## CHAPTER 9 ANALYSIS

Andrés Gutiérrez (Economic Commission for Latin America and the Caribbean)

Pedro Luis do Nascimento Silva (Sociedade para o Desenvolvimento da Pesquisa Científica)

**CHAPTER 9  
ANALYSIS**

### 9.1. Introduction

1. 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 - *Accountability and Transparency*: 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 - *Prevention of Misuse*: Statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

2. 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 mostly *descriptive purposes*, such as estimating population means or obtaining estimates of population frequency distributions. Other times, however, the data is used for *analytical or inferential purposes*, involving the testing of hypotheses 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.

3. Household surveys play a critical role not only in tracking national progress but also 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 monitor progress and achieve sustainable development targets.

4. What makes survey data special or challenging for analytical purposes is that they are collected through complex sampling methods that often involve:

* Clustering: Households are grouped into geographic or administrative units (e.g., blocks, enumeration areas or census tracts). Instead of selecting households randomly across the entire population, clusters are first selected, and then households within those clusters are sampled.
* Stratification: Dividing the population into non-overlapping subgroups—called strata—based on characteristics such as region, urban/rural status, or socioeconomic level. Samples are then drawn independently within each stratum to ensure representation and improve the precision of survey estimates.
* Unequal probabilities of selection: Different units (e.g., clusters, households or individuals) may have varying chances of being selected, often to ensure adequate representation of small or important subpopulations.
* Weighting adjustments: Correcting initial sampling weights to account for nonresponse, undercoverage, or to calibrate the sample with known population totals. These adjustments help reduce bias and improve the precision and accuracy of survey estimates.

Sampling design aspects are described in Chapter 5, while sample weighting is described in Chapter 8.

5. Traditional data analysis methods and software ignore these features, leading to biased estimates of both the target parameters and their associated standard errors. This chapter presents the effect of such simplifications and explains the necessary adjustments that have to be made in order to appropriately incorporate the survey design into the analysis. Also, this chapter relies on the fundamental concepts of the design-based paradigm for survey design and analysis and aims to empower users to analyse household survey data accurately and effectively by presenting relevant models, methods, and software that enable the data analyst to understand key steps in the data analysis process and to incorporate complex designs into their analyses.

6. Different readers of the chapter may find some of its parts more useful than others. Here, the contents of each section are described, so that readers may direct their attention to the topics of relevance to them. After a short introduction, Section 9.2 provides guidance on preparing a household survey data analysis plan to help users apply appropriate methods and procedures effectively.

7. Section 9.3 provides a brief discussion on the fundamental principles of design-based inference, emphasizing that conclusions drawn from probability household surveys should be based on a pair: the point estimate and its standard error (or another relevant quality measure). Section 9.4 begins the journey with the key tools for descriptive analysis for means, proportions, ratios, and other typical descriptive parameters.

8. Section 9.5 is devoted to more complex parameters that allow comparisons of the phenomenon of interest between subgroups for continuous and discrete variables. It also presents standard tests to compare estimates and measure the degree of correlation and association between variables.

9. Section 9.6 focuses on modelling survey outcomes. It starts with a discussion on the role of weighting when estimating regression coefficients, followed by presentation of proper approaches to estimate complex parameters in linear regression models. Section 9.7 discusses preparation and presentation of tables, while Section 9.8 presents a summary of tools for survey data visualization. Finally, Section 9.9 presents some notes on challenges ahead for household survey data analysis arising from situations such as non-probability sampling, small area estimation, linked data. References are provided at the end.

10. Throughout the chapter, practical examples are provided to illustrate how to conduct various types of data analysis using a typical household survey dataset. By the end of the chapter, readers will be equipped with the knowledge and tools needed to analyse household survey data effectively by accounting for the complexities of typical household survey designs.

### 9.2. Planning and preparation for analysis

11. As stated in Chapters 2 and 3, planning the analysis stage of a survey is an essential part of the overall survey planning process. Following the Generic Statistical Business Process Model (United Nations Economic Commission for Europe, 2014), this step corresponds to the subprocess labelled as 2.1 - Design Outputs.

12. 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. For primary data producers, creating a comprehensive tabulation plan ensures alignment with survey objectives. For secondary users, clear research questions and attention to survey metadata enable accurate and meaningful analyses.

#### 9.2.1. Primary data producers/users

13. They are responsible for planning and executing the survey to collect the intended data. They are also the first group of data users. For them, planning the analysis typically involves preparing a *tabulation 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). It also helps those designing the survey questionnaire(s) to ensure that all items needed are included, and possibly, help avoid adding items / questions that are not really needed for the planned analysis (see Chapter 4).

14. Preparing a *tabulation 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 administrative divisions in national household surveys. A typical list of domain defining variables would include geographic levels (provinces, etc.); type of region (urban, rural); sex; age groups; education; race / ethnicity; etc.
3. Response or Survey Variables: These are the variables being analysed to understand how they vary across domains of interest. For instance, continuous survey variables (like income) might be used to produce summaries like means, medians, quantiles or other statistics. Categorical survey variables (like labour force status) will generate a column for each category where the corresponding cross-classified frequencies or proportions will appear, with one row for each of the classes of the domain defining variables.

15. As discussed throughout Chapter 5, an important consideration when defining domains is related to sample design, particularly when defining strata and sample sizes. For example, most national household surveys require estimates at the state or province level, making stratification by state or province essential. Moreover, if precision is required at the provincial level, sample sizes should be computed for each province to meet those requirements and then summed to obtain the total national sample size.

16. When domains are defined by characteristics unavailable in the sampling frame (say age groups in area-based household surveys), sample size calculations must ensure that the sample is large enough for estimates of the rarest group to meet precision requirements. 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 older adults (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, 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.

#### 9.2.2. Secondary data users

17. Secondary data users are all those who were not involved in data production and will access and analyse the survey microdata after its release, typically using only public datasets and documentation provided by the data producers or curators. Their first task is to define clear research questions and locate the relevant survey data and documentation (metadata). High-quality survey metadata must describe the sampling design and estimation methods used (including details on stratification, clustering, and survey weights), for both descriptive parameters and their corresponding precision measures. Similar approaches may be applied to data accessed through secure services offered by some NSOs or other producing agencies.

18. For example, a well-defined research question posed by a secondary data user might be: “Do rural households face digital exclusion compared to urban households?”. This type of question directs the analysis towards estimating relevant parameters and their associated precision measures. The question might involve assessing whether the proportion of rural households with internet access is significantly lower than that of urban households. Clearly stating this question ensures a more direct and accurate analysis. In this case, the user starts with a null hypothesis, which assumes no difference between the two types of areas, and an alternative hypothesis, which suggests that the proportion of households with internet access is lower in rural areas.

19. A related but different hypothesis could involve testing whether the average cost of broadband internet access differs between rural and urban areas. In this case, the null hypothesis would state that the average costs are equal in both areas, while the alternative hypothesis would suggest that the average cost is higher in rural areas. Listing the research questions in this way helps the analyst focus on estimating the relevant parameters from the survey data, along corresponding precision measures, both of which are required to compute test statistics that would provide the evidence required to answer them.

20. For such estimation to be carried out, the secondary data user must first find out how the sampling was conducted and incorporate the sampling design variables (provided in the survey database) during the analysis stage. As we will discuss throughout this chapter, it is necessary to account for the survey weights (see Chapter 8) when computing point estimates of both descriptive and model parameters, and to account for the structural components of the sampling design and estimation process (stratification, clustering, unequal inclusion probabilities, nonresponse adjustment, and calibration of survey weights, if any) when estimating standard errors or other measures of precision for the point estimates.

21. It is strongly recommended that data producers include detailed information about the sampling design in the metadata released with the survey microdata, enabling secondary data users to account for these aspects in their analyses. Users that disregard including the sampling design in their analyses do so at their own risk, potentially producing biased estimates that may lead to incorrect inferences and misguided decisions.

#### 9.2.3. Quality control for secondary users

22. A standard quality control process for secondary data users would consist of the following steps:

1. Load the data and metadata: Ensure that the survey microdata is properly linked to metadata describing the key aspects of the sampling design (strata and cluster identifiers) and estimation processes (final survey weights, replicate weights if applicable, non-response adjustments, calibration, etc.). If the design metadata is not entirely available, refer to paragraphs 54-60 for guidance.
2. Replicate published estimates: Reproduce selected estimates from the primary analysis, including measures of precision, to verify that the data and sampling design variables have been correctly interpreted and properly loaded.
3. Compare scenarios: When the users can replicate published estimates, they can proceed with the new analysis for which no previous results are available. This analysis should be conducted under two scenarios: one accounting for the sampling design, and one that ignores it. Comparing the results help assess the impact of not incorporating the sample design.
4. Finalize the analysis: Base the final interpretation solely on the results that account for the sampling design, ensuring that point estimates can be appropriately extended to confidence intervals and that conclusions are drawn accordingly.

### 9.3. Accounting for the sampling design

23. To draw valid conclusions about the population from survey data, it is essential to adopt a design-based inference approach. This framework treats the sampling design as the basis for inference, assuming that the sample has been selected through a well-defined probability sampling design in which every unit in the population has a calculable, non-zero chance of being sampled. Design-based inference ensures that estimates are unbiased (or nearly unbiased) under the sampling design, regardless of the underlying distribution of the study variable.

24. Sampling weights, indicating how many population units each sampled unit represents, are essential to this approach. They allow users to produce estimates that account for the sampling design and yield results that are representative of the population. A well-documented sampling 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.

25. When analysing household survey data, ignoring the sampling design undermines the representativeness, accuracy, and credibility of survey-based findings, which may lead to incorrect decisions. Korn and Graubard (1995) provide illustrative examples demonstrating how weighted and unweighted survey estimates can yield substantially different results. Accounting for the sampling design is essential when analysing household survey data to ensure valid and unbiased estimates. As seen in the previous chapters, regular household surveys have two major characteristics:

* They use *complex sampling designs* (e.g., stratification, clustering, and unequal probabilities of selection) to enhance the efficiency and accuracy of data collection.
* They define *sampling weights* for each sampling unit (primary, and remaining ones) to properly represent the population.

26. To illustrate the problem faced by ignoring the sampling design, we provide a simple example. Suppose a country has two regions: Region A with 100 people, and Region B with 900 people. Wealthy people live in Region A, with an average income of $10,000, while less wealthy people live in Region B, with average income of $2,000. The true population average income is $2,800, because:

|  |  |  |
| --- | --- | --- |
|  |  | (9-1) |

27. Suppose a survey is conducted in which 50 people are sampled at random 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 corresponding population sizes. This way, the average income is estimated with severe bias as:

|  |  |  |
| --- | --- | --- |
|  |  | (9-2) |

This result arises even though both sample means within regions are perfect estimates of the corresponding population averages.

28. When the sampling design is properly incorporated, weights are applied proportional to the population sizes of each region. This way, units in Region A would receive a weight of , while units in Region B would receive a weight of . Under this scenario, the average income is unbiasedly estimated as:

|  |  |  |
| --- | --- | --- |
|  |  | (9-3) |

Ignoring the sampling design (and corresponding weights) causes the Region A (smaller but wealthier) estimate to dominate the overall estimate, even though its population represents only 10% of the country’s population. This creates a bias, making the average population income seem higher than it actually is. By considering the sampling design (the sample was stratified by region, with equal allocation), and applying weights, the corrected estimate reflects the true contribution of each Region to the global average, thereby eliminating the bias.

29. Another crucial aspect to consider is clustering. Most household surveys select cluster samples in the first sampling stage, where clusters are defined by census tracts, enumeration areas or similar groups of households. Within each sampled cluster, a sub-sample of households is then selected for the survey. While such designs are cost efficient for in-person surveys, to mitigate the loss of precision it is customary to increase the final sample size (see Chapter 5).

30. Consider the following example where clusters are city blocks, and a sample of 10 households is selected within each cluster. Suppose the target parameter is the proportion of households which have electricity provided by the city grid. Assuming that each cluster either is included or excluded from the grid, a sample of 1,000 households obtained from such a design would be as precise as a sample of only 100 households, if they were selected by simple random sampling. The reason is that once a single household in each cluster is observed, the remaining nine households in that cluster’s sample do not contribute additional information. A naïve analysis of the full sample of 1,000 households could mislead the user into believing that the estimates are as precise as those from a simple random sample of 1,000 households from the entire population, which was not the case; consequently, standard errors would be severely underestimated.

31. To enable adequate analysis that accounts for the sampling design, it is strongly recommended that survey datasets must contain identifiers for strata and primary sampling units (PSU), as well as weights for relevant units of analysis (e.g., households or individuals). Alternatively, if such information is not available, the dataset should at least contain replicate weights, and the user should have clear guidance on how to compute both point estimates and corresponding standard errors.

#### BOX – ESS4 sampling design

The 2018/2019 Ethiopian Socioeconomic Survey (ESS4) is part of the Living Standards Measurement Study (LSMS), an initiative led by the World Bank in collaboration with the Central Statistical Agency of Ethiopia (CSA). Its objective is to provide high-quality data for the analysis of poverty, food security, agricultural development, and other key socioeconomic aspects. The survey provides information on household structure, employment, income, consumption, education, health, access to infrastructure, and characteristics of the agricultural sector. One of its key features is the combination of household survey data with administrative records and remote sensing technologies, such as satellite imagery, to improve the accuracy of economic and social indicator estimates.

The sampling design of ESS4 is based on a two-stage stratified sampling, with different strategies for rural and urban areas. In the first stage, enumeration areas (EA) in rural zones were selected using a simple random sampling (SRS), while in urban areas, EAs were selected using systematic sampling with probability proportional to size (PPS) within each region. This approach ensured a proportional allocation of the sample between urban and rural areas.

In the second stage, households within each selected EA were chosen following a systematic simple random sampling. In rural areas, 10 agricultural households were selected, along with 2 non-agricultural households. In urban areas, 15 households per EA were selected, regardless of economic activity. The final number of completed household questionnaires was 6,770 across 535 EAs. This design ensures representative coverage of socioeconomic conditions in both rural and urban areas.

For more details on the methodology, sampling design, and data access, you can refer to the official source on the [World Bank Microdata Library](https://microdata.worldbank.org/index.php/catalog/3823/study-description#study_desc1674579234511).

#### 9.3.1. Parameters and estimators

32. When analyzing survey data, the first step is to define the *parameter* of interest which is a fixed numerical value that describes a specific characteristic of the entire *population* (*)*. Common examples include the *population total,* and the *population mean*. These values would only be known if data were observed for every unit in the population. As this is not possible, household sample surveys are carried out in order to make inferences about these population parameters from the *sample (*).

33. The design-based approach for inference and estimation relies on the known probability structure defined by the sampling design. All statistical properties of the estimators (such as unbiasedness, precision or consistency) are evaluated with respect to the randomization distribution induced by the design. In probability sampling, every unit in the population has a calculable chance of being included in the sample. As stated in Chapter 8, these *inclusion probabilities* are used to compute the *basic* *sampling weights*.

