IASSC Technical Reports

The Italian Lives Survey

Sample Design, Weighting, Variance Estimation, and Data Analysis

Maurizio Pisati

IASSC
INSTITUTE FOR ADVANCED STUDY OF SOCIAL CHANGE

IASSC Technical Reports No. 2
April 2023

The Italian Lives Survey

Sample Design, Weighting, Variance Estimation, and Data Analysis

Maurizio Pisati

IASSC
INSTITUTE FOR ADVANCED STUDY
OF SOCIAL CHANGE

The **Institute for Advanced Study of Social Change (IASSC)** is a permanent observatory on social change based in the Department of Sociology and Social Research of the University of Milan-Bicocca (Italy).

The **IASSC Technical Reports** series presents technical and scientific information from projects of the Institute for Advanced Study of Social Change.

Suggested citation:
Pisati, M. (2023) *The Italian Lives Survey: Sample Design, Weighting, Variance Estimation, and Data Analysis*. Milano: Institute for Advanced Study of Social Change.

Published by the Institute for Advanced Study of Social Change (IASSC), Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy, https://iassc.unimib.it/en.

Typeset in LaTeX using the suftesi class by Ivan Valbusa.

ISBN (e-book): 979-12-80999-00-9

CONTENTS

LIST OF FIGURES v

LIST OF TABLES vii

PREFACE ix

ABOUT THE AUTHOR xi

1 SAMPLE DESIGN 1

1.1 Introduction 1
1.2 Overview 2
1.3 Primary Sampling Stage 3
1.4 Secondary Sampling Stage 5
1.5 Tertiary Sampling Stage 7
1.6 Ultimate Sampling Stage 9
1.7 Sample Design Implementation 9
1.8 Sample Representativeness 15

2 SURVEY WEIGHTING 25

2.1 Introduction 25
2.2 Overview 26
2.3 Primary Sampling Stage 28
2.4 Secondary Sampling Stage 28
2.5 Tertiary Sampling Stage 30
2.6 Ultimate Sampling Stage 33
2.7 Weighted Sample Representativeness 36
2.8 Cohort Survival Rates 38

3 VARIANCE ESTIMATION 41

3.1 Introduction 41
3.2 Balanced Repeated Replication 46
3.3 Implementation of BRR 48
3.4 Quality of Variance Estimates 49

4 THE ANALYSIS OF ITA.LI DATA: A BRIEF GUIDE FOR STATA USERS 59

4.1 Introduction 59
4.2 Setting Up Data for Analysis 60
4.3 Univariate Analysis 74
4.4 Bivariate Analysis 89
4.5 Multiple Regression Analysis 105
4.6 Treatment Effect Estimation 129
4.7 Event History Analysis 137

REFERENCES 165

LIST OF FIGURES

1.1 Schematic representation of the ITA.LI sample design 2
1.2 Stratification plan for the primary sampling stage of the ITA.LI sample design 4
1.3 Spatial distribution of the PSUs selected for the ITA.LI sample design 5
1.4 Designated SSU sample size for each selected PSU of the ITA.LI sample design 8
1.5 Implementation of the ITA.LI sample design 11
1.6 Average number of screened residential addresses per responding household, by administrative region and degree of urbanization . . . 13
1.7 Household Response Rate 1/Within-household individual response rate, by administrative region and degree of urbanization 14
1.8 Household-level representativeness of the ITA.LI realized sample with respect to variable *Region of residence* 18
1.9 Household-level representativeness of the ITA.LI realized sample with respect to variable *Household size* 19
1.10 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Sex* 19
1.11 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Age group* 20
1.12 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Region of residence* 21
1.13 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Educational degree* 22
1.14 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Occupational status* 22
1.15 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Marital status* 23

1.16 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Citizenship* 23

2.1 Diagrammatic representation of the process of construction of the ITA.LI survey weights 27
2.2 Individual-level representativeness of the ITA.LI realized sample with respect to a select set of variable distributions 38
2.3 Cohort survival rates 39

LIST OF TABLES

2.1 Population joint frequency distribution of variables *Area of residence* and *Household size* for household weight calibration 32
2.2 Population marginal frequency distributions for individual weight calibration 35

3.1 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Number of valid cases used in the analysis, by quantity of interest and target subpopulation 52
3.2 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Estimation method effect, by quantity of interest and target subpopulation 53
3.3 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Design effect, by quantity of interest and target subpopulation 54
3.4 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Coefficient of variation, by quantity of interest and target subpopulation 55
3.5 Evaluation of the quality of ITA.LI variance estimates for the regression coefficients of a linear model: Estimation method effects and design effects 56

4.1 Stata data files comprising the public-use version of the ITA.LI database: Structure 60
4.2 Stata data files comprising the public-use version of the ITA.LI database: Contents 61
4.3 Approximate design degrees of freedom for select subpopulations defined by combinations of *Age group*, *Area of residence*, and *Sex* ... 86

4.4 Approximate design degrees of freedom for select subpopulations defined by combinations of *Age group* and *Area of residence* 87

PREFACE

In 2017, the Department of Sociology and Social Research of the University of Milano-Bicocca (hereafter, the Department) was included by the then Ministry of Education, University and Research (MIUR) in the list of the six "Departments of Excellence" operating in Italy in the field of political and social sciences.

As a result of this recognition, the Department received a five-year grant to create a research center dedicated to the study of social change. This center, called the *Institute for Advanced Study of Social Change* (IASSC, https://iassc.unimib.it/en), was established at the Department in 2018 with the first goal of launching a large-scale longitudinal quantitative/qualitative survey on the life courses of Italians. The first wave of this survey, called *Italian Lives* (ITA.LI), was conducted between the last quarter of 2019 and the first quarter of 2021. The second wave is currently underway.

The purpose of this volume is to illustrate some aspects of the *Italian Lives* survey that are particularly relevant for the proper analysis of the data collected during the survey itself. Specifically, the volume is divided into four chapters. The first illustrates the sample design of the first wave of *Italian Lives*, providing a detailed description of each sampling stage. The second chapter describes the procedure used to construct the survey weights and provides an assessment of the representativeness of the weighted realized sample of the survey. The third chapter presents the variance estimation method chosen for the survey and illustrates its implementation. Finally, the fourth chapter provides a brief guide to the analysis of the *Italian Lives* data using the Stata statistical software.

I would like to thank Frauke Kreuter (University of Maryland and Ludwig-Maximilians-Universität München) and Richard Valliant (University of Michigan and University of Maryland) for their advice on using the sample replicates approach. I am also very grateful to Ben Jann (University of Bern) for clar-

ifying some aspects of his excellent Stata command `dstat` and for making some changes to the command that improved its use for design-based estimation and inference. Finally, many thanks to Jeff Pitblado, Executive Director of Statistical Software at StataCorp LLC, who, in addition to answering my questions about survey data analysis in Stata, suggested the procedure for calculating the approximate number of degrees of freedom to be used in subpopulation analysis presented in Section 4.3.

ABOUT THE AUTHOR

Maurizio Pisati is a professor of Social Research Methods at the University of Milano-Bicocca (Italy).

1

SAMPLE DESIGN

1.1. *Introduction*

The basic purpose of *survey research* is to gather information about the units of a finite population of interest: voters of Massachusetts's 7th Congressional District in 2020; registered supporters of the UK's Labour Party on January 1, 2022; students enrolled in a university course in Paris in the 2021/22 academic year; people living in Italy on December 31, 2021; and so on. Since it is generally impractical or impossible – given existing time and resource constraints – to study the target population in its entirety, survey researchers typically limit themselves to examining a subset of it, called a *sample*.

The ideal requirement of any survey sample is that it be *representative* of the target population (Lohr 2022). This means that the conclusions drawn from the sample must be able to be generalized – within estimable margins of error – to the entire population of interest. However, there is no single, one-size-fits-all way to meet this requirement. That is, a sample that aims to be representative of the target population can be selected in many different ways. The way the sample of a given survey is selected from the target population – i.e., the set of procedures that are carried out to choose the units of the target population to be contacted for interview – is called the *sample design* of the survey (Groves *et al.* 2009).

In general, the choice of the sample design for a given survey depends on the survey goals, as well as on the availability of information about the target population. In any case, the sample design must be chosen to ensure that the information of interest can be collected in a manner appropriate to the research objectives (Shapiro 2008a).

The purpose of this chapter is to illustrate the sample design of the first wave of the *Italian Lives* survey (henceforth ITA.LI). The next section provides a general overview of the sample design. Four more sections follow, presenting

a detailed description of the key features of each sampling stage. The seventh section discusses the field implementation of the sample design. Finally, the last section offers an assessment of the representativeness of the realized sample prior to survey weighting adjustment.

1.2. *Overview*

The *target population* of ITA.LI is defined as all persons aged 16 and over residing in private households in Italy at the time of interview.

Given the available financial resources, the *target sample size* of the survey was set at approximately 5,000 households. With an estimated average number of two eligible persons per household (Istat 2019), this corresponds to around 10,000 targeted individuals.

To reach this target, ITA.LI used a four-stage area probability sample design with stratification at the first stage (Levy and Lemeshow 2008; Lohr 2022). Figure 1.1 provides a stylized illustration of the sampling process. First, a random sample of Italian municipalities was drawn, stratified by region, degree of urbanization, and population size. Second, within each selected municipality, a random sample of residential addresses was chosen. Third, a single household was picked at random from each selected address. Finally, all household members aged 16 and over were considered eligible for interview.

The choice of such complex design was dictated by two main reasons. On the one hand, since CAPI was adopted as the interview mode, limiting the fieldwork to a relatively small set of geographical clusters of households (the first-stage sample of Italian municipalities) increased the cost-efficiency of the design (Potter 2008). On the other hand, the constraints on available sampling frames – specifically, the impossibility of accessing exhaustive and updated

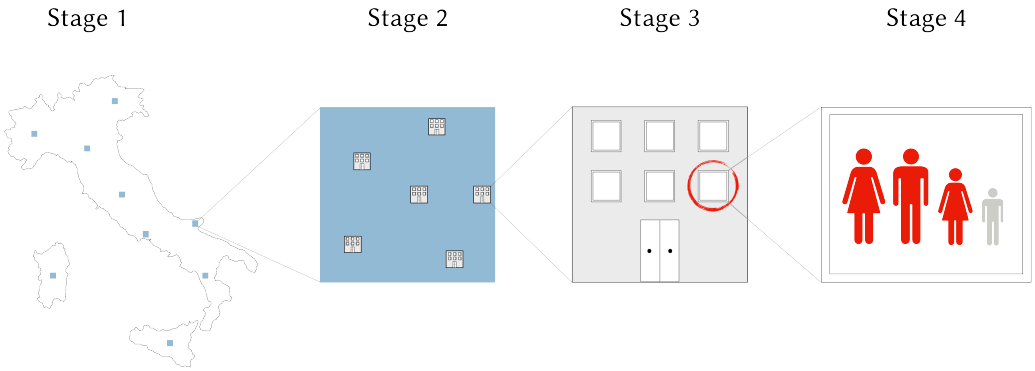


Figure 1.1 Schematic representation of the ITA.LI sample design.

lists of private households resident in Italy – left no other choice than going through lists of residential addresses (available at the municipality level) to reach households.

The following four sections provide a detailed description of the key features of each sampling stage.

1.3. *Primary Sampling Stage*

The *primary sampling units* (PSUs) of the ITA.LI sample design are the 8,000 or so municipalities into which Italy is divided.

Based on both the available financial resources and the target sample size (see Section 1.2), a sample of 280 PSUs was set for the study. To ensure widespread coverage of the national territory – and hence a good representation of the large socio-economic and cultural heterogeneity that characterizes Italy – prior to selection, PSUs were stratified following a two-step procedure.

First, PSUs were partitioned into 46 *preliminary strata* defined by a reduced combination of the 20 administrative regions that make up Italy and the three degrees or urbanization defined by Eurostat (2021).[1] Second, within each of these preliminary strata, PSUs were further stratified based on population size, resulting in a total of 150 strata.

Of these 150 strata, 20 contained only a single *self-representing* PSU, i.e., a PSU that is large enough to be selected with certainty and is assigned its own stratum (Groves *et al.* 2009; Hall 2008). The remaining 130 strata contained two or more *non-self-representing* PSUs; from each of these strata, two PSUs were selected without replacement and with probability proportional to size, using the number of private households registered to reside in the PSUs as the measure of size.

The probability of selection of PSUs is then defined as follows:

$$\pi_{l|h} = \begin{cases} 1, & \text{if } h \in \text{self-representing} \\ \dfrac{M_{lh}}{M_h} \times 2, & \text{if } h \in \text{non-self-representing} \end{cases} \tag{1.1}$$

where l indexes PSUs; h indexes strata; $\pi_{l|h}$ denotes the probability of selection of PSU l from stratum h; M_{lh} denotes the number of private households registered to reside in PSU l of stratum h; and M_h denotes the total number of private households registered to reside in stratum h.

1 Of the 60 possible combinations of administrative region and degree of urbanization, 25 were merged into 11 super-classes because they were either empty or too sparsely populated (see Figure 1.2 below).

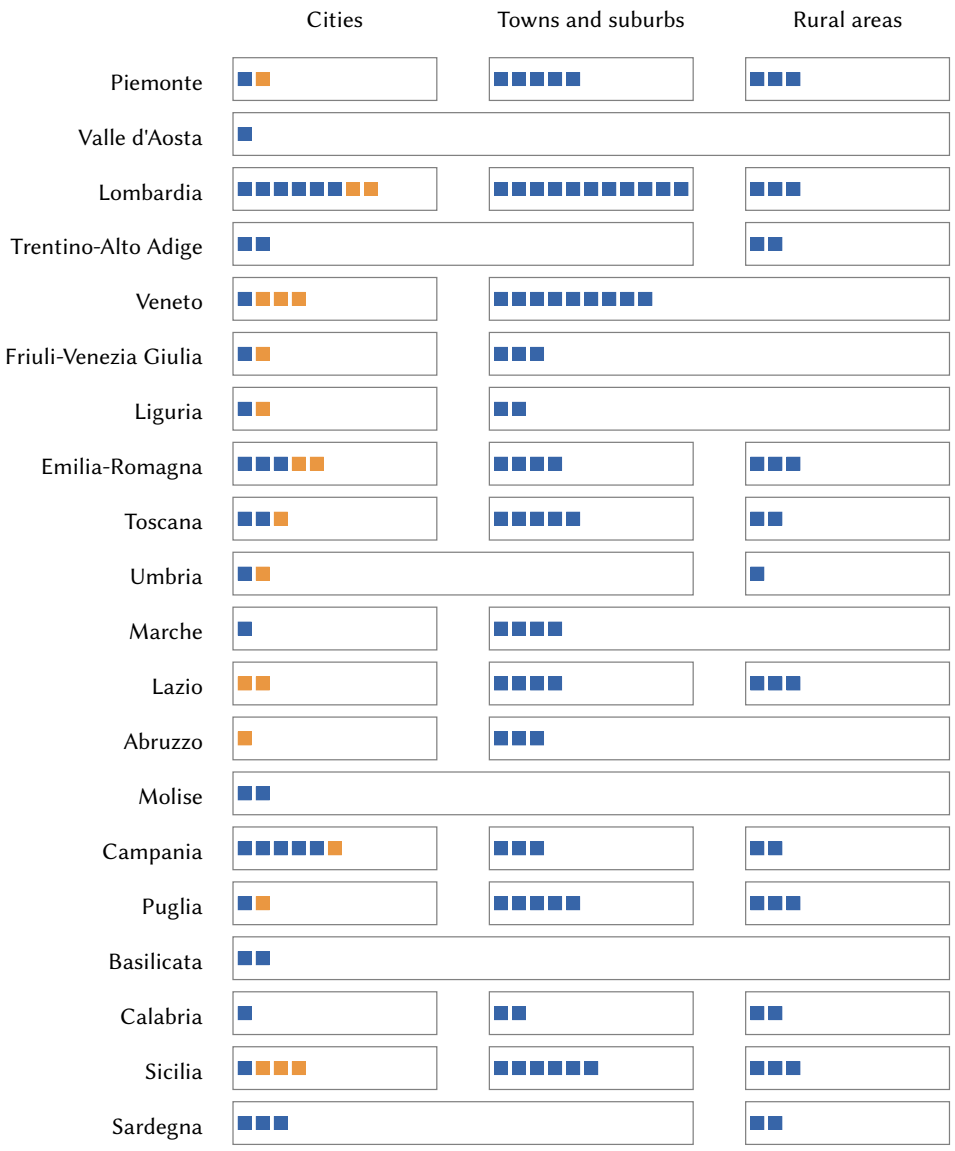


Figure 1.2 Stratification plan for the primary sampling stage of the ITA.LI sample design: Self-representing (orange squares) and non-self-representing (blue squares) strata, by administrative region and degree of urbanization.

Figure 1.2 offers a graphical summary of the stratification plan for the primary sampling stage of the ITA.LI sample design. Figure 1.3, in turn, shows the spatial distribution of the 280 PSUs selected for the study.

Figure 1.3 Spatial distribution of the self-representing (orange) and non-self-representing (blue) PSUs selected for the ITA.LI sample design.

1.4. *Secondary Sampling Stage*

Given the target population of ITA.LI (see Section 1.2), the ideal secondary sampling units of the survey would be the private households residing in the selected PSUs. Exhaustive and updated lists of such households, however, are neither publicly available nor usable by anyone other than authorized public institutions. Therefore, an intermediate sampling unit was introduced

between the PSUs and the households, namely residential addresses.

Specifically, the *secondary sampling units* (SSUs) of the ITA.LI sample design are all the residential addresses located in the 280 municipalities selected at the primary sampling stage, i.e., the addresses in each sampled PSU at which one or more private households were officially registered to reside at the beginning of the study.

Official lists of residential addresses are kept by the Registry Office of each municipality. Although such lists are not publicly available, the Italian National Institute of Statistics – our partner in designing the sample (Lucchini *et al.* 2023) – is entitled to access them and, thus, was able to use them as sampling frames to draw the desired sample of SSUs from each selected PSU.

Within each PSU, the SSU sample size was determined as follows. First, the target household sample was distributed among the 150 first-stage strata using proportional allocation (Larsen 2008). Formally:

$$\overline{m}_h \approx M_h \times f \tag{1.2}$$

where $\overline{m}_h$ denotes the target number of private households allocated to stratum h; $f = \overline{m}/M$ denotes the sampling fraction; $\overline{m}$ denotes the target household sample size; M denotes the total number of private households registered to reside in Italy; and all other symbols are defined as above.

Second, the target number of households allocated to each stratum was equally distributed among the sampled PSUs belonging to that stratum. Formally:

$$\overline{m}_{lh} = \frac{\overline{m}_h}{n_{1h}} \tag{1.3}$$

where $\overline{m}_{lh}$ denotes the target number of households allocated to PSU l of stratum h; n_{1h} denotes the number of sampled PSUs belonging to stratum h; and all other symbols are defined as above.

Since – as anticipated above (see Section 1.2) – the third stage of the ITA.LI sample design requires that a single household be randomly drawn from each selected SSU, in principle the SSU sample size for each PSU l of stratum h should equal $\overline{m}_{lh}$. In survey research, however, it is common experience that, for various reasons (e.g., unavailability, ineligibility, nonresponse), a number of sample units are lost in the data collection process.

The most common way of tackling this problem is to inflate the designated sample size to compensate for anticipated loss of sample units. To be effective, though, this strategy requires that accurate estimates of study outcome rates be available (Valliant *et al.* 2018). An alternative – adopted in this study – is to use the *sample replicates* approach (Lavrakas 2008b). This strategy entails "to

randomly select a large number of sample cases under a 'worst-case scenario,' randomly subdivide the full sample into data collection subsamples (sometimes called sample replicates), and release only the number of replicates necessary to meet the analytic objectives." (Valliant *et al.* 2018, p. 183). To ensure that the released sample can be considered a simple random sample of the designated sample, "once a replicate has been released for data collection, all cases in that replicate must be worked and given a disposition code." (Valliant *et al.* 2018, p. 184).

The implementation of the sample replicates approach in the ITA.LI sample design was as follows. First, the designated SSU sample size for each PSU l of stratum h was set at $n_{2lh} = \overline{m}_{lh} \times 8$, so as to allow for very low contact/eligibility/response rates. Second, to make the pursuit of the target sample size as flexible as possible, each designated SSU sample was partitioned into small sample replicates of size 4.

Within each PSU, SSUs were selected without replacement and with probability proportional to size, using the number of private households registered to reside in the SSUs as the measure of size. The probability of selection of SSUs is then defined as follows:

$$\pi_{k|lh} = \frac{M_{klh}}{M_{lh}} \times n_{2lh} \tag{1.4}$$

where $\pi_{k|lh}$ denotes the probability of selection of SSU k from PSU l of stratum h; M_{klh} denotes the number of private households registered to reside in SSU k of PSU l and stratum h; and all other symbols are defined as above.

Figure 1.4 displays the designated SSU sample size for each PSU of the ITA.LI sample design.

1.5. *Tertiary Sampling Stage*

The *tertiary sampling units* (TSUs) of the ITA.LI sample design are all the private households that, at the time of interview, were regularly residing at the addresses selected at the secondary sampling stage.

Since no official list of all private households residing at each selected SSU was available a priori, the construction of suitable sampling frames for TSU selection was deferred to the fieldwork stage. Specifically, following the practice known as *field listing* (Kalton *et al.* 2014), it was planned to assign interviewers the task of constructing a list of all private households apparently residing at each selected SSU, and then randomly drawing one of these households from the list (for details, see Lucchini *et al.* 2023).

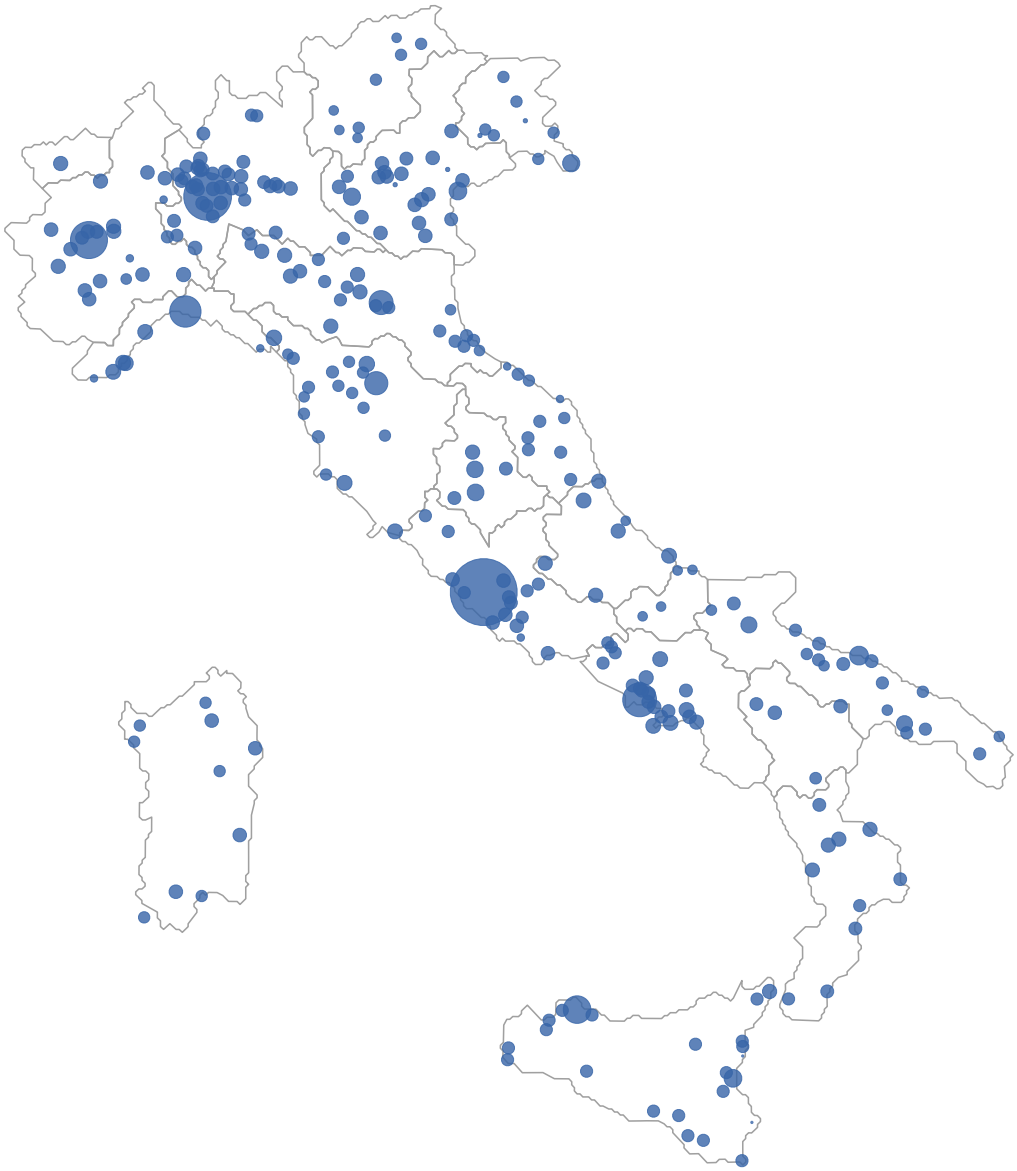


Figure 1.4 Designated SSU sample size for each selected PSU of the ITA.LI sample design. The solid circles representing PSUs are drawn with size proportional to the corresponding SSU sample size.

