CHAPTER 9: ANALYSIS OF HOUSEHOLD SURVEY DATA

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

2024-09-11

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

[2]SCIENCE, pedronsilva@gmail.com

Contents

Abstract **9**

Introduction **11**

1 The golden pair: sample design and estimator **13**

2 Descriptive parameters **15**

3 Comparisons and association **17**

4 Regression: modelling survey data **19**

5 Data visualization **21**

6 Other modeling scenarios **23**

List of Figures

List of Tables

Abstract

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

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

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

Introduction

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

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

Chapter 1

The golden pair: sample design and estimator

Defining the survey design is crucial for analyzing complex survey data. We must make sure that PSU, strata, and weights are available in the database. When not available, the database should contain replicate weights, or the researcher should have a valid expression to compute variance estimates. Defining the sample design is crucial for design-based inference and analysis, ensuring accuracy, precision, and consistency. A well-defined survey design facilitates statistical analysis, supports effective data interpretation, and enables meaningful insights into complex phenomena.

1.1 Parameters and estimators

Based on the design-based approach this section will discuss the basic principles of inductive inference and how, using the sampling weights (from chapter 8), we can get unbiased estimators for population parameters, with a special focus on totals and means.

1.2 Uncertainty in household surveys

As the sample is a small subset of the population, we show the importance of generating not only point estimates, but also related uncertainty measures. In this subsection we will present some approaches for variance estimation: exact and approximate formulas (ultimate cluster) for variances of totals; and Taylor linearization for means, ratios and other parameters. We also introduce replication methods and generalized variance functions, which are essential in the cases when PSU or strata are missing from the database.

1.3 Using software to generate valid inferences

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

Chapter 2

Descriptive parameters

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

2.1 Frequencies

Considering survey weights assigned to each respondent helps provide estimates that are representative of the target population.

2.2 Means, proportions, and ratios

These measures of central tendency provide insights into the average or typical response within the data.

2.3 Percentiles and inequality measures

For continuous data, these measures divide the data into intervals, indicating the proportion of data points below a certain value, and are particularly useful for understanding distributions and identifying outliers. Also, the section will cover the estimation of the Gini inequality index.

2.4 NSO – Practical example

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

Chapter 3

Comparisons and association

Elaborate analyses of household survey data must be adjusted for the complex survey design to account for clustering, stratification, and weighting. This section will introduce the reader on the main methods currently used to compare subgroups and make conclusions based on a valid inferential context.

3.1 Cross-tabulations

We first examine the relationship between two or more variables by creating contingency tables, which reveal how responses vary across different categories.

3.2 Tests for group comparisons

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

3.3 Tests of Independence

To measure the level of association between categorical variables we present the Rao-Scott correction for the Pearson Chi-squared test of independence.

3.4 Correlation

To conclude on the degree of association between variables, we show the proper approach to include sampling weights and complex sampling design.

3.5 NSO – Practical example

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

Chapter 4

Regression: modelling survey data

Modelling survey data is a common task among researcher; some of them include the features of the sampling design in computing standard error of the estimated regression parameters. In this section we will deal with the problem of weighting in regression models and present a parsimonious solution.

4.1 To weight or not to weight?

We present the pros and cons of including the complex design features in the estimation of regression parameters and their associated standard errors. We present some adjustment to the sampling weights to fit these kind of models (senate sampling weights, normalized sampling weights, Pfeffermann model weights).

4.2 Some inferential approaches to modelling data

When modelling survey data, one should deal with two sources of variability: the one devoted to the complex sampling design and the one that comes from the very model. Combining these sources into a valid set up requires of some advanced methods. We will mention some of them: pseudo likelihood, combined inference.

4.3 Linear models

We present a primer of linear models and estimation of regression coefficients along with their standard errors. Also, this subsection will introduce basic tools to model diagnosis.

4.4 Logistic models

To model the probability of discrete variables, we apply the principles of design-based inference.

4.5 NSO – Practical example

In this subsection, we will share the experience of an NSO in applying models to household survey data, and the results they present in terms of significance of models and relations among variables.

Chapter 5

Data visualization

In this section we delve deeper on how to present the results of the analysis of household surveys using graphics. This part is important because household surveys estimates are subject to error and the researcher should face this challenge on presenting not only point estimates, but also standard errors.

5.1 Weighted Histograms

To visualize the distribution of continuous variables, adjusted for survey weights.

5.2 Bar Charts

To display the distribution of categorical variables with standard errors.

5.3 Box Plots

To show the distribution of continuous variables, including measures of central tendency, variability, and outliers, across different groups or strata.

5.4 Scatter Plots

To explore the relationship between two continuous variables, potentially revealing patterns or trends in survey data.

5.5 Maps

To display the behavior of the interest variable across geographical domains.

5.6 NSO – Practical example

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

Chapter 6

Other modeling scenarios

In this section we indicate the literature and software supporting the fitting of some other models to complex household survey data, including:

6.1 Multilevel models

6.2 Survival models

6.3 Loglinear models for contingency tables