34. Estimators used to make inferences of population parameters apply the sampling weights to create weighted sums of the survey data, producing estimates for the population parameters. When weights are correctly incorporated, the resulting estimates are consistent with the true population values, ensuring that these estimates are unbiased, meaning that, on average, the estimates will equal the parameter if the survey were repeated multiple times under the same conditions.

35. As stated in Chapter 8, in most surveys, the basic sampling weights need adjustments to improve the accuracy of survey estimates. Adjustments may account for survey nonresponse, 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 align more closely with known population distributions, such as age or sex distributions, typically derived from recent population censuses or from population projections. Note also that comparing survey estimates of these distributions obtained before weight-adjustment is useful in assessing differential coverage or nonresponse.

36. The *population total* and *mean* of a survey variable can be respectively estimated by weighted estimators given by and , where . The basic design weights are given by the reciprocals of the inclusion probabilities of the sampled units (denoted ), for all .

37. When survey weights are calibrated and/or adjusted for non-response, the above expressions should be modified accordingly by replacing the basic design weights with the adjusted or calibrated weights , for all . Recall that denotes the set of units in a sample selected from the population using a *probability sampling design* , that ensures strictly positive first order inclusion probabilities . These inclusion probabilities are assumed known , at least to the data producers.

38. 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 estimates. This variability is measured using either the *sampling variance*, the *standard error (se)* or the *coefficient of variation* (*cv*) of the estimates. The latter two are functions of the first and serve to indicate by how much the estimate might differ from the true population value in absolute (*se*) or relative (*cv*) terms. Under the design-based framework and assuming full response, is unbiased for and its sampling variance can be estimated unbiasedly by

|  |  |  |
| --- | --- | --- |
|  |  | (9-4) |

where and . This result assumes that the sampling design is such that .

39. Accounting for the sampling design is crucial for the computation of variance. For example, consider a population of size , and a simple random sample of size . Assume that the sample is selected without replacement and that the following sample values were observed: . Under this sampling design, the estimated variance of the Horvitz-Thompson estimator becomes:

|  |  |  |
| --- | --- | --- |
|  |  | (9-5) |

where is the sample variance of the observed values. Substituting appropriately, the formula becomes . In contrast, when ignoring the sampling design, a naive analyst might incorrectly compute the variance using the simplified formula , overestimating the variance by failing to account for the sampling design.

40. Notice that the estimate for the population total is . The correct standard error, based on the sampling design, is . Besides, if the variance were instead estimated using a naive method that ignores the sampling design, the resulting confidence interval would be wider and misaligned, potentially producing misleading inferences. This example clearly demonstrates the importance of considering the sampling design when estimating variances, standard errors and confidence intervals.

41. While the formula for variance estimation is general and applies to most sample designs used in household surveys, it is not used in practice because the second-order inclusion probabilities (and corresponding pairwise weights ) are generally unknown to secondary survey data users. In fact, even data producers do not compute such pairwise weights, as simpler and more efficient methods for variance estimation exist that do not require them. These methods are often used in practice, allowing users to quantify uncertainty without requiring such detailed information.

#### 9.3.2. Approaches to variance estimation

42. As mentioned earlier, when working with household surveys, it is essential to provide not only point estimates but also quantify the uncertainty around these estimates. Understanding and estimating uncertainty is a critical part of analysing household survey data. By applying appropriate methods, users can measure the precision of their estimates. Various methods are available to estimate precision, and with the aid of modern software, these approaches can be implemented efficiently to support accurate and meaningful analysis.

* *Estimating Equations* provides a flexible framework for estimating totals, means, ratios, and other parameters as well as their corresponding variances ([Binder, 1983](#ref-Binder1983)) comprising a unifying idea of sampling theory.
* *Taylor Linearization* is an approach that relies on approximating complex non-linear statistics by linear ones and then estimating corresponding variances of the linear approximating quantity.
* The *Ultimate Cluster* method is often used in surveys that collect data through stratified multi-stage sampling and relies on computing the variance between quantities calculated at the level of the primary sampling units (PSU). It is often combined with *Taylor Linearization* for obtaining estimates of variances of non-linear statistics, such as means, ratios, etc.
* The *Bootstrap* and other replication methods rely on sampling repeatedly from the observed sample, computing estimates from each replicate, and then using the variability between the replicate estimates to estimate corresponding variance of the main estimate.

##### 9.3.2.1. Estimating equations and Taylor linearization

43. Many population parameters can be expressed as solutions of *population estimating equations*, involving population totals. While technical details can be complex, the key idea is that the same principles used to estimate totals can also be applied to estimate variances. This general framework makes the method simple and versatile, allowing for efficient implementation in specialized software.

44. A generic population *estimating equation* is given by , where is an *estimating function* evaluated for unit ; and is the 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 estimate many different types of parameters.

* For the case of the population total, take . The corresponding population estimation equation is given by . Solving the estimating equation for gives the *population total,* .
* Similarly, for population means, take . The corresponding population estimation equation is given by . By solving for in the estimating equation, we obtain the *population mean*, .
* For ratios of population totals, taking , the corresponding population estimating equation is given by . Solving the estimating equation for gives the *population ratio,* .

45. The idea of defining population parameters as solutions to population *estimating equations* leads to a general method for obtaining corresponding sample estimators. The approach involves using *sampling estimating equations* of the form . Under *probability sampling*, assuming full response, 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.

46. Taylor Linearization is a widely used method for approximating the variance of non-linear estimators; the method involves approximating the non-linear estimator with a first-order Taylor expansion around the estimated parameter, thereby transforming the problem into one that can be treated using linear techniques. This approach is valuable because it allows for the estimation of variances in situations where direct formulas may not exist or be easily computed.

47. A consistent estimator for the variance of non-linear estimators obtained as solutions of sample estimating equations, derived using Taylor Linearization, is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-6) |

where . This approach enables the estimation of many population parameters and their variances using well-known methods for estimating totals. Its simplicity and generality have enabled its integration into specialized software.

##### 9.3.2.2. Ultimate cluster

48. *The Ultimate Cluster method* is a straightforward and powerful approach for estimating the variance of totals in surveys that use stratified multi-stage cluster sampling designs. This method, proposed by Hansen, Hurwitz, and Madow ([1953](#ref-Hansen1953)), simplifies the complex nature of multi-stage designs by focusing only on the variation between the PSUs. It assumes that, within each sampling stratum, PSUs were sampled independently with replacement (potentially with unequal probabilities), even if they were actually selected without replacement in the real survey sampling process.

49. The method considers the variation between statistics computed at the PSU level. When properly applied, it implicitly reflects any subsampling conducted within the PSUs, enabling for simpler yet reliable variance estimates. This approach is particularly powerful, as it accommodates a wide range of complex 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 the following:

* Unbiased estimates of totals for the variable(s) of interest are available for each sampled PSU.
* Data are available for at least two sampled PSUs within each stratum (if the sample is stratified in the first stage).
* The survey dataset contains complete information on PSUs, strata and weights.

50. Consider a multi-stage sampling design, in which PSUs are selected in stratum , for . Let denote an estimate of the population total of PSU *i* in stratum . An unbiased estimator of the population total is given by where . Notice that and represent the population and sample sets of PSUs in stratum , respectively. The *Ultimate Cluster* estimator of the corresponding variance is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-7) |

For more details, see Hansen, Hurwitz, and Madow (1953, vol. I, p. 257) or Wolter (2007).

51. Although the method was originally proposed for estimation of variances of total estimators, it can also be applied in combination with *Taylor Linearization* and *Estimating Equations* approaches to obtain variance estimates for estimators of many other population quantities that can be expressed as a solution to an estimating equation. This makes the method versatile and useful for a wide range of applications in household survey analysis.

52. A key assumption of the method is that, within each stratum, PSUs are selected independently and with replacement. In practice, however, most household surveys select PSUs without replacement, leading to more efficient designs. Consequently, variance estimates based on the independence assumption are approximations of the true sampling variances. When the sampling fraction is small (e.g., less than 5%), these approximations are generally adequate and sufficiently accurate to be used by NSOs or secondary analysts.

53. The *Ultimate Cluster* method is particularly appealing due to its simplicity. Survey practitioners often prefer it over more complex approaches that explicitly account for all stages of the sampling design. Although these detailed methods may provide slightly more accurate variance estimates, they are harder to implement and require more extensive information about the sampling process. In contrast, this method offers a reasonable approximation that performs well in most practical purposes, especially for estimating totals or means. A discussion about the quality of this approximation and alternatives can be found in [Särndal, Swensson, and Wretman (1992](#ref-SSW92), p. 153).

##### 9.3.2.3. Bootstrap

54. To protect confidentiality, publicly available survey microdata often excludes essential design information (e.g., strata or PSU identifiers), which restricts users’ ability to obtain valid variance estimates. In these cases, it is advisable for NSOs to supply replication weights, allowing analysts to compute appropriate standard errors. Without either design variables or replication weights, secondary users cannot replicate published standard errors or adequately account for the survey’s complex design in variance estimation.

55. Replication methods for variance estimation are based on the idea of resampling from the available sample, computing the estimates from each replicate, and using the variability between the estimates across replicates to estimate the variance of the original estimate. They are particularly useful when information on stratum and/or PSUs identifiers is unavailable in the database, and the *Ultimate Cluster* method cannot be used.

56. The *bootstrap* method is a powerful and flexible replication method for estimating variances in surveys and many other contexts. Originally proposed by Efron ([1979](#ref-Efron1979)) for non-survey data, the version commonly applied to household surveys is the Rao-Wu-Yue Rescaling Bootstrap ([Rao, Wu, and Yue 1992](#ref-Rao1992)), which is well-suited to stratified multi-stage sampling designs and is widely used for variance estimation in complex survey data.

57. This method relies on creating many replicated datasets, which are slightly different versions of the original sample, that simulate repeatedly drawing samples from the population. In practice, this is done by creating multiple columns of replicate weights within the original sample dataset, each reflecting the process of re-sampling from the available sample. These replicates yielding multiple columns of weights should be generated by the survey provider (NSO) following the steps outlined below:

1. First, generate a new sample for each stratum by randomly selecting PSUs from the original sample with replacement (allowing PSUs to be selected more than once). Each selected PSU is included in the new dataset along with *all* its associated observations. Assuming that the first-stage sample size in stratum *h* is greater than two (i.e., ), the size of this random sample with replacement will consist of PSUs in each of the design strata (one fewer than the actual number in stratum *h*).
2. This process of creating new samples is repeated many times, usually hundreds, to generate multiple replicated datasets (that is, Step 1 is repeated times). Let denote the number of times PSU of stratum is selected in replicate . Note that can range from to .
3. For each replicate , new Bootstrap weights are calculated for each unit, reflecting how often each PSU appears in that replicate. These weights ensure that the replicated datasets remain representative for the original population. The Bootstrap weight for unit within PSU of stratum in replicate is given by .

Whenever the original sampling weights are adjusted for non-response or calibrated, the same adjustments must be applied to each set of Bootstrap weights to adequately account for additional uncertainty introduced by non-response or calibration adjustment

58. If the NSO does not release strata or PSU identifiers in the database but provides Bootstrap replicate weights, secondary users can still estimate appropriate standard errors. For each replicate , the target parameter can be estimated as using the Bootstrap weights instead of the original weights . The variability of these replicate estimates is then used to approximate the variance of the original estimator, with the dispersion across replicates reflecting its uncertainty. The bootstrap variance estimator is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-8) |

Where is the average of the replicate estimates.

59. The Bootstrap method offers several advantages. Although computationally intensive, requiring many replicates to be generated and processed, 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 practical solution for variance estimation when other methods are not available or impractical. It is particularly helpful for users analysing a survey dataset lacking design variables (strata and PSUs identifiers) but that does include a set of replicated weights.

60. Due to its simplicity, the Bootstrap method enables users to estimate variances even without specialized software. Nonetheless, many modern statistical packages support Bootstrap replication and variance estimation, making it accessible to a wide range of users, producing robust variance estimates even for complex parameters and remaining one of the most flexible tools for analysing survey data. However, it is less suitable for repeated surveys with overlapping samples, or in case of large sampling fractions and small sample sizes (Bruch, 2011).

#### 9.3.3. The design effect

61. As defined by Kish ([1965, 258](#ref-Kish_1965)), the design effect ( is the ratio of the variance of an estimator under the a complex sampling design to the variance of the same estimator under a simple random sampling (SRS) of the same size. It is estimated as:

|  |  |  |
| --- | --- | --- |
|  |  | (9-9) |

where represents the estimated variance of an estimator under a complex sampling design , and denotes the estimated variance of under an SRS of the same size. As claimed in Chapter 5, the design effect quantifies the increase in variance due to clustering and other complexities in the sampling design compared to SRS. According to United Nations ([2008, 49](#ref-United_Nations_2008)), the design effect can be interpreted in three ways: (1) as the factor by which variance increases under a complex design compared to SRS, (2) as a measure of the relative loss of precision, or (3) as a reflection of how much larger the sample size would need to be under the complex design to achieve the same variance as with SRS.

62. Park et al. ([2003](#ref-Park_2003)) propose that the design effect of a survey can be decomposed into three multiplicative components:

1. Effect due to unequal weighting: This component tends to slightly increase variance when sampling weights are unequal across units. Using uniform weights avoids such increase, making self-weighting designs desirable in household surveys.
2. Effect due to stratification: Stratification can reduce variance when implemented effectively, though the reduction is usually modest in household sample surveys.
3. The effect due to multi-stage sampling: This component typically increases variance because units within the same cluster tend to be more similar to each other than to those in different clusters.

63. In practice, is especially useful for evaluating the quality of survey estimates and planning the design of future surveys. A large indicates that the complex design introduces significant inefficiencies, leading to inflated variances and reduced precision of estimates. Conversely, a close to one suggests that the design has minimal impact on variance. By understanding these effects, researchers can assess whether adjustments to weighting, stratification, or subsampling size are needed to improve efficiency in future surveys.

64. Users should be cautious when interpreting a large *DEFF*, as it does not necessarily indicate a poor sampling design. The survey context should be considered; for example, a *DEFF* greater than three may seem alarming, but it is often the result of practical constraints, such as budget limitations, logistical challenges, or the need to maintain respondent cooperation. In some household surveys, it may be necessary to subsample eligible individuals within selected households rather than include all of them. Additionally, field challenges like noncoverage and nonresponse can lead to greater variability in weights (see Chapter 8) which may contribute to higher *DEFF* values.

#### 9.3.4. Using software to generate valid inferences

65. The design and analysis of household surveys must make extensive use of existing computational tools. This section reviews some computational approaches within statistical software that are used for each of the statistical processes required to publish official figures with appropriate levels of accuracy and reliability. Key processes that analysts should focus on include modelling nonresponse and statistical imputation, estimating standard errors for each indicator of interest to be included in the production tables, and analysing multivariate relationships between survey variables.