Within each SSU, the TSU to be contacted for interview was selected by simple random sampling. Thus, the probability of selection of TSUs is:

$$\pi_{j|klh} = \frac{1}{M_{klh}} \tag{1.5}$$

where $\pi_{j|klh}$ denotes the probability of selection of TSU j from SSU k in PSU l of stratum h; and all other symbols are defined as above.

1.6. *Ultimate Sampling Stage*

The *ultimate sampling units* (USUs) of the ITA.LI sample design are all persons aged 16 and over who, at the time of interview, were members of the private households selected at the tertiary sampling stage.

Since all eligibles individuals were selected in the sample with certainty, the probability of selection of each USU is equal to 1. Formally:

$$\pi_{i|jklh} = 1 \tag{1.6}$$

where $\pi_{i|jklh}$ denotes the probability of selection of USU i from TSU j in SSU k and PSU l of stratum h.

It is worth noting that the ITA.LI sample design is *epsem*, i.e., all units in the target population have the same a priori probability of being selected in the designated sample (Battaglia 2008). Specifically, for self-representing strata, the a priori probability of each unit i in the target population to be selected in the designated sample is defined as follows:

$$\begin{aligned}
\pi_{i(\mathrm{sr})} &= \pi_{l|h} \cdot \pi_{k|lh} \cdot \pi_{j|klh} \cdot \pi_{i|jklh} \\
&= 1 \cdot \frac{M_{klh}}{M_{lh}} \times n_{2lh} \cdot \frac{1}{M_{klh}} \cdot 1 \\
&= 1 \cdot \frac{M_{klh}}{M_{lh}} \times \frac{M_h \times f \times 8}{n_{1h}} \cdot \frac{1}{M_{klh}} \cdot 1 \\
&= f \times 8
\end{aligned} \tag{1.7}$$

since $n_{1h} = 1$ and $M_{lh} \equiv M_h$. For self-representing strata, then, the a priori probability of each unit i in the target population to be selected in the designated sample equals the sampling fraction f adjusted for the sample-replicate inflation factor. The same holds true for non-self-representing strata:

$$\begin{aligned}
\pi_{i(\mathrm{nsr})} &= \pi_{l|h} \cdot \pi_{k|lh} \cdot \pi_{j|klh} \cdot \pi_{i|jklh} \\
&= \frac{M_{lh}}{M_h} \times 2 \cdot \frac{M_{klh}}{M_{lh}} \times n_{2lh} \cdot \frac{1}{M_{klh}} \cdot 1 \\
&= \frac{M_{lh}}{M_h} \times 2 \cdot \frac{M_{klh}}{M_{lh}} \times \frac{M_h \times f \times 8}{n_{1h}} \cdot \frac{1}{M_{klh}} \cdot 1 \\
&= f \times 8
\end{aligned} \tag{1.8}$$

since $n_{1h} = 2$.

1.7. *Sample Design Implementation*

Typically, in survey research, the implementation of any complex sample design is far from perfect. There are several things that can go wrong at each

sampling stage. Overall, we can classify them into three main categories: issues related to the quality of the sampling frames, issues related to the quality of the fieldwork, and issues related to the participation in the survey (Biemer and Lyberg 2003; Groves *et al.* 2009; Lavrakas 2008a).

First, sampling frames may exclude one or more units in the target population (undercoverage) and/or include ineligible units (i.e., units that do not belong in the target population). Second, interviewer performance during fieldwork may undermine the successful implementation of the sample design in several ways; most notably, interviewers might generate incorrect or incomplete field-listed sampling frames, devote more time and effort to locating or contacting some kinds of sampling units rather than others, circumvent field sampling rules, and achieve comparatively low cooperation rates. Finally, regardless of interviewers' efforts, designated respondents may be unwilling or unable to participate in the survey. All in all, these issues result in what are known as *coverage error* and *nonresponse error*.

To a greater or lesser extent, all the above issues affected the implementation of the ITA.LI sample design. As regards, in particular, sampling frames, although the Registry Office lists of residential addresses used to select SSUs are generally expected to be of good quality, their degree of update and completeness – and, therefore, their coverage of the target population – is likely to vary across municipalities. Also, albeit interviewers were given strict guidelines on how to construct the list of private households residing at each sampled SSU and how to select the designated household from this list (Lucchini *et al.* 2023), evidence from fieldwork monitoring and ex-post data quality analysis suggests that the task was not always properly executed.

Like in most current survey research, however, the main issue in the implementation of the ITA.LI sample design was the limited propensity of the designated respondents to participate in the survey. In this regard, it is worth noting that, in addition to the general unwillingness to cooperate that characterizes today's sample surveys (Luiten *et al.* 2020), ITA.LI suffered the adverse effects of the COVID-19 pandemic, which arose when fieldwork was about halfway through[2] and is known to have sharply reduced the response rate of many sample surveys (see, for instance, Rothbaum and Hokayem 2021; U.S. Bureau of Labor Statistics 2022).

Figure 1.5 documents the implementation of the ITA.LI sample design by reporting the final dispositions (AAPOR 2016) assigned to the sampling units selected at each sampling stage of the survey, along with the corresponding absolute frequencies.

2 For details on the ITA.LI fieldwork, see Lucchini *et al.* (2023).

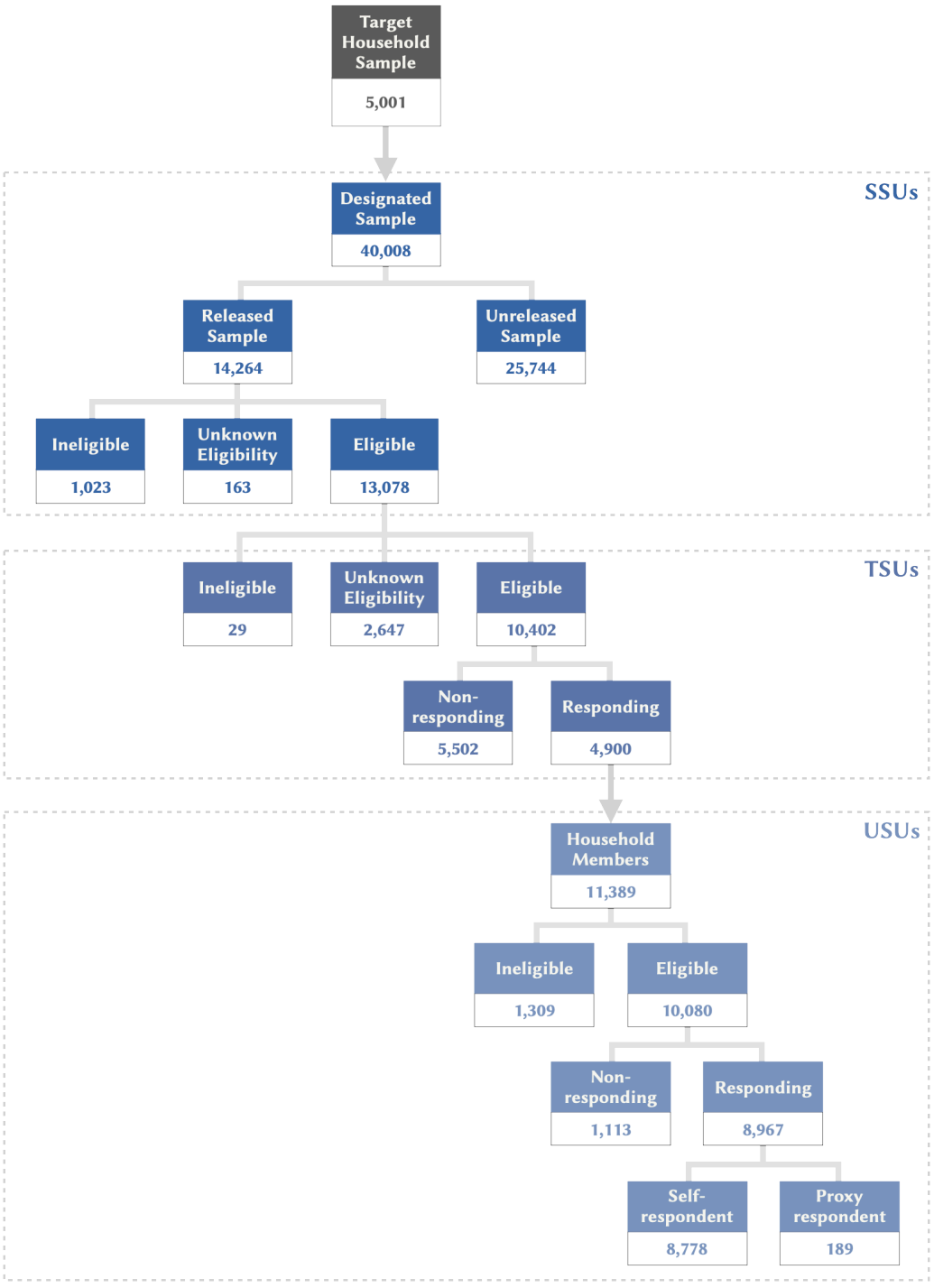


Figure 1.5 Implementation of the ITA.LI sample design: Final dispositions and corresponding absolute frequencies.

First, we can see that the realized household sample size ($m = 4{,}900$) is very close to the target ($\overline{m} = 5{,}001$). To achieve this result, it was necessary to release 14,264 residential addresses, or 37% of the designated sample of SSUs. This means that, on average, some 2.9 addresses had to be screened to reach one eligible responding household. Figure 1.6 shows that such input-to-output ratio varies considerably across macro-strata, taking values as low as 1.2 (in the small northern region of Valle d'Aosta) and as high as 5.8 (in the cities of the southern region of Abruzzo). However, the observed variation exhibits no consistent pattern by administrative region or degree of urbanization.

Back to Figure 1.5, we can see that of the 14,264 residential addresses making up the released SSU sample, 7.2% were classified as ineligible,[3] 1.1% were attributed the unknown eligibility status,[4] and the other 91.7% were found eligible.

Proceeding to the tertiary sampling stage, Figure 1.5 shows that of the 13,078 households selected from the eligible residential addresses, 0.2% were classified as ineligible, 20.2% received the unknown eligibility status,[5] and the remaining 79.6% were considered eligible. Of the latter, 47.1% agreed to participate in the survey.

A total of 11,389 individuals were found to belong to the 4,900 responding households. Of these, 10,080 were 16 years of age or older at the time of contact and, therefore, declared eligible for interview. Some of them, however, refused (10.6%) or were unable (0.4%) to cooperate, so that in the end only 8,967 interviews were completed, corresponding to a within-household individual response rate of 89%. Most interviews were administered to self-respondents, but a small proportion (2.1%) were answered by proxy respondents.[6]

Based on the information reported in Figure 1.5 – and putting proxy interviews on the same par with self-respondent interviews – the household Response Rate 1 (AAPOR 2016) of ITA.LI amounts to 37.1%. By comparison, the sixth Italian wave of the Survey of Health, Ageing and Retirement in Europe, carried out in 2014, had – for its refreshment sample – a household Response Rate 1 of 44.9% and a within-household individual response rate of 92.9% (Bergmann *et al.* 2019).

3 Specifically, 4% were assigned AAPOR disposition code "4.50 Not a housing unit"; 0.3% code "4.51 Business, government office, other organization"; 2.1% code "4.61 Regular vacant residences"; and 0.8% code "4.62 Seasonal/Vacation/Temporary residence".

4 Specifically, 0.4% were assigned AAPOR disposition code "3.11 Not attempted or worked"; 0.1% code "3.17 Unable to reach/unsafe area"; and 0.6% code "3.18 Unable to locate address".

5 Specifically, 1.7% were assigned AAPOR disposition code "320 Unknown if eligible household", while 18.5% were classified as "3.21 No screener completed".

6 For rules on the use of proxy respondents in ITA.LI, see Lucchini *et al.* (2023).

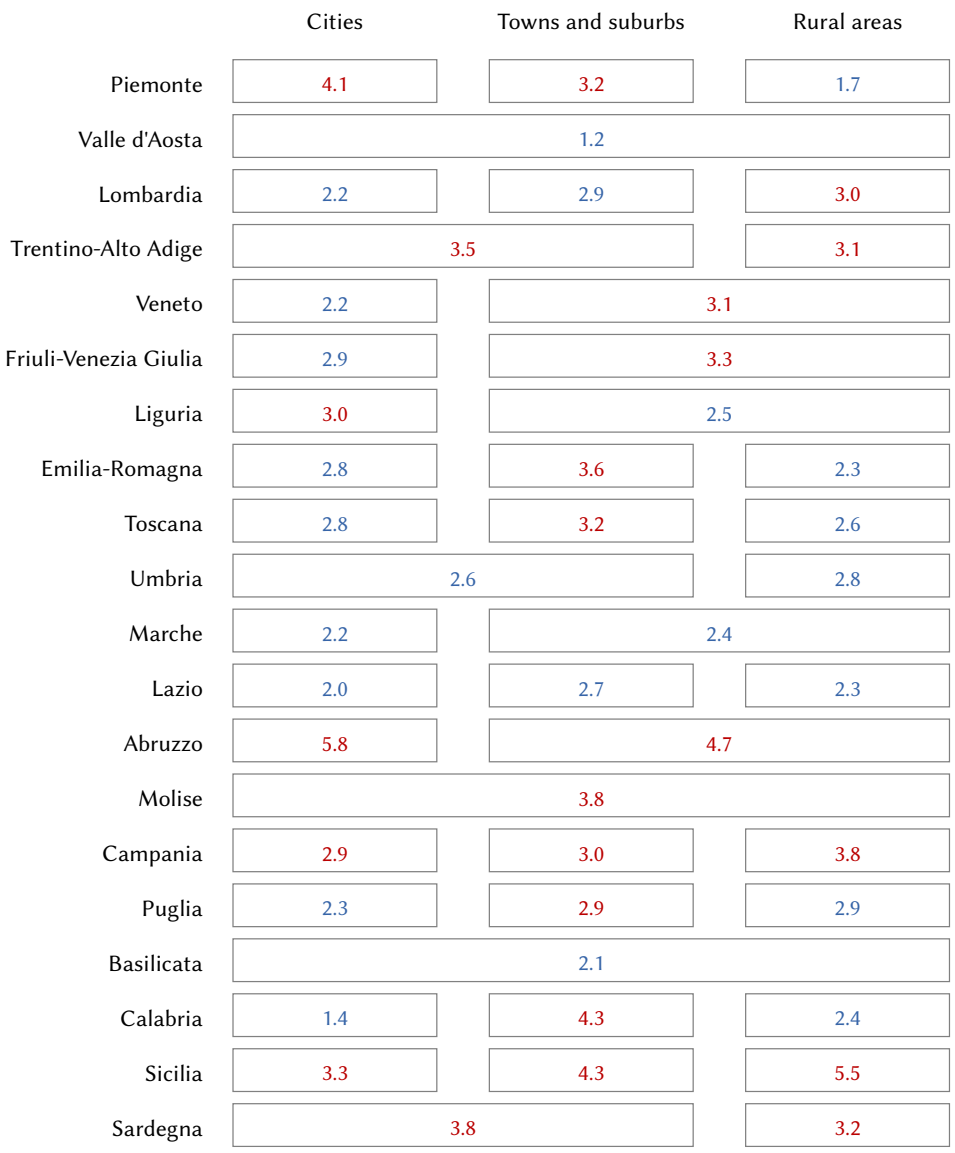


Figure 1.6 Average number of screened residential addresses per responding household, by administrative region and degree of urbanization (blue = values below the mean; red = values above the mean).

Figure 1.7 shows how both the household Response Rate 1 and the within-household individual response rate of ITA.LI differ across macro-strata. As we can see, the household Response Rate 1 varies widely between a low of 19.7%, recorded in the cities of Abruzzo, and a high of 83%, observed in Valle d'Aosta. The within-household individual response rate, on the other hand,

Figure 1.7 Household Response Rate 1 / Within-household individual response rate, by administrative region and degree of urbanization (blue = values above the mean; red = values below the mean).

is less variable, ranging from 75.9% in the cities of Abruzzo to 100% in the cities of Marche. It is worth noting that the two response rates do not covary systematically, nor do they exhibit any consistent pattern of variation by administrative region or degree of urbanization.

1.8. *Sample Representativeness*

As mentioned above (see Section 1.1), in order for a sample survey to produce valid estimates of the quantities of interest, it is required that its realized sample be representative of the target population, i.e., faithfully reproduce – within estimable margins of error – its attributes, or at least those of direct interest to the study.

When implemented perfectly, an epsem sample design – such as the one adopted for ITA.LI (see Section 1.6) – is expected to produce representative samples without any adjustment (Daniel 2012). However, as we saw in the previous section, the implementation of the ITA.LI sample design was far from perfect. On one hand, it is possible that the sampling frames used to select the SSUs and the TSUs failed to include all the units in the target population, resulting in undercoverage. On the other hand, and more importantly, the execution of the sampling process was characterized by fairly high nonresponse rates, especially at the household level.

Together, undercoverage and nonresponse make up the *nonobservation error*, which is the exclusion of a subset of the target population from the survey (Bethlehem *et al.* 2011; Kish 1965). Such exclusion may undermine the representativeness of the realized sample of a survey. This occurs when the composition of the excluded population differs from the composition of the included population with respect to one or more variables of interest. In general, the greater this difference – called the *contrast* – the lower the representativeness of the sample. Moreover, the negative impact of the contrast increases with the relative size of the excluded population (Bethlehem *et al.* 2011; Groves 2006). In symbols:

$$B(Y) = K_Y \tilde{P} \tag{1.9}$$

where $B(Y)$ denotes the deviation from perfect representativeness with respect to variable Y in a given survey sample; K_Y denotes the contrast between the included and excluded populations with respect to Y; and $\tilde{P}$ denotes the relative size of the excluded population.

As for ITA.LI, its high nonresponse rate clearly indicates that the relative size of the excluded population $\tilde{P}$ – whether or not there is undercoverage – is fairly large and, therefore, could have a substantial bearing on the representativeness of the realized sample of the survey. It remains to be determined, however, whether the contrast K_Y is large enough to realize these potential detrimental effects on representativeness.

The estimation of K_Y requires knowledge of the distribution of Y among both survey respondents and nonrespondents. With ITA.LI, this knowledge is not available with respect to nonrespondents and, therefore, $B(Y)$ must be estimated directly, without going through the calculation of K_Y.

One of the most common ways to achieve this goal is the *benchmark comparison approach*, which involves comparing the distribution of Y among survey respondents with the corresponding distribution from a benchmark data source, usually a population census (Groves 2006; Lohr 2022). The dissimilarity between these two distributions – variously measured – corresponds exactly to $B(Y)$ (Bethlehem *et al.* 2011).

In the following, the benchmark comparison approach is used to assess the representativeness of the ITA.LI realized sample with respect to some key variables. In this respect, a few remarks are in order:

- The data source used as the benchmark is the Italian Permanent Census of Population and Housing, carried out by Istat since 2018. Specifically, the data used herein refer to December 31, 2019.[7]
- Only a small set of standard sociodemographic variables are examined, the ones in common between ITA.LI and the available census data.
- Currently, census data are only available in aggregate form, which places some limitations on the comparison with ITA.LI. In particular, while the target population of ITA.LI is limited to persons aged 16 and over residing in private households (see Section 1.2), the available census data are for the whole resident population, including the *institutional residents*.[8] This means that there is a mismatch between the two sources of data to be compared, in terms of both age and residence type.
- In the analyses that follow, the age mismatch is addressed by proper subsetting of the data. The mismatch in residence type, on the other hand, is accommodated by subtracting the estimated number of institutional residents from each census count of interest. These estimated numbers were obtained based on the available information on the subject: the distribution by region of institutional residents in the year 2019; and the distribution by region, sex, and age of institutional residents in the year 2011.[9]

7 Data were taken from https://www.istat.it/it/censimenti/popolazione-e-abitazioni/risultati.

8 Institutional residents are all persons who are not members of private households but live in institutional collective dwellings, such as military installations, correctional and penal institutions, religious institutions, hospitals, and so forth.

9 Data for 2019 were taken from the Italian Permanent Census of Population and Housing (see footnote 7 above). Data for 2011 were taken from http://dati-censimentopopolazione.istat.it/Index.aspx.

- Survey data are considered in their raw form, so as to assess the representativeness of the ITA.LI realized sample before survey weighting adjustment.

Figures 1.8 to 1.16 display the results of our assessment. For each variable Y considered in the analysis, five pieces of information are reported: (a) the percent distribution of Y in the realized sample; (b) the percent distribution of Y in the benchmark population; (c) the ratio (with associated 95% confidence interval) of the sample relative distribution to the population relative distribution of Y; (d) the index of dissimilarity (D) between the sample and population distributions of Y;[10] and (e) the Pearson's X^2 statistic (with associated p-value) for a goodness-of-fit test – with survey design correction – of whether the sample distribution of Y differs significantly from the corresponding population distribution.[11]

To begin, Figures 1.8 and 1.9 assess the representativeness of the ITA.LI realized sample by using households as the units of analysis. Specifically, Figure 1.8 shows the extent to which the distribution of variable *Region of residence* as observed in the sample differs from that in the population. As we can see, the index of dissimilarity takes value 4.5, meaning that, overall, about one sampled household out of 20 should have come from a different region to exactly reproduce the population distribution of the variable under scrutiny. Although this value is not exceedingly large,[12] the results of the goodness-of-fit test ($X^2 = 184.4$, $p = 0.000$) clearly show that the sample relative distribution of variable *Region of residence* is significantly different from the corresponding distribution in the population. The two graphs that make up Figure 1.8 suggest that much of this difference comes from the over-representation, within the sample, of some small regions – in particular Valle d'Aosta and Molise – which is offset by the under-representation of Lombardia and Sicilia.

Figure 1.9 shows that the discrepancy between the ITA.LI realized sample and the benchmark population at the household level is even larger when a second variable is considered: *Household size*, defined as the number of stable household members. In this case, the index of dissimilarity rises up to 9.8, more than twice the previous value. Of course, the goodness-of-fit test ($X^2 = 231.5$, $p = 0.000$) confirms that the sample relative distribution of

10 The index of dissimilarity D (Duncan and Duncan 1955) was computed using the user-written Stata command `reldist` (Jann 2021).

11 The test was performed using the user-written Stata command `mgof` (Jann 2008).

12 As a way of comparison, the average index of dissimilarity between the sample and population distributions of two large British surveys – BHPS Wave 1 and Understanding Society Wave 1 – in terms of seven key variables is 2.7 (own calculations based on data reported in Lynn and Borkowska 2018). More generally, according to Agresti (2013) two distributions should be considered quite close when $D < 3$.

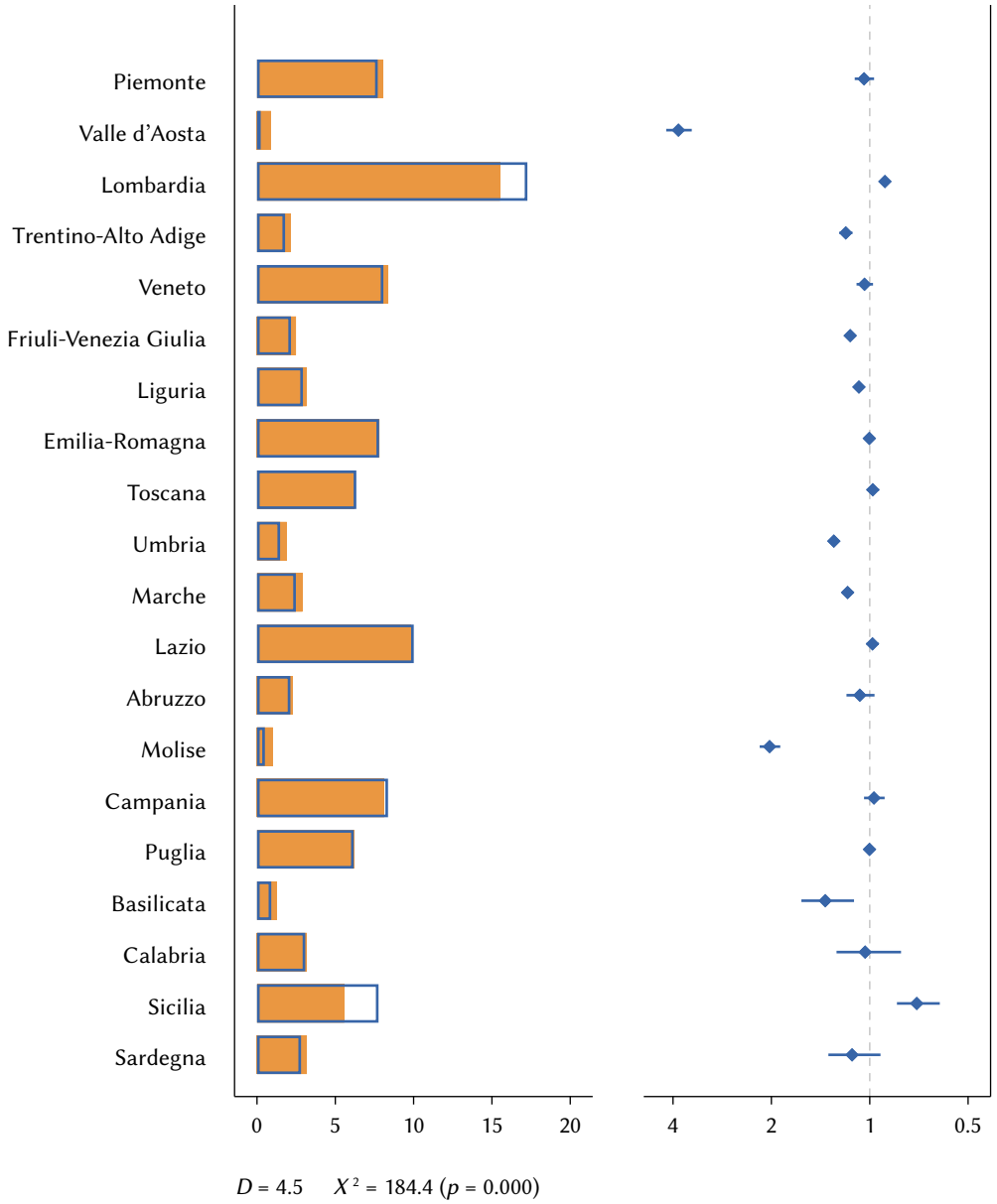


Figure 1.8 Household-level representativeness of the ITA.LI realized sample with respect to variable *Region of residence*. The left graph compares the percent distribution of the variable in the sample (orange filled bar) with the corresponding distribution in the benchmark population (blue empty bar). The right graph displays the ratio (with associated 95% confidence interval) of the sample relative distribution to the population relative distribution of the variable (y-axis in logarithmic scale). The bottom of the graph reports the index of dissimilarity (D) between the sample and population distributions of the variable, and the Pearson's X^2 statistic (with associated p-value) for a goodness-of-fit test of whether the sample distribution of the variable differs significantly from the corresponding population distribution.

