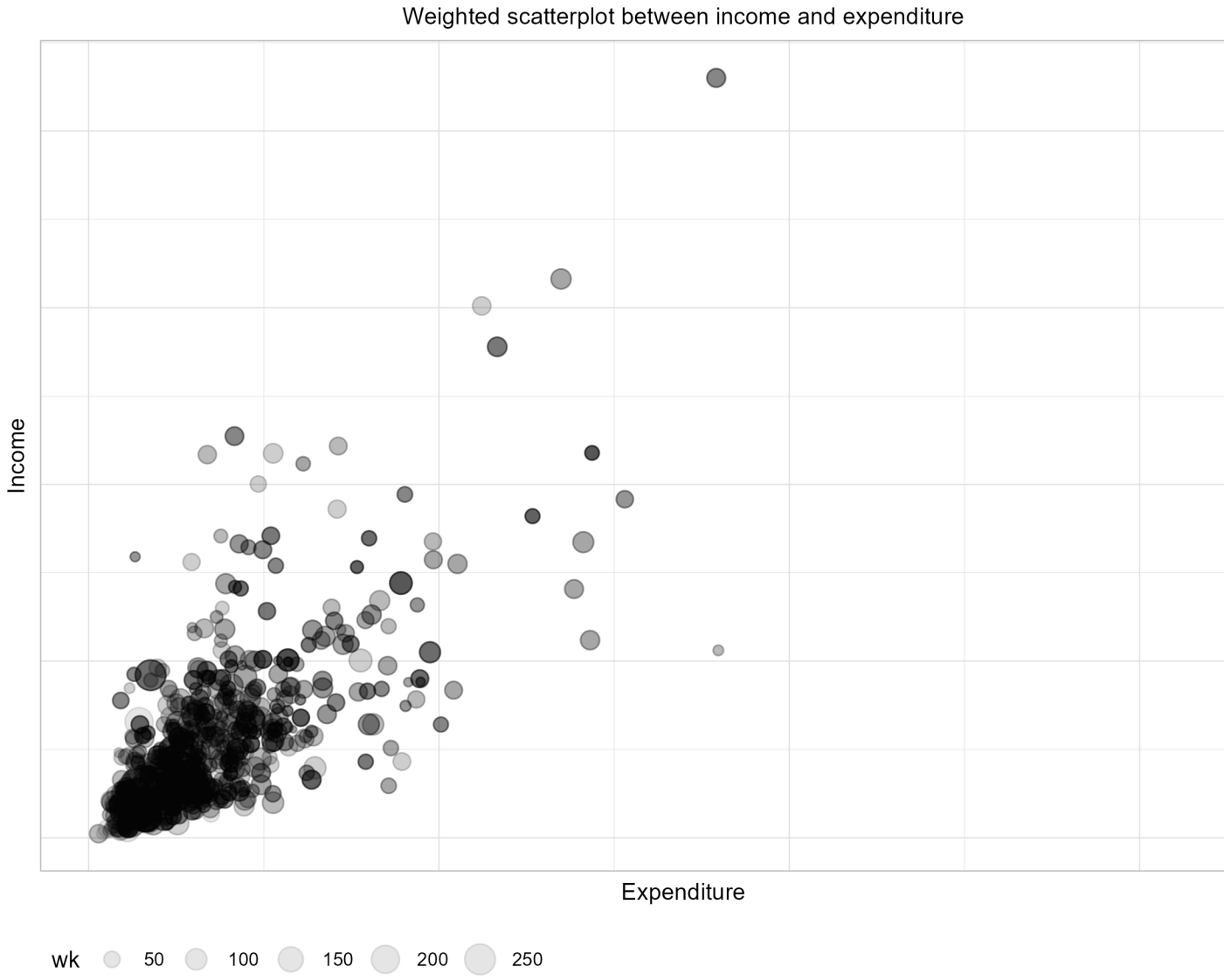


Figure 7.4: Weighted scatterplot between income and expenditure

References

Bibliography

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

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

Binder, D. A. and Kovacevic, M. S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equations approach. *Survey Methodology*, 21(2):137–145.

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.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Fay, R. E. (1979). On adjusting the pearson chi-square statistic for clustered sampling. *ASA Proceedings of the Social Statistics Section*, pages 402–408.

Fellegi, I. P. (1980). Approximate joint estimation of the parameters of multinomial distributions in the analysis of data from complex surveys. *Journal of the American Statistical Association*, 75(370):261–268.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankyha, Series C*, 37:117–132.

Fuller, W. A. (2002). Regression estimation for survey samples (with discussion). *Survey Methodology*, 28(1):5–23.

Gutiérrez, H. A. (2015). *TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population*. R package version 3.2.2.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory*, volume 1 and 2. John Wiley and Sons, New York.

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

Heeringa, S. G., West, B. T., and Berglund, P. A. (2017b). *Applied Survey Data Analysis, second edition*. Chapman and Hall - CRC, 2nd edition edition.

IBM (2017). *IBM SPSS Complex Samples*.

Jacob, G., Damico, A., and Pessoa, D. (2024). *Poverty and Inequality with Complex Survey Data*.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, 36:1–37.

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.

Langel, M. and Tillé, Y. (2013). Variance estimation of the gini index: revisiting a result several times published: Variance estimation of the gini index. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):521–540.

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

Lumley, T. (2016). survey: analysis of complex survey samples. R package version 3.32.

Miller, J. E. (2004). *The Chicago Guide to Writing About Numbers*. University of Chicago Press, Chicago.

Nations, U. (2005). *Household Surveys in Developing and Transition Countries*. United Nations, New York, NY.

Neter, J., Wasserman, W., and Kutner, M. H. (1996). *Applied Linear Statistical Models*. McGraw-Hill.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality. *Journal of the European Survey Research Association*, 3(3):167–195.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? why is it a problem? how can we approach it? *Survey Methodology*, 37(2):115–136.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. N., Wu, C. F. J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18:209–217.

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12:46–60.

Rojas, H. A. G. (2020). *samplesize4surveys: Sample Size Calculations for Complex Surveys*. R package version 4.1.1.

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

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

SAS (2010). *SAS/STAT 9.22 User's Guide - Survey Sampling and Analysis Procedures*.

Shah, B. V., Folsom, R. E., LaVange, L., Wheeless, S. C., Boyle, K. E., and Williams, R. L. (1993). Statistical methods and mathematical algorithms used in sudaan.

Shah, B. V., Holt, M. M., and Folsom, R. F. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 41(3):43–57.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. John Wiley and Sons, New York.

STATA (2017). *STATA Survey Data*.

Thomas, D. R. and Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82:630–636.

Tillé, Y. and Matei, A. (2016). *sampling: Survey Sampling*. R package version 2.8.

Westat (2007). *WesVar 4.3. Users guide.*