66. United Nations ([2005, sec. 7.8](#ref-United_Nations_2005)) 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, West, and Berglund ([2017](#ref-Heeringa_West_Berglund_2017), Appendix A).

67. In general, statistical software such as R, Stata, SAS, and SPSS provide packages and libraries designed to enhance the efficiency of variance estimation methods for complex samples, including replication techniques for design-based variance estimation. The R software is free to use, while the others are licensed products requiring paid licenses. In addition to generating descriptive statistics (such as means, totals, proportions, percentiles, and ratios), these programs also allow for the fitting of regression models, all while accounting for the survey design. Software tailored for survey analysis automatically reports the design effect, helping researchers with a better interpretation of the variability in their data.

68. All these packages and computational tools require users to input specific variables about the survey design, including sampling weights, strata, and cluster identifiers. Below, we provide a non-comprehensive overview of features and capabilities available offered by major statistical packages.

##### 9.3.4.1. R

69. R is a free, open-source software increasingly used in survey processing and social research. It is often the preferred tool for implementing the latest scientific developments and procedures available for survey data analysis ([R Core Team 2024](#ref-R_2024)). As an open-source platform, researchers can contribute with their own packaged computational functions to Comprehensive R Archive Network (CRAN), making them available to the broader community. The *samplesize4surveys* package ([Gutiérrez, 2020](#ref-ss4s)) helps to compute the sample size for individuals and households in repeated, panel, and rotational household surveys. The *sampling* ([Tillé and Matei 2016](#ref-Yves)) and *TeachingSampling* ([Gutiérrez, 2015](#ref-TS)) packages support the selection of probability samples from sampling frames under a wide range of designs and algorithms. The *survey* package ([Lumley 2024](#ref-TL)) allows for analysing household survey data and obtaining appropriate standard error estimates, once the survey design is predefined using the *svydesign*() function. For analysing inequality measures, the *convey* package (Pessoa, et. al., 2024) can be used. For regression modeling, the *svydiags* package (Valliant, 2024) provides diagnostic tools such as residual analysis, leverage values, variance inflation factors, and collinearity diagnostics for survey data. The package *PracTools* (Valliant et al, 2025) offers tools for sample size calculations, sample design, design effects, and variance components for multi-stage designs.

##### 9.3.4.2. Stata

70. The *svy* environment in Stata provides tools for valid inference from household surveys ([Stata 2017](#ref-STATA_2017)). The *svyset* command specifies variables identifying survey design features, such as sampling weights, clusters, and strata. The *svydescribe* command generates tables describing the structure of strata and sampling units at a given survey stage. Once survey design is specified, models can be estimated using standard commands, and the resulting statistics will properly account for the complex survey design. The *svy* environment also supports regression modelling commands, allowing for prediction of the response outcome based on covariates.

##### 9.3.4.3. SPSS

71. The *Complex Samples* module in SPSS ([IBM 2017](#ref-IBM_2017)) supports the selection of complex samples using user-defined sampling schemes. Once the sampling plan is defined, an analysis plan must be created by assigning design variables, estimation methods, and sample unit sizes. When this setup is complete, the module enables the estimation of frequencies, descriptive statistics, and crosstabulations. It also supports the estimation of ratios and regression coefficients in linear models, along with associated hypothesis tests. Additionally, the module allows for the estimation of nonlinear models, such as logistic, ordinal, and Cox regressions.

##### 9.3.4.4. SAS

72. This statistical software includes a procedure for selecting probability samples called *SURVEYSELECT*, which integrates sample selection methods such as simple random sampling, systematic sampling, probability proportional to size sampling, and stratified allocation tools ([SAS 2010](#ref-SAS_2017)). To analyse data from complex samples, SAS offers specialized procedures: *SURVEYMEANS* estimates totals, means, proportions, and percentiles, along with standard errors, confidence intervals, and hypothesis tests; *SURVEYFREQ* produces descriptive statistics in one- and two-way tables, provides sampling error estimates, and performs tests for goodness-of-fit, independence, risk measures, and odds ratios; *SURVEYREG* and *SURVEYLOGISTIC* fit linear and logistic regression models, respectively, estimating regression coefficients and associated errors while offering detailed diagnostics; *SURVEYPHREG* fits survival models using pseudo-maximum likelihood estimation techniques.

### 9.4. Descriptive parameters

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

74. 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 behaviours and trends. Frequencies indicate the number of households/people in a specific category (e.g., number of poor people), while proportions reflect the share of households/people with a particular attribute (e.g., poverty rate).

75. 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., median household income). There are also metrics for particular types of analysis, such as poverty (FGT indices), inequality (Gini, Theil, Atkinson), polarization (Wolfson, DER), among others (Jacob, Damico, and Pessoa, [2024](#ref-Jacob2024)).

#### 9.4.1. Frequencies and proportions

76. One of the most fundamental tasks in household survey analysis is estimating the size of subpopulations, for example, 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.

77. For example, valuable insights come from understanding how many people live below the poverty line, how many are unemployed, or how many have completed a certain level of education. These insights help address inequalities, support the design of targeted interventions, and promote equitable development across communities. Understanding the distribution across categories provides essential information for tackling disparities and advancing inclusive development.

78. 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 income quintiles, labour force 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. These estimates are calculated by combining survey responses with sampling weights, which 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:

|  |  |  |
| --- | --- | --- |
|  |  | (9-10) |

where is the sample of households or individuals in PSU of stratum ; is the sample of PSUs within stratum ; and is the weight (expansion factor) of unit within PSU in stratum .

79. Subpopulation size estimates work similarly but focus on a subset of the population defined by a specific characteristic. For example, to estimate the number of people in a particular category, we identify the relevant group in the survey data and sum up their sampling weights. This approach allows analysts to estimate not only the total population size but also the size of specific groups of interest. To do this, a binary variable should be defined, . It takes the value one if unit from PSU in stratum belongs to category (which was not controlled for during the weighting adjustment stage) in the discrete variable , and zero otherwise. A sampling estimator for this parameter is given by the following expression:

|  |  |  |
| --- | --- | --- |
|  |  | (9-11) |

If *d* is a category that was controlled for in the weighting, then will equal the external control value and should not be treated as an estimate.

80. 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 a binary variable. This approach ensures that the estimate accurately reflects the population distribution. As noted by Heeringa, West, and Berglund ([2017](#ref-Heeringa2017)), by recoding the original response categories into simple indicator variables with values of 1 and 0 (e.g., 1=Yes, 0=No), the estimator for a proportion is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (9-12) |

As this defines a nonlinear estimator, Taylor linearization can be applied to approximate its variance by defining the corresponding estimating function as . Many statistical packages provide proportion estimates and standard errors expressed on a percentage scale.

81. When target proportions are close to 0 or 1, larger sample sizes may be necessary to obtain statistically valid results. Additionally, alternative methods should be employed to ensure confidence intervals remain within the [0,1] range. Note that traditional symmetric normal-based confidence intervals may fall outside this range, rendering them uninterpretable. To address this issue, alternative confidence interval estimates, such as those proposed by Rust and Hsu ([2007](#ref-Rust2007ConfidenceIF)), and Dean and Pagano ([2015](#ref-DeanPagano2015)) are available. One alternative based on using the logit transformation of the estimated proportion is:

|  |  |  |
| --- | --- | --- |
|  |  | (9-13) |

#### Note that stands for the margin of error of the estimator; i.e., ; denotes the quantile of a t-student distribution with degrees of freedom, leaving an area of *α/2* to its right. The degrees of freedom are determined by the number of PSUs minus the number of strata (i.e., ). Several other alternatives are available, including those that can be computed even for proportions equal to zero or one.

#### 9.4.2. Totals, means and ratios

82. In household surveys, analysing numerical data often involves estimating key descriptive statistics such as means, totals, and ratios. These measures summarize important characteristics of the population and support decision-making. Estimates may be derived for the entire population or targeted subgroups, depending on the research objectives. As mentioned by Heeringa, West, and Berglund ([2017](#ref-Heeringa_West_Berglund_2017)), the estimation of population totals or means, along with their corresponding variances, has played a crucial role in the development of probability sampling theory.

83. Estimating population totals is fundamental to survey analysis. Population means, proportions and ratios are all functions of population totals. A total represents the sum of a specific variable (e.g., total income or total expenditure) across the entire population. For example, 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 of totals and means. Ratios compare two numeric variables. Estimates for such parameters can be obtained either for the entire population or disaggregated by domains of interest, depending on research needs.

84. After defining the sampling design, the estimation process for the parameters of interest is carried out. For surveys with complex sampling designs that include stratification and subsampling in PSUs (assumed to be within stratum ) indexed by , the estimator for the population total can be written as:

|  |  |  |
| --- | --- | --- |
|  |  | (9-14) |

Under full response, the Ultimate Cluster variance estimator for was provided in Section 9.2.

85. The confidence interval for the population total is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-15) |

When the degrees of freedom are large, the Student’s t-distribution converges to the Normal distribution. Thus, it is common for NSOs to report confidence intervals using the Normal approximation. However, these intervals rely on assumptions that may not hold for estimates based on small sample sizes but generally perform well in large-scale household surveys.

86. Population means, or averages, are also very important for understanding the variable’s central tendency. For instance, average household expenditure can reflect economic performance of a population. The mean is calculated as the total of a variable divided by the population size. As mean estimation depends on both totals and population sizes, its accuracy relies on precise estimation of both components. The population mean estimator is expressed as a ratio of two totals:

|  |  |  |
| --- | --- | --- |
|  |  | (9-16) |

Because is a nonlinear estimator, its exact variance has no closed form. Thus, either resampling methods such as the bootstrap or Taylor series approximations must be used. The latter may be achieved recalling that its sampling estimating equation is ; therefore, we can apply the variance estimator given in Section 9.2 with . These computations are automatically done when using appropriate software to handle complex survey data.

87. Ratios measure relationship between two variables. For example, the ratio of household expenditure to income reveals spending patterns. 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. For instance, SDG indicator 2.1.1 (prevalence of undernourishment) can be estimated using a ratio of two continuous variables: food consumption (calories or energy intake) and dietary energy requirements (based on age, sex, and physical activity level). The point estimator for a ratio is the quotient of the estimators for the totals:

|  |  |  |
| --- | --- | --- |
|  |  | (9-17) |

For variance estimation, it is needed to specify the estimating function as , when and are the numerator and denominator variables, respectively, and apply the variance estimator given in Section 9. 2.

#### BOX – ESS4 expenditure descriptive statistics

Section 7b of the household questionnaire of the ESS4 is related to expenditures on non-food items. After appropriately accounting for the sampling design, the table below presents the estimated percentage of households that reported spending on these items, the estimated average expenditure, and the expenditure ratio relative to total non-food expenditures. Standard errors are provided in parentheses.

|  |  |  |  |
| --- | --- | --- | --- |
| **Nonfood item** | **Percentage** | **Mean** | **Ratio** |
| Clothes for MEN | 58 (1.8 ) | 739 (38 ) | 17.3 (0.6 ) |
| Clothes for WOMEN | 63.1 (1.8 ) | 617 (28 ) | 14.4 (0.5 ) |
| Clothes for BOYS | 52.5 (1.2 ) | 411 (16 ) | 9.6 (0.4 ) |
| Clothes for GIRLS | 51.7 (1.2 ) | 353 (14 ) | 8.3 (0.3 ) |
| Kitchen equipment | 17.1 (1.5 ) | 77 ( 7 ) | 1.8 (0.2 ) |
| Linens | 28.3 (1.6 ) | 197 (15 ) | 4.6 (0.3 ) |
| Furniture | 7.9 (0.8 ) | 160 (23 ) | 3.7 (0.5 ) |
| Lamp | 27.9 (1.7 ) | 68 ( 9 ) | 1.6 (0.2 ) |
| Ceremonial expenses | 61.3 (2.2 ) | 1338 (83 ) | 31.4 (1.3 ) |
| Informal social security | 44.8 (2.2 ) | 146 (13 ) | 3.4 (0.3 ) |
| Religious donations | 40.7 (2.1 ) | 161 (16 ) | 3.8 (0.4 ) |

From the table clothing appears to be a high-priority expenditure, with around 63% of households reporting spending on women’s clothing and 58% on men’s clothing. Ceremonial expenses have the highest mean expenditure, significantly surpassing all other categories. Clothing expenditures seems to highlight some gender differences where men's and boy’s clothing receives higher spending per household compared to women’s and girls. Not surprisingly, the highest expenditure ratio is for ceremonial expenses, reinforcing its dominant role in nonfood expenditures.

#### 9.4.3. Correlations

88. Correlation analysis is a useful method for understanding the relationship between two numeric variables in survey data. For example, an analyst might be interested in knowing whether household income and expenditure are related, and if so, how strongly. The Pearson correlation coefficient (which ranges from –1 to 1) is commonly used to measure this relationship, as it quantifies the strength and direction of a linear relationship between two numeric variables. A positive coefficient value indicates that as one variable increases, the other also tends to increase; a negative value indicates the opposite, as one variable increases, the other tends to decrease; and a value close to zero suggests little or no linear relationship between the variables.

89. When analysing survey data, correlation is estimated using the survey weights. These weights ensure that the correlation estimate reflects the relationships in the entire population, not just the sample. Weighted correlations adjust for complex survey design, accounting for stratification, clustering, and unequal probabilities of selection. The correlation coefficient is computed by comparing how the variables vary together (covariance) and normalizing this by their individual variations. This normalization ensures the correlation is unaffected by the units of measurement of the variables, making it easier to interpret.

90. The Pearson correlation coefficient between two numeric survey variables, say and , can be estimated as:

|  |  |  |
| --- | --- | --- |
|  |  | (9-18) |

For non-numeric variables (e.g., categorical or ordinal), alternative correlation measures are available. Additionally, models can be fitted to ascertain relationships between a specified response and covariates – see subsection 9.5 for details.

#### 9.4.4. Percentiles and inequality measures

91. Percentiles and quantiles are useful tools for analysing 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. An estimator of the CDF in a complex sampling design is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-19) |

Where is an indicator variable taking the value 1 if is less than or equal to a specific value , and 0 otherwise. Following Heeringa, West, and Berglund ([2017](#ref-Heeringa2017)), to estimate quantiles, one first considers the order statistics denoted as and finds the value of such that:

|  |  |  |
| --- | --- | --- |
|  |  | (9-20) |

Hence, the estimator of the -th quantile is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-21) |

Quantiles are inherently nonlinear measures, making their variance estimation more complex. Kovar, Rao, and Wu ([1988](#ref-kovar1988bootstrap)) present results from a simulation study where they recommend using replication methodswhen estimating standard errors for quantiles in survey data.