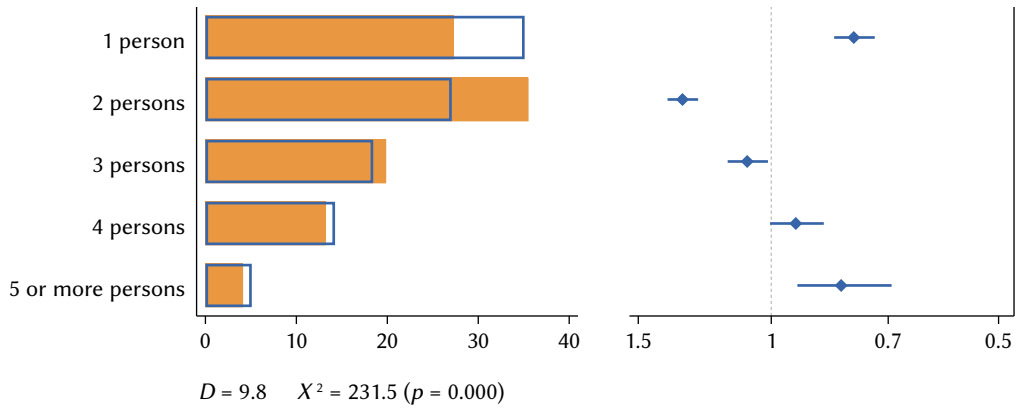


Figure 1.9 Household-level representativeness of the ITA.LI realized sample with respect to variable *Household size*. For details on graph content, see Figure 1.8 caption and text.

the variable under examination differs significantly from the corresponding population distribution. The charts reveal that the main sources of such large discrepancy are the under-representation of single-person households (27.3% in the sample versus 35.1% in the population) and the corresponding over-representation of two-person households (35.5% versus 27.1%).

Let us now turn to evaluating the representativeness of the ITA.LI realized sample by taking individuals as the units of analysis. Figures 1.10 through 1.16 describe the difference between such sample and the benchmark population with respect to the relative distribution of seven variables: sex, age group, region of residence, educational degree, occupational status, marital status, and citizenship.

Starting with variable *Sex*, Figure 1.10 shows that the discrepancy between its sample and population relative distributions is fairly small ($D = 2.5$), albeit statistically significant ($X^2 = 22.2, p = 0.000$), and takes the form of a mild

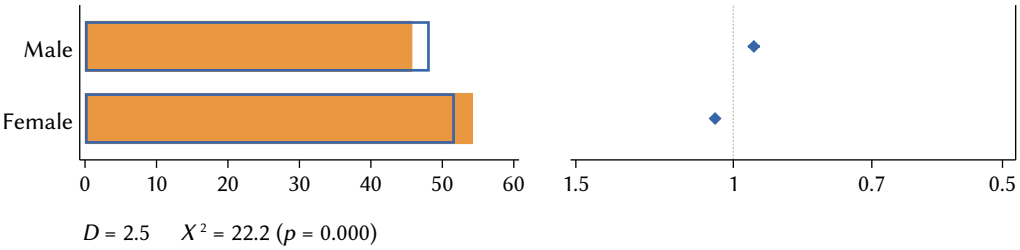


Figure 1.10 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Sex* (persons aged 16 years and over). For details on graph content, see Figure 1.8 caption and text.

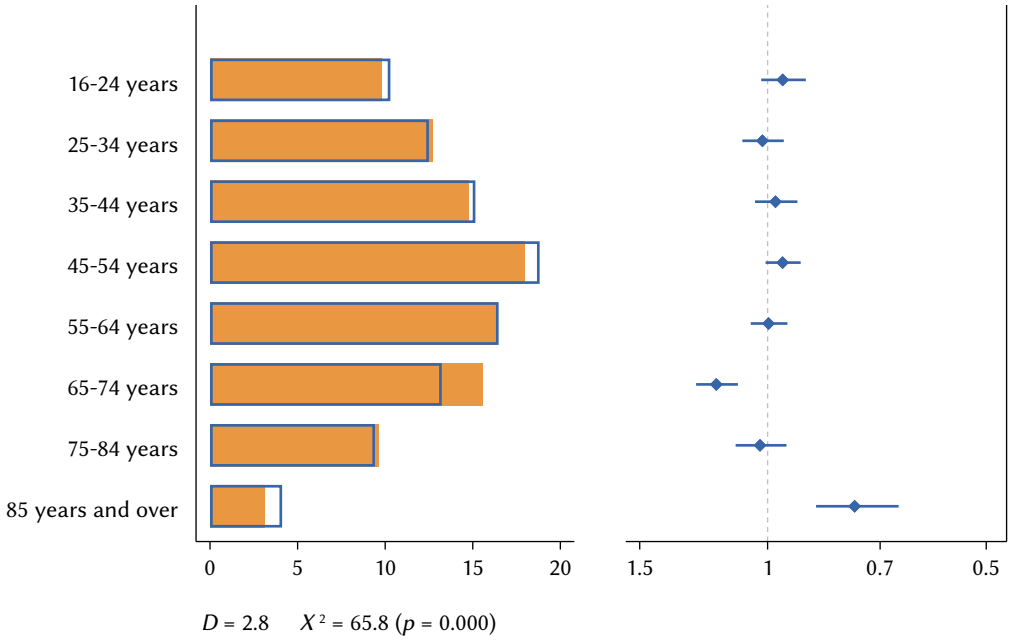


Figure 1.11 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Age group* (persons aged 16 years and over). For details on graph content, see Figure 1.8 caption and text.

under-representation of males, offset by an equivalent over-representation of females.

The case of variable *Age group*, depicted in Figure 1.11, is very similar. Again, the lack of representativeness of the ITA.LI realized sample with respect to this variable is statistically significant ($X^2 = 65.8$, $p = 0.000$), but the index of dissimilarity takes a rather low value ($D = 2.8$). The largest discrepancies between the sample and the benchmark population are due to older individuals. Specifically, there is a clear over-representation of 65-74 year olds and, on the other hand, a significant under-representation of people aged 85 years and over.

Turning to variable *Region of residence*, represented in Figure 1.12, the situation is unsurprisingly similar to that already observed at the household level (see Figure 1.8). On the one hand, the dissimilarity between the sample and population distributions of the variable at hand is noticeable ($D = 5.4$) and statistically significant ($X^2 = 444.2$, $p = 0.000$). On the other hand, this dissimilarity is characterized by the over-representation of the smaller regions – particularly evident with Valle d'Aosta and Molise – paralleled by the under-representation of two of the largest Italian regions: Lombardia and Sicilia.

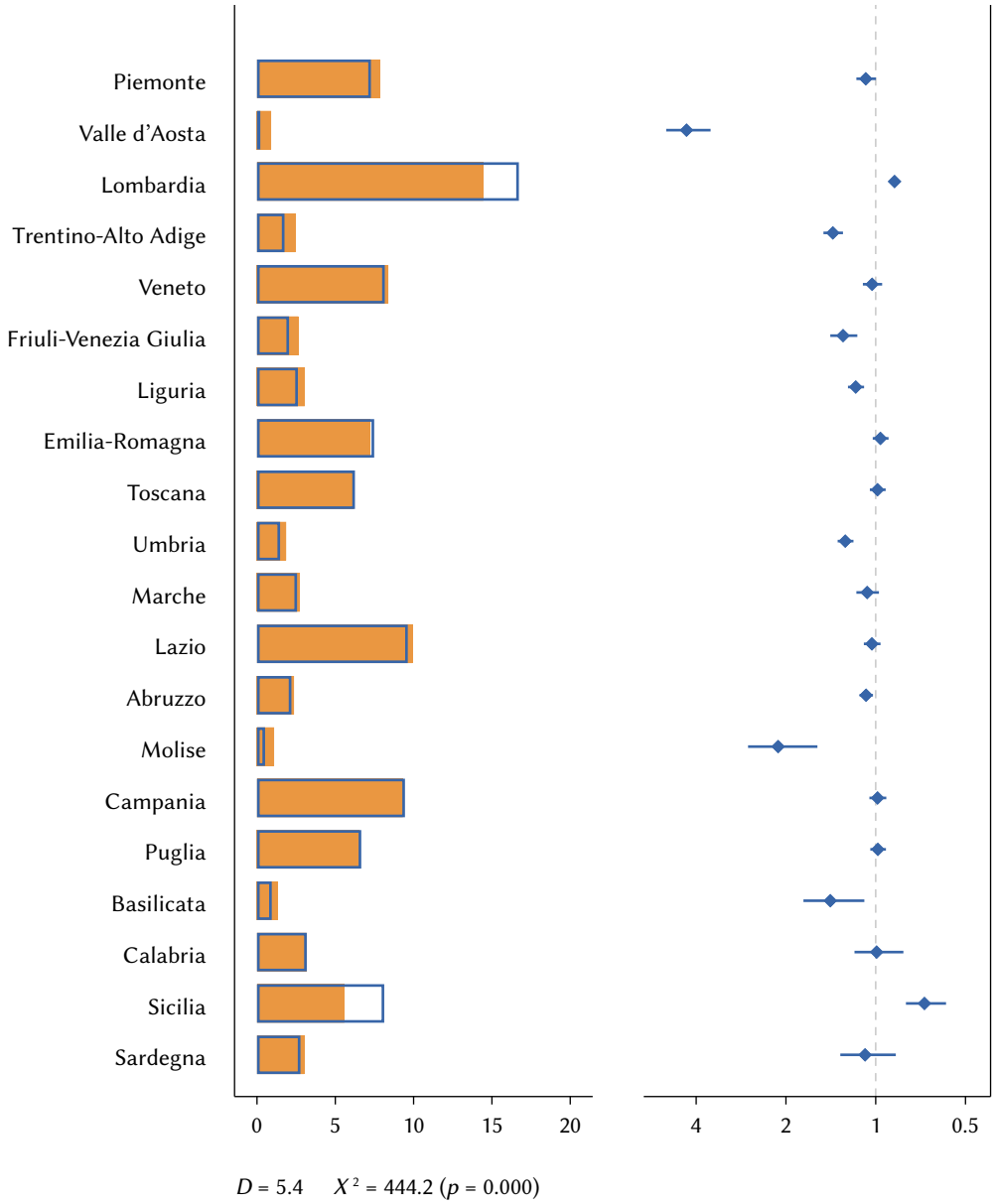


Figure 1.12 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Region of residence* (persons aged 16 years and over). For details on graph content, see Figure 1.8 caption and text.

With variable *Educational degree*, the discrepancy between the ITA.LI realized sample and the benchmark population rises substantially. As seen in Figure 1.13, in this case the dissimilarity index has a value of $D = 7.6$ and the goodness-of-fit test appears highly significant ($X^2 = 219.5, p = 0.000$). Going into detail, the charts show that this deficit in representativeness is attributable to the under-representation of the educational degrees placed

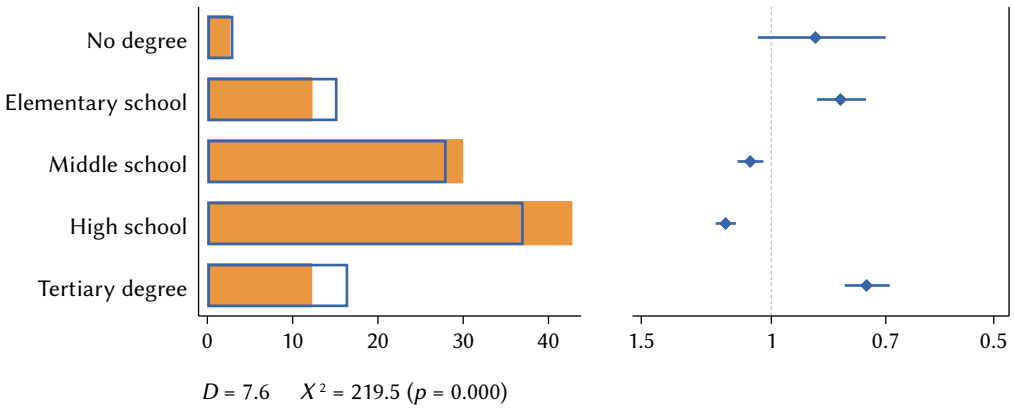


Figure 1.13 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Educational degree* (persons aged 25 years and over). For details on graph content, see Figure 1.8 caption and text.

at the opposite ends of the range (elementary school certificate and university degrees), coupled with a parallel over-representation of the intermediate educational degrees (middle school certificate and high school diploma).

The representativeness of the ITA.LI realized sample becomes acceptable again with respect to variable *Occupational status*. Figure 1.14 shows that its sample relative distribution differs only slightly – though significantly ($X^2 = 41.1$, $p = 0.000$) – from the corresponding population distribution ($D = 2.7$). Such difference takes the form of a modest over-representation of both job seekers and retired people, matched by an equally low under-representation of homemakers and persons in other status (mainly unable to work).

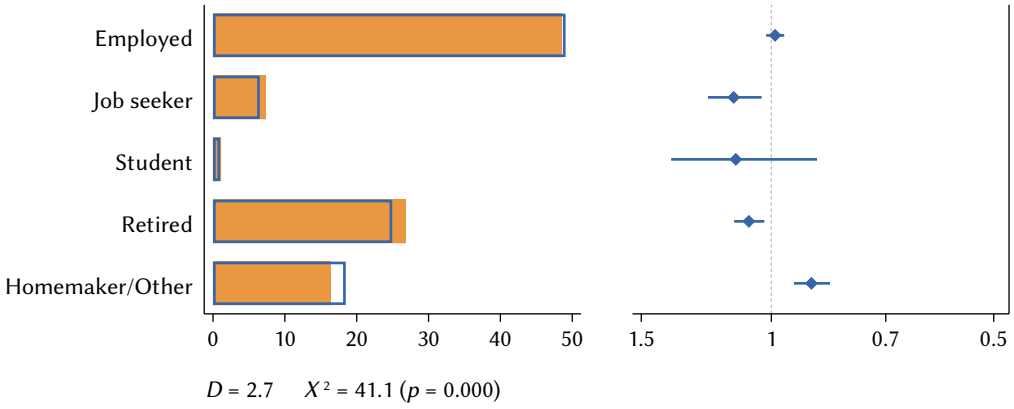


Figure 1.14 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Occupational status* (persons aged 25 years and over). For details on graph content, see Figure 1.8 caption and text.

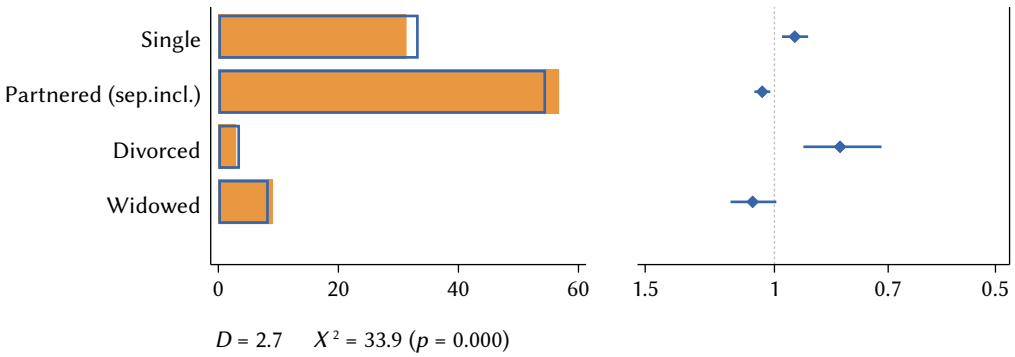


Figure 1.15 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Marital status* (persons aged 16 years and over). For details on graph content, see Figure 1.8 caption and text.

As we can see from Figure 1.15, a low level of discrepancy between the sample and the population relative distributions also characterizes variable *Marital status* ($D = 2.7$), in the form of a slight over-representation of partnered and widowed persons, together with an analogous mild under-representation of single and divorced people.

Finally, Figure 1.16 reveals a substantial under-representation of non-Italian residents in the ITA.LI realized sample, within which they constitute a share that is less than half of that in the benchmark population (3.6% versus 8%).

In summary, our analysis shows that the ITA.LI realized sample has both positive and negative aspects when it comes to its representativeness. At the household level, the sample's representativeness exhibits its most glaring shortcoming with respect to variable *Household size*, where two-person households are substantially over-represented at the expense of single-person households. At the individual level, on the other hand, while variables *Sex*, *Age group*, *Occupational status* and *Marital status* reveal relatively small discrepancies between the sample and the benchmark population, variables *Educational degree* and

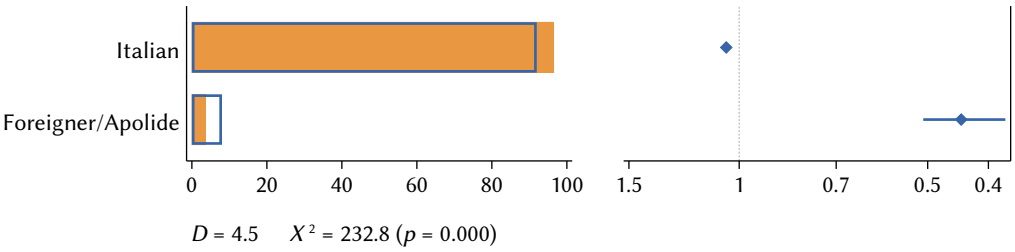


Figure 1.16 Individual-level representativeness of the ITA.LI realized sample with respect to variable *Citizenship* (persons aged 20 years and over). For details on graph content, see Figure 1.8 caption and text.

Citizenship turn out to be rather off the mark. Specifically, on one side people holding an elementary school certificate or a university degree are largely under-represented in favor of those who have a middle school certificate or a high school diploma; on the other side, non-Italian residents are only half of those who should be in a perfectly representative sample. Overall, these findings suggest the need to adjust the ITA.LI realized sample through appropriate weighting, a topic that will be covered in the next chapter.

2

SURVEY WEIGHTING

2.1. *Introduction*

A survey weight w_i is a positive numerical value assigned to each unit i of a survey's realized sample, such that it expresses the number of units in the target population represented by i (Heeringa *et al.* 2017). Survey weights are a critical component of any survey data analysis since – when calculated and applied correctly – they allow for population inference, i.e., they help obtain (approximately) unbiased estimators for the quantities of interest in the target population (Heeringa *et al.* 2017; Valliant *et al.* 2018). Specifically, survey weights do this by compensating for "unequal selection probabilities, nonresponse, noncoverage, and sampling fluctuations from known population values" (Kalton and Flores-Cervantes 2003, p. 81).

Generally, survey weights are developed in three stages (Haziza and Beaumont 2017). First, each unit in the designated sample is assigned a *base weight* (or design weight) equal to the inverse of its probability of selection in the sample; if – as in the case of ITA.LI – the sample replicates approach was used and the released sample is smaller than the designated sample, then the base weight will need to be adjusted accordingly (Valliant *et al.* 2018). Second, the base weights assigned to the eligible responding units are adjusted to compensate for the removal from the sample of both units with unknown eligibility and nonresponding units. Finally, most often the resulting weights from the previous two stages are subjected to *calibration*, i.e., they are further adjusted "so that survey weighted estimates agree with known population totals available from external sources (e.g., the census or administrative data) for important variables" (Haziza and Beaumont 2017, p. 207); survey weight calibration has two main purposes: to compensate for nonobservation error and, possibly, to improve the precision of survey estimators (Heeringa *et al.* 2017; Kalton and Flores-Cervantes 2003). It should be noted that, in case of

multistage sample designs, base weights are computed separately for each sampling stage, while their adjustments for unknown eligibility and nonresponse are calculated incrementally as one moves from one sampling stage to the next (Valliant and Dever 2018).

The aim of this chapter is to describe the procedure used to construct the ITA.LI survey weights. The next section provides a general overview of the procedure, while the subsequent four sections illustrate its implementation for each of the sampling stages of the ITA.LI sample design. The chapter concludes with a section dedicated to an assessment of the representativeness of the ITA.LI weighted realized sample. The exposition that follows draws extensively on two reference texts that are mentioned here once and for all: Valliant *et al.* (2018) and Valliant and Dever (2018). Any additional relevant literature will be cited as usual.

2.2. *Overview*

As mentioned in the previous section, the process of survey weight construction relies on two basic operations: the calculation of base weights and their adjustment. In the case of a multistage sample design such as the one used for ITA.LI, this process runs sequentially through the various sampling stages. Specifically, at each stage the corresponding *conditional base weights* are first calculated. Then – from the second stage onwards – the conditional base weights are multiplied by the weights calculated at the previous stage, thus obtaining the *unconditional base weights* for the current stage. Finally, these weights are adjusted as needed, resulting in the *adjusted base weights* for the stage at hand. When the units of a given sampling stage (e.g., households or individuals) are units of analysis for the survey, the adjusted base weights for that stage constitute the *final survey weights* for the respective units, possibly after calibration.

Figure 2.1 offers a diagrammatic representation of the process of construction of the ITA.LI survey weights. As we can see, this is a rather complex procedure that begins with the calculation of the base weights assigned to the PSUs and results in two sets of final weights: that for households, obtained at the tertiary sampling stage, and that for individuals, which is the terminal output of the entire process.

The following four sections provide a complete description of the procedure followed to calculate the ITA.LI survey weights at each sampling stage of the survey.

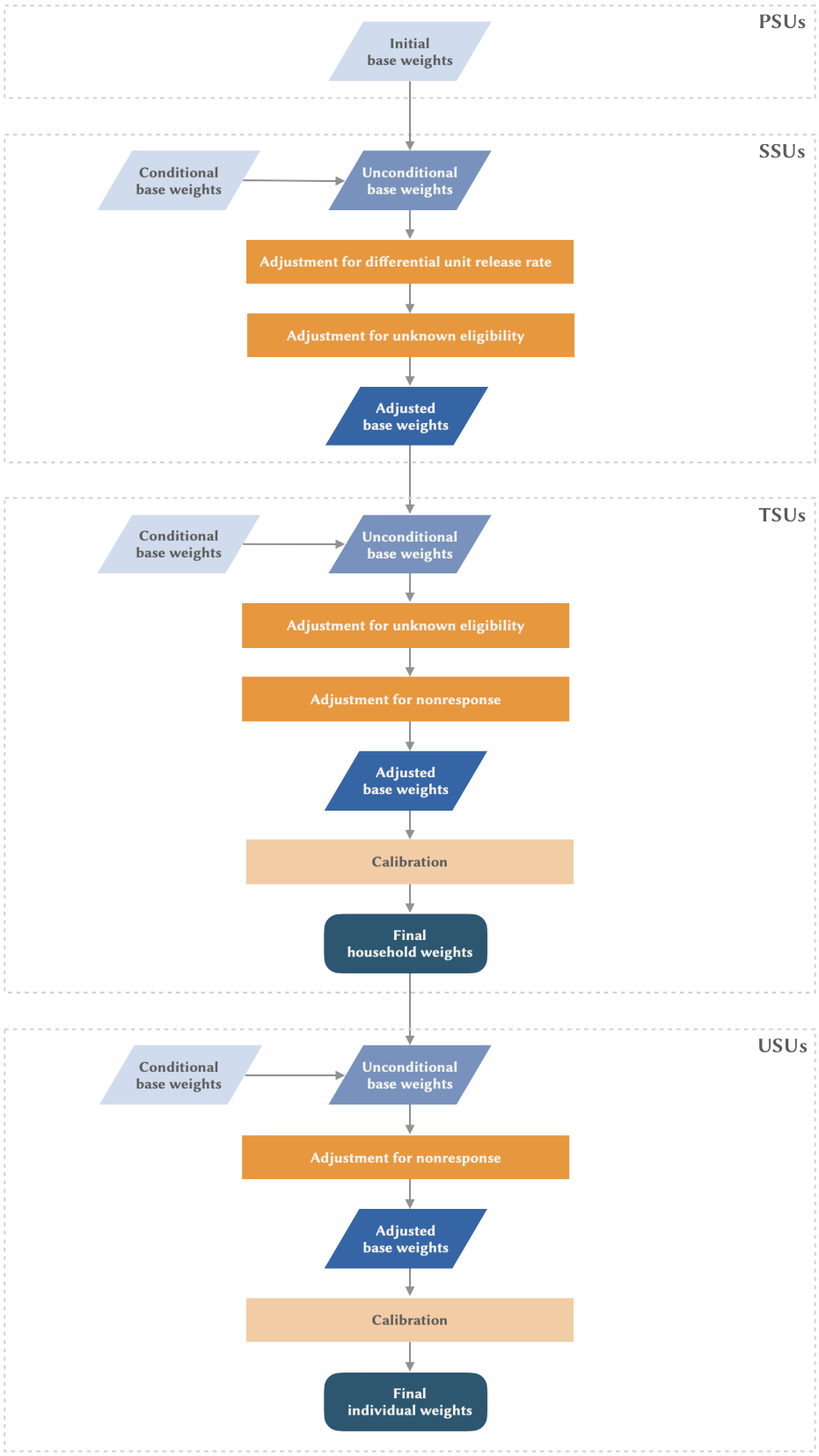


Figure 2.1 Diagrammatic representation of the process of construction of the ITA.LI survey weights.