92. Economic inequality is a critical area of focus for governments and international organizations. The Gini coefficient is a widely used measure to quantify income or wealth inequality. It compares the income distribution of the target population to that of a perfectly equal distribution. In household surveys, it is estimated using weights that account for the survey design. Normalized weights are often used to simplify calculations. The Gini coefficient ranges from 0 to 1, where 0 indicates perfect equality (equal income for all) and higher values indicate greater inequality. This metric coefficient is critical for tracking changes in income distribution over time and comparing inequality between regions or countries.

93. Following the estimating equation proposed by Binder and Kovacevic ([1995](#ref-binder1995estimating)), the Gini coefficient estimator is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-22) |

where is the normalized sampling weight, defined as

|  |  |  |
| --- | --- | --- |
|  |  | (9-23) |

and is the estimated CDF for individual in cluster of stratum . Osier ([2009](#ref-osier2009variance)) and Langel and Tillé ([2013](#ref-Langel_Tille_2013)) provide important computational details for estimating the variance of this complex estimator.

#### BOX – ESS4 income correlations and inequalities indicators

The ESS4 dataset reveals intriguing correlations among various types of household expenditures, and the researcher can get initial estimates to assess the strength of the relationship between some of them. This can be done at the national level, capturing the overall association, as well as disaggregating to showcase differences (for example between urban and rural, or by region). At the national level, the correlation between educational spending and utility expenses is estimated at 0.19 (0.03), suggesting a moderate positive association. In contrast, the association between educational spending and food away from home is notably weaker at 0.06 (0.02).

When the data is disaggregated by area type, it is observed that in rural areas the education-utilities relationship remains moderate at 0.19, albeit with higher uncertainty (standard error = 0.07). Meanwhile, the correlation between educational spending and food away from home in rural areas is negligible at 0.009. In urban areas, the positive correlation with utilities is slightly lower at 0.16 (standard error = 0.02), and the correlation with food away from home is marginally higher at 0.039 (standard error = 0.02), though it remains weak.

A regional breakdown of the data provides further insights. Regions such as Dire Dawa (0.33, standard error = 0.07), Somali (0.31, standard error = 0.05), Amhara (0.24, standard error = 0.05), and Harar (0.23, standard error = 0.05) exhibit moderate, positive correlations between education and utilities, suggesting that higher investment in education tends to coincide with increased utility spending. In contrast, the association with food away from home varies more widely. Tigray shows a stronger positive relationship at 0.25 (standard error = 0.06), whereas Benishangul Gumuz (-0.07, standard error = 0.02) and Dire Dawa (-0.04, standard error = 0.02) show negative correlations, indicating different spending trade-offs or priorities.

Furthermore, when overall household expenditures are considered, including both food and non-food items, the Gini coefficient of per capita expenditure for Ethiopia is estimated at 0.58 (0.01). Benishangul Gumuz emerges as the region with the highest inequality, with an estimated Gini coefficient of 0.67 (0.03) while, Somali is estimated to be the region with less inequality with an estimated Gini of 0.49 (0.02). These findings underscore the substantial variability in expenditure patterns and economic inequality across the country.

### 9.5. Associations between categorical variables

#### 9.5.1. Motivation and concepts

94. Household surveys often collect data on categorical variables (which classify populations into distinct groups or categories). Examples include labour force status (e.g., “employed,” “unemployed,” “outside the labour force”), educational attainment (e.g., “primary,” “secondary,” “tertiary”), or access to services (“Yes”, “No”). Determining whether two categorical variables are related, or associated, is an important aspect of survey analysis. Analysing associations between categorical variables is useful in various 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) to inform targeted interventions, and social research (studying connections between demographic factors and access to services) to provide insights into societal trends.

95. When analysing associations between two categorical variables, we examine whether the distribution of one variable depends on the other. To assess this relationship, analysts compare the frequency of different category combinations. For example, they might count how many individuals fall into each combination of labour force status and educational attainment. These counts can be used to calculate proportions, showing the relative frequency of each pairing within the population. As illustrated in the following sections, this analysis often starts with a contingency table (a grid that presents the counts or proportions of units in each combination of categories for two variables). For example, one axis of the table might list labour force statuses, while the other lists educational attainment levels.

96. We start by defining some notation. Let and denote two categorical variables, with and classes respectively. To formulate hypothesis tests for the independence between and , we consider a *superpopulation model*, assuming that the pairs correspond to observations from identically distributed random vectors , which have a joint distribution specified by

|  |  |  |
| --- | --- | --- |
|  |  | (9-24) |

with .

97. If a census were conducted collecting data on and from every unit in the population, we could calculate the population counts of units in classes for :

|  |  |  |
| --- | --- | --- |
|  |  | (9-25) |

and the corresponding population proportions as , where denotes the total number of units in the population (previously denoted as *N*). Under the superpopulation model, the population proportions could be used to estimate (or approximate) the unknown probabilities . Since we will have one sample, not a census, the population proportions must be estimated using weighted estimators described in the previous sections.

#### 9.5.2. Cross-tabulations

98. Cross-tabulations, also known as contingency tables, are a fundamental tool in survey analysis. They organize data into tables, 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 clear way to explore variables’ interactions. For example, a contingency table might examine how labour force status varies by educational attainment, or how access to the internet differs between urban and rural households. In the specialized literature, cross-tabulations are also referred to as *contingency tables*. Visualizations like stacked bar charts enhance data comprehension by revealing patterns and differences, complementing tables to effectively communicate findings. (See Section 9.8 for more details.)

99. A contingency table is a is a two-dimensional array with rows indexed by and columns indexed by . Rows and columns represent the categories of two variables, and each cell contains the frequency or proportion of observations for the corresponding combination of categories. For example, the rows might represent categories of a domain-defining variable such as educational attainment, and the columns might represent categories of another variable, such as labour force status. The table can also include marginal totals (summarizing data for each row or column) and a grand total (representing the overall population).

100. In household surveys, frequencies in contingency tables are estimated using survey weights. For each cell, the weighted frequency represents the estimated number of individuals in the population within the corresponding combination of categories. For most household sample surveys, a typical two-way contingency table output comprises the weighted frequencies that estimate the population frequencies, as shown in Figure 9.1:

**Figure 9.1.** Contingency table with the weighted frequencies that estimate the population frequencies.

|  |  | | | |
| --- | --- | --- | --- | --- |
|  | 1 |  |  | row marg. |
| 1 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| col. marg. |  |  |  |  |

where the estimated frequency in cell is obtained as

|  |  |  |
| --- | --- | --- |
|  |  | (9-26) |

and , and .

101. Weighted frequencies can also be converted into proportions, which indicate the relative size of each group compared to the total population, or to the totals of rows or columns in the table. 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:

|  |  |  |
| --- | --- | --- |
|  |  | (9-27) |

and and . Two-way tables can also display the estimates of population relative frequencies (Figure 9.2).

***Figure 9.2.*** *Contingency table with estimates of population relative frequencies.*

|  |  | | | |
| --- | --- | --- | --- | --- |
|  | 1 |  |  | row marg. |
| 1 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| col. marg. |  |  |  |  |

#### BOX – ESS4 religious composition

The ESS4 questionnaire allows for the estimation of religious affiliation across regions in Ethiopia. For instance, the following table presents a cross-tabulation estimating the percentage of people aged 10 years or older who identify with one of the listed religions, with standard errors shown in parentheses.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Region** | **Orthodox** | **Catholic** | **Muslim** | **Protestant** | **Other** | **Row Marginal** |
| Addis Ababa | 74.9 (2.5) | 0.5 (0.2) | 15.0 (2.1) | 9.3 (1.5) | 0.3 (0.1) | 4.6 (0.2) |
| Afar | 8.2 (2.7) | 0 (0.0) | 83.5 (3.2) | 2.2 (1.2) | 6.1 (1.0) | 0.8 (0.1) |
| Amhara | 71.7 (5.2) | 0.2 (0.1) | 21.1 (5.1) | 0.5 (0.2) | 6.6 (0.9) | 23.9 (0.6) |
| Benishangul Gumuz | 36.3 (6.9) | 0.7 (0.3) | 45.5 (7.9) | 11.2 (3.7) | 6.3 (1.3) | 1.0 (0.0) |
| Dire Dawa | 35.5 (4.8) | 0.5 (0.2) | 58.1 (5.2) | 4.2 (0.8) | 1.8 (0.8) | 0.5 (0.0) |
| Gambela | 29.8 (5.3) | 6.1 (2.8) | 8.8 (2.0) | 51.1 (6.9) | 4.2 (1.1) | 0.4 (0.0) |
| Harar | 23.6 (3.2) | 0.2 (0.1) | 69.4 (3.7) | 3.8 (0.9) | 3 (0.7) | 0.3 (0.0) |
| Oromia | 27.8 (4.2) | 1.8 (1.2) | 36.6 (5.3) | 24.9 (4.7) | 9 (1.5) | 37 (0.8) |
| SNNP | 21.6 (4.1) | 0.7 (0.2) | 15.7 (4.5) | 54.6 (5.3) | 7.4 (1.2) | 20.2 (0.6) |
| Somali | 0.6 (0.6) | 0 (0.0) | 98.0 (1.0) | 0.1 (0.1) | 1.3 (0.7) | 5.0 (0.2) |
| Tigray | 92.2 (1.9) | 2.1 (1.8) | 1.7 (0.6) | 0 (0.0) | 4.0 (0.6) | 6.3 (0.2) |
| **Col. Marginal** | 41.9 (2.2) | 1.0 (0.5) | 29.1 (2.5) | 21.2 (2.1) | 6.8 (0.6) | 100 (0.0) |

As the table presents the religious composition of different regions, the results reveal strong regional religious concentrations: Tigray, Amhara, and Addis Ababa have predominantly Orthodox populations, while Afar, Somali, Harar, and Dire Dawa are mostly Muslim. Oromia, SNNP, and Gambela have significant Protestant communities. Some regions, like Benishangul Gumuz, display a more balanced mix of Orthodox, Protestant, and Muslim populations. The presence of Catholicism is minimal across all regions.

#### 9.5.3. Testing for independence

102. A hypothesis test is a statistical procedure used to evaluate evidence in favour of or against a statement or assumption about a population. In this process, a null hypothesis () is proposed, representing the statement that needs to be tested, and an alternative hypothesis (), which opposes the null hypothesis. These statements are often based on belief or experience and tested using the evidence gathered from the sample data. One of the two hypotheses will be considered true only if the statistical evidence obtained from the sample supports it. The decision is based on the statistical evidence gathered from the data. This process is called hypothesis testing.

103. 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, analysts might ask: “Is there a relationship between educational level and labour force 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 unrelated.

104. To perform these tests, analysts rely on models that assume the data comes from a larger, hypothetical population (a *superpopulation*). The observed survey data 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.

105. Cross-tabulations can be presented using either absolute values or percentages, both of which may be useful for different and complementary purposes. To test the hypothesis of independence, observed proportions in a contingency table are compared with the expected ones 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. Following Heeringa, West, and Berglund ([2017](#ref-Heeringa2017)), the null hypothesis that and are independent is defined as:

|  |  |  |
| --- | --- | --- |
|  |  | (9-28) |

106. Hence, to test independence, we compare the estimated proportions with the estimated expected population proportions under the null hypothesis, . If there is a large difference between them, the independence hypothesis is not supported by the data. In other words, if the observed data differs from the expected values under the null hypothesis, then there is evidence against the validity of .

107. Testing for independence in survey data is more complex than in traditional tests, as it requires adjustments to account for the sampling design. The Rao-Scott adjustment modifies traditional chi-square tests to incorporate these effects. The traditional test statistic is adjusted using the generalized design effect (GDEFF), which accounts for the sampling complexity. The adjusted test statistic follows a chi-square distribution under the null hypothesis. The Rao-Scott adjusted test statistic ([Rao and Scott 1984](#ref-Rao1984)) is defined by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-29) |

where estimates the cell frequencies under the null hypothesis and is an estimate of the generalized design effect (Heeringa, West, and Berglund ([2017](#ref-Heeringa2017)) p. 177). Under , the large sample distribution of is .

108. For small sample sizes or limited degrees of freedom, F-distribution-based adjustments to the improve accuracy. As mentioned by Heeringa, West, and Berglund ([2017](#ref-Heeringa2017)), it was Fay ([1979](#ref-Fay1979)), along with Fellegi ([1980](#ref-Fellegi1980)), who began proposing corrections to Pearson’s chi-square statistic based on a generalized design effect. Rao and Scott ([1984](#ref-Rao1984)) and Thomas and Rao ([1987](#ref-thomas1987small_sample)) later expanded these methods. The Rao-Scott adjustment requires the calculation of generalized design effects, which are more complex analytically than Fellegi’s approach. Nevertheless, Rao-Scott-adjusted statistics are the standard for analysing categorical survey data in specialized software.

#### BOX – ESS4: Education and Residential Zone

A deeper dive into the ESS4 data unveils interesting insights into how educational attainment among parents varies by residential zone. The survey data allow for an estimation of the percentage of biological fathers and mothers reaching one of four distinct education levels, segmented by two different area types. The following table, constructed from cross-tabulations, offers a swift snapshot of these patterns.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Education** | **Father/Rural** | **Father/Urban** | **Mother/Rural** | **Mother/Urban** |
| No education | 52.4 (2.2) | 29.7 (1.9) | 77.3 (1.9) | 42.6 (2.6) |
| Primary | 41.6 (2.0) | 35.6 (1.9) | 21.1 (1.9) | 32.9 (2.2) |
| Secondary | 4.2 (0.8) | 19.5 (1.7) | 1.3 (0.3) | 15.4 (1.6) |
| Above secondary | 1.8 (0.4) | 15.2 (1.4) | 0.3 (0.1) | 9.2 (1.1) |

The data shows both significant gender and area-type disparities in educational attainment among parents in Ethiopia. In both rural and urban areas, mothers consistently have lower levels of education compared to fathers. For instance, 77.3% of rural mothers have no education, which is notably higher than the 52.4% observed for rural fathers; similarly, in urban settings, 42.6% of mothers have no education versus 29.7% of fathers. Additionally, the percentages of mothers achieving secondary and above secondary education are considerably lower than those of fathers in both environments.

The adjusted Rao & Scott chi-square test reveals a highly significant association between maternal education and residential zone (F = 100.29, p < 0.001), rejecting the null hypothesis of independence. This result effectively suggests that a mother’s educational level is intricately linked to whether she resides in a rural or urban area. A similar pattern emerges for paternal education, with the chi-square test showing a significant association (F = 76.82, p < 0.001). Although the F statistic for fathers is slightly lower than that for mothers, the relationship remains robust, underscoring that fathers in urban environments tend to have higher educational levels compared to those in rural settings. These findings vividly illustrate how residential context plays a crucial role in shaping educational outcomes for both mothers and fathers.

#### 9.5.4. Tests for group comparisons

109. Comparing group means is a common goal in household survey analysis. For example, analysts might ask: “Is there a significant difference in average income between male and female headed households?” To answer such questions, statistical tests are used, accounting for the complexities of survey data. This section explains methods for testing differences in means, adjusted for survey design, with examples to illustrate their application, introducing t-tests and contrasts that are appropriately adjusted for the sampling design.