2.3. *Primary Sampling Stage*

As shown in Figure 2.1, the starting point of the procedure was the calculation of the PSU base weights, defined as the inverse of the probability of each sampled PSU to be selected in the study. Formally:

$$d_{0l} = \frac{1}{\pi_{l|h}} \tag{2.1}$$

where $\pi_{l|h}$ is defined as in Equation 1.1.

2.4. *Secondary Sampling Stage*

Figure 2.1 shows that going from the PSU base weights to the SSU adjusted base weights involved four steps.

First, we computed the SSU conditional base weights, defined as the inverse of the probability of selection of each sampled SSU from its parent PSU. Formally:

$$d_{0k|l} = \frac{1}{\pi_{k|lh}} \tag{2.2}$$

where $\pi_{k|lh}$ is defined as in Equation 1.4.

Second, the SSU conditional base weights were multiplied by the PSU base weights to obtain the SSU unconditional base weights. Formally:

$$d_{0k} = d_{0k|l} \times d_{0l} \tag{2.3}$$

The selection probabilities used in Equation 2.2 refer to the designated sample. However, as we know from Section 1.7, only a random subsample of the designated SSUs was actually fielded. To account for this subsampling, the SSU unconditional base weights were adjusted as follows:

$$d_{1k} = d_{0k} \times a_{1k} \tag{2.4}$$

where a_{1k} denotes the subsampling weight adjustment factor for SSU k. This element is defined below:

$$a_{1k} = \begin{cases} \dfrac{n_{2lh}}{r_{2lh}}, & \text{if } k \in \text{released sample} \\ 0, & \text{if } k \in \text{unreleased sample} \end{cases} \tag{2.5}$$

where n_{2lh} denotes the designated SSU sample size for PSU l of stratum h; and r_{2lh} denotes the number of SSUs that were actually released for fieldwork in PSU l of stratum h.

Finally, the SSU unconditional base weights were further adjusted to account for the 163 SSUs with unknown eligibility status (see Figure 1.5). The rationale behind this adjustment is simple and intuitive: units of unknown eligibility are zeroed out and the weights associated with them are redistributed to units for which the eligibility status is known. If we assume that the probability of having the unknown eligibility status is not constant but varies depending on some properties of the SSUs, then the magnitude of weight adjustment will need to vary accordingly. To this end, typically SSUs are divided into a number of distinct classes based on variables that are assumed to be associated with eligibility status, and the weight adjustment factor is calculated separately for each class.

Most commonly, there is very limited information on units with unknown eligibility status. Therefore, the definition of adjustment classes is usually based on simple criteria. In our case, we partitioned the SSUs into 46 classes corresponding to the first-stage preliminary strata defined in Section 1.3. Then, for each class $c = 1, \ldots, 46$, we calculated the unknown eligibility weight adjustment factor as follows:

$$f_{c(\text{unk-ssu})} = \frac{\sum_{k \in ssu_c} d_{1k}}{\sum_{k \in ssu_{c,\text{kn}}} d_{1k}} \tag{2.6}$$

where ssu_c denotes the complete set of SSUs belonging to class c; $ssu_{c,\text{kn}}$ denotes the subset of ssu_c containing only the SSUs of known eligibility; and d_{1k} is defined as in Equation 2.4.

Based on $f_{c(\text{unk-ssu})}$, we derived the unknown eligibility weight adjustment factor for each SSU k using the following rule:

$$a_{2k} = \begin{cases} f_{c(\text{unk-ssu})}, & \text{if } k \in ssu_{c,\text{kn}} \\ 0, & \text{if } k \in ssu_{c,\text{unk}} \end{cases} \tag{2.7}$$

where $ssu_{c,\text{unk}}$ denotes the subset of ssu_c containing the SSUs of unknown eligibility; and all other symbols are defined as above.

Lastly, by applying the adjustment factor a_{2k} to weights d_{1k}, we obtained the SSU adjusted base weights:

$$d_{2k} = d_{1k} \times a_{2k} \tag{2.8}$$

After all these steps were completed, 14,101 SSUs resulted in being assigned a non-zero weight. Of these, the ineligible units were discarded, so that only the 13,078 eligible SSUs were passed to the next stage.

2.5. *Tertiary Sampling Stage*

The steps taken to arrive at the TSU adjusted base weights closely mimic those used for generating the SSU adjusted base weights. In the case of TSUs, however, the resulting weights were further fine-tuned by calibration, so as to obtain the *final household weights*.

Specifically, first of all, the TSU conditional base weights were calculated, defined as the inverse of the probability of selection of each sampled TSU from its parent SSU. Formally:

$$d_{0j|kl} = \frac{1}{\pi_{j|klh}} \tag{2.9}$$

where $\pi_{j|klh}$ is defined as in Equation 1.5.

Second, the TSU conditional base weights were multiplied by the SSU adjusted base weights to obtain the TSU unconditional base weights. Formally:

$$d_{0j} = d_{0j|kl} \times d_{2k} \tag{2.10}$$

Then, two adjustments were applied in sequence to the TSU unconditional base weights: one for unknown eligibility and the other for nonresponse. Regarding the former, the same procedure as described in the previous section was followed. Precisely, the TSUs were partitioned into the 46 classes corresponding to the first-stage preliminary strata of the ITA.LI sample design, and for each class c the unknown eligibility weight adjustment factor was calculated as follows:

$$f_{c(\text{unk-tsu})} = \frac{\sum_{j \in tsu_c} d_{0j}}{\sum_{j \in tsu_{c,\text{kn}}} d_{0j}} \tag{2.11}$$

where tsu_c denotes the complete set of TSUs belonging to class c; $tsu_{c,\text{kn}}$ denotes the subset of tsu_c containing only the TSUs of known eligibility; and d_{0j} is defined as in Equation 2.10.

Based on $f_{c(\text{unk-tsu})}$, we derived the unknown eligibility weight adjustment factor for each TSU j using the following rule:

$$a_{1j} = \begin{cases} f_{c(\text{unk-tsu})}, & \text{if } j \in tsu_{c,\text{kn}} \\ 0, & \text{if } j \in tsu_{c,\text{unk}} \end{cases} \tag{2.12}$$

where $tsu_{c,\text{unk}}$ denotes the subset of tsu_c containing the TSUs of unknown eligibility; and all other symbols are defined as above.

Finally, by applying the adjustment factor a_{1j} to weights d_{0j}, we obtained the following adjusted weights:

$$d_{1j} = d_{0j} \times a_{1j} \tag{2.13}$$

A second adjustment was applied to the TSU unconditional base weights to account for the non-responding eligible TSUs. With proper adaptation, the approach taken closely mirrors that just described for calculating the unknown eligibility weight adjustment factor. First, for each of the 46 classes defined above, the nonresponse weight adjustment factor was computed:

$$f_{c(\text{nr-tsu})} = \frac{\sum_{j \in tsu_{c,\text{el}}} d_{1j}}{\sum_{j \in tsu_{c,\text{er}}} d_{1j}} \tag{2.14}$$

where $tsu_{c,\text{el}}$ denotes the subset of tsu_c containing all the eligible TSUs; $tsu_{c,\text{er}}$ denotes the subset of tsu_c containing only the responding eligible TSUs; and d_{1j} is defined as in Equation 2.13.

Then, based on $f_{c(\text{nr-tsu})}$, the nonresponse weight adjustment factor for each TSU j was determined as follows:

$$a_{2j} = \begin{cases} f_{c(\text{nr-tsu})}, & \text{if } j \in tsu_{c,\text{er}} \\ 1, & \text{if } j \in tsu_{c,\text{in}} \\ 0, & \text{if } j \in tsu_{c,\text{enr}} \cup tsu_{c,\text{unk}} \end{cases} \tag{2.15}$$

where $tsu_{c,\text{in}}$ denotes the subset of tsu_c containing the ineligible TSUs; $tsu_{c,\text{enr}}$ denotes the subset of tsu_c containing the non-responding eligible TSUs; and all other symbols are defined as above.

Lastly, the adjustment factor a_{2j} was applied to weights d_{1j}, resulting in the TSU adjusted base weights:

$$d_{2j} = d_{1j} \times a_{2j} \tag{2.16}$$

Of the 13,078 TSUs considered here, only 4,929 received a non-zero weight. Twenty-nine of these were discarded as ineligible, leaving us with the 4,900 responding eligible TSUs.

As mentioned at the opening of the current section, the weights assigned to these units were further adjusted by *calibration* to obtain the final household weights. This particular type of weight tuning aims on the one hand at correcting for coverage errors and, possibly, the residual nonresponse error from previous adjustments; and on the other hand at increasing the precision of some survey estimators. Such dual goal is sought by "calibrating" the

Table 2.1 Population joint frequency distribution of variables *Area of residence* and *Household size* for household weight calibration.

	Area of residence					
	North-West	North-East	Center	South	Islands	*Total*
Household size						
1 person	2,744,015	1,839,639	1,947,347	1,630,563	912,288	9,073,852
2 persons	2,074,408	1,451,536	1,391,822	1,382,163	703,425	7,003,354
3 persons	1,264,417	912,302	971,821	1,091,094	530,600	4,770,234
4 persons	894,341	670,919	693,731	989,444	436,441	3,684,876
5+ persons	295,843	250,595	243,642	379,018	149,708	1,318,806
Total	7,273,024	5,124,991	5,248,363	5,472,282	2,732,462	25,851,122

Source: Own elaboration of data from the Italian Permanent Census of Population and Housing (2019).

realized sample to the target population with respect to a certain set of variables $\mathbf{X}$; this means to adjust survey weights so that the distribution of $\mathbf{X}$ in the weighted sample exactly matches the corresponding distribution in the target population. The achievement of the goals of calibration is more likely the stronger the correlation of $\mathbf{X}$ with nonobservation and the outcomes of interest (Caughey *et al.* 2020).

Ideally, the match between the weighted sample and the target population should be with respect to the full *joint distribution* of $\mathbf{X}$, as required by the weight calibration method known as *poststratification*. When the number of cells making up the joint distribution of $\mathbf{X}$ is relatively large, however, poststratification may become impractical. In this case, a viable option is *raking*, which requires that the match between the weighted sample and the target population be with respect to a proper set of lower-order *marginal distributions* of $\mathbf{X}$.[1] Raking is based on *iterative proportional fitting* (Deming and Stephan 1940), a procedure "which involves iteratively adjusting the [base] weights to match each margin in succession until the weights stabilize" (Caughey *et al.* 2020, p. 15, f.n. 11).

In this study, the TSU adjusted base weights defined in Equation 2.16 were calibrated by poststratification with respect to the joint distribution of two variables: *Area of residence* and *Household size*. The target population counts for such distribution, reported in Table 2.1, come from the Italian Permanent Census of Population and Housing;[2] specifically, the data used herein refer to December 31, 2019.

1 A lower-order marginal distribution of $\mathbf{X}$ is a joint distribution over a subset of $\mathbf{X}$.

2 https://www.istat.it/it/censimenti/popolazione-e-abitazioni/risultati.

Now, let $s = 1, \ldots, 25$ index the poststrata defined by the cross-classification of the two selected calibration variables. For each poststratum s, the poststratification weight adjustment factor was computed as follows:

$$f_{s(\text{ps-tsu})} = \frac{N_s}{\sum_{j \in tsu_{s,\text{er}}} d_{2j}} \tag{2.17}$$

where N_s denotes the target population count for poststratum s; $tsu_{s,\text{er}}$ denotes the set of responding eligible TSUs belonging to poststratum s; and d_{2j} is defined as in Equation 2.16.

Then, based on $f_{s(\text{ps-tsu})}$, the poststratification weight adjustment factor for each TSU j was determined as follows:

$$a_{3j} = \begin{cases} f_{s(\text{ps-tsu})}, & \text{if } j \in tsu_{s,\text{er}} \\ 0, & \text{otherwise} \end{cases} \tag{2.18}$$

where all symbols are defined as above.

Finally, the adjustment factor a_{3j} was applied to weights d_{2j} to obtain the final weight for each responding eligible household j:

$$w_{\text{H}j} = d_{2j} \times a_{3j} \tag{2.19}$$

2.6. *Ultimate Sampling Stage*

Since the probability of selection of each eligible household member is equal to one (see Equation 1.6), the final household weights derived in the previous section correspond to the USUs unconditional base weights. Formally:

$$d_{0i} = w_{\text{H}j} \qquad \text{if } \; i \in j \tag{2.20}$$

As shown in Figure 2.1, these weights were first adjusted for nonresponse and then calibrated to arrive at the *final individual weights*.

The nonresponse weight adjustment factor was computed in three steps. First, a logistic regression model was used to estimate the probability of response of the 10,080 eligible household members. Specifically, the probability of response was modeled as a function of the main effects of six categorical predictors: *Sex*, *Age group* (15 categories), *Region of residence* (20 categories), *Educational degree* (five categories), *Occupational status* (five categories), and *Citizenship* (two categories). Second, the estimated probabilities were sorted and divided into 10 classes of approximately equal size based on the deciles of

their distribution. Finally, for each class c, the nonresponse weight adjustment factor was computed as follows:

$$f_{c(\text{nr-usu})} = \frac{\sum_{i \in usu_{c,\text{el}}} d_{0i}}{\sum_{i \in usu_{c,\text{er}}} d_{0i}} \tag{2.21}$$

where $usu_{c,\text{el}}$ denotes the eligible USUs belonging to class c; $usu_{c,\text{er}}$ denotes the responding eligible USUs belonging to class c; and d_{0i} is defined as in Equation 2.20.

Based on $f_{c(\text{nr-usu})}$, the nonresponse weight adjustment factor for each USU i was determined as follows:

$$a_{1i} = \begin{cases} f_{c(\text{nr-usu})}, & \text{if } i \in usu_{c,\text{er}} \\ 1, & \text{if } i \in usu_{c,\text{in}} \\ 0, & \text{if } i \in usu_{c,\text{enr}} \end{cases} \tag{2.22}$$

where $usu_{c,\text{in}}$ denotes the ineligible USUs belonging to class c; $usu_{c,\text{enr}}$ denotes the non-responding eligible USUs belonging to class c; and all other symbols are defined as above.

By applying the adjustment factor a_{1i} to weights d_{0i}, the USU adjusted base weights were obtained:

$$d_{1i} = d_{0i} \times a_{1i} \tag{2.23}$$

As the final step in the process of constructing the ITA.LI survey weights, the USU adjusted base weights were calibrated so as to generate the final individual weights. The set of variables $\mathbf{X}$ used to this aim is similar to that employed for nonresponse adjustment: *Sex*, *Birth cohort*, *Region of residence*, *Educational degree*, *Occupational status*, and *Citizenship*. Since poststratification with respect to the joint distribution of $\mathbf{X}$ was not feasible, calibration was carried out by raking with respect to a relevant set of second-order marginal distributions of $\mathbf{X}$: *Sex* × *Birth cohort*, *Sex* × *Region of residence*, *Sex* × *Educational degree*, *Sex* × *Occupational status*, and *Sex* × *Citizenship*.

The population counts for calibration, reported in Table 2.2, are from the Italian Permanent Census of Population and Housing and refer to December 31, 2019. It should be noted that these counts represent not the ITA.LI target population, but rather the *total* population residing in private households in Italy, thus including individuals aged 0-15 years. This is because the population counts of interest were available only in aggregate form and it was not possible to separate out individuals not in the ITA.LI target population. Accordingly, calibration was performed on both the ineligible and the responding eligible members of the cooperating households – with the sole exception of 26 infants born in 2020, who were not counted.

Table 2.2 Population marginal frequency distributions for individual weight calibration.

	Sex		
	Male	Female	*Total*
Birth cohort			
1921-1929	205,596	524,910	730,506
1930-1939	1,411,446	2,125,447	3,536,893
1940-1944	1,165,517	1,437,776	2,603,293
1945-1949	1,543,207	1,756,110	3,299,317
1950-1954	1,639,938	1,805,017	3,444,955
1955-1959	1,865,390	2,012,912	3,878,302
1960-1964	2,201,749	2,318,104	4,519,853
1965-1969	2,391,344	2,467,660	4,859,004
1970-1974	2,344,266	2,388,214	4,732,480
1975-1979	2,073,920	2,093,310	4,167,230
1980-1984	1,779,839	1,775,024	3,554,863
1985-1989	1,640,585	1,620,281	3,260,866
1990-1994	1,587,953	1,519,884	3,107,837
1995-1999	1,545,986	1,406,090	2,952,076
2000-2004	1,482,553	1,384,822	2,867,375
2005-2009	1,457,572	1,373,841	2,831,413
2010-2014	1,349,569	1,274,998	2,624,567
2015-2019	1,165,630	1,122,961	2,288,591
Region			
Piemonte	2,075,095	2,194,349	4,269,444
Valle d'Aosta	60,571	63,502	124,073
Lombardia	4,881,615	5,078,854	9,960,469
Trentino-Alto Adige	524,660	540,447	1,065,107
Veneto	2,372,478	2,466,508	4,838,986
Friuli-Venezia Giulia	581,654	614,043	1,195,697
Liguria	722,380	787,160	1,509,540
Emilia-Romagna	2,154,856	2,273,754	4,428,610
Toscana	1,771,133	1,898,552	3,669,685
Umbria	416,954	447,792	864,746
Marche	731,720	773,074	1,504,794
Lazio	2,758,172	2,955,552	5,713,724
Abruzzo	628,310	660,285	1,288,595
Molise	146,388	151,926	298,314
Campania	2,772,704	2,919,958	5,692,662
Puglia	1,916,509	2,024,352	3,940,861
Basilicata	270,075	280,356	550,431
Calabria	922,405	964,101	1,886,506
Sicilia	2,356,925	2,495,981	4,852,906
Sardegna	787,456	816,815	1,604,271

(Continued on next page)

Table 2.2 (Continued).

	Sex		
	Male	Female	*Total*
Educational degree			
Age 0-8	2,225,496	2,107,263	4,332,759
No degree	1,035,614	1,473,779	2,509,393
Elementary school	3,651,852	5,122,712	8,774,564
Middle school	8,633,787	7,583,211	16,216,998
High school	9,841,286	9,735,915	19,577,201
Tertiary degree	3,464,025	4,384,481	7,848,506
Occupational status			
Age 0-14	3,968,042	3,749,562	7,717,604
Employed	13,534,829	10,003,542	23,538,371
Job seeker	1,777,939	1,781,571	3,559,510
Student	1,937,400	2,138,937	4,076,337
Retired	5,622,292	5,792,930	11,415,222
Homemaker/Other	2,011,558	6,940,819	8,952,377
Citizenship			
Italian	26,435,456	27,806,344	54,241,800
Foreigner	2,416,604	2,601,017	5,017,621
Total	28,852,060	30,407,361	59,259,421

Source: Own elaboration of data from the Italian Permanent Census of Population and Housing (2019).

If we denote by a_{2i} the weight calibration factor computed for individual i, the final individual weights w_{1i} were obtained as follows:

$$w_{1i} = d_{1i} \times a_{2i} \tag{2.24}$$

where d_{1i} is defined as in Equation 2.23.

2.7. *Weighted Sample Representativeness*

As mentioned at the beginning of this chapter, the main goal of survey weights is to help obtain (approximately) unbiased estimators for the quantities of interest in the target population. In practice, this amounts to improving the extent to which the survey sample is representative of the target population. The aim of this section is to assess whether the survey weights we constructed go in this direction, i.e., increase the representativeness of the ITA.LI realized sample.

To make our assessment, we might be tempted to determine whether the lack of representativeness found in the unweighted sample with respect to a basic set of variables (see Section 1.8) is remedied – at least partially – in the weighted sample. This approach, however, would be tautological, since almost all the variables used to assess the representativeness of the ITA.LI unweighted sample were also employed to calibrate the survey weights.

Expressly, survey weight calibration ensures that a weighted sample is perfectly representative of the target population with respect to the variable distributions used for calibration. As far as we are concerned, this means that the ITA.LI weighted realized sample is perfectly representative of the target population with respect to (a) the joint distribution of variables *Area of residence* and *Household size* at the household level (see Table 2.1); and (b) a select subset of second-order marginal distributions of $\mathbf{X} = \{$*Sex*, *Birth cohort*, *Region of residence*, *Educational degree*, *Occupational status*, *Citizenship*$\}$ at the individual level (see Table 2.2). Therefore, using any of these variables – either individually or combined as indicated above – to assess the representativeness of the ITA.LI weighted realized sample would be pointless, as the fit between the sample and the target population would be perfect by design.

All this considered – and given the information available about the target population – we will base our assessment on just a limited set of elements: (a) the univariate distribution of variable *Marital status*; and (b) a set of second-order marginal distributions of $\mathbf{X}$ not already employed for survey weight calibration.[3]

Figure 2.2 reports the results of our assessment. The graph displays two versions of the index of dissimilarity (D) between the sample and population variable distributions: one for the unweighted sample (blue empty bar) and the other for the weighted sample (orange filled bar).[4] As we can see, weighting improves the representativeness of the ITA.LI realized sample in all cases considered. The reduction in dissimilarity between the sample and the target population that is gained by switching from the unweighted to the weighted

3 Because of the limited data available to the purpose, the assessment of the representativeness of the ITA.LI weighted realized sample is carried out only at the individual level.

4 The index of dissimilarity D was computed using the user-written Stata command `reldist` (Jann 2021). All variables are defined as in Section 1.8 except for variable *Birth cohort*, whose definition varies according to the variable with which it is cross-tabulated. Specifically, when combined with variable *Region of residence*, variable *Birth cohort* is divided in eight categories as follows: Up to 1939, 1940-49, 1950-59, 1960-69, 1970-79, 1980-89, 1990-99, 2000-03; when combined with variables *Educational degree* and *Occupational status*, it is divided in three categories: Up to 1954, 1955-69, 1970-94; and when combined with variable *Citizenship*, it is divided in seven categories: Up to 1939, 1940-49, 1950-59, 1960-69, 1970-79, 1980-89, 1990-99.

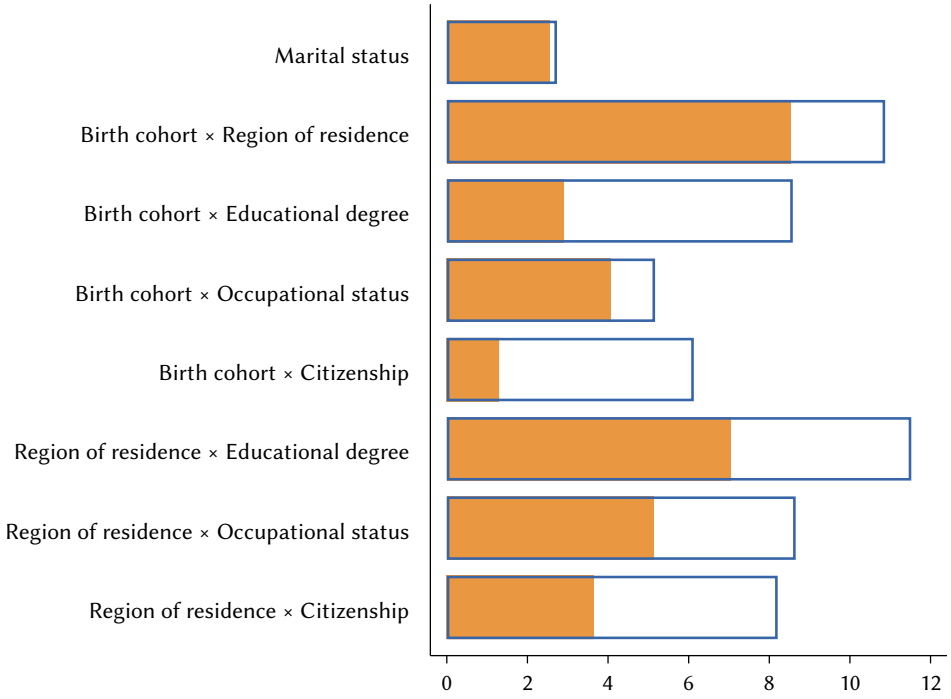


Figure 2.2 Individual-level representativeness of the ITA.LI realized sample with respect to a select set of variable distributions. The graph displays two versions of the index of dissimilarity (D) between the sample and population variable distributions: one for the unweighted sample (blue empty bar) and the other for the weighted sample (orange filled bar).

sample is smallest for the univariate distribution of variable *Marital status* (−7% in relative terms), while it takes its maximum value for the joint distribution of variables *Birth cohort* and *Citizenship* (−79%). On average, the relative reduction in the index of dissimilarity D due to survey weighting is about −40%. Given that the remaining dissimilarity could largely fall within the margins of random estimation error (see Chapter 3), this is a good result.

2.8. *Cohort Survival Rates*

One of the defining goals of a longitudinal study such as ITA.LI is to analyze the ways and extent to which several phenomena of interest vary across birth cohorts. This comparative analysis, however, typically suffers from the fact that, at the time the study is carried out, the members of the various cohorts eligible to participate in the study are only a subset of all people born in those cohorts, as some have since emigrated and others have died. Now, since the

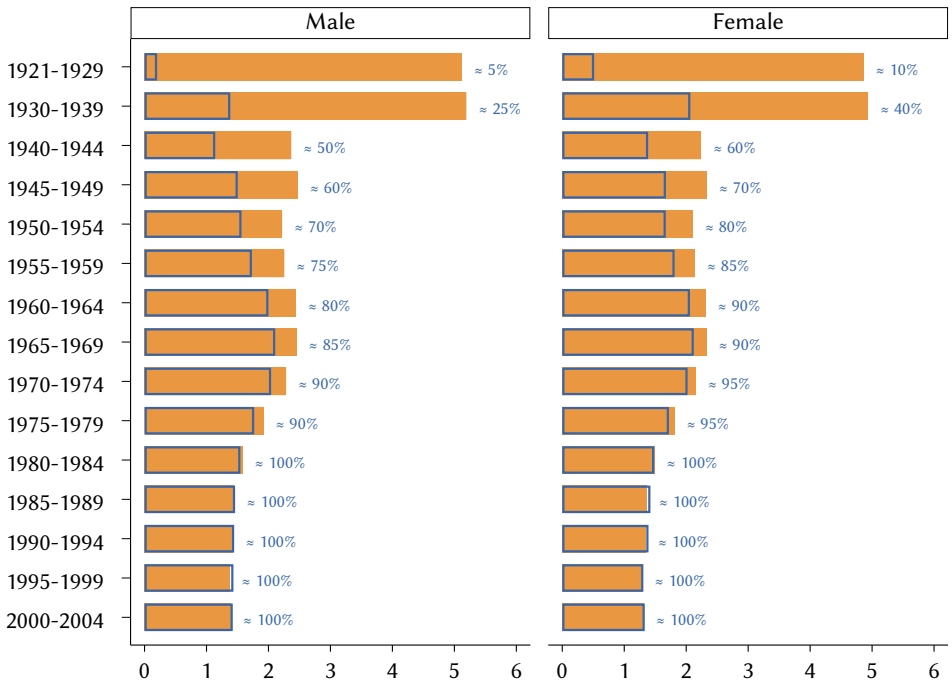


Figure 2.3 Registered number of births in Italy (in millions, orange filled bar), number of Italy-born individuals residing in Italy at study time estimated by the ITA.LI weighted realized sample (in millions, blue empty bar), and approximate cohort survival rate, by birth cohort and sex.

propensity to emigrate and life expectancy are socially differentiated, it is reasonable to expect that the "surviving" members of the various cohorts at study time will not be socially representative of all individuals born in the respective cohorts. This representativeness deficit will be substantially irrelevant when the cohort survival rate is high, but will become increasingly critical as the survival rate decreases. Therefore, to properly assess the potential and limitations of a cohort analysis, it may be useful to have an estimate of the survival rate of each cohort under examination.

Figure 2.3 displays the approximate cohort survival rates estimated for present-day Italy using the ITA.LI weighted realized sample.[5] As can be seen,

5 The registered numbers of births were taken from the *Human Mortality Database*, University of California at Berkeley and Max Planck Institute for Demographic Research (https://www.mortality.org). The numbers of Italy-born individuals residing in Italy at study time were estimated using the ITA.LI weighted realized sample and auxiliary information on place of birth taken from the 2011 Italian Census of Population (http://dati-censimentopopolazione.istat.it/Index.aspx).

the male cohorts since 1960 and the female cohorts since 1950 are characterized by fairly high survival rates ($\geq$ 80%). On the other hand, men and women born in the 1920s and 1930s are decidedly underrepresented in today's Italy. The remaining birth cohorts (1940-59 for men, 1940-49 for women) are in an intermediate situation. Overall, these data suggest that any cohort analysis covering the period between the immediate post-World War II era and the early 2000s is unlikely to be affected by any severe selection bias. Conversely, birth cohorts from before the end of World War II should be omitted from the analysis or, if included, treated with great caution.

3

VARIANCE ESTIMATION

3.1. *Introduction*

As noted in the first chapter, the basic purpose of survey research is to learn about a given population of interest through the study of a subset of it, called a *sample*. By definition, generalization from a part (here, the sample) to the whole (here, the target population) is an imperfect process, as it relies on incomplete data about the object of study (Copi *et al.* 2019). This implies that all results of sample survey research will always be affected by some degree of *uncertainty*, whose magnitude must be estimated if such results are to be interpretable (King *et al.* 1994). This is the task of *statistical inference*.

In survey research, the inferential problem – i.e., gauging the uncertainty that surrounds sample results – is usually approached from one of two main perspectives: the design-based theory of inference and the model-based theory of inference (Särndal 1978, 1985). Briefly, in the *design-based* approach, the objects of inference are the true values of the quantities of interest (means, proportions, regressions coefficients, and so on) in a finite target population. In the *model-based* approach, on the other hand, the objects of inference are the properties of a random process – or, equivalently, the parameters of a stochastic model – that is assumed to have generated the target population. Here we will only consider the design-based approach.

The starting point of design-based inference[1] is a *finite population* $\mathcal{U}$, defined as a set of N units of analysis circumscribed in time and space. Within $\mathcal{U}$, a given quantity of interest Q takes a fixed value θ. Suppose we are interested in knowing θ. To this end, first, using a certain sample design, we select a subset of $\bar{n}$ units from $\mathcal{U}$, which we denote by $\mathcal{S}$. Then, we collect the relevant data on the sampled units. Third, we define an *estimator* of Q, that is, a procedure – which we denote by $\hat{\Theta}$ – to calculate Q from the available

1 The following discussion is largely based on Lohr (2022, pp. 34-49).

data. Finally, we apply the chosen estimator to the collected sample data, thus obtaining a *sample estimate* of Q, which we denote by $\hat{\theta}_{\mathcal{S}}$.

The problem with $\hat{\theta}_{\mathcal{S}}$ is that, in general, it deviates to some extent from the true value of Q in the target population. Formally:

$$\hat{\theta}_{\mathcal{S}} = \theta + \epsilon_{\mathcal{S}} \tag{3.1}$$

where $\epsilon_{\mathcal{S}}$ denotes the *estimation error*, i.e., the difference between the sample estimate of Q and its true population value.

The estimation error $\epsilon_{\mathcal{S}}$ summarizes the action of a number of factors that affect the life cycle of any sample survey – first of all the fact that the data used to calculate $\hat{\theta}_{\mathcal{S}}$ are incomplete, as they pertain only to a subset of the target population (Groves 1989; Groves *et al.* 2009). If it were possible to quantify exactly the impact of all these factors on the sample estimate of Q, then statistical inference would be a deterministic process, for in this case the value of $\epsilon_{\mathcal{S}}$ would be known and, therefore, we could tell the true population value of Q by simply subtracting from its sample estimate $\hat{\theta}_{\mathcal{S}}$ the estimation error $\epsilon_{\mathcal{S}}$. The value of $\epsilon_{\mathcal{S}}$, however, can never be determined with certainty, so that exact knowledge of θ is precluded to sample survey research.

While finding the exact value of $\epsilon_{\mathcal{S}}$ in a given sample $\mathcal{S}$ is unfeasible, according to the design-based approach to statistical inference it is nonetheless possible to estimate the *probabilities of occurrence* of the different values of $\epsilon_{\mathcal{S}}$ – and, therefore, of $\hat{\theta}_{\mathcal{S}}$ – over hypothetical *repeated sampling*. Knowledge of such probabilities would allow one to quantify how likely it is that a given sample is a "good" one, that is, a sample whose estimate of the quantity of interest Q is relatively close to the true population value of Q. For this possibility to be realized, however, a basic condition must be met: samples must be selected by a well-defined *random mechanism*, that is, a *probability sample design* whereby each sample has a known, nonzero probability of being selected.

To illustrate the concept of "repeated sampling" and its implications for design-based statistical inference, let us go back to the previous example and suppose that our research goal is still to learn about θ – that is, the true value of the quantity of interest Q in the target population $\mathcal{U}$ – using sample data and a given estimator $\hat{\Theta}$. To this purpose, we opt for a sample of size $\bar{n}$ selected from $\mathcal{U}$ using a probability sample design $\mathbb{D}$.

Any given probability sample design $\mathbb{D}$ can generate a finite number of samples of size $\bar{n}$ from $\mathcal{U}$. Let $\Omega_{\mathbb{D}(\bar{n})}$ denote the finite set of all samples of size $\bar{n}$ permissible under $\mathbb{D}$, and let S denote the cardinality of the set.

Now, suppose that we (a) select sample $\mathcal{S} = 1$ from $\Omega_{\mathbb{D}(\bar{n})}$; (b) collect the relevant data on the responding subset of $\mathcal{S} = 1$, which we denote by $\mathcal{R}_1$;

(c) apply the estimator $\hat{\Theta}$ to the collected data; and (d) obtain a sample estimate of Q defined as in Equation 3.1, that is: $\hat{\theta}_1 = \theta + \epsilon_1$. Let us repeat steps (a)-(d) for all the remaining $S - 1$ samples in $\Omega_{\mathbb{D}(\bar{n})}$. In the end, we will have S sample estimates $\hat{\theta}_{\mathcal{S}}$, one for each sample of size $\bar{n}$ permissible under $\mathbb{D}$.

Since the action of the factors determining $\epsilon_{\mathcal{S}}$ tends to fluctuate stochastically with each sample $\mathcal{S}$, the value of the sample estimate $\hat{\theta}_{\mathcal{S}}$ generated by the estimator $\hat{\Theta}$ will vary accordingly across the S samples. The set of all possible values of $\hat{\theta}_{\mathcal{S}}$ obtained from our hypothetical repeated-sampling procedure, together with their probabilities of occurrence, form the *probability distribution* of the estimator $\hat{\Theta}$, also known as its *sampling distribution*.

For inferential purposes, the essential characteristics of the sampling distribution of an estimator $\hat{\Theta}$ are three: expected value, variance and shape. The *expected value* is defined as follows:

$$E(\hat{\Theta}) = \frac{\sum_{\mathcal{S}=1}^{S} \hat{\theta}_{\mathcal{S}}}{S} \tag{3.2}$$

where all symbols are defined as above. Basically, the expected value represents the average of all sample estimates $\hat{\theta}_{\mathcal{S}}$.

A key property of the estimator of any quantity of interest Q is its ability to generate sample estimates of Q that, *on average*, match the true population value of Q. The extent to which the estimator deviates from this optimal performance amounts to its *bias*, formally defined as follows:

$$B(\hat{\Theta}) = E(\hat{\Theta}) - \theta \tag{3.3}$$

If $E(\hat{\Theta}) = \theta$, then the bias is zero and we say that the estimator is *unbiased* – i.e., hits the estimation target on average. Conversely, if $E(\hat{\Theta}) \neq \theta$, then we say that the estimator is *biased* – i.e., off the estimation target on average – by an amount equal to $B(\hat{\Theta})$.

The ability of an estimator to be "correct" on average – that is, unbiased – is certainly important, but it is not sufficient to determine its quality. There may in fact be estimators that are unbiased but, at the same time, poorly informative because they generate highly variable sample estimates. Along with unbiasedness, then, the second major property of an estimator is *precision*, that is, its ability to generate sample estimates that, *on average*, are close to each other. This property is represented by the *variance* of the estimator, defined as follows:

$$V(\hat{\Theta}) = \frac{\sum_{\mathcal{S}=1}^{S} \left(\hat{\theta}_{\mathcal{S}} - E(\hat{\Theta})\right)^2}{S} \tag{3.4}$$

where all symbols are defined as above. As we can see, the variance equals the average (squared) deviation of sample estimates from their expected value. The smaller such deviation, the higher the precision of the estimator.

Combined together, unbiasedness and precision express the *accuracy* of an estimator, which, in mathematical terms, corresponds to its *mean squared error*. Formally:

$$\begin{aligned} \text{MSE}(\hat{\Theta}) &= \frac{\sum_{S=1}^{S} (\hat{\theta}_S - \theta)^2}{S} \\ &= V(\hat{\Theta}) + B(\hat{\Theta})^2 \end{aligned} \tag{3.5}$$

where all symbols are defined as above. This expression shows that the mean squared error equals the average (squared) deviation of sample estimates from the true population value of the quantity of interest. The smaller such deviation, the higher the accuracy of the estimator and, ultimately, the lower the overall uncertainty that surrounds sample estimates.

If we assume that an estimator is asymptotically unbiased, then its mean squared error will correspond to the sampling variance which, consequently, will account for all of the uncertainty of survey results. In this case, uncertainty will be best represented by confidence intervals. A *confidence interval* is a range of contiguous values built around a sample estimate of the quantity of interest Q, in which the true population value of Q is expected to lie with a certain probability. The length of this range is a function of the variance of the estimator of Q: the smaller the variance, the shorter the confidence interval, and the lower the uncertainty surrounding the estimate of Q.

To build an appropriate confidence interval, it is necessary to know the *shape* of the sampling distribution of the estimator of Q. Statistical theory shows that if the sample size is sufficiently large and $\hat{\Theta}$ is an asymptotically unbiased estimator of Q, the sampling distribution of $\hat{\Theta}$ will be asymptotically *normal*, with mean (approximately) equal to the true population value of Q and variance as defined in Equation 3.4 (Heeringa *et al.* 2017). Formally:

$$\hat{\Theta} \overset{a}{\sim} \text{N}\left(\theta, V(\hat{\Theta})\right) \tag{3.6}$$

where N denotes the normal probability distribution; and all other symbols are defined as above.

When (3.6) holds, an approximate $100(1-\alpha)\%$ Wald confidence interval for the true population value of Q may be constructed as follows:

$$\hat{\theta}_S \pm z_{1-\alpha/2}\sqrt{V(\hat{\Theta})} \tag{3.7}$$

where $100(1-\alpha)\%$ denotes the confidence level; $z_{1-\alpha/2}$ denotes the $100(1-\alpha/2)^{\text{th}}$ percentile of the standard normal distribution; and all other symbols are defined as above. Sometimes, $z_{1-\alpha/2}$ is replaced with $t_{1-\alpha/2,df}$, that is, the $100(1-\alpha/2)^{\text{th}}$ percentile of the t distribution with df degrees of freedom. When df is sufficiently large, the two percentiles are approximately equal; in smaller samples, $t_{1-\alpha/2,df} > z_{1-\alpha/2}$, so that the use of the t-distribution percentile will yield a wider (more conservative) confidence interval.

From a design-based perspective, the confidence interval (3.7) can be given the following interpretation: if such a confidence interval were constructed for each of the S samples in $\Omega_{\mathbb{D}(\bar{n})}$, approximately $100(1-\alpha)\%$ of the resulting intervals would include the true population value of Q.

The problem with Equation 3.7 is that the repeated-sampling procedure described above is just hypothetical. In real-world survey research, only one sample is selected, which gives no chance to calculate the full sampling distribution of $\hat{\Theta}$ and its variance. In a real research setting, therefore, the variance of estimators must itself be *estimated* using the single-sample data at hand (Wolter 2007). When – as in the case of ITA.LI – data come from a complex sample survey, variance estimation is far from straightforward, for in principle it should fully account for all sample design elements (like stratification and clustering) and survey weighting (Valliant and Dever 2018).

There exist several alternative methods of variance estimation to deal with complex survey data under the design-based inferential framework, the most popular ones being *linearization* and *replication* (Wolter 2007). The latter, in turn, includes three main approaches: the *jackknife*, *balanced repeated replication*, and the *bootstrap*. In large samples, all these methods are asymptotically equivalent and, therefore, equally suitable (see, *inter alia*, Heeringa *et al.* 2017; Kolenikov 2010; Lohr 2022; Pedlow 2008; Rust 1985; Rust and Rao 1996; Wolter 2007). However, some practical differences between methods may affect the choice of the one best suited for a specific study.

Compared with linearization, replication methods offer two main advantages (Valliant and Dever 2018). On the one hand, they allow to account for all steps in the procedure of weight construction and, therefore, to fully capture its impact on the variance of the estimators of interest. On the other hand, since they do not require the explicit use of stratum and sampling unit identifiers for the correct estimation of the variance of the estimators of interest, they help protect the anonymity of respondents.

Within replication methods, balanced repeated replication and the bootstrap are more versatile than the jackknife since, unlike the latter, they can be used also for nonsmooth estimators, such as the median and other percentiles

(Rust and Rao 1996). Balanced repeated replication and the bootstrap are similar under several respects (Kolenikov 2010; Rust and Rao 1996), but balanced repeated replication may be less computationally demanding (Shao 1996).

For all these reasons, balanced repeated replication was chosen as the variance estimation method for ITA.LI. The next section is devoted to an overview of the basic features of this method. The third section illustrates how the method was implemented in ITA.LI. The chapter ends with a brief assessment of the quality of ITA.LI variance estimates.

3.2. *Balanced Repeated Replication*

The *balanced repeated replication* (BRR) method of variance estimation was proposed by McCarthy (1966, 1969) to deal with sample designs in which exactly two PSUs are selected from each of H strata – or designs that can be reformulated into two PSUs per stratum (Lohr 2022).

The basic idea behind this method is to treat the original sample of the survey of interest – which we will refer to as the *full sample* and denote by $\mathcal{S}$ – as if it were a population from which R subsamples, called *replicates*, are systematically drawn. Each replicate is then used to calculate an estimate of the quantity of interest Q. The variance of the resulting R estimates is taken as an estimator of the variance of the estimator of Q (Wolter 2007).

Let us take a closer look at the whole procedure, starting with the generation of replicates. The BRR method requires that each replicate be a half-sample formed by simply selecting one PSU from each stratum. In a design with H strata and two PSUs per stratum, it is possible to generate 2^H different replicates of this kind. In principle, all of them should be used to compute the BRR estimate of the variance of interest without loss of information (Wolter 2007). When H is large, however, the generation of the complete set of half-sample replicates becomes unfeasible. McCarthy (1966, 1969) showed that, when facing this case, it is still possible to obtain a maximally efficient estimator of the variance of interest by analyzing a relatively small subset of the complete set of half-sample replicates. Such a subset, called the *balanced set*, must include R half-sample replicates, where R is the smallest multiple of 4 that is greater than or equal to H.

The R half-sample replicates that form the balanced set are identified by means of a Hadamard matrix of order R, that we will denote by $\mathbf{B}$. The latter is a square matrix of order R whose elements are either $+1$ or -1, and whose rows and columns are pairwise orthogonal. In the context of BRR, the rows of $\mathbf{B}$ represent replicates, while columns represent strata. Whenever $R > H$,

only H columns are retained, so that $\mathbf{B}$ reduces to an $R \times H$ matrix (Wolter 2007).

The elements of matrix $\mathbf{B}$ specify which PSU should be selected from each stratum to form each half-sample replicate. If matrix element $b_{rh} = +1$, then the first PSU of stratum h should be selected for replicate r. Conversely, if matrix element $b_{rh} = -1$, then the second PSU of stratum h should be selected for replicate r.

Once the balanced set of half-sample replicates has been defined, the next move is to use such replicates to calculate R estimates of the quantity of interest Q. To this aim, for each replicate r, a three-step procedure is used. First, to compensate for the fact that only half of the original PSUs are retained, the base weights of the selected PSUs are multiplied by two. These inflated weights are then adjusted and, possibly, calibrated using exactly the same procedure used for the full sample. Finally, the resulting weights – called the *replicate weights* – are used to compute an estimate of Q, which we will call the *replicate estimate* and denote by $\hat{\theta}_{r(\text{BRR})}$ (Valliant and Dever 2018).

In the end, R replicate estimates are obtained. Their variance can be taken as a consistent estimator of the variance of the estimator of Q (Wolter 2007). Formally:

$$\widehat{V}(\hat{\Theta})_{\text{BRR}} = \frac{\sum_{r=1}^{R} (\hat{\theta}_{r(\text{BRR})} - \hat{\theta}^*)^2}{R} \tag{3.8}$$

where $\hat{\theta}^*$ can be either the full-sample estimate of Q or the average of the R replicate estimates.

Equation 3.8 can be used to express the inferential uncertainty that surrounds the full-sample estimate of Q by computing an appropriate confidence interval. Specifically, an approximate $100(1 - \alpha)\%$ Wald confidence interval for θ may be constructed as follows:

$$\hat{\theta}_S \pm t_{1-\alpha/2,df} \sqrt{\widehat{V}(\hat{\Theta})_{\text{BRR}}} \tag{3.9}$$

where $\hat{\theta}_S$ denotes the full-sample estimate of Q; $df = H$; and all other symbols are defined as above.

The standard BRR method of variance estimation suffers one major drawback: removing half of the PSUs from each replicate halves the information available to calculate each replicate estimate, making it less precise or, in some cases, even unfeasible (Judkins 1990; Valliant and Dever 2018). A solution to this problem was devised by Fay (Dippo *et al.* 1984; Fay 1989), who proposed forming the replicates not by completely dropping one PSU from each stratum, but rather by retaining both PSUs and multiplying their base weights by

different adjustment factors. The latter are defined as follows (Judkins 1990; Valliant and Dever 2018):

$$\begin{aligned} f_{1hr(\text{Fay})} &= 1 + b_{rh}(1-\rho) \\ f_{2hr(\text{Fay})} &= 1 - b_{rh}(1-\rho) \end{aligned} \tag{3.10}$$

where $f_{1hr(\text{Fay})}$ denotes the base weight adjustment factor for the first PSU of stratum h in replicate r; $f_{2hr(\text{Fay})}$ denotes the base weight adjustment factor for the second PSU of stratum h in replicate r; $1-\rho$ is the *perturbation factor* ($0 \leq \rho < 1$); and b_{rh} is defined as above. It should be noted that when $\rho = 0$, Equation 3.10 describes the standard BRR.

Once the PSU base weights have been adjusted by factors $f_{1hr(\text{Fay})}$ and $f_{2hr(\text{Fay})}$, the procedure for calculating the R replicate estimates of the quantity of interest Q runs as indicated above for the standard BRR. In the end, such estimates – that we will denote by $\hat{\theta}_{r(\text{Fay})}$ – are plugged into the formula for the Fay variance estimator, defined as follows:

$$\widehat{V}(\hat{\Theta})_{\text{Fay}} = \frac{\sum_{r=1}^{R} (\hat{\theta}_{r(\text{Fay})} - \hat{\theta}^*)^2}{R(1-\rho)^2} \tag{3.11}$$

where all symbols are defined as above.

An approximate $100(1-\alpha)\%$ Wald confidence interval for θ can then be computed by replacing $\widehat{V}(\hat{\Theta})_{\text{BRR}}$ with $\widehat{V}(\hat{\Theta})_{\text{Fay}}$ in Equation 3.9:

$$\hat{\theta}_S \pm t_{1-\alpha/2,df} \sqrt{\widehat{V}(\hat{\Theta})_{\text{Fay}}} \tag{3.12}$$

where all symbols are defined as above.

3.3. *Implementation of BRR*

As mentioned above, balanced repeated replication was chosen as the variance estimation method for ITA.LI. Specifically, we opted for Fay's variant of BRR.[2] This section illustrates how the method was implemented in ITA.LI.

2 Fay's variant of the BRR variance estimation method is used in many large-scale surveys, such as the *American Community Survey* (U.S. Census Bureau), the *Current Population Survey's Annual Social and Economic Supplement* (U.S. Census Bureau and Bureau of Labor Statistics), the *Current Population Survey's Tobacco Use Supplement* (U.S. Census Bureau and National Cancer Institute), the *Medical Expenditure Panel Survey* (U.S. Agency for Healthcare Research and Quality), the *National Agricultural Workers Survey* (U.S. Department of Labor), the *Programme for International Student Assessment* (OECD), the *Survey of Income and Program Participation* (U.S. Census Bureau), and the *Teaching and Learning International Survey* (OECD).