##### 9.5.4.1. Hypothesis Test for the Difference of Parameters

110. Many parameters of interest, such as differences in means or weighted sums of means, can be expressed as linear combinations of descriptive statistics. These combinations are often used to construct economic indices or compare population means. The variance of these combinations is important for understanding the precision of the estimate. The most common cases include differences in means, weighted sums of means used in economic indices, etc. Consider a function that is a linear combination of descriptive statistics, as shown below:

|  |  |  |
| --- | --- | --- |
|  |  | (9-30) |

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

|  |  |  |
| --- | --- | --- |
|  |  | (9-31) |

and its variance is calculated as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (9-32) |

The variance expression requires both the variances of the individual estimators and the covariances between them.

111. When comparing two populations, several potential hypotheses can be tested. For example, the null hypothesis may state that the parameters (such as total, means, proportions, ratios, among others) of the two populations are equal, while the alternative hypothesis could suggest that they differ, or that one is greater than or less than the other.

112. Of particular interest is analysing 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 correspond to observations from identically distributed random variables having means if unit belongs to domain , with . Then we can define the difference in population means between domains 1 and 2 as . As an example, consider that is the average household income for households with male heads of household, and is the average household income for households with female heads.

113. This difference in means can be consistently estimated by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-33) |

where is the sample estimator of (). The hypotheses to test are typically:

* Null hypothesis (): No difference between the means.
* Alternative hypothesis (): There is a difference, which could be in either direction (greater or less).

114. To test one of these hypotheses, the following test statistic is used

|  |  |  |
| --- | --- | --- |
|  |  | (9-34) |

This statistic follows a Student’s-*t* distribution with degrees of freedom (, where is calculated as the difference between the number of PSUs in the sample and the number of strata . Also, its associated standard error is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-35) |

115. A confidence interval for the difference in means is constructed using the estimated difference, its associated standard error, and the appropriate critical t-value. This interval provides a range of plausible values for the true difference in means, offering a more complete understanding of the data. This confidence interval is constructed as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (9-36) |

116. The same general approach can be extended to other parameters (e.g., proportions, totals, ratios, and more broadly, any differentiable function of totals). Since their sampling distribution also approximates a *t*-distribution (Valliant & Rust, 2010), confidence intervals and hypothesis tests can be constructed in the same way: using the point estimate of the difference, its standard error, and the appropriate critical value from the *t*-distribution.

#### BOX – ESS4: Differences on per-capita expenditure patterns across regions and areas

A deeper analysis of ESS4 data uncovers a vivid picture of two contrasting economic realities. In rural areas, households average a per-capita expenditure of about 762 Ethiopian Birr (ETB), with a relatively modest standard error of ETB 46, suggesting steady and limited spending patterns. Urban areas, however, tell a different story—with average expenditures ascending to ETB 1923 and a standard error of about ETB 123, highlighting higher spending. This disparity hints at urban environments offering more diverse goods, services, and income opportunities, while rural regions may be bound by more constrained economic conditions.

The 95% confidence interval for the mean expenditure difference range from ETB 904 to ETB 1,419, with a point estimated difference of ETB 1,161—strongly favouring urban households. A design-adjusted *t*-test confirms this disparity, yielding a *t*-statistic of 8.85 and an extremely low p-value (< 0.001), leaving no doubt that the difference is statistically significant. These findings provide compelling evidence of an economic difference, where urban residents experience far greater spending power than their rural counterparts, underscoring the stark inequality in per capita expenditures based on residential location.

A closer look at regional disparities reveals that significant differences in household expenditures between urban and rural areas persist not only at the national level but also within regions. For instance, in the Tigray region, the gap is particularly striking, with urban households spending an average of ETB 1,712 more than their rural counterparts. A highly significant design-adjusted t-statistic of 4.38 (p < 0.0001) confirms this difference, with a 95% confidence interval ranging from ETB 927 to ETB 2,498, suggesting a substantial and well-defined urban advantage. In the Afar region, the pattern persists but with slightly more variability. Urban households outspend rural ones by an average of ETB 1,017, with a statistically significant design-adjusted t-statistic of 2.93 (p = 0.00554). The 95% confidence interval, spanning from ETB 316 to ETB 1,719, indicates a broader range of possible differences. These findings underscore how regional economic conditions influence spending patterns, reinforcing the broader trend of urban financial dominance across Ethiopia.

### 9.6. Regression: Modelling survey data

117. Regression modelling is a powerful tool for analysing relationships between variables in survey data. It allows researchers to estimate how one or more dependent variables (outcomes) are associated with one or more independent variables. Nolan and Speed (2000) along with Freedman (2005) discuss proper model formulation. For example, a r esearcher might model household income (dependent variable) as a function of education attainment and labour force status (independent variables) using household survey data. As the data is assumed to come from complex sampling designs, traditional regression methods must be adapted to this more complex situation.

118. In this section, we explore how survey weights and sampling design features can be incorporated into regression model specification and fitting. We also discuss a parsimonious solution to weighting challenges. Modelling 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.

#### 9.6.1. To weight or not to weight?

119. Heeringa, West, and Berglund ([2017](#ref-Heeringa_West_Berglund_2017)) 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 dealing with weights into regression models when working with complex survey data:

* 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. The 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.
* The model-based approach argues that weights are unnecessary if the model is correctly specified for the sample, and the sampling is non-informative (in the sense that the model holding for the sample would be the same as the model holding for the population). This approach assumes that the relationships between the variables are well-represented by the model, regardless of the sampling design, and that using weights can increase the variability of the estimates, leading to unnecessarily larger standard errors.

120. The choice between these two approaches depends on the context and the sensitivity of the inferences to the inclusion of weights. Skinner, Holt, and Smith ([1989](#ref-skinner1989analysis)), Pfeffermann ([2011](#ref-pfeffermann2011modelling)) discuss whether survey weights should be included in the estimation of regression parameters and their associated standard errors. 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 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 includes key survey design variables (e.g., strata or clusters). However, if unweighted models are to be reported, ensure that this is well documented and justified, since such models require stronger model assumptions when compared to the weighted models.

121. Incorporating survey weights ensures population representativeness by correcting for over or under sampling of certain groups, ensuring the regression model reflects the true population distribution. Additionally, they provide accurate variance estimates by adjusting for stratification, clustering, and unequal selection probabilities, resulting in valid standard errors, confidence intervals and test statistics. The design-based approach estimates population-level estimating equations for regression coefficients allowing weighted estimates to approximate unbiased census values — that is, the regression coefficients that would be obtained if a full census were conducted. This holds true even if the model is misspecified.

122. Sampling weights can inflate the variance of parameter estimates, particularly if they vary widely. Extreme or highly variable weights may also lead to unstable estimates, where certain observations disproportionately influence the model fitting. For explanatory or analytical purposes (e.g., understanding relationships between variables), unweighted models can sometimes provide more efficient and stable estimates.

123. Unweighted regression may fail to produce meaningful estimates if the model is misspecified. To ensure correct specification, analysts are encouraged to make every effort to include the appropriate variables in the model. However, even when the model is correctly specified, it remains important to account for design stratification and clustering when estimating standard errors under the unweighted approach. It is also important to conduct proper diagnostic analysis of fitted models – see Subsection 9.6.4.

#### 9.6.2. Some inferential approaches to modelling data

124. 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 one arises from the model itself, which is used to analyse the data and make inferences about the population. To combine both sources of variability into a coherent framework, advanced inferential methods are required. These methods aim to respect the structure of the sampling design while also accounting for the assumptions and uncertainties within the model. Two main approaches used for this purpose are pseudo-likelihood (Molina & Skinner, 1992) and combined inference (Binder, 2011).

125. The pseudo-likelihood approach modifies the traditional likelihood methods to account for the complexities of the survey design. In this framework, inference is based on the repeated sampling distribution induced by the design, while the model distribution remains secondary (although pseudolikelihood estimates may be model-unbiased or consistent if the model is correctly specified). In simpler terms, this approach adjusts standard modelling techniques to properly reflect how the sample was drawn. Such adjustments are crucial, as ignoring the sampling design can lead to biased estimates and incorrect conclusions about the population.

126. 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.

#### 9.6.3. Linear models

127. A regression model explains how 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.

##### 9.6.3.1. Basic definitions

128. A simple linear regression model is defined as , where represents the dependent variable, is the independent variable, and are the model parameters. The variable is known as the *random error* of the model.

129. For more complex situations, multiple linear regression models allow for the inclusion of several independent variables, accounting for the simultaneous effects of multiple factors on the outcome. In these models, each independent variable has a coefficient, which indicates the strength and direction of its relationship with the dependent variable. A positive coefficient suggests that as the independent variable increases, the dependent variable also increases. Generalizing the simple model, multiple linear regression is presented below:

|  |  |  |
| --- | --- | --- |
|  |  | (9-37) |

Another way to write the multiple regression model is:

|  |  |  |
| --- | --- | --- |
|  |  | (9-38) |

where, and . The subscript refers to the population element.

130. Regression models rely on several assumptions. Heeringa, West, and Berglund (2017) outline these as follows:

* . The average error for any given value of the independent variable is zero, meaning that the model does not systematically over- or under-predict outcomes.
* . The variability of errors is constant across all levels of the independent variables, a property known as homoscedasticity (homogeneity of variance).
* : The errors conditioned on the covariates follow a normal distribution. This property also extends to the response variable .
* : Errors for different observations are independent, meaning that the outcome for one observation does not influence another. This way, the errors in different observed elements are not correlated with the values given by their predictor variables.

131. When these assumptions hold, regression models can provide accurate and unbiased estimates of variable relationships. The predicted values from the model represent the expected outcomes based on the independent variables, making regression a useful tool for understanding patterns in data and making informed predictions. Shah, Holt, and Folsom ([1977](#ref-shah1977inference)) discuss some aspects related to the violation of these assumptions and provide appropriate methods for making inferences about the parameters of linear regression models using survey data.

132. Once the linear regression model and its assumptions are defined, the best unbiased linear estimator is the expected value of the dependent variable conditioned on the independent variables , as:

|  |  |  |
| --- | --- | --- |
|  |  | (9-39) |

133. As noted by Heeringa, West, and Berglund ([2017](#ref-Heeringa_West_Berglund_2017)), the first authors to empirically discuss the impact of complex sampling designs on regression model inferences were Kish and Frankel ([1974](#ref-kish1974inference)), who highlighted the challenges posed by complex sampling designs. Later, Fuller ([1975](#ref-fuller1975regression)) developed a variance estimator for regression model parameter estimators based on Taylor linearization with unequal weighting of observations under stratified and two-stage sampling designs. Similarly, Binder ([1983](#ref-binder1983variances)) derived sampling distributions of estimators for regression parameters in finite populations, along with variance estimators for complex samples. Skinner, Holt, and Smith ([1989](#ref-skinner1989analysis)) studied the properties of variance estimators for regression coefficients under complex sample designs. Fuller ([2002](#ref-fuller2002regression)) summarized estimation methods for regression models in complex samples, and Pfeffermann ([2011](#ref-pfeffermann2011modelling)) discussed various approaches to fitting linear regression models to complex survey data.

##### 9.6.3.2. Estimation of parameters

134. Traditional statistical methods assume that sample data are independently and identically distributed and follow a specific probability distribution (e.g., Binomial, Poisson, Exponential, Normal, etc.). However, these assumptions are often violated in complex survey designs, which may involve clustering, stratification, and unequal selection probabilities. When fitting regression models with complex survey data, conventional estimators (e.g., maximum likelihood or least squares) can produce biased estimates.

135. For illustration, consider the estimation of the slope ) in a simple linear regression model. From the estimating equations method, this estimator is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-40) |

Note that the essential difference between this estimator and the traditional one is the inclusion of sampling weights. Extending these methods to multiple regression is algebraically complex and lies beyond the scope of this chapter; see Heeringa, West, and Berglund (2017) for a detailed discussion.

136. 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. As a generalization, according to Kish and Frankel ([1974](#ref-kish1974inference)), the variance estimation of coefficients in a multiple linear regression model requires weighted totals for the squares and cross-products of all combinations of and .

##### 9.6.3. Estimation and prediction

137. After confirming that the model fits well and satisfies its assumptions, the next step is to assess whether the independent variables significantly explain 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.

138. Given the distributional properties of the regression coefficient estimators, a natural test statistic for evaluating the significance of these parameters is based on the Student’s-t distribution as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (9-41) |

where represents the degrees of freedom (number of PSUs minus number of strata) and is the number of predictor variables in the fitted model.

139. This test statistic evaluates the hypotheses versus the alternative . Similarly, a confidence interval of for can be constructed, as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (9-42) |

140. According to Neter, Wasserman, and Kutner ([1996](#ref-neter1996applied)), linear regression models are essentially used for two purposes. One is to explain the variable of interest in terms of predictors, and the other 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 section, and the second is achieved as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (9-43) |

This equation gives the predicted value of given the observed explanatory variables. These predictions can be used for inference and practical purposes, such as imputing missing data, forecasting, or generating synthetic populations.

#### BOX – ESS4: a model to describe per capita expenditures in relation with other covariates

To explore the factors shaping household spending, a linear model was applied to the logarithm of per capita expenditure using the ESS4 data. A log transformation is useful because expenditure data is often skewed, with some households spending significantly more than others, stretching the distribution and obscuring meaningful patterns. By transforming the response variable, the data becomes more balanced, allowing for clearer and more interpretable relationships between predictors and spending. Moreover, this approach enables the results to be framed in relative terms, highlighting how different factors influence spending in percentage changes rather than absolute amounts — an essential perspective in economic analysis. The fitted model is as follows:

According to the available data, this model includes some demographic and religious variables that may influence per capita expenditure. The urban variable captures differences in expenditure between urban and rural residents, where urban areas typically exhibit higher costs of living and economic opportunities. Religion is included to analyze potential cultural or socioeconomic differences in spending behavior among Catholics, Protestants, Muslims, and individuals of other religions. Gender is an important factor, as expenditure patterns often differ between males and females due to labor market participation and household roles. Age groups are also considered since expenditure may vary across different life stages, with older individuals potentially spending less due to retirement. Interaction terms between urban residence and religion assess whether the effect of religion on expenditure differs in urban areas. The estimated coefficients and corresponding p values are found in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Coefficient estimate** | **Std. Error** | **t value** | **p value** |
| (Intercept) | 6.75 | 0.06 | 105.88 | 0.00 |
| Urban | 0.62 | 0.09 | 6.92 | 0.00 |
| Catholic | -0.84 | 0.34 | -2.48 | 0.01 |
| Protestant | -0.38 | 0.10 | -3.93 | 0.00 |
| Muslim | -0.53 | 0.09 | -6.11 | 0.00 |
| Other religion | -0.28 | 0.11 | -2.57 | 0.01 |
| Female | -0.06 | 0.01 | -4.28 | 0.00 |
| 31 to 45 years | -0.08 | 0.03 | -3.16 | 0.00 |
| 46 to 65 years | -0.09 | 0.03 | -2.89 | 0.00 |
| 66 or more years | -0.17 | 0.06 | -2.82 | 0.00 |
| Urban: Catholic | 0.90 | 0.39 | 2.29 | 0.02 |
| Urban: Protestant | 0.25 | 0.13 | 1.90 | 0.06 |
| Urban: Muslim | 0.17 | 0.13 | 1.32 | 0.19 |
| Urban: Other Religion | -0.66 | 0.29 | -2.29 | 0.02 |

#### BOX – ESS4: a model to describe per capita expenditures in relation with other covariates

The model estimates the logarithm of per capita expenditure as a function of urban residence, religion, gender, age group, and interactions between urban residence and religious affiliation. The coefficients indicate the expected percentage change in expenditure for each characteristic, holding other factors constant. From the table, it can be stated that:

* Living in an urban area is associated with 62% higher expenditure (β1=0.62, p<0.01), conditional on the values of other covariates being held constant. This conditional interpretation applies similarly to the remaining results presented below.
* Compared to the reference group (Orthodox), all religious groups exhibit lower expenditures. The strongest negative effect is for Catholics (β2=−0.84, p=0.01) and Muslims (β4=−0.53 p<0.01).
* Women (β6=−0.06, p<0.01) and elder people (β9=−0.17, p<0.01) tend to have lower expenditures.
* Urban Catholics have 90% higher expenditure than their rural Catholic counterparts (β11=0.90, p=0.02).
* Urban residents identifying as "Other Religion" show lower expenditure compared to rural counterparts (β14=−0.66, p=0.02).
* The interaction terms for Protestants and Muslims are not statistically significant (p>0.05), suggesting similar effects in both urban and rural areas.