First, it was necessary to rearrange PSUs. As described in Chapter 1, the ITA.LI sample consists of 280 PSUs drawn from 150 strata. From 130 of these strata, exactly two PSUs per stratum were selected. The remaining 20 strata contain a single self-representing PSU each. While the strata of the former type fit BRR perfectly, the latter do not and, therefore, had to be rearranged to mimic the two-PSU-per-stratum logic. The rearrangement consisted in randomly partitioning the SSUs belonging to each self-representative PSU into two equal-sized groups, which were then treated – for the purposes of BRR – as the two (pseudo) PSUs of the corresponding stratum (Chen and Parker 2016; Korn and Graubard 1999; Potter *et al.* 2003). As a result, all 150 strata ended up containing exactly two (actual or pseudo) PSUs each.

Given $H = 150$, the balanced set of half-sample replicates was identified by means of a Hadamard matrix of order $R = 152$ – the latter being the smallest multiple of 4 greater than or equal to 150.[3] Since R exceeds H by two units, the first two columns of the Hadamard matrix were dropped, so as to obtain the desired 152×150 matrix $\mathbf{B}$ of ones and minus ones.

Finally, the two variable elements of Equation 3.11 were defined as follows: (a) the value of the perturbation factor was set to 0.5 (equivalently, $\rho = 0.5$), as suggested by Judkins (1990, see also Rao and Shao 1999); and (b) $\hat{\theta}^*$ was defined to be $\hat{\theta}_S$, that is, the full-sample estimate of Q.

The above setup was used to generate $R = 152$ sets of household-level replicate weights w^r_{Hj} $(r = 1, \ldots, R)$ and an equal number of sets of individual-level replicate weights w^r_{Ii} $(r = 1, \ldots, R)$. These weights (either household-level or individual-level) will be used to calculate $R = 152$ replicate estimates of the quantity of interest Q, which, in turn, will form the basis for estimating the variance of the estimator of Q as per Equation 3.11.

3.4. *Quality of Variance Estimates*

As noted in the opening of this chapter, the variance of survey estimators plays a major role in survey data analysis, since – assuming no or negligible bias – it conveys the precision of the sample estimates of the quantities of interest and, therefore, the degree of inferential uncertainty surrounding them. In any sample survey, then, it is important to obtain variance estimates that are as accurate as possible, so as to properly gauge the ability of the sample estimates of the quantities of interest to provide a reliable – i.e., close to reality – description of the phenomenon under study in the target population.

3 The Hadamard matrix was taken from http://neilsloane.com/hadamard/had.152.pal.txt.

This section provides a brief assessment of the quality of ITA.LI variance estimates. In the first part of the procedure, a small set of simple quantities of interest – means and proportions – was selected and, for each of them, an estimate of the variance of the corresponding estimator was computed on the ITA.LI weighted realized sample using the Fay's variant of BRR (see Section 3.3). The variance estimates, which we denote by $\hat{v}(\hat{\Theta})_{\text{Fay}}$, were then converted into the corresponding *standard errors* by taking their square root. Formally:

$$\widehat{se}(\hat{\Theta})_{\text{Fay}} = \sqrt{\hat{v}(\hat{\Theta})_{\text{Fay}}} \tag{3.13}$$

Finally, such standard error estimates were evaluated using three measures: the estimation method effect, the design effect, and the coefficient of variation. In the second part of the procedure, the same type of evaluation was applied to somewhat more elaborate quantities of interest: the regression coefficients of a linear model. In this case, however, the assessment was based on only two measures: the estimation method effect and the design effect.

The three measures used in our evaluation tap different dimensions of the quality of variance estimates. The first has to do with the quality of the *method* used to estimate variances. As explained in detail in the previous sections, the Fay's variant of BRR was chosen as the variance estimation method for ITA.LI. According to the relevant literature, this method has the same asymptotic properties as all other available variance estimation methods (Wolter 2007). In finite samples, however, this uniformity may not hold and, therefore, it is useful to subject Fay's variance estimation method to a comparative evaluation. To this aim, we confronted Fay's method with *Taylor series linearization*, the standard variance estimation method for complex sample survey data which is generally regarded as fully efficient (Kim and Wu 2013; Lohr 2022). Specifically, we computed the *estimation method effect*, here defined as follows:

$$esteff = \frac{\widehat{se}(\hat{\Theta})_{\text{Fay}}}{\widehat{se}(\hat{\Theta})_{\text{TSL}}} \tag{3.14}$$

where $\widehat{se}(\hat{\Theta})_{\text{Fay}}$ denotes the sample estimate of the standard error of the estimator of a given quantity of interest computed on the ITA.LI weighted realized sample using the Fay's variant of BRR; and $\widehat{se}(\hat{\Theta})_{\text{TSL}}$ denotes the sample estimate of the same standard error computed on the ITA.LI weighted realized sample using Taylor series linearization. On average, $esteff$ should approach unity, attesting to the validity of Fay's method (Rao and Shao 1999).

The second measure we used in our evaluation has to do with the *efficiency* of the ITA.LI sample design. As explained in the previous chapters,

ITA.LI adopted a complex multistage sample design, and its base weights went through an elaborate process of adjustment and calibration resulting in a highly variable set of final weights. Compared with a simpler sample design, a complex design like this tends to be significantly less efficient – i.e., to yield larger variance estimates (Heeringa *et al.* 2017). To quantify this loss of efficiency, we computed the *design effect*, here defined as follows (Kish 1995):

$$deft = \frac{\widehat{se}(\hat{\Theta})_{\text{Fay}}}{\widehat{se}(\hat{\Theta})_{\text{SRSWR}}} \tag{3.15}$$

where $\widehat{se}(\hat{\Theta})_{\text{Fay}}$ denotes again the sample estimate of the standard error of the estimator of a given quantity of interest computed on the ITA.LI weighted realized sample using the Fay's variant of BRR; and $\widehat{se}(\hat{\Theta})_{\text{SRSWR}}$ denotes the estimate of the same standard error that would be obtained if the same data had been collected using the most basic sample design possible: *simple random sampling with replacement* (SRSWR). As such, *deft* expresses the effect of the ITA.LI sample design, in combination with weight adjustment and calibration, on the precision of the estimators of the quantities of interest. If $deft > 1$, then we can conclude that the ITA.LI sample design, combined with weight adjustment and calibration, inflate the standard error of the estimator of interest by $100(deft - 1)\%$ compared to a hypothetical SRSWR sample design.

The third measure we used in our evaluation – limited to means and proportions – has to do with the *relative precision* of the sample estimates of the quantities of interest and corresponds to the *coefficient of variation*, here defined as follows:

$$\text{CV} = \frac{\widehat{se}(\hat{\Theta})_{\text{Fay}}}{\hat{\theta}_S} \times 100 \tag{3.16}$$

where $\widehat{se}(\hat{\Theta})_{\text{Fay}}$ is defined as above; and $\hat{\theta}_S$ denotes the full-sample estimate of the quantity of interest Q. The coefficient of variation can be regarded as an indicator of the reliability of a sample estimate $\hat{\theta}_S$ (Shapiro 2008b; Valliant *et al.* 2018). In general, the smaller the value of a CV, the higher is the level of relative precision of the corresponding estimate and, therefore, the closer is that estimate to the true population value of Q. Currently, there is no agreement on what the minimum acceptable level of estimates' relative precision should be. Based on the evaluation of a number of U.S. Census Bureau publications, the U.S. Office of Financial Management (Gardner *et al.* 2015) recommends the following classification of estimates' reliability: good ($\text{CV} \leq 15\%$), fair ($15\% < \text{CV} \leq 30\%$), and poor ($\text{CV} > 30\%$). Statistics Canada (2020), in turn, uses the following reliability categories: if $\text{CV} \leq 16.5\%$, estimates are regarded as sufficiently reliable; if $16.5\% < \text{CV} \leq 33.3\%$, it is

Table 3.1 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Number of valid cases used in the analysis, by quantity of interest and target subpopulation. Table rows denote the quantities of interest considered in the analysis, while table columns denote the subpopulations within which the quantities of interest have been computed.

	Sex		*Level of education*			
	Male	Female	Low	Medium	High	*Total*
Age						
Mean	4,010	4,768	4,382	3,382	1,014	8,778
Life satisfaction						
Mean	3,995	4,752	4,366	3,375	1,006	8,747
Weight						
Mean	3,436	3,930	3,620	2,907	839	7,366
Height						
Mean	3,731	4,342	4,003	3,142	928	8,073
Has children						
% Yes	2,262	3,052	2,993	1,821	500	5,314
% No	1,685	1,642	1,304	1,521	502	3,327
Self-reported health						
% Excellent	705	601	457	628	221	1,306
% Good	2,260	2,704	2,126	2,192	646	4,964
% Satisfactory	815	1,099	1,329	467	118	1,914
% Poor	159	271	351	64	15	430
% Bad	56	62	95	16	7	118
Home internet access						
% Yes	2,770	3,087	2,223	2,743	891	5,857
% No, can't afford	156	195	281	64	6	351
% No, other	1,084	1,486	1,878	575	117	2,570

recommended that estimates be considered with caution; if CV $>$ 33.3%, estimates are deemed unreliable. Finally, according to the U.S. Census Bureau Statistical Quality Standards (U.S. Census Bureau 2022), in most cases CV should not exceed 30%.

Table 3.1 lists the means and proportions used in the first part of the evaluation procedure. As can be seen, the means of four quantitative variables (*Age*, *Life satisfaction*, *Weight*, and *Height*) and the percent distributions of three categorical variables (*Has children*, *Self-reported health*, and *Home internet access*) were considered. These quantities were calculated both within the entire target population and within subpopulations defined either by sex or by level

Table 3.2 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Estimation method effect, by quantity of interest and target subpopulation. Table rows denote the quantities of interest considered in the analysis, while table columns denote the subpopulations within which the quantities of interest have been computed. For the definition of estimation method effect, see Equation 3.14.

	Sex		*Level of education*			
	Male	Female	Low	Medium	High	*Total*
Age						
Mean	1.01	1.01	1.03	1.08	1.05	1.03
Life satisfaction						
Mean	1.05	1.04	1.03	1.04	1.05	1.05
Weight						
Mean	1.00	1.00	1.05	1.06	0.95	1.00
Height						
Mean	0.98	1.04	1.00	1.02	1.05	1.04
Has children						
% Yes	0.88	0.85	0.98	0.94	1.07	0.79
% No	0.88	0.85	0.98	0.94	1.07	0.79
Self-reported health						
% Excellent	1.01	1.03	1.07	1.06	0.99	1.01
% Good	1.06	1.09	1.03	1.03	1.00	1.03
% Satisfactory	1.01	1.02	1.02	0.98	0.86	1.00
% Poor	1.01	0.96	0.99	1.11	1.03	1.00
% Bad	0.97	0.93	0.93	0.96	1.01	0.96
Home internet access						
% Yes	1.00	1.03	1.00	1.03	0.93	1.02
% No, can't afford	1.00	1.02	1.00	1.33	1.07	1.03
% No, other	0.99	1.01	0.99	0.98	1.02	1.01

of education. The table reports the number of valid cases for each combination of quantity and subpopulation. The range of situations is very wide: it varies from proportions calculated on fewer than 10 cases, to means obtained from thousands of observations, allowing us a comprehensive evaluation of the ITA.LI variance estimates.

The estimation method effects reported in Table 3.2 confirm our expectations: Fay's variant of BRR and Taylor series linearization tend to yield very similar variance estimates. Indeed, there is little variation in the distribution of the estimation method effects considered in the analysis, most of which have values between 0.9 and 1.1, and whose geometric mean is exactly 1.

Table 3.3 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Design effect, by quantity of interest and target subpopulation. Table rows denote the quantities of interest considered in the analysis, while table columns denote the subpopulations within which the quantities of interest have been computed. For the definition of design effect, see Equation 3.15.

	Sex		*Level of education*			
	Male	Female	Low	Medium	High	*Total*
Age						
Mean	0.31	0.28	0.77	1.09	1.35	0.31
Life satisfaction						
Mean	1.98	1.73	1.96	1.73	1.67	2.37
Weight						
Mean	1.57	1.40	1.43	1.14	1.04	1.30
Height						
Mean	1.32	1.42	1.05	1.07	1.06	1.17
Has children						
% Yes	1.06	0.95	1.41	1.10	1.58	1.11
% No	1.06	0.95	1.41	1.10	1.58	1.11
Self-reported health						
% Excellent	1.47	1.41	1.66	1.56	1.55	1.74
% Good	1.49	1.41	1.47	1.49	1.59	1.65
% Satisfactory	1.25	1.16	1.40	1.31	1.11	1.28
% Poor	1.65	1.08	1.55	1.27	1.42	1.50
% Bad	1.88	1.09	1.67	1.19	1.23	1.66
Home internet access						
% Yes	1.41	1.24	1.50	1.61	1.39	1.63
% No, can't afford	1.64	1.40	1.77	1.73	2.13	1.84
% No, other	1.48	1.27	1.53	1.57	1.50	1.71

Table 3.3 shows the value taken by the design effect on each combination of quantity of interest and subpopulation. Just as we would expect (Heeringa *et al.* 2017), the efficiency of the ITA.LI sample design tends to be better than that of the benchmark sample design for those combinations involving only variables used for weight calibration (*Age*, *Sex*, and *Level of education*). In most cases, however, the design effect is greater than 1, thus confirming that clustering and (unequal) weighting push the variance estimates upward. The geometric mean of all design effects is 1.32, suggesting that, on average, the ITA.LI sample design inflates the standard errors of the corresponding estimators by slightly more than 30% compared to a hypothetical SRSWR sample design.

Table 3.4 Evaluation of the quality of ITA.LI variance estimates for means and proportions: Coefficient of variation, by quantity of interest and target subpopulation. Table rows denote the quantities of interest considered in the analysis, while table columns denote the subpopulations within which the quantities of interest have been computed. For the definition of coefficient of variation, see Equation 3.16.

	Sex		*Level of education*			
	Male	Female	Low	Medium	High	*Total*
Age						
Mean	0.18	0.16	0.40	0.71	1.24	0.12
Life satisfaction						
Mean	0.67	0.56	0.74	0.56	0.79	0.55
Weight						
Mean	0.41	0.40	0.45	0.44	0.59	0.29
Height						
Mean	0.09	0.08	0.09	0.10	0.15	0.07
Has children						
% Yes	1.44	1.04	1.43	1.86	4.31	0.96
% No	1.86	1.96	3.19	2.20	4.33	1.50
Self-reported health						
% Excellent	4.92	5.34	7.42	5.84	7.68	4.38
% Good	2.10	1.89	2.39	2.05	3.29	1.61
% Satisfactory	3.84	3.10	3.17	6.05	8.42	2.58
% Poor	11.07	6.17	7.19	16.26	32.55	6.51
% Bad	16.59	13.95	13.05	28.68	41.52	12.06
Home internet access						
% Yes	1.55	1.44	2.40	1.47	1.54	1.30
% No, can't afford	10.39	9.43	8.95	20.77	52.86	8.49
% No, other	3.63	2.67	2.52	6.26	10.51	2.73

Table 3.4 reports the coefficient of variation for each combination of quantity of interest and subpopulation. These values are definitely reassuring, as they testify that, in almost all cases considered here, the relative precision of sample estimates is fully satisfactory. Specifically, out of 84 estimates considered, 77 have a coefficient of variation less than 15%, four assume a value between 15% and 30%, and only three exhibit a coefficient of variation exceeding the critical threshold of 30%. It should be noted that the three estimates found to be unreliable refer to proportions calculated on a very low number of cases, equal to or less than 15. Our analysis, therefore, provides good grounds for concluding that, within ITA.LI, Fay's method of variance estimation is able

Table 3.5 Evaluation of the quality of ITA.LI variance estimates for the regression coefficients of a linear model: Estimation method effects (*esteff*) and design effects (*deft*). The model regresses variable *Life satisfaction* on one quantitative variable (*Age*) and five categorical variables (*Sex*, *Level of education*, *Has children*, *Self-reported health*, and *Home internet access*). Number of valid cases used in the analysis: $n = 8{,}575$. For the definition of estimation method effect and design effect, see respectively Equation 3.14 and Equation 3.15.

	esteff	*deft*
Age	0.98	1.34
Sex		
Male	– [a]	– [a]
Female	0.99	1.06
Level of education		
Low	– [a]	– [a]
Medium	1.08	1.27
High	1.07	1.52
Has children		
Yes	– [a]	– [a]
No	0.98	1.52
Self-reported health		
Excellent	– [a]	– [a]
Good	0.96	1.42
Satisfactory	1.00	1.39
Poor	1.03	1.34
Bad	0.99	1.82
Home internet access		
Yes	– [a]	– [a]
No, can't afford	1.03	1.74
No, other	1.03	1.52
Constant	1.01	1.53

[a] Reference category.

to generate sufficiently small variance estimates even when the number of cases is as low as $n = 20$.

Finally, Table 3.5 displays the estimation method effects and the design effects associated with the regression coefficients of a linear model. These values fully confirm our previous observations: on one hand, Fay's variance estimation method and Taylor series linearization tend to yield very similar variance estimates (geometric mean of estimation method effects = 1.01); on the other hand, the ITA.LI sample design is significantly less efficient than a

comparable SRSWR sample design (geometric mean of design effects = 1.44).

In summary, our assessment of the quality of ITA.LI variance estimates highlighted two main points. First, Fay's method of variance estimation yields variance estimates quite similar to those produced by Taylor series linearization, i.e., the standard variance estimation method for complex sample survey data, generally regarded as fully efficient. Second, although the ITA.LI sample design, due to clustering and (unequal) weighting, is significantly less efficient than an equivalent SRSWR sample design, its survey estimates still tend to be sufficiently reliable – i.e., close to the corresponding population values – even when based on a relatively small number of cases.

4

THE ANALYSIS OF ITA.LI DATA: A BRIEF GUIDE FOR STATA USERS

4.1. *Introduction*

To generate accurate estimates of the quantities of interest and proper measures of the uncertainty surrounding them, the analysis of complex survey data must be carried out so as to give due consideration to all relevant features of the sample design, most notably stratification, clustering and weighting. Failure to do so may lead to substantially biased estimation and inference, thus making survey results unreliable (Lohr 2022). Specifically, a "failure to account for sampling weights in estimation can substantially bias population estimates of key descriptive parameters, and a failure to account for complex sampling features when estimating the variances of estimates can lead to incorrect statements regarding sampling variability" (West *et al.* 2016). It is very important, then, that complex survey data users receive all the instruction they need to perform the analyses of interest correctly.

The purpose of this chapter is to provide guidance for proper design-based analysis of ITA.LI data using the Stata statistical software (StataCorp 2021c), adopted by the ITA.LI research team as the package of choice for data management and analysis.[1] The next section illustrates the structure and contents of the public-use version of the ITA.LI database, as well as the procedure for setting it up for analysis. The following five sections present – for illustrative purposes only – several example analyses of ITA.LI data: univariate analysis, bivariate analysis, multiple regression analysis, treatment effect estimation, and event history analysis.

1 For a recent comparative review of Stata's capabilities for survey data analysis, see West *et al.* (2018).

Table 4.1 Stata data files comprising the public-use version of the ITA.LI database: Structure.

File name	*Record type*	*No. of records*	*Subsample*	*No. of units*
`household_grid.dta`	Individuals	11,389	All members of responding eligible households	11,389
`personal_data.dta`	Individuals	8,778	All self-respondents	8,778
`residential_mobility.dta`	Episodes	22,779	All self-respondents	8,778
`education.dta`	Episodes	23,959	Self-respondents who ever enrolled in school	8,731
`job_history.dta`	Episodes	23,764	All self-respondents	8,778
`partnership_history.dta`	Episodes	6,523	Self-respondents who ever married or cohabited	6,149
`caring.dta`	Episodes	1,118	Self-respondents who ever cared for a relative	896
`financial_resources.dta`	Households	4,789[a]	All responding eligible households	4,789[a]
`proxy.dta`	Individuals	189	All proxy respondents	189
`brr-weights-hh.dta`	Households	4,900	All responding eligible households	4,900
`brr-weights-ind.dta`	Individuals	10,250	Self-respondent, proxy-respondent and ineligible members of responding eligible households	10,250

[a] One hundred eleven responding eligible households were excluded for providing insufficient data.

The discussion in this chapter assumes that the reader has a good working knowledge of statistics, at the level of Agresti (2018) or Knoke *et al.* (2002), and at least a basic familiarity with Stata, as provided for example by Kohler and Kreuter (2012) or Mehmetoglu and Jakobsen (2022).

4.2. *Setting Up Data for Analysis*

The public-use version of the ITA.LI database consists of 11 Stata data files, briefly described in Tables 4.1 and 4.2.[2] The files are of two types – rectangular and hierarchical – and are linked together by two common ID variables (keys): `W19HID` (Household ID) and `W19CID` (Individual ID).

To set up the data for analysis, the following four-step procedure is recommended. First, the variables of interest should be extracted from the existing

2 For details on the data collected by ITA.LI, see Lucchini *et al.* (2023).

Table 4.2 Stata data files comprising the public-use version of the ITA.LI database: Contents.

File name	*Contents*
`household_grid.dta`	Household composition, socio-demographic characteristics of household members
`personal_data.dta`	Pre-school education, military/civil service, family background, reproductive history, quality of life, Internet access and use, personality traits, health status, political preferences
`residential_mobility.dta`	Residential history
`education.dta`	Education history
`job_history.dta`	Labor force participation history
`partnership_history.dta`	Marriage and cohabitation history
`caring.dta`	Experience of caring for relatives
`financial_resources.dta`	Household financial status, household material resources, subsidies
`proxy.dta`	Proxy-respondent questionnaire
`brr-weights-hh.dta`	Household-level full-sample and replicate weights
`brr-weights-ind.dta`	Individual-level full-sample and replicate weights

data files and stored in one or more temporary data files. Note that this selection should always include the appropriate key variable: `W19HID`, if the analysis is at the household level; or `W19CID`, if the analysis is at the individual level. Second, the newly created data files should be merged with the file containing the appropriate weights: `brr-weights-hh.dta`, if the analysis is at the household level; or `brr-weights-ind.dta`, if the analysis is at the individual level. Third, the Stata command `svyset` should be used to specify all the information required for design-based estimation and inference. Finally, the resulting data file should be saved for subsequent use.

To see this procedure in action, let us create the individual-level data file that we will use for most of the examples presented in the next sections. First, we extract the variables of interest from two of the data files that comprise the public-use version of the ITA.LI database: `household_grid.dta` and `personal_data.dta`. Here is the relevant Stata code, followed by the corresponding output:

```
/* Select variables of interest from file "household_grid.dta" */
use "household_grid.dta", clear
keep W19CID W19AREA W19SEX W19BIRTH_Y W19INTSTR_Y W19OCC W19EDU
save "tempfile1.dta", replace
```

```
/* Select variables of interest from file "personal_data.dta" */
use "personal_data.dta", clear
keep W19CID W19PD101 W19PD501 W19PD800 W19PD810 W19PD811 W19PD812
save "tempfile2.dta", replace
```

```
. /* Select variables of interest from file "household_grid.dta" */
. use "household_grid.dta", clear
. keep W19CID W19AREA W19SEX W19BIRTH_Y W19INTSTR_Y W19OCC W19EDU
. save "tempfile1.dta", replace
(file tempfile1.dta not found)
file tempfile1.dta saved
.
. /* Select variables of interest from file "personal_data.dta" */
. use "personal_data.dta", clear
. keep W19CID W19PD101 W19PD501 W19PD800 W19PD810 W19PD811 W19PD812
. save "tempfile2.dta", replace
(file tempfile2.dta not found)
file tempfile2.dta saved
```

Second, using variable `W19CID` as the key, we merge the newly created data files `tempfile1.dta` and `tempfile2.dta` with the file containing the individual weights:

```
/* Open individual weights file and select relevant variables */
use "brr-weights-ind.dta", clear
keep W19CID subsample fiw brr_iw_*

/* Merge information from temporary files */
merge 1:1 W19CID using "tempfile1.dta"
drop _merge
erase "tempfile1.dta"
merge 1:1 W19CID using "tempfile2.dta"
drop _merge
erase "tempfile2.dta"
```

```
. /* Open individual weights file and select relevant variables */
. use "brr-weights-ind.dta", clear
. keep W19CID subsample fiw brr_iw_*
.
. /* Merge information from temporary files */
. merge 1:1 W19CID using "tempfile1.dta"

    Result                      Number of obs
    ----------------------------------------
    Not matched                         1,139
        from master                         0  (_merge==1)
        from using                      1,139  (_merge==2)

    Matched                            10,250  (_merge==3)
    ----------------------------------------
```

```
. drop _merge
. erase "tempfile1.dta"
. merge 1:1 W19CID using "tempfile2.dta"
```

Result	Number of obs	
Not matched	2,611	
from master	2,611	(_merge==1)
from using	0	(_merge==2)
Matched	8,778	(_merge==3)

```
. drop _merge
. erase "tempfile2.dta"
```

The resulting dataset contains 11,389 records, one for each member of the 4,900 eligible households included in the ITA.LI realized sample (see Figure 1.5). Here is their distribution by status:

```
/* Display raw and weighted counts of individuals  */
generate raw = 1
label variable raw "Raw counts"

generate weighted = fiw
label variable weighted "Weighted counts"

table subsample, statistic(sum raw) statistic(sum weighted) missing
```

```
. /* Display raw and weighted counts of individuals  */
. generate raw = 1
. label variable raw "Raw counts"
.
. generate weighted = fiw
(1,139 missing values generated)
. label variable weighted "Weighted counts"
.
. table subsample, statistic(sum raw) statistic(sum weighted) missing
```

	Raw counts	Weighted counts
Subsample		
Ineligibles	1,283	8,120,488
Proxy respondents	189	946,337
Self-respondents	8,778	50,192,596
.	1,139	0
Total	11,389	59,259,421

The 1,139 missing cases are those without a weight, amounting to the 1,113 non-responding eligible household members, plus the 26 infants born in 2020

who were excluded from weight construction (see Section 2.6). The remaining 10,250 cases represent the 59,259,421 individuals reported to be residing in private households in Italy as of December 31, 2019 (see Table 2.2). Of these, 1,283 – corresponding to a population of 8,120,488 individuals – are ineligible because they were under age 16 at the time of interview. We are left with the 8,967 responding eligible individuals, representing the 51,138,933 members of the target population. Since proxy respondents were administered a very short questionnaire, these individuals provide very little information and, therefore, will be excluded from the working sample. The analyses, then, will focus on the 8,778 self-respondent individuals, representing around 98% of the target population:[3]

```
/* Select self-respondents */
keep if (subsample == 3)
```

```
. /* Select self-respondents */
. keep if (subsample == 3)
(2,611 observations deleted)
```

Now, we can svyset the dataset in memory:

```
/* svyset data in memory */
svyset [pw = fiw], vce(brr) brrweight(brr_iw_*) fay(0.5) dof(150) mse
```

```
. /* svyset data in memory */
. svyset [pw = fiw], vce(brr) brrweight(brr_iw_*) fay(0.5) dof(150) mse
Sampling weights: fiw
             VCE: brr
             MSE: on
     BRR weights: brr_iw_1 .. brr_iw_152
Fay's adjustment: .5
       Design df: 150
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: <observations>
           FPC 1: <zero>
```

The instruction [pw = fiw] specifies the name of the variable containing the full-sample weights, used by Stata to compute the point estimates of the quantities of interest. Option vce(brr) sets the variance estimation method to balanced repeated replication (BRR). Option brrweight(brr_iw_*) specifies

3 Overall, the removal of proxy respondents does not substantially affect the relative composition of the ITA.LI realized sample, since on the one hand the removed individuals make up a very small share of the total (less than 2%), and on the other hand proxy respondents are a near-random subset of the total – they are significantly overrepresented only among individuals over 80 and those unable to work.

the names of the variables containing the $R = 152$ sets of replicate weights. Option `fay(0.5)` requests that Fay's variant of BRR be adopted, with $\rho = 0.5$. Option `dof(150)` sets to $H = 150$ the design degrees of freedom for confidence intervals and null hypothesis significance testing, where H denotes the number of sampling strata. Finally, option `mse` requests that the reference value of the Fay variance estimator be set to the full-sample estimate of the quantity of interest. For complete details on BRR and its implementation in ITA.LI, see Section 3.3.

As the final step, we save the dataset in memory for subsequent use:

```
/* Save data file */
drop subsample raw weighted
save "itali.dta", replace
```

```
. /* Save data file */
. drop subsample raw weighted

. save "itali.dta", replace
file itali.dta saved
```

Here are the contents of the newly created data file:

```
/* Open and describe data file contents */
use "itali.dta", clear
describe
```

```
. /* Open and describe data file contents */
. use "itali.dta", clear

. describe

Contains data from itali.dta
 Observations:         8,778
    Variables:           166                  10 Jan 2023 20:01
```

Variable name	Storage type	Display format	Value label	Variable label
W19CID	str7	%9s		Individual ID
fiw	double	%10.0g		Full-sample individual weight
brr_iw_1	double	%10.0g		Individual replicate weight #1
brr_iw_2	double	%10.0g		Individual replicate weight #2
brr_iw_3	double	%10.0g		Individual replicate weight #3
brr_iw_4	double	%10.0g		Individual replicate weight #4
brr_iw_5	double	%10.0g		Individual replicate weight #5
brr_iw_6	double	%10.0g		Individual replicate weight #6
brr_iw_7	double	%10.0g		Individual replicate weight #7
brr_iw_8	double	%10.0g		Individual replicate weight #8
brr_iw_9	double	%10.0g		Individual replicate weight #9
brr_iw_10	double	%10.0g		Individual replicate weight #10
brr_iw_11	double	%10.0g		Individual replicate weight #11
brr_iw_12	double	%10.0g		Individual replicate weight #12
brr_iw_13	double	%10.0g		Individual replicate weight #13

brr_iw_14	double	%10.0g	Individual replicate weight #14
brr_iw_15	double	%10.0g	Individual replicate weight #15
brr_iw_16	double	%10.0g	Individual replicate weight #16
brr_iw_17	double	%10.0g	Individual replicate weight #17
brr_iw_18	double	%10.0g	Individual replicate weight #18
brr_iw_19	double	%10.0g	Individual replicate weight #19
brr_iw_20	double	%10.0g	Individual replicate weight #20
brr_iw_21	double	%10.0g	Individual replicate weight #21
brr_iw_22	double	%10.0g	Individual replicate weight #22
brr_iw_23	double	%10.0g	Individual replicate weight #23
brr_iw_24	double	%10.0g	Individual replicate weight #24
brr_iw_25	double	%10.0g	Individual replicate weight #25
brr_iw_26	double	%10.0g	Individual replicate weight #26
brr_iw_27	double	%10.0g	Individual replicate weight #27
brr_iw_28	double	%10.0g	Individual replicate weight #28
brr_iw_29	double	%10.0g	Individual replicate weight #29
brr_iw_30	double	%10.0g	Individual replicate weight #30
brr_iw_31	double	%10.0g	Individual replicate weight #31
brr_iw_32	double	%10.0g	Individual replicate weight #32
brr_iw_33	double	%10.0g	Individual replicate weight #33
brr_iw_34	double	%10.0g	Individual replicate weight #34
brr_iw_35	double	%10.0g	Individual replicate weight #35
brr_iw_36	double	%10.0g	Individual replicate weight #36
brr_iw_37	double	%10.0g	Individual replicate weight #37
brr_iw_38	double	%10.0g	Individual replicate weight #38
brr_iw_39	double	%10.0g	Individual replicate weight #39
brr_iw_40	double	%10.0g	Individual replicate weight #40
brr_iw_41	double	%10.0g	Individual replicate weight #41
brr_iw_42	double	%10.0g	Individual replicate weight #42
brr_iw_43	double	%10.0g	Individual replicate weight #43
brr_iw_44	double	%10.0g	Individual replicate weight #44
brr_iw_45	double	%10.0g	Individual replicate weight #45
brr_iw_46	double	%10.0g	Individual replicate weight #46
brr_iw_47	double	%10.0g	Individual replicate weight #47
brr_iw_48	double	%10.0g	Individual replicate weight #48
brr_iw_49	double	%10.0g	Individual replicate weight #49
brr_iw_50	double	%10.0g	Individual replicate weight #50
brr_iw_51	double	%10.0g	Individual replicate weight #51
brr_iw_52	double	%10.0g	Individual replicate weight #52
brr_iw_53	double	%10.0g	Individual replicate weight #53
brr_iw_54	double	%10.0g	Individual replicate weight #54
brr_iw_55	double	%10.0g	Individual replicate weight #55
brr_iw_56	double	%10.0g	Individual replicate weight #56
brr_iw_57	double	%10.0g	Individual replicate weight #57
brr_iw_58	double	%10.0g	Individual replicate weight #58
brr_iw_59	double	%10.0g	Individual replicate weight #59
brr_iw_60	double	%10.0g	Individual replicate weight #60
brr_iw_61	double	%10.0g	Individual replicate weight #61
brr_iw_62	double	%10.0g	Individual replicate weight #62
brr_iw_63	double	%10.0g	Individual replicate weight #63
brr_iw_64	double	%10.0g	Individual replicate weight #64
brr_iw_65	double	%10.0g	Individual replicate weight #65
brr_iw_66	double	%10.0g	Individual replicate weight #66
brr_iw_67	double	%10.0g	Individual replicate weight #67
brr_iw_68	double	%10.0g	Individual replicate weight #68
brr_iw_69	double	%10.0g	Individual replicate weight #69
brr_iw_70	double	%10.0g	Individual replicate weight #70
brr_iw_71	double	%10.0g	Individual replicate weight #71

```
brr_iw_72       double  %10.0g                Individual replicate weight #72
brr_iw_73       double  %10.0g                Individual replicate weight #73
brr_iw_74       double  %10.0g                Individual replicate weight #74
brr_iw_75       double  %10.0g                Individual replicate weight #75
brr_iw_76       double  %10.0g                Individual replicate weight #76
brr_iw_77       double  %10.0g                Individual replicate weight #77
brr_iw_78       double  %10.0g                Individual replicate weight #78
brr_iw_79       double  %10.0g                Individual replicate weight #79
brr_iw_80       double  %10.0g                Individual replicate weight #80
brr_iw_81       double  %10.0g                Individual replicate weight #81
brr_iw_82       double  %10.0g                Individual replicate weight #82
brr_iw_83       double  %10.0g                Individual replicate weight #83
brr_iw_84       double  %10.0g                Individual replicate weight #84
brr_iw_85       double  %10.0g                Individual replicate weight #85
brr_iw_86       double  %10.0g                Individual replicate weight #86
brr_iw_87       double  %10.0g                Individual replicate weight #87
brr_iw_88       double  %10.0g                Individual replicate weight #88
brr_iw_89       double  %10.0g                Individual replicate weight #89
brr_iw_90       double  %10.0g                Individual replicate weight #90
brr_iw_91       double  %10.0g                Individual replicate weight #91
brr_iw_92       double  %10.0g                Individual replicate weight #92
brr_iw_93       double  %10.0g                Individual replicate weight #93
brr_iw_94       double  %10.0g                Individual replicate weight #94
brr_iw_95       double  %10.0g                Individual replicate weight #95
brr_iw_96       double  %10.0g                Individual replicate weight #96
brr_iw_97       double  %10.0g                Individual replicate weight #97
brr_iw_98       double  %10.0g                Individual replicate weight #98
brr_iw_99       double  %10.0g                Individual replicate weight #99
brr_iw_100      double  %10.0g                Individual replicate weight #100
brr_iw_101      double  %10.0g                Individual replicate weight #101
brr_iw_102      double  %10.0g                Individual replicate weight #102
brr_iw_103      double  %10.0g                Individual replicate weight #103
brr_iw_104      double  %10.0g                Individual replicate weight #104
brr_iw_105      double  %10.0g                Individual replicate weight #105
brr_iw_106      double  %10.0g                Individual replicate weight #106
brr_iw_107      double  %10.0g                Individual replicate weight #107
brr_iw_108      double  %10.0g                Individual replicate weight #108
brr_iw_109      double  %10.0g                Individual replicate weight #109
brr_iw_110      double  %10.0g                Individual replicate weight #110
brr_iw_111      double  %10.0g                Individual replicate weight #111
brr_iw_112      double  %10.0g                Individual replicate weight #112
brr_iw_113      double  %10.0g                Individual replicate weight #113
brr_iw_114      double  %10.0g                Individual replicate weight #114
brr_iw_115      double  %10.0g                Individual replicate weight #115
brr_iw_116      double  %10.0g                Individual replicate weight #116
brr_iw_117      double  %10.0g                Individual replicate weight #117
brr_iw_118      double  %10.0g                Individual replicate weight #118
brr_iw_119      double  %10.0g                Individual replicate weight #119
brr_iw_120      double  %10.0g                Individual replicate weight #120
brr_iw_121      double  %10.0g                Individual replicate weight #121
brr_iw_122      double  %10.0g                Individual replicate weight #122
brr_iw_123      double  %10.0g                Individual replicate weight #123
brr_iw_124      double  %10.0g                Individual replicate weight #124
brr_iw_125      double  %10.0g                Individual replicate weight #125
brr_iw_126      double  %10.0g                Individual replicate weight #126
brr_iw_127      double  %10.0g                Individual replicate weight #127
brr_iw_128      double  %10.0g                Individual replicate weight #128
brr_iw_129      double  %10.0g                Individual replicate weight #129
```

```
brr_iw_130      double  %10.0g                Individual replicate weight #130
brr_iw_131      double  %10.0g                Individual replicate weight #131
brr_iw_132      double  %10.0g                Individual replicate weight #132
brr_iw_133      double  %10.0g                Individual replicate weight #133
brr_iw_134      double  %10.0g                Individual replicate weight #134
brr_iw_135      double  %10.0g                Individual replicate weight #135
brr_iw_136      double  %10.0g                Individual replicate weight #136
brr_iw_137      double  %10.0g                Individual replicate weight #137
brr_iw_138      double  %10.0g                Individual replicate weight #138
brr_iw_139      double  %10.0g                Individual replicate weight #139
brr_iw_140      double  %10.0g                Individual replicate weight #140
brr_iw_141      double  %10.0g                Individual replicate weight #141
brr_iw_142      double  %10.0g                Individual replicate weight #142
brr_iw_143      double  %10.0g                Individual replicate weight #143
brr_iw_144      double  %10.0g                Individual replicate weight #144
brr_iw_145      double  %10.0g                Individual replicate weight #145
brr_iw_146      double  %10.0g                Individual replicate weight #146
brr_iw_147      double  %10.0g                Individual replicate weight #147
brr_iw_148      double  %10.0g                Individual replicate weight #148
brr_iw_149      double  %10.0g                Individual replicate weight #149
brr_iw_150      double  %10.0g                Individual replicate weight #150
brr_iw_151      double  %10.0g                Individual replicate weight #151
brr_iw_152      double  %10.0g                Individual replicate weight #152
W19AREA         byte    %13.0g     W19AREA_en
                                              Area
W19SEX          byte    %27.0g     W19SEX_en
                                              Sex
W19BIRTH_Y      int     %9.0g                 Date of birth: year
W19OCC          byte    %32.0g     W19OCC_en
                                              Current job situation
W19EDU          byte    %347.0g    W19EDU_en
                                              Highest educational degree obtained
W19INTSTR_Y     int     %9.0g      W19INTSTR_Y_en
                                              Interview start date: year
W19PD101        byte    %59.0g     W19PD101_en
                                              [Education] Pre-primary
W19PD501        byte    %29.0g     W19PD501_en
                                            [Quality of life] Satisfaction: general
W19PD800        byte    %29.0g     W19PD800_en
                                           [Health] Self-reported health (SF12)
W19PD810        int     %29.0g     W19PD810_en
                                              [Health] Weight
W19PD811        int     %29.0g     W19PD811_en
                                              [Health] Height
W19PD812        byte    %29.0g     W19PD812_en
                                              [Health] Insomnia
------------------------------------------------------------------------------
Sorted by: W19CID
```

As can be seen, the data file includes the individual ID (`W19CID`), one variable containing the full-sample individual weights (`fiw`), 152 variables containing as many sets of replicate weights (`brr_iw_1` to `brr_iw_152`), and 12 variables representing various properties of the respondents. Weight variables – as specified by `svyset` – are required for design-based analysis of any kind, so they should always be included in any working data file. All other variables,

on the other hand, will be added according to the objectives of the analysis.

Before moving on to the next sections, let us create some new variables from the existing ones, and save the resulting data file:

```
/* Open data file */
use "itali.dta", clear

/* Create variable "region" */
generate region = W19AREA
label variable region "Region of residence"
label define l_region 1 "North-West", modify
label define l_region 2 "North-East", modify
label define l_region 3 "Center", modify
label define l_region 4 "South", modify
label define l_region 5 "Islands", modify
label values region l_region

/* Create variable "sex" */
generate sex = W19SEX - 1
label variable sex "Sex"
label define l_sex 0 "Male", modify
label define l_sex 1 "Female", modify
label values sex l_sex

/* Create variable "age" */
generate age = W19INTSTR_Y - W19BIRTH_Y
label variable age "Age"

/* Create variable "agegroup" */
generate agegroup = irecode(age,24,34,44,54,64,74) + 1
label variable agegroup "Age group"
label define l_agegroup 1 "16-24 years", modify
label define l_agegroup 2 "25-34 years", modify
label define l_agegroup 3 "35-44 years", modify
label define l_agegroup 4 "45-54 years", modify
label define l_agegroup 5 "55-64 years", modify
label define l_agegroup 6 "65-74 years", modify
label define l_agegroup 7 "75 years and over", modify
label values agegroup l_agegroup

/* Create variable "educ" */
generate educ = W19EDU
recode educ (1/3 = 1) (4 = 2) (5/6 = 3) (7/11 = 4)
label variable educ "Educational degree"
label define l_educ 1 "None/Elementary school", modify
label define l_educ 2 "Middle school", modify
label define l_educ 3 "High school", modify
```

```
label define l_educ 4 "Tertiary degree", modify
label values educ l_educ

/* Create variable "empstat" */
generate empstat = W19OCC
recode empstat (1 = 1) (2/3 = 2) (5 = 3) (7 = 4) (4 6 = 5)
label variable empstat "Employment status"
label define l_empstat 1 "Employed", modify
label define l_empstat 2 "Job seeker", modify
label define l_empstat 3 "Student", modify
label define l_empstat 4 "Retired", modify
label define l_empstat 5 "Homemaker/Other", modify
label values empstat l_empstat

/* Create variable "preschool" */
generate preschool = (W19PD101 == 1) if !missing(W19PD101)
label variable preschool "Attended pre-primary school 1+ years"
label define l_preschool 0 "No", modify
label define l_preschool 1 "Yes", modify
label values preschool l_preschool

/* Create variable "lifesat" */
generate lifesat = W19PD501 if !missing(W19PD501)
label variable lifesat "Overall life satistaction"

/* Create variable "srh" */
generate srh = 6 - W19PD800
label variable srh "Self-reported health"
label define l_srh 1 "Bad", modify
label define l_srh 2 "Poor", modify
label define l_srh 3 "Satisfactory", modify
label define l_srh 4 "Good", modify
label define l_srh 5 "Excellent", modify
label values srh l_srh

/* Create variable "insomnia" */
generate insomnia = W19PD812>1 if !missing(W19PD812)
label variable insomnia "Suffered from insomnia in past 4 weeks"
label define l_insomnia 0 "No", modify
label define l_insomnia 1 "Yes", modify
label values insomnia l_insomnia

/* Create variable "weight" */
generate weight = W19PD810 if !missing(W19PD810)
label variable weight "Self-reported weight (kilos)"
```

```
/* Create variable "height" */
generate height = W19PD811 if !missing(W19PD811)
label variable height "Self-reported height (cm)"

/* Create variable "bmi" */
replace weight = . if (weight < 40)
replace height = . if (height < 140)
generate bmi = weight / (height / 100)^2
replace bmi = . if (bmi > 50)
label variable bmi "Body mass index"

/* Select relevant variables */
keep W19CID fiw brr_iw_* region-bmi

/* Save data file */
compress
save "itali.dta", replace
```

```
. /* Open data file */
. use "itali.dta", clear

.
. /* Create variable "region" */
. generate region = W19AREA
. label variable region "Region of residence"
. label define l_region 1 "North-West", modify
. label define l_region 2 "North-East", modify
. label define l_region 3 "Center", modify
. label define l_region 4 "South", modify
. label define l_region 5 "Islands", modify
. label values region l_region
.
. /* Create variable "sex" */
. generate sex = W19SEX - 1
. label variable sex "Sex"
. label define l_sex 0 "Male", modify
. label define l_sex 1 "Female", modify
. label values sex l_sex
.
. /* Create variable "age" */
. generate age = W19INTSTR_Y - W19BIRTH_Y
. label variable age "Age"
.
. /* Create variable "agegroup" */
. generate agegroup = irecode(age,24,34,44,54,64,74) + 1
. label variable agegroup "Age group"
. label define l_agegroup 1 "16-24 years", modify
```

```
. label define l_agegroup 2 "25-34 years", modify
. label define l_agegroup 3 "35-44 years", modify
. label define l_agegroup 4 "45-54 years", modify
. label define l_agegroup 5 "55-64 years", modify
. label define l_agegroup 6 "65-74 years", modify
. label define l_agegroup 7 "75 years and over", modify
. label values agegroup l_agegroup
.
. /* Create variable "educ" */
. generate educ = W19EDU
. recode educ (1/3 = 1) (4 = 2) (5/6 = 3) (7/11 = 4)
(8746 changes made to educ)
. label variable educ "Educational degree"
. label define l_educ 1 "None/Elementary school", modify
. label define l_educ 2 "Middle school", modify
. label define l_educ 3 "High school", modify
. label define l_educ 4 "Tertiary degree", modify
. label values educ l_educ
.
. /* Create variable "empstat" */
. generate empstat = W19OCC
. recode empstat (1 = 1) (2/3 = 2) (5 = 3) (7 = 4) (4 6 = 5)
(4514 changes made to empstat)
. label variable empstat "Employment status"
. label define l_empstat 1 "Employed", modify
. label define l_empstat 2 "Job seeker", modify
. label define l_empstat 3 "Student", modify
. label define l_empstat 4 "Retired", modify
. label define l_empstat 5 "Homemaker/Other", modify
. label values empstat l_empstat
.
. /* Create variable "preschool" */
. generate preschool = (W19PD101 == 1) if !missing(W19PD101)
(76 missing values generated)
. label variable preschool "Attended pre-primary school 1+ years"
. label define l_preschool 0 "No", modify
. label define l_preschool 1 "Yes", modify
. label values preschool l_preschool
.
. /* Create variable "lifesat" */
. generate lifesat = W19PD501 if !missing(W19PD501)
(31 missing values generated)
. label variable lifesat "Overall life satistaction"
.
. /* Create variable "srh" */
```

```
. generate srh = 6 - W19PD800
(46 missing values generated)
. label variable srh "Self-reported health"
. label define l_srh 1 "Bad", modify
. label define l_srh 2 "Poor", modify
. label define l_srh 3 "Satisfactory", modify
. label define l_srh 4 "Good", modify
. label define l_srh 5 "Excellent", modify
. label values srh l_srh
.
. /* Create variable "insomnia" */
. generate insomnia = W19PD812>1 if !missing(W19PD812)
(103 missing values generated)
. label variable insomnia "Suffered from insomnia in past 4 weeks"
. label define l_insomnia 0 "No", modify
. label define l_insomnia 1 "Yes", modify
. label values insomnia l_insomnia
.
. /* Create variable "weight" */
. generate weight = W19PD810 if !missing(W19PD810)
(1,412 missing values generated)
. label variable weight "Self-reported weight (kilos)"
.
.
. /* Create variable "height" */
. generate height = W19PD811 if !missing(W19PD811)
(705 missing values generated)
. label variable height "Self-reported height (cm)"
.
. /* Create variable "bmi" */
. replace weight = . if (weight < 40)
(3 real changes made, 3 to missing)
. replace height = . if (height < 140)
(10 real changes made, 10 to missing)
. generate bmi = weight / (height / 100)^2
(1,458 missing values generated)
. replace bmi = . if (bmi > 50)
(9 real changes made, 9 to missing)
. label variable bmi "Body mass index"
.
. /* Select relevant variables */
. keep W19CID fiw brr_iw_* region-bmi
.
. /* Save data file */
. compress
  variable region was float now byte
  variable sex was float now byte
  variable age was float now byte
  variable agegroup was float now byte
```

```
  variable educ was float now byte
  variable empstat was float now byte
  variable preschool was float now byte
  variable lifesat was float now byte
  variable srh was float now byte
  variable insomnia was float now byte
  variable weight was float now int
  variable height was float now int
  variable W19CID was str7 now str6
  (307,230 bytes saved)
. save "itali.dta", replace
file itali.dta saved
```

Now we are ready for data analysis, starting with the distribution of single variables. For the purpose of our discussion, in the following we will distinguish between two types of variables: *qualitative*, whose values represent an exhaustive and mutually exclusive set of ordered or unordered categories, and *quantitative*, whose values represent numerical measurements or counts.

4.3. *Univariate Analysis*

The purpose of *univariate analysis* is to describe the distribution of variables taken one at a time. Let us begin with `srh`, a qualitative variable representing respondents' self-reported health status. This information was collected using a single question asking respondents to rate their overall health at the time of interview on a five-point ordinal scale, ranging from "Excellent" to "Poor". Forty-six subjects did not answer, so that the number of valid cases for the analysis is 8,732.

The standard Stata command for calculating and reporting the *percent distribution* of a variable is `tabulate`. By prefixing this command with `svy`, all quantities of interest (point estimates, standard errors, and confidence intervals) are calculated according to the rules of design-based estimation and inference:

```
/* Open data file */
use "itali.dta", clear

/* Percent distribution of variable "srh" */
svy : tabulate srh, percent se ci format(%5.1f)
```

```
. /* Open data file */
. use "itali.dta", clear

.
. /* Percent distribution of variable "srh" */
. svy : tabulate srh, percent se ci format(%5.1f)
(running tabulate on estimation sample)
```

```
                                      Number of obs    =      8,732
                                      Population size  = 49,983,951
                                      Replications     =        152
                                      Design df        =        150
```

Self-reported health	percentage	se	lb	ub
Bad	2.1	0.3	1.7	2.7
Poor	5.7	0.4	5.0	6.5
Satisfac	21.9	0.6	20.8	23.1
Good	54.9	0.9	53.1	56.6
Excellen	15.4	0.7	14.1	16.7
Total	100.0			

```
Key: percentage = Cell percentage
             se = Brr standard error of cell percentage
             lb = Lower 95% confidence bound for cell percentage
             ub = Upper 95% confidence bound for cell percentage
```

It should be noted that the confidence intervals reported by `tabulate` are not of the Wald type described in Chapter 3 (see Equation 3.12); rather, they are calculated using a logit transform so that the endpoints always lie between zero and one (StataCorp 2021b).