##### 9.6.3.3. Working with weights

141. When analysing data from complex surveys, a critical question arises: how should survey weights be used in regression models? To address the challenges of using raw sampling weights, several adjustments have been proposed to balance accuracy and efficiency. Some include:

[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 weight variability, particularly useful in large samples with excessively variable raw weights:

|  |  |  |
| --- | --- | --- |
|  |  | (9-44) |

[2] Normalized Weights: These rescale raw weights to sum to one, preventing unnecessary variance inflation. This is useful when comparing models with different subsets of data or when variance inflation is a concern.

|  |  |  |
| --- | --- | --- |
|  |  | (9-45) |

142. In these approaches, weights are defined as simple multiplicative functions of the standard raw weights. As such, they cannot be used to estimate totals, nor do they alter the coefficient of variation of the weights. Moreover, they do not affect estimates of ratios, such as proportions or means. These re-scaling options are avoidable if one uses the appropriate software such as those mentioned in Subsection 9.3.4.

#### 9.6.4. Model diagnostics

143. When using statistical models with household survey data, assessing model adequacy is essential to ensure valid conclusions. Common issues in traditional regression models (such as multicollinearity, or influential observations) also arise in the case of complex survey data. To address these challenges, we recommend conducting diagnostic checks to evaluate model assumptions and performance. These checks help confirm whether the model adequately represents the data and whether its results are reliable. This involves examining several key 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.

##### 9.6.4.1. Coefficient of determination

144. The coefficient of determination, also known as , is a common measure of goodness-of-fit in regression models. It 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 is given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9-46) |

where is the weighted sum of squared errors given by

|  |  |  |
| --- | --- | --- |
|  |  | (9-47) |

and is the total weighted sum of squares given by

|  |  |  |
| --- | --- | --- |
|  |  | (9-48) |

##### 9.6.4.2. Standardized residuals

145. Residuals are the differences between observed and predicted values. Analysing residuals is critical for diagnosing whether the model violates key assumptions. Residuals should show no specific pattern when plotted against predicted values or independent variables; otherwise, it may suggest non-constant variance (heteroscedasticity) or a non-linear relationship.

146. Graphical analysis is often used to detect issues, with residuals plotted against predicted values serving as a common diagnostic tool. A careful examination of the residuals can help the researcher determine if the fitting process has met the assumptions or if one or more assumptions have been violated, requiring a review of the model specification or the fitting procedure. These residuals are defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (9-49) |

where is the predicted value of under the fitted model, and is the survey weight for unit in the complex sample dataset, and is the variance function of the outcome. These residuals are used to assess normality and constant variance.

147. If the assumption of constant variance is not met, the estimators remain unbiased and consistent, but they lose efficiency (they no longer have the smallest variance among all unbiased estimators). One way to analyse the assumption of constant variance in the errors is through graphical analysis, by plotting the model residuals against or against . If these plots reveal any pattern other than a random scatter, it suggests the error variance is not constant.

##### 9.6.4.3. Influential observations

148. Another set of techniques used for model diagnostic analysis involves examining influential observations. Certain data points can have a disproportionately large impact on the model fit. These influential points may not be outliers but could still significantly affect model parameter estimates. An observation is deemed influential if its removal from the data set causes a substantial change in the model fit. Importantly, 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. Common techniques include:

1. Cook’s Distance: Measures the impact of removing a data point on the overall model fit.
2. Statistic: Assesses the effect of removing a data point on individual regression coefficients, measuring the change in the estimated coefficient vector.
3. Statistic: Evaluates the influence of a data point on the overall model fit by measuring the change in the model fit when a particular observation is removed.

These methods are described in detail in traditional regression textbooks but must be adapted for use with complex surveys. Specialized survey software can be used to implement these necessary adjustments. The R package *svydiags* provides most of the tools needed for diagnostics – see Valliant (2024).

### 9.7. Tables

149. Tables are a fundamental tool for disseminating household survey estimates. They 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. Therefore, we outline some core principles for preparing tables with survey results.

150. It is important to distinguish three main types of tables that can be used for presenting the results of a survey:

* **Presentation Tables**: 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**: Provide comprehensive details for users requiring deeper analysis; they are typically larger, covering a wide range of variables and subgroups or domains.
* **Long Tables**: Structured for use in databases or data systems; they contain raw or minimally processed data, organized for further analysis or integration with other datasets.

151. Regardless of the type of table, certain principles should guide their design to ensure they are effective and user-friendly. According to Miller ([2004](#ref-miller2004chicago)), two fundamental principles should be considered:

* **Principle 1**. Make it easy for the reader to find and understand the numbers presented in the 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 clearly, helping to focus attention on the results, using logical and intuitive ordering of rows and columns, grouping related variables or categories together and minimizing clutter by avoiding unnecessary lines, colours, or decorations.

#### 9.7.1. Presentation tables

152. The primary goal of *presentation tables* is to communicate key results clearly and effectively. They organize data to emphasize significant patterns, trends, or insights revealed by the survey, supporting the accompanying text. These tables help readers quickly grasp the main findings without being overwhelmed by excessive detail. They are generally small, highlighting key survey results for press releases, executive summaries, scientific articles, reports, or webpages showcasing survey outputs. They are not expected to provide all results on a topic but focus on key results that capture the reader’s attention to the most important insights the data have produced.

153. Presentation tables should sort rows (and columns, if applicable) in a way that helps the reader perceive patterns, such as high or low values in the estimates. These tables prioritize readability over detail. Numbers should not exceed 3 or 4 digits. For population counts, use thousands; for percentages, use no more than one decimal place, or none, if the precision of the estimates does not warrant providing decimals (e.g. margins of error larger than 1%).

154. The following Box shows an example of a presentation table that highlights key findings on education levels for fathers and mothers of individuals under 18. The accompanying text provides context, emphasizing the fact that mothers are far more likely to have no education attainment than fathers.

155. The table in the Box concisely displays the percentage of fathers and mothers in four educational attainment categories (No education, Primary, Secondary, Above Secondary). It shows how a well-designed presentation table can summarize key findings and complement textual analysis. By organizing data clearly and emphasizing critical patterns, these tables enhance the readability and impact of survey results.

#### BOX – ESS4: example of presentation table and corresponding text

The following table summarizes the educational attainment of biological mothers and fathers of individuals under 18. It reveals significant disparities between fathers and mothers, highlighting a gender gap in access to education. A small fraction of parents has completed secondary or higher education. These low education levels, especially among mothers, may contribute to intergenerational cycles of limited educational attainment. Addressing these disparities through adult education programs, particularly for women, could enhance educational opportunities for future generations.

TABLE: Parental Education Levels for Children Under 18 (%).

|  |  |  |
| --- | --- | --- |
| **Education** | **Father** | **Mother** |
| No education | 47.4 | 69.7 |
| Primary | 40.3 | 23.6 |
| Secondary | 7.5 | 4.3 |
| Above secondary | 4.7 | 2.2 |

Note: The classification of primary, secondary, and above secondary education levels was established based on the corresponding ISCED (International Standard Classification of Education) categories across educational cycles in Ethiopia.

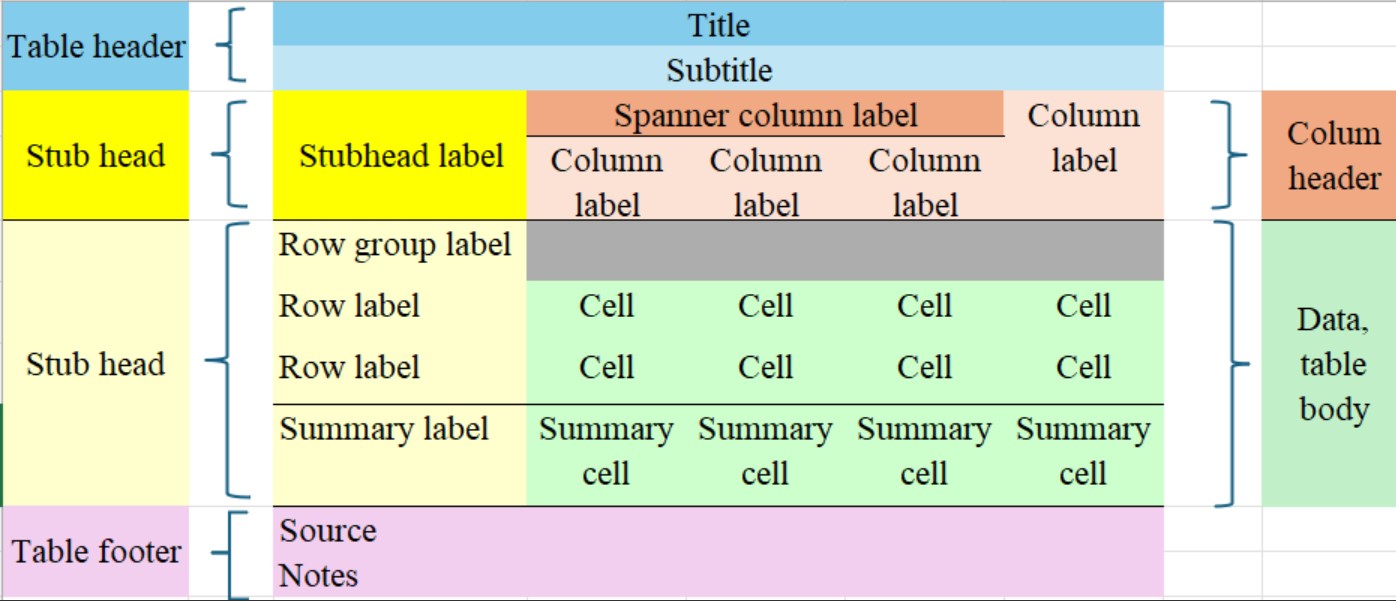
#### 9.7.2. Reference tables

156. Reference tables are longer tables designed to present comprehensive results from statistical studies. These tables are typically included in reports to provide detailed information but 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, which are often intended to be read by machines rather than directly by people, should be disseminated as database-like tables, accessible via downloads or interactive websites, whereas presentation and reference tables are designed for straightforward interpretation in a visual review.

157. Reference tables usually use core classification, domain definition or explanatory variables to define the rows, while outcome classification or output variables define the columns. In both cases, sorting should help readers locate the data they need, using alphabetic sorting or well-known classifications. In many cases, these tables have been replaced by interactive databases, allowing users to obtain the tables they want from a website.

158. All tables (regardless of type) should be self-sustaining, meaning they include all necessary metadata, to retain meaning even when removed from their original context. A self-sustaining table presents all relevant information so that its data can be interpreted without consulting the main text. Such tables are easier to understand, whether viewed within the original report or separately. Figure 9.3 presents the essential components of a table.

**Figure 9.3.** Anatomy of a table.



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

* Title (and optional subtitle): A table’s title is mandatory, and must give a clear, precise indication of the data presented, answering the questions about what, where and when regarding the data to be presented inside the table. It must be concise and avoid using verbs.
* Column headers: They should label the data in each column, providing relevant metadata (e.g., unit of measurement, time period, geographical area).
* Row headers and stub: They are provided as the first column in the table, identifying the data that is displayed in each row of the table.
* Source: It must always be provided at the bottom, including the responsible organization and the name of the survey or study that produced the results. Omitting the citation of the source prevents readers from finding more information about the data presented.
* Notes (optional): They provide additional context to understand and use the table correctly. Avoid using long texts, which if needed, would be better placed in a separate document. If multiples notes are needed, number them sequentially and use 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 the table to avoid confusion.
* Data: It is the table’s core content and the most important piece of information. Present it for easy extraction of relevant information. For clarity, consider whether rows or columns prioritize the message, as this determines portrait or landscape orientation. Visual cues (dividing or dotted lines, shading, spacing) can guide the reader in the proper direction.

160. When designing tables to present statistical data, clarity and consistency are crucial. Maintain uniform spacing across columns to enhance readability and avoid unnecessary text or excessive width. Time series data should always be organized chronologically, preferably ascending for reference tables, to ensure a logical progression. Categorize data using standard classifications to facilitate understanding and comparison across different datasets.

161. Arrange rows and columns logically, aligning data to the right for decimal point consistency. Decimal places should be limited to what is necessary for precision, rounding to 3–4 significant digits to simplify the data while preserving its integrity. Avoid blank cells; instead, use symbols for missing or “not applicable” values to keep the table informative and complete.

162. These practices improve table usability, helping readers to analyse and interpret the data efficiently. By adhering to these guidelines, analysts create professional and accessible presentations that effectively communicate statistical insights. These recommendations also apply to longer tables provided as databases, that can have additional resources if embedded on websites. For example, there may be support for users to sort tables by column values, which would be useful to navigate the table contents in a more efficient way.

#### 9.7.3. Dissemination of estimates

163. National Statistical Offices routinely produce descriptive statistics, such as totals, averages, proportions, and ratios from 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 various dissemination channels, including:

* Public Reports: Comprehensive summaries of key household survey findings.
* 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

164. These efforts aim to make the data understandable and actionable for policymakers, researchers, and the public. When publishing results tables, NSOs strive for clarity and usability, organizing them to highlight trends, comparisons, and key variable distributions. Common features include:

* Aggregated Data: Grouped by domain-defining variables (e.g., age, sex, region, or socioeconomic status).
* Uncertainty Measures: Including standard errors, coefficient of variation or confidence intervals to contextualize estimates.
* Metadata: Detailed explanations of the data collection methods, concepts, definitions, and limitations.

165. By presenting data in a user-friendly format, NSOs ensure accessibility for diverse audiences. Chapter 10 provides more detailed discussion on presentation of survey findings. However, 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 large standard errors and variances, or are otherwise unreliable. NSOs apply established criteria to determine when suppression is necessary, ensuring data credibility. To address this, the following approaches can be used:

1. Quality thresholds: Predefined thresholds for sample sizes and/or quality measures such as the coefficient of variation or standard errors.
2. Flagging and suppression: Estimates with precision below thresholds are either flagged with reliability warnings or omitted entirely from published tables.
3. Transparency: Clear documentation explains suppressed estimates, maintaining transparency, and trust.

### 9.8. Data visualization

166. In this section, we discuss how to present data and estimates 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 be used to show distributions and associations from the raw (unweighted) sample data, they can be misleading for the corresponding population distributions and associations. Therefore, modified plots accounting for survey weights should be used instead.

167. For example, a bar chart showing income distribution should incorporate weights to properly represent the estimated distribution for the entire population. Similarly, scatter plots exploring associations between variables should use weighted markers or density adjustments to ensure accurate depictions. In addition, when displaying survey estimates, which are subject to sampling error, it is important to convey this message by presenting both point estimates along with standard errors or confidence intervals.

168. Incorporating design features into visualizations helps ensure that viewers understand the inherent uncertainty in survey estimates, fostering more informed interpretations. When the survey units have different sampling weights, these must be accounted for in graphs. The main reason is that weights can be seen as the number of population units each sample unit represents.

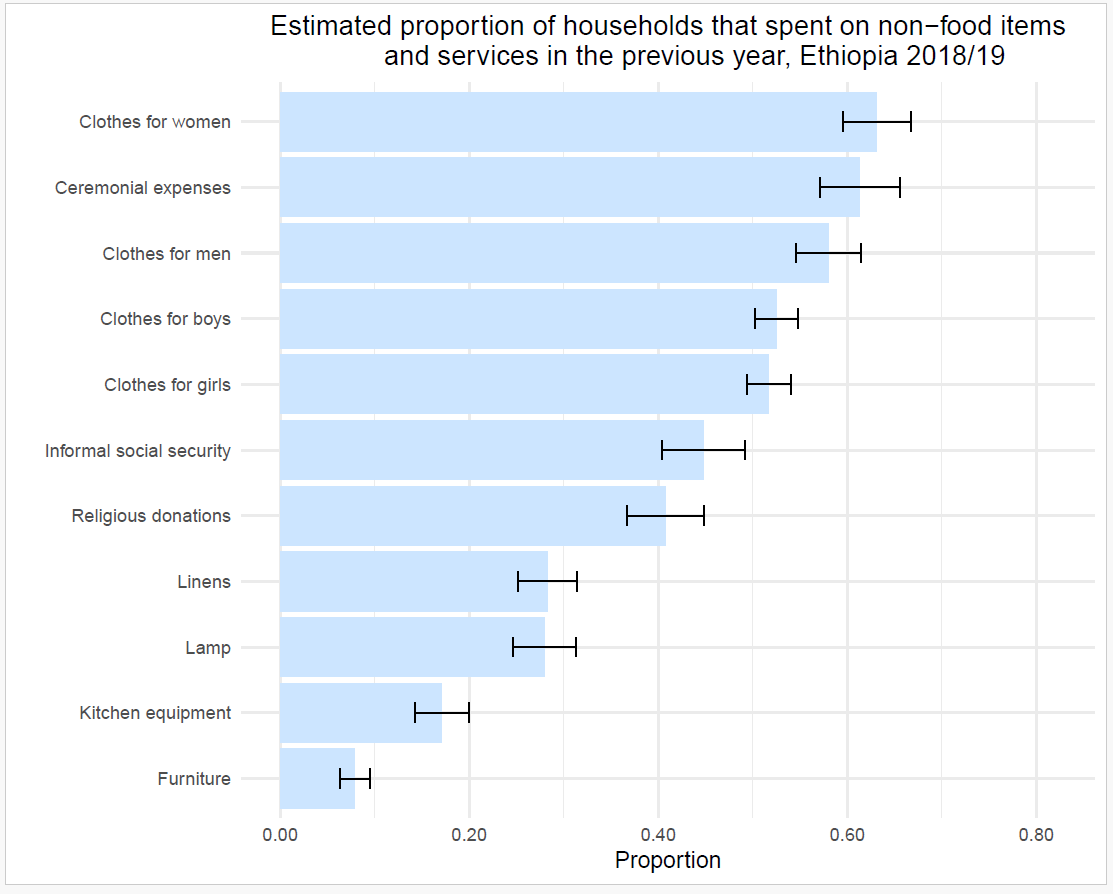
169. If graphs are created without considering weights, the visual representation will reflect the sample rather than the population. This discrepancy can distort distributions, proportions, or variable relationships. Incorporating weights ensures that the graphs provide an accurate representation of the population.

#### 9.8.1. Bar charts

170. Bar charts are commonly used to visualize categorical data from contingency tables, summarizing weighted counts or proportions, and ensuring the results reflect population-level characteristics rather than just sample data. Error lines should be overlaid on bars to indicate confidence interval widths, conveying the uncertainty of point estimates.

171. As an example, the bar chart in Figure 9.4 compares the estimated proportion of households that reported spending money on non-food items, with error lines indicating the 95% confidence intervals for each estimate. Furniture and kitchen equipment had the lowest levels of consumption, while women’s clothes and ceremonial expenses were the most important items in terms of expenditure.

**Figure 9.4.** Proportion (bars) of households reporting some expense in non-food items and related 95% confidence interval (error lines).



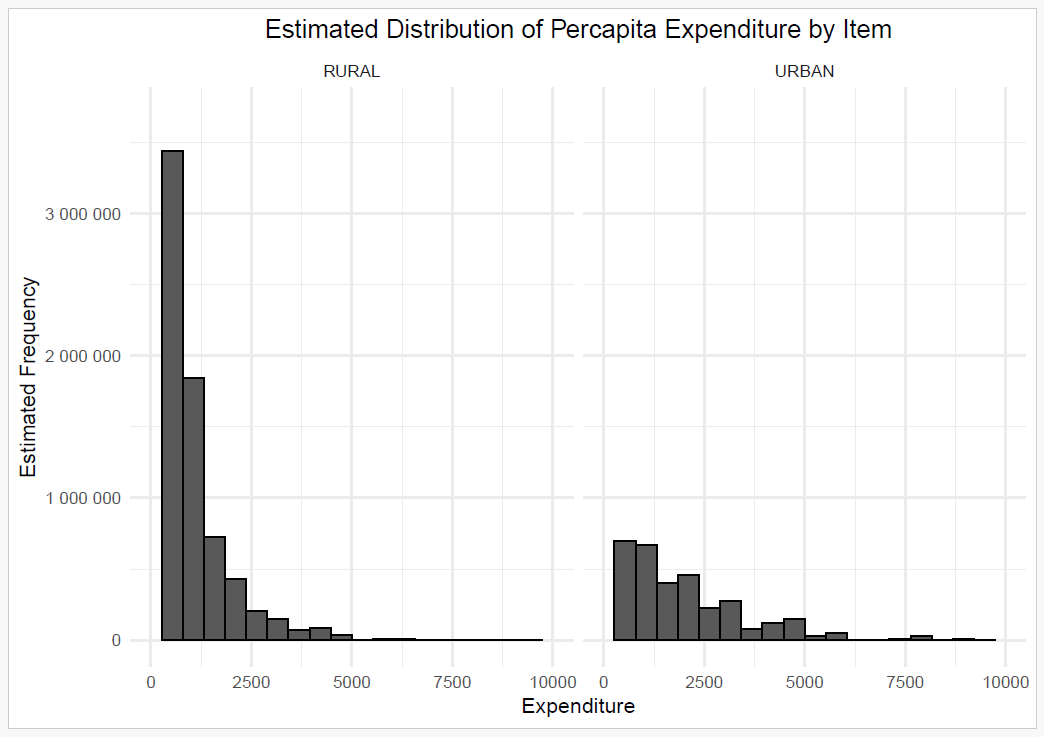
#### 9.8.2. Histograms

172. Histograms show how the values of a single continuous numeric survey variable are distributed. The horizontal axis is divided into intervals (bins), with each bar representing the number of observations that fall within that range. In a histogram, the area of each bar corresponds to the number or proportion of observations in the bin. The height is obtained by dividing this value by the width of the interval. This differs from a bar chart, where the height alone represents the value and the bar widths have no analytical meaning. When displaying sample data, the sampling weights must be incorporated to show frequencies or relative frequencies of population units within histogram bins. A common application is the use of histograms to represent the income or spending distributions. These visualizations allow researchers to observe the expanded population’s distribution including its shape, dispersion, and trends.

173. Histograms can be used to compare subgroups, such as geographic areas (urban and rural) or demographic characteristics like sex (male and female). This helps identify key differences, for instance, by examining expenditure distribution between rural and urban areas. Such comparisons help visualize potential disparities between subgroups, making results more accessible, by giving a visual representation for audiences unfamiliar with technical details of estimation methods. Combined with smoothed density estimates, histograms provide a more comprehensive and accurate population-level view.

174. As an example, Figure 9.5 compares histograms of clothing expenditure in rural (left) an urban (right) areas. Both are highly right-skewed, indicating most people have low expenditures, while a few have much higher ones. The rural distribution is concentrated at very low values, with a sharp decline as expenditure increases. The urban distribution, however, shows greater spread, reflecting higher expenditures on average.

**Figure 9.5.** Estimated distribution of per capita expenditures on clothes in rural (left) and urban (right) areas.



#### 9.8.3. Scatter plots

175. Scatter plots are the tool of choice to explore relationships between two continuous variables, potentially revealing patterns or trends in the data. It is necessary to 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’s sampling weight. Plots like these can be obtained using standard survey software. As stated by Lumley ([2010](#ref-Lumley2010)), when dealing with large datasets, displaying all the data points in a scatter plot can be overwhelming and cluttered, and several strategies can help address this issue:

1. Subsampling: Select a smaller, manageable random subsample from the full dataset. The subsample should be selected with probabilities proportional to the sampling weights, ensuring that it behaves approximately like a simple random sample from the population. This approach maintains representativeness while improving interpretability.
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.
3. Smoothed Scatter Plots: Avoid plotting individual points altogether and instead estimate and display trends. For example, calculate specific quantiles 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.

176. 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.

177. Figure 9.6 illustrates the weighted relationship between income and expenditure on a log scale. In this plot, the size of the points represents the final sampling weight assigned to each observation. A linear trend is observed, suggesting a strong association between income and expenditure. There is an upward trend, indicating that income and expenditure tend to increase together, although for some few households, the dispersion of points reveals that income is not always associated with proportionally higher expenditure. 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.

**Figure 9.6.** Weighted scatterplot between income and expenditure.

A graph of a graph showing a number of points

AI-generated content may be incorrect.

### 9.9. Challenges ahead

178. Throughout this chapter we have emphasized the importance of including the sampling design in the analysis stage, to obtain valid, unbiased and reliable estimates of population parameters. This approach, known as design-based inference, has dominated the way to make inferences from a sample to the population. Under this paradigm, valid inferences can be drawn without relying on model assumptions, which explains its widespread use by NSOs for analysing household survey data and producing official estimates. However, it generally demands large sample sizes to ensure reliable results.

179. However, as the demand for more detailed, frequent and diverse estimates grows, the design-based approach reveals certain limitations that prevent it from addressing emerging challenges. In such cases, the inferential paradigm is shifting to allow for new perspectives and approaches that offer appropriate technical solutions to these emerging issues. In this context, NSOs and data analysts should prepare to incorporate new paradigms into the production and analysis of household survey data. Next, we mention some of the challenges that should be considered in the following years and indicate the literature being developed on how to address them.

#### 9.9.1. Model-based inference

180. Under the model-based approach, population values are considered to be generated from a random variable following an underlying probability distribution or statistical model. In this approach, the sampling design plays a secondary role, as the validity of the analyses depends on proper model specification. Its flexibility allows for prediction of units outside the sample, which is useful when dealing with complex problems on small sample sizes or rare populations. Under this approach, it is assumed that the random variables are governed by a working model for all . For instance, one can assume a general linear model, such that , with and , where and denote the expectation and variance under the model , respectively.

181. Denoting by the sample selected from the population, and , its complement, the population total can be expressed as a random variable that can be decomposed as: . Under the model-based approach, the Best Linear Unbiased Predictor (BLUP) for this parameter is given by . Valliant, et. al. (2018) argue that the resulting estimators are unbiased under the model used to construct them, but can be biased if the model is misspecified or if the model that fits the sample is different from the one that describes the population as a whole.

182. In some cases, this approach is the only viable option for survey data analysis; for example, if the sample lacks a probability sampling design, or if the sample is too small for reliable design-based inferences on certain subpopulations or targets a hard-to-reach population. Under this model-based framework, researchers may also choose between Bayesian or Frequentist approaches; the details and theory of model-based inference are widely covered in Valliant et al. (2000) and Brewer (2002).

#### 9.9.2. Inference with non-probabilistic sample surveys

183. The increasing adoption of nonprobability methods for sampling reflects a paradigm shift in contemporary survey research, motivated by several factors. First, deteriorating response rates in traditional surveys have raised concerns about selection bias and representativeness even when the initial sampling was carried out using probability sampling methods. Second, the substantial financial costs and respondent burden associated with probabilistic sampling designs have rendered them less feasible for many research applications. Third, a growing demand for timely statistical indicators derived from nonprobability data sources (web panels, administrative records, and social media analytics).

184. While the proliferation of alternative data sources enables researchers to obtain larger samples at reduced costs, the cost-effectiveness of nonprobability data collection does not ensure their analytical utility. Researchers must rigorously evaluate methodological appropriateness before employing such datasets for statistical inference. Valliant (2020) discusses three approaches to estimation from nonprobability samples, namely: *superpopulation modelling*, *quasi-randomization*, and *doubly robust estimation*. The first one was explained in the previous paragraphs, and the remaining ones are described next.

185. The *superpopulation approach* is derived from the design-based framework, where a model estimated from the sample and is used to project it to the population. Notice that, after some algebra, the predictor in (9-50) can be written as , which is a weighted sum of sample values, with model-based weights , where , , and is the matrix of observed covariate values in the sample. Since can be accessed through external population control data, this approach can be used when is not known for the entire population, but control totals are available from an auxiliary data source.

186. The *quasi-randomization approach* treats the nonprobability sample as if it were obtained via an unknown probability mechanism. Under this approach, units in are assumed to be missing at random (MAR) from the sample, and inclusion probabilities can be estimated using external data () that covers all units in the population. Given these pseudo-inclusion probability estimates for all , the estimator of the population total will be given by , with quasi-randomization weights defined as . If covariate values cannot be obtained for the entire population, a reference probabilistic sample can be used to estimate pseudo-weights. However, estimates under this approach are sensitive to model misspecification.

187. The *doubly robust estimation approach* combines *quasi-randomization* and *superpopulation modelling*. Under this approach, estimators are constructed to be approximately unbiased and consistent if either the pseudo-inclusion model or the superpopulation model are correctly specified. First, pseudo-inclusion probabilities are estimated; then, a working model is specified for the variable of interest. This allows quasi-randomization weights to be used in computing model-based weights. Assuming a general linear model , the estimator of the population total takes the form , with model-based weights .

#### 9.9.3. Small Area Estimation

188. The Sustainable Development Goals (SDGs) highlight the importance of disaggregated data across different population groups—such as by sex, age, ethnicity, geographic area, among others. In this context, subnational disaggregation of indicators has gained increasing interest among NSOs. Small area estimation (SAE) methods have been proposed to address the disaggregation problem, aiming to produce precise and reliable estimates for domains with small samples by combining multiple data sources. These approaches achieve efficiency gains when auxiliary data and survey variables are sufficiently correlated (Rao & Molina, 2015).

189. When the analysis aims to produce disaggregated estimates for domains not considered in the survey’s sampling design, direct (design-based) methods often fail to provide reliable results. Direct estimators require sufficiently large samples to yield precise estimates; however, non-planned domains may be under sampled or even have zero sample size in the survey. In such cases, the limitations of direct estimators become evident, necessitating alternative (indirect) approaches to address these challenges. SAE models can be classified according to the level of aggregation of the auxiliary variables. The most well-known categories in the related literature are area-level (or aggregated-level) models and unit-level models that have been widely applied in poverty mapping, unemployment rate estimation, and for other socioeconomic indicators.

190. The Fay-Herriot model (Fay and Herriot, 1979) is a well-known area-level model that combines direct survey estimates with auxiliary information from related sources, such as administrative records, satellite imagery or census data, to improve the precision of estimates for small domains. This model assumes a linear relationship between the direct estimates and area-level covariates while accounting for random area-specific effects and sampling errors. This model relates a direct survey estimate for area to the parameter of interest through the following sampling model: , where . In addition, a linking model describes the relationship between and area-level covariates: , where . The Empirical BLUP for is a weighted average of the direct estimate and the regression-synthetic estimate: , with . By borrowing strength across areas through a random effects structure, the model produces more stable and reliable estimates, particularly for domains with small or null sample sizes.

191. Molina and Rao (2010) proposed a unit-level nested error model to derive the Empirical Best Linear Unbiased Predictor (EBLUP) for nonlinear poverty indicators. This model assumes that a transformed welfare variable for the -th household belonging to area 𝑑, given household covariates , can be written as , where the represents area-level random effects and denotes household-level errors. Let denote the population in area 𝑑, the sample of households in area 𝑑, and . The EBLUP for the total income in area *d* is given by , where are the predicted values for the out-of-sample households, and are estimators for the regression coefficients and random effects, respectively.

192. Estimating the Mean Squared Error (MSE) for small area predictors involves distinct approaches for unit-level and area-level models. For unit-level models, MSE estimation typically combines analytical approximations with resampling methods (parametric bootstrap is preferred for nonlinear indicators) simulating replicates to empirically quantify variability. For area-level models, MSE estimation often relies on Taylor linearization or jackknife resampling. In general, computationally intensive methods like MCMC are often required.

#### 9.9.4. Linked survey data

193. The attempt to answer many research questions requires linking data collected in one household sample survey with data from another source, be it another household survey, a census, a register or even a non-traditional data source. Such data or record linkage may have been planned before conducting the survey, in which case there might be good keys to link the survey records to those in the additional data source. But often the record linking may take place with surveys where the record linkage was not planned from the start (say when carried out by secondary analysts). As an example, consider the case when survey data on the vaccination status of children obtained from a large household sample survey need to be linked to mortality records for a period of say 12 – 24 months after the survey was conducted, to test the hypothesis that vaccination reduces child mortality.

194. Record linkage techniques (Fellegi & Sunter, 1969; Christen, 2012) are increasingly used in household surveys to enhance data quality and expand analytical scope. Applications of this technique include conducting post-enumeration surveys (PES) to evaluate census coverage, addressing survey nonresponse by imputing missing data, and enriching microdata for small area estimation. These approaches can produce richer statistics but require careful treatment of linkage error and survey design effects. Chambers & Silva (2020) review non-deterministic linkage methods and propose analytical frameworks to mitigate bias from imperfect matches.

195. Post-enumeration surveys (United Nations, 2010) are conducted after a census to assess its coverage and accuracy, aiming to identify undercounts, overcounts, and misclassifications. Achieving this requires linking PES records to the corresponding census records for the same individuals or households, which involves establishing a one-to-one match between the two sources. Deterministic or probabilistic record linkage methods are typically used, relying on identifiers such as names, dates of birth, sex, addresses, or other relevant variables. This example represents one of the earliest and most established applications of record linkage in official statistics, demonstrating the feasibility of integrating independent data sources and paving the way for broader applications in surveys, censuses, and administrative data integration.

196. Record linkage can also be an effective tool for addressing nonresponse in household surveys by providing auxiliary information to compensate for missing responses (Rothbaum & Bee, 2021). When survey participants fail to answer specific questions, their records can be linked to administrative data to recover relevant information. This approach allows researchers to impute missing values, adjust survey weights, or model nonresponse mechanisms more accurately, thereby reducing potential bias. For example, linking income or employment data from tax or social security registers to survey nonrespondents can help estimate their likely contributions without additional data collection. By integrating linked data, statistical offices can preserve the precision of survey estimates while minimizing the impact of nonresponse on key indicators.

197. In the case of small area estimation models, survey microdata can be enriched with auxiliary information from administrative sources, censuses, or other registers (Chambers, Fabrizi, & Salvati, 2019). By connecting survey responses to administrative or census records at the individual or household level, researchers can construct a consistent set of auxiliary variables that enhance the explanatory power of SAE models. These variables can serve as predictors for key outcomes, improving the precision and reliability of estimates for small geographic areas or subpopulations. For example, linking household surveys with tax, social security, or education records allows the creation of predictors that capture employment, income, or poverty-related characteristics at local levels.

### 9.10. References

Binder, David A. 1983. “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” *International Statistical Review* 51 (3): 279–92. <https://doi.org/10.2307/1402588>.

Binder, David. A. 2011. Estimating Model Parameters from a Complex Survey under a Model-Design Randomization Framework. *Pakistan Journal of Statistics*, *27*(4), 371–390.

Binder, David A., and Milojica S. Kovacevic. 1995. “Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach.” *Survey Methodology* 21 (2): 137–45.

Brewer, K. R. W. (2002). *Combined Survey Sampling Inference: Weighing Basu’s Elephants*. Arnold.

Bruch, C., Munnich, R., & Zins, S. (2011). *Variance Estimation for Complex Surveys*. Advanced Methodology for European Laeken Indicators - European Commission.

Chambers, R., Fabrizi, E., & Salvati, N. (2019). Small area estimation with linked data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(1), 78–98.

Chambers, R. L. & Silva, A. D. D. (2020). *Improved secondary analysis of linked data: a framework and an illustration*. Journal of the Royal Statistical Society, Series A, 183 (1): 37-59.

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.

Dean, Natalie, and Marcello Pagano. 2015. “Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys.” *Journal of Survey Statistics and Methodology* 3 (4): 484–503. <https://doi.org/10.1093/jssam/smv024>.

Efron, Bradley. 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*, 402–8.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data*. Journal of the American Statistical Association*, 74(366), 269–277.

Fellegi, Ivan 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): 261–68.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Freedman, D.A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press, New York.

Fuller, Wayne A. 1975. “Regression Analysis for Sample Survey.” *Sankyha, Series C* 37: 117–32.

———. 2002. “Regression Estimation for Survey Samples (with Discussion).” *Survey Methodology* 28 (1): 5–23.

Gutiérrez, Hugo Andrés. 2015. *TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population*. R package. <https://CRAN.R-project.org/package=TeachingSampling>.

Gutiérrez, Hugo Andrés. 2020. *Samplesize4surveys: Sample Size Calculations for Complex Surveys*. R package. <https://CRAN.R-project.org/package=samplesize4surveys>

Hansen, Morris H., William N. Hurwitz, and William G. Madow. 1953. *Sample Survey Methods and Theory*. Vol. 1 and 2. New York: John Wiley; Sons.

Heeringa, Steven G., Brady T. West, and Patricia A. Berglund. 2017. *Applied Survey Data Analysis*. Chapman and Hall CRC Statistics in the Social and Behavioral Sciences Series. CRC Press.

IBM. 2017. *IBM SPSS Complex Samples*. <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS_Complex_Samples.pdf>.

Jacob, Guilherme, Anthony Damico, and Djalma Pessoa. 2024. *Poverty and Inequality with Complex Survey Data*. <https://www.convey-r.org/>.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Kish, Leslie, and Martin R Frankel. 1974. “Inference from Complex Samples.” *Journal of the Royal Statistical Society, Series B* 36: 1–37.

Korn, E.G., and Graubard, B. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *American Statistician*, 49(3), 291-295.

Kovar, J. G., J. N. K. Rao, and C. F. J. Wu. 1988. “Bootstrap and Other Methods to Measure Errors in Survey Estimates.” *Canadian Journal of Statistics* 16 (Suppl.): 25–45.

Langel, Matti, and Yves Tillé. 2013. “Variance Estimation of the Gini Index: Revisiting a Result Several Times Published.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (2): 521–40. <https://doi.org/10.1111/j.1467-985X.2012.01048.x>.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using r*. *Wiley Series in Survey Methodology*. John Wiley; Sons.

———. 2024. “*survey: analysis of complex survey samples*.” (Version 4.4). R package. <https://cran.r-project.org/package=survey>

Miller, Jane E. 2004. *The Chicago Guide to Writing about Numbers*. Chicago: University of Chicago Press.

Molina, E. A. and Skinner, C. J. 1992. Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes, *Computational Statistics and Data Analysis,* Volume 13, Issue 4. Pages 395-405.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38(3), 369–385.

Neter, John, William Wasserman, and Michael H. Kutner. 1996. *Applied Linear Statistical Models*. McGraw-Hill.

Noland, D., and Speed, T. (2000). *Stat Labs: Mathematical Statistics through Applications.* Springer, New York.

Osier, Guillaume. 2009. “Variance Estimation for Complex Indicators of Poverty and Inequality.” *Journal of the European Survey Research Association* 3 (3): 167–95. <http://ojs.ub.uni-konstanz.de/srm/article/view/369>.

Park, Inho, Marianne Winglee, Jay Clark, Keith Rust, Andrea Sedlak, and David Morganstein. 2003. “Design Effects and Survey Planning.” *Proceedings of the 2003 Joint Statistical Meetings - Section on Survey Research Methods*, 8.

Pessoa, D., Damico, A., & Jacob, G. (2024). *convey: Estimation of indicators on social exclusion and poverty and its linearization, variance estimation* (Version 1.0.1) [R package]. <https://github.com/ajdamico/convey/>

Pfeffermann, Danny. 2011. “Modelling of Complex Survey Data: Why Model? Why Is It a Problem? How Can We Approach It?” *Survey Methodology* 37 (2): 115–36.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Rao, J. N. K., & Molina, I. (2015). *Small-Area Estimation*. 2nd edition. John Wiley & Sons, Ltd.

Rao, J. N. K., and A. J. Scott. 1984. “On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data.” *The Annals of Statistics* 12: 46–60.

Rao, J. N. K., C F J Wu, and K. Yue. 1992. “Some Recent Work on Resampling Methods for Complex Surveys.” *Survey Methodology* 18: 209–17.

Rothbaum, J., & Bee, A. (2021). Addressing nonresponse bias in household surveys using linked administrative data. American Economic Association.

Rust, Keith F., and Valerie Hsu. 2007. “Confidence Intervals for Statistics for Categorical Variables from Complex Samples.” In. <https://api.semanticscholar.org/CorpusID:195852485>.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SAS. 2010. *SAS/STAT 9.22 User’s Guide - Survey Sampling and Analysis Procedures*. <https://support.sas.com/documentation/cdl/en/statugsurveysamp/63778/PDF/default/statugsurveysamp.pdf>.

Shah, B. V., M. M. Holt, and R. F. Folsom. 1977. “Inference about Regression Models from Sample Survey Data.” *Bulletin of the International Statistical Institute* 41 (3): 43–57.

Skinner, Chris J, Daniell Holt, and Tom M F Smith. 1989. *Analysis of Complex Surveys*. New York: John Wiley; Sons.

Stata 2023. *Stata 18 documentation.* [*https://www.stata.com/features/documentation/*](https://www.stata.com/features/documentation/)*.*

Thomas, D. R., and J. N. K. Rao. 1987. “Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling.” *Journal of the American Statistical Association* 82: 630–36.

Tillé, Yves, and Alina Matei. 2016. *Sampling: Survey Sampling*. <https://CRAN.R-project.org/package=sampling>.

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

United Nations. 2008. *Designing Household Survey Samples: Practical Guidelines*. Studies in Methods / Department of Economic and Social Affairs, Statistics Division Series f. New York, NY: United Nations.

United Nations. 2010. *Post-enumeration surveys: Operational guidelines* (Series F, No. 98). United Nations. <https://unstats.un.org/unsd/publication/seriesf/seriesf_98e.pdf>

United Nations Economic Commission for Europe. (2014). *Generic Statistical Business Process Model* (ECE/CES/2014/1). United Nations. <https://unstats.un.org/unsd/nationalaccount/workshops/2015/gabon/bd/GSBPM-ENG.pdf>​[UNECE+8](https://statswiki.unece.org/display/GSBPM/GSBPM%2Bv5.1?utm_source=chatgpt.com)

Valliant, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231–263. <https://doi.org/10.1093/jssam/smz003>

Valliant, R. (2024). *svydiags: Regression Model Diagnostics for Survey Data* (Version 0.7). R package. <https://cran.r-project.org/web/packages/svydiags/index.html>.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-93632-1>

Valliant R., Dorfman A. H., Royall R. M. (2000). Finite Population Sampling and Inference: A Prediction Approach. John Wiley & Sons, Inc., New York.

Valliant, R., & Rust, K. F. (2010.). Degrees of Freedom Approximations and Rules-of-Thumb. *Journal of Ofﬁcial Statistics*, *26*(4), 585–602.

Westat. 2007. *WesVar 4.3. Users Guide.* <http://users.nber.org/~jroth/chap1.pdf>.

Wolter, K.M. (2007). *Introduction to Variance Estimation.* New York: Springer.