As an alternative to `tabulate`, we can use the command `proportion`:

```
/* Percent distribution of variable "srh" */
svy, dots(10) : proportion srh, percent cformat(%5.1f)
```

```
. /* Percent distribution of variable "srh" */
. svy, dots(10) : proportion srh, percent cformat(%5.1f)
(running proportion on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Percent estimation          Number of obs    =      8,732
                                    Population size  = 49,983,951
                                    Replications     =        152
                                    Design df        =        150
```

	Percent	BRR * std. err.	Logit [95% conf.	interval]
srh				
Bad	2.1	0.3	1.7	2.7
Poor	5.7	0.4	5.0	6.5
Satisfactory	21.9	0.6	20.8	23.1
Good	54.9	0.9	53.1	56.6
Excellent	15.4	0.7	14.1	16.7

The use of `proportion` has two main benefits over `tabulate`. First, it enables access to several postestimation statistics, such as the design effect (Equation 3.15), the coefficient of variation (Equation 3.16), and the misspecification effect. The latter, in particular, measures the extent to which we would underestimate the standard error of the estimator of interest if we did not account for stratification, clustering and weighting – in practice, if we omitted the `svy` prefix. To obtain misspecification effects, after running the estimation command we type the following:

```
/* Misspecification effects after -svy : proportion- */
estat effects, meft
```

```
. /* Misspecification effects after -svy : proportion- */
. estat effects, meft
```

	Proportion	BRR * std. err.	MEFT
srh			
Bad	.0211589	.0025518	2.06515
Poor	.0573518	.0037348	1.61284
Satisfact~y	.219319	.0056667	1.2799
Good	.5485315	.0088138	1.66278
Excellent	.1536388	.006731	1.76351

For illustration, consider the proportion of those who feel they are in bad health. For this quantity, the misspecification effect is equal to 2.07. This means that had we calculated the standard error of the proportion without considering the survey design features, its value would have been underestimated by $100(1 - (1/2.06515)) = 51.6\%$, with similar consequences for the width of the corresponding confidence interval.

A second benefit of using `proportion` is that it allows different types of confidence intervals to be calculated. The standard Wald interval described in Chapter 3 assumes that the sampling distribution of the proportion of interest is asymptotically normal (see discussion in Section 3.1). Normality, however, is hardly achieved when the proportion is based on a small number of cases and/or takes values close to zero or one (Dean and Pagano 2015; Korn and Graubard 1999). To overcome this problem, `proportion` by default calculates confidence intervals using the logit transform mentioned above. Alternatively, the user can request that the intervals of interest be calculated using other methods, including the Agresti-Coull, Clopper-Pearson, Jeffreys, and Wilson methods (Dean and Pagano 2015; Franco *et al.* 2019). For example:

```
/* Percent distribution of variable "srh" with Jeffreys CIs */
svy, dots(10) : proportion srh, percent cformat(%5.1f) citype(jeffreys)
```

```
. /* Percent distribution of variable "srh" with Jeffreys CIs */
. svy, dots(10) : proportion srh, percent cformat(%5.1f) citype(jeffreys)
(running proportion on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Percent estimation          Number of obs    =      8,732
                                    Population size = 49,983,951
                                    Replications     =        152
                                    Design df        =        150
```

	Percent	BRR * std. err.	Jeffreys [95% conf.	interval]
srh				
Bad	2.1	0.3	1.7	2.7
Poor	5.7	0.4	5.0	6.5
Satisfactory	21.9	0.6	20.8	23.1
Good	54.9	0.9	53.1	56.6
Excellent	15.4	0.7	14.1	16.7

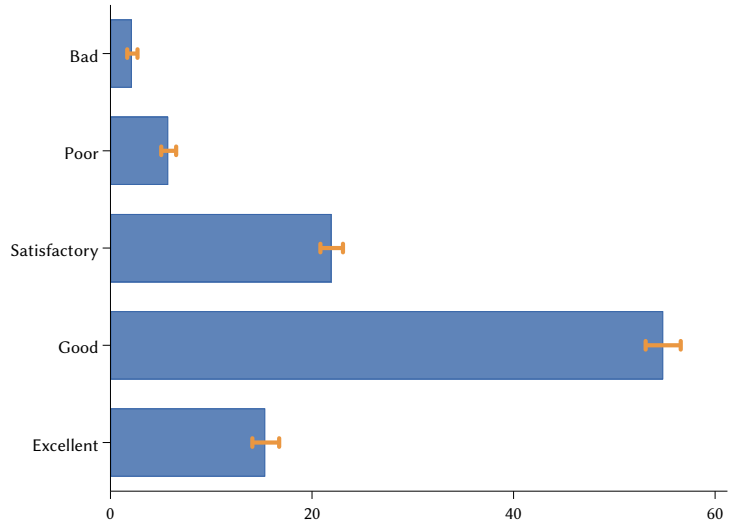
As can be seen, in this case there is no noticeable difference between the logit interval and the Jeffreys interval.

Yet another tool for depicting the percent distribution of qualitative variables is the excellent user-written command `dstat` (Jann 2020). In addition to doing all that `proportion` does, `dstat` – among other things – provides the ability to graphically represent the distribution of interest. Here is an example:

```
/* Percent distribution of variable "srh" : Table + Graph */
dstat proportion srh, percent cformat(%5.1f) table      ///
    vce(svy, dots(10)) graph( p1(color("55 101 168"))   ///
    ciopts(color("234 151 65") lwidth(*5)) )
```

```
. /* Percent distribution of variable "srh" : Table + Graph */
. dstat proportion srh, percent cformat(%5.1f) table      ///
>     vce(svy, dots(10)) graph( p1(color("55 101 168"))   ///
>     ciopts(color("234 151 65") lwidth(*5)) )
(running dstat_svyr to obtain evaluation grid)
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Percent                     Number of obs    =      8,732
                                    Population size = 49,983,951
                                    Replications     =        152
                                    Design df        =        150
```

srh	Coefficient	BRR * std. err.	logit transformed [95% conf.	interval]
srh				
Bad	2.1	0.3	1.7	2.7
Poor	5.7	0.4	5.0	6.5
Satisfactory	21.9	0.6	20.8	23.1
Good	54.9	0.9	53.1	56.6
Excellent	15.4	0.7	14.1	16.7



Although it is mainly considered in the analysis of quantitative variables, the degree of *variability*, or *dispersion*, of a distribution is also relevant to qualitative variables. The user-written command `dstat` allows design-based estimation and inference for several measures of qualitative variability, including the Gini mutability index (natural and normalized) and Shannon entropy (Budescu and Budescu 2012). Here is the Stata code to compute these measures for variable `srh`, followed by the associated output:

```
/* Measures of variability for variable "srh" */
dstat (gimp gimpn entropy) srh, vce(svy, dots(10))
```

```
. /* Measures of variability for variable "srh" */
. dstat (gimp gimpn entropy) srh, vce(svy, dots(10))
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Summary statistics          Number of obs    =      8,732
                                    Population size = 49,983,951
                                    Replications     =        152
                                    Design df        =        150
```

srh	Coefficient	BRR * std. err.	[95% conf.	interval]
gimp	.6236705	.0068981	.6100405	.6373005
gimpn	.7795882	.0086226	.7625507	.7966257
entropy	1.19547	.0127888	1.1702	1.220739

where `gimp` and `gimpn` denote, respectively, the natural and normalized forms of the Gini mutability index, while `entropy` denotes Shannon entropy.

The Gini mutability index and Shannon entropy apply to the analysis of all qualitative variables alike. When dealing specifically with *ordinal variables* like `srh`, however, it may be preferable to use more specialized indices. The user-written command `ineqord` (Jenkins 2020) allows many such indices to be calculated. Although `ineqord` does not directly support design-based inference, it can still be used for this purpose by taking advantage of the flexibility of the `svy` prefix. If, for example, we are interested in estimating the $1 - I^2$ index of Blair and Lacy (2000) and Jenkins's inequality indices J_d and J_u (Jenkins 2021), we can run the following Stata code:

```
/* Indices of ordinal variation for variable "srh" */
svy brr Blair_Lacy=r(blairlacy) Jd=r(Jd) Ju=r(Ju), dots(10) :   ///
    ineqord srh if (srh < .)
```

```
. /* Indices of ordinal variation for variable "srh" */
. svy brr Blair_Lacy=r(blairlacy) Jd=r(Jd) Ju=r(Ju), dots(10) :   ///
>     ineqord srh if (srh < .)
(running ineqord on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

BRR results                                     Number of obs    =      8,732
                                                Population size = 49,983,951
                                                Replications     =        152
                                                Design df        =        150

      Command: ineqord srh if (srh < .)
   Blair_Lacy: r(blairlacy)
           Jd: r(Jd)
           Ju: r(Ju)
```

	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
Blair_Lacy	.4322191	.007348	58.82	0.000	.4177002	.446738
Jd	.4635812	.0041745	111.05	0.000	.4553327	.4718297
Ju	.433616	.0052219	83.04	0.000	.4232981	.443934

Let us now examine *Body mass index* (BMI), a quantitative variable defined as body weight (in kilograms) divided by height squared (in meters). One thousand four hundred sixty-seven subjects did not report their weight or height, so that the number of valid cases for the analysis is 7,311.

Although Stata provides some official commands for design-based analysis of quantitative distributions, this type of analysis can be done most easily using a third-party tool, the user-written command `dstat` mentioned above. Specifically, `dstat` supports design-based estimation and inference for a wide array of distribution functions – including the probability density, cumulative distribution, and quantile functions – and summary statistics.

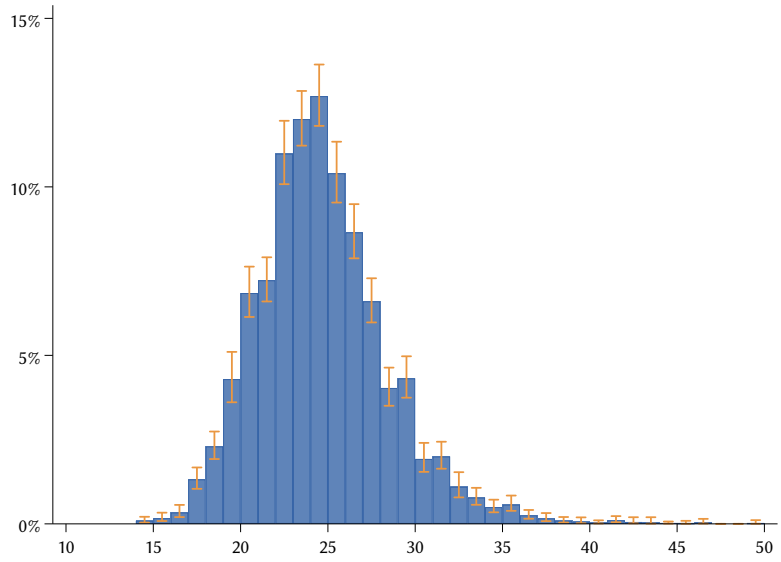
To explore the analytical capabilities of `dstat`, let us begin with a standard *histogram*, for which the command can provide both a tabular and a graphical representation:

```
/* Histogram of variable "bmi" */
dstat histogram bmi, at(14(1)50) percent cformat(%5.2f)      ///
    vce(svy, dots(10)) table graph(p1(color("55 101 168"))   ///
        ciopts(color("234 151 65") lwidth(*2))               ///
        ylabel(0 "0%" 5 "5%" 10 "10%" 15 "15%")              ///
        xlabel(10(5)50)                                      ///
    )
```

```
. /* Histogram of variable "bmi" */
. dstat histogram bmi, at(14(1)50) percent cformat(%5.2f)      ///
>     vce(svy, dots(10)) table graph(p1(color("55 101 168"))   ///
>         ciopts(color("234 151 65") lwidth(*2))               ///
>         ylabel(0 "0%" 5 "5%" 10 "10%" 15 "15%")              ///
>         xlabel(10(5)50)                                      ///
>     )
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Histogram (percent)          Number of obs    =      7,311
                                     Population size = 41,317,101
                                     Replications     =        152
                                     Design df        =        150
```

bmi	Coefficient	BRR * std. err.	logit transformed [95% conf.	interval]
14	0.09	0.04	0.04	0.21
15	0.17	0.06	0.08	0.33
16	0.34	0.09	0.20	0.56
17	1.32	0.16	1.04	1.67
18	2.30	0.21	1.92	2.74
19	4.29	0.38	3.61	5.10
20	6.85	0.38	6.14	7.63

21	7.23	0.33	6.60	7.91
22	10.98	0.48	10.08	11.96
23	12.01	0.41	11.22	12.85
24	12.69	0.46	11.81	13.63
25	10.40	0.46	9.53	11.34
26	8.65	0.41	7.88	9.49
27	6.60	0.33	5.98	7.29
28	4.03	0.29	3.50	4.64
29	4.32	0.31	3.74	4.97
30	1.93	0.22	1.54	2.40
31	2.00	0.20	1.64	2.44
32	1.10	0.19	0.78	1.53
33	0.78	0.13	0.57	1.07
34	0.50	0.09	0.34	0.72
35	0.57	0.11	0.39	0.84
36	0.25	0.06	0.15	0.41
37	0.16	0.06	0.08	0.32
38	0.09	0.04	0.04	0.20
39	0.06	0.04	0.02	0.19
40	0.03	0.02	0.01	0.10
41	0.10	0.04	0.05	0.23
42	0.05	0.04	0.02	0.20
43	0.04	0.03	0.01	0.20
44	0.02	0.01	0.01	0.07
45	0.01	0.01	0.00	0.09
46	0.04	0.03	0.01	0.15
47	0.00	(omitted)		
48	0.00	(omitted)		
49	0.02	0.02	0.00	0.11
50	0.02	0.02	0.00	0.11



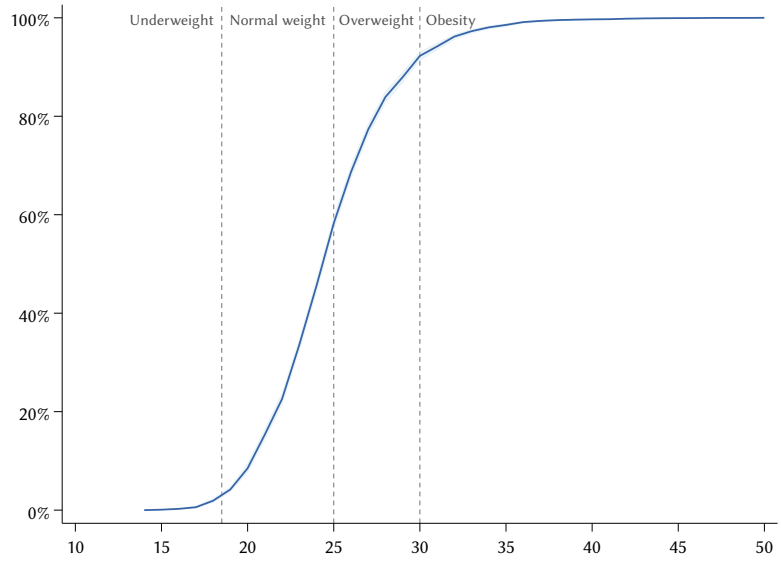
Another useful representation of the entire distribution of a quantitative variable is provided by the *cumulative distribution function*, which, in this case, also gives us a clear picture of the prevalence of the standard BMI categories (Flegal *et al.* 2014):

```
/* Cumulative distribution function of variable "bmi" */
dstat cdf bmi, at(14(1)50) percent vce(svy, dots(10)) table      ///
    cformat(%5.2f) graph(                                        ///
        p1(color("55 101 168") lwidth(*2))                       ///
        ylabel(0 "0%" 20 "20%" 40 "40%" 60 "60%" 80 "80%"        ///
               100 "100%")                                       ///
        xlabel(10(5)50)                                          ///
        xline(18.5 25 30, lpattern(shortdash) lcolor(gs7))       ///
        text(100 15.6 "Underweight", size(*0.9) color(gs5))      ///
        text(100 21.75 "Normal weight", size(*0.9) color(gs5))   ///
        text(100 27.5 "Overweight", size(*0.9) color(gs5))       ///
        text(100 31.8 "Obesity", size(*0.9) color(gs5))          ///
    )
```

```
. /* Cumulative distribution function of variable "bmi" */
. dstat cdf bmi, at(14(1)50) percent vce(svy, dots(10)) table      ///
>     cformat(%5.2f) graph(                                        ///
>         p1(color("55 101 168") lwidth(*2))                       ///
>         ylabel(0 "0%" 20 "20%" 40 "40%" 60 "60%" 80 "80%"        ///
>                100 "100%")                                       ///
>         xlabel(10(5)50)                                          ///
>         xline(18.5 25 30, lpattern(shortdash) lcolor(gs7))       ///
>         text(100 15.6 "Underweight", size(*0.9) color(gs5))      ///
>         text(100 21.75 "Normal weight", size(*0.9) color(gs5))   ///
>         text(100 27.5 "Overweight", size(*0.9) color(gs5))       ///
>         text(100 31.8 "Obesity", size(*0.9) color(gs5))          ///
>     )
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: CDF in percent                Number of obs    =      7,311
                                      Population size = 41,317,101
                                      Replications     =        152
                                      Design df        =        150
```

bmi	Coefficient	BRR * std. err.	logit transformed [95% conf.	interval]
14	0.00	(omitted)		
15	0.09	0.04	0.04	0.21
16	0.26	0.07	0.14	0.45
17	0.59	0.12	0.40	0.88
18	1.91	0.23	1.50	2.43
19	4.21	0.29	3.67	4.81
20	8.50	0.51	7.55	9.56

21	15.35	0.62	14.15	16.62
22	22.57	0.69	21.25	23.96
23	33.56	0.70	32.18	34.96
24	45.57	0.78	44.04	47.11
25	58.26	0.80	56.68	59.82
26	68.66	0.77	67.13	70.16
27	77.31	0.70	75.90	78.66
28	83.92	0.64	82.62	85.14
29	87.95	0.58	86.75	89.05
30	92.26	0.49	91.24	93.18
31	94.19	0.40	93.35	94.93
32	96.19	0.31	95.52	96.76
33	97.28	0.23	96.79	97.71
34	98.06	0.20	97.62	98.42
35	98.56	0.18	98.16	98.87
36	99.13	0.14	98.81	99.36
37	99.38	0.11	99.11	99.56
38	99.54	0.10	99.30	99.69
39	99.63	0.09	99.41	99.76
40	99.69	0.08	99.48	99.81
41	99.72	0.08	99.52	99.84
42	99.82	0.06	99.65	99.91
43	99.88	0.05	99.73	99.95
44	99.91	0.04	99.80	99.96
45	99.94	0.03	99.83	99.98
46	99.95	0.03	99.84	99.98
47	99.98	0.02	99.89	100.00
48	99.98	0.02	99.89	100.00
49	99.98	0.02	99.89	100.00
50	100.00	.	100.00	100.00



In terms of summary statistics, `dstat` enables design-based estimation and inference of several measures of location, variability, skewness, and kurtosis.

For illustration, let us use `dstat` to estimate only a selection of these quantities: the arithmetic mean (`mean`); the 10th, 25th, 50th, 75th, and 90th percentiles (`p10 p25 p50 p75 p90`); the standard deviation (`sd`); the interquartile range (`iqr`); the coefficient of skewness (`skewness`); and the coefficient of kurtosis (`kurtosis`). Here is the relevant Stata code, followed by the corresponding output:

```
/* Summary statistics for variable "bmi" */
dstat summarize (mean p10 p25 p50 p75 p90 sd iqr skewness   ///
    kurtosis) bmi, vce(svy, dots(10)) cformat(%5.2f)
```

```
. /* Summary statistics for variable "bmi" */
. dstat summarize (mean p10 p25 p50 p75 p90 sd iqr skewness   ///
>     kurtosis) bmi, vce(svy, dots(10)) cformat(%5.2f)
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Summary statistics          Number of obs    =      7,311
                                    Population size = 41,317,101
                                    Replications     =        152
                                    Design df        =        150

------------------------------------------------------------------------------
             |                BRR *
         bmi |  Coefficient  std. err.     [95% conf. interval]
-------------+----------------------------------------------------------------
        mean |        24.67       0.07        24.54        24.80
         p10 |        20.24       0.09        20.06        20.43
         p25 |        22.20       0.08        22.05        22.36
         p50 |        24.34       0.10        24.14        24.53
         p75 |        26.67       0.13        26.42        26.92
         p90 |        29.39       0.01        29.37        29.41
          sd |         3.76       0.07         3.63         3.89
         iqr |         4.47       0.13         4.22         4.72
    skewness |         0.85       0.08         0.69         1.00
    kurtosis |         5.16       0.39         4.40         5.93
------------------------------------------------------------------------------
```

All the analyses illustrated so far have focused on the entire target population and, therefore, have been carried out – net of missing values – on the full sample. Researchers, however, are often interested in focusing their analysis on subsets of the target population, variously referred to as subpopulations, subclasses, or domains (West *et al.* 2008). In these cases the overall analytical approach does not change, but the number of degrees of freedom used to build confidence intervals or to perform null hypothesis significance testing must be adjusted.

The standard formula for calculating the (approximate) number of design degrees of freedom in the analysis of complex survey data is $(n_{\text{PSU}} - H)$, where

n_{PSU} denotes the number of PSUs used for variance estimation and H denotes the number of sampling strata (West *et al.* 2008). In our case, since $n_{\text{PSU}} = 300$ and $H = 150$, the number of design degrees of freedom is $300 - 150 = 150$. This is exactly the value we specified with option `dof(150)` when we `svyset` the working dataset (see Section 4.2). In subpopulation analysis, however, the standard formula can lead to a substantial overestimation of the true number of degrees of freedom, when the members of the subpopulation of interest are not uniformly distributed among all sampling strata and PSUs (Rust and Rao 1996). In this situation, "one rule of thumb is that the number of degrees of freedom for a [subpopulation] is unlikely to exceed $(n' - H')$, where H' is the number of strata that contain at least one sample member from the [subpopulation], and n' is the total number of PSUs selected that contain at least one sample member from the [subpopulation]." (Rust and Rao 1996, p. 303).

The $(n' - H')$ formula, however, cannot be implemented by users of the public-use version of the ITA.LI database, which, for preserving the privacy of respondents, does not include PSU and stratum identifiers. As a partial remedy to this limitation, Tables 4.3 and 4.4 report the (approximate) design degrees of freedom for a number of subpopulations of potential interest, obtained by applying the $(n' - H')$ formula to the full ITA.LI database. Table 4.3, for example, suggests that if we wanted to focus our analysis on men aged 80 and older, the number of design degrees of freedom should be set at 46. Likewise, according to Table 4.4, the (approximately) correct number of degrees of freedom for the subpopulation of 16-24 year olds residing in the North West regions is 29.

To conduct a subpopulation analysis in Stata, one must supplement the usual estimation commands with (a) the specification of the subpopulation of interest, and (b) the appropriate number of design degrees of freedom for that subpopulation. For illustration, suppose we want to replicate the previous analysis of the distribution of variable *Self-reported health*, but considering only women aged 80 and older. According to Table 4.3, the number of design degrees of freedom for this subpopulation should be set at 51. Here, then, is the Stata code to be run, followed by the corresponding output:

```
/* Subpopulation analysis of variable "srh" */
svy, dots(10) subpop(if sex==1 & age>=80) dof(51) :   ///
    proportion srh, percent cformat(%5.1f)
```

Table 4.3 Approximate design degrees of freedom for select subpopulations defined by combinations of *Age group*, *Area of residence*, and *Sex*.

	Sex		
	Male	Female	*Total*
Age group			
16-19 years	29	27	58
16-24 years	73	65	103
16-49 years	140	144	146
16-64 years	145	145	146
20-24 years	39	42	78
25-29 years	40	43	73
25-54 years	138	144	145
25-64 years	143	145	146
25-74 years	144	147	147
25 years and over	145	147	147
30-34 years	53	46	80
35-39 years	51	50	85
40-44 years	48	68	95
40 years and over	143	145	146
45-49 years	59	77	104
50-54 years	62	74	102
50 years and over	140	143	144
55-59 years	64	78	109
55-64 years	99	113	127
60-64 years	60	70	101
60 years and over	134	136	140
65-69 years	55	72	98
65-74 years	99	102	123
65 years and over	126	128	139
70-74 years	55	67	91
70 years and over	110	118	134
75-79 years	26	50	76
80 years and over	46	51	80
Area of residence			
North-West	35	35	35
North-East	34	34	34
Center	27	27	27
South	34	34	34
Islands	16	17	17

Table 4.4 Approximate design degrees of freedom for select subpopulations defined by combinations of *Age group* and *Area of residence*.

	Area of residence				
	North West	North East	Center	South	Islands
Age group					
16-19 years	17	19	9	10	3
16-24 years	29	25	15	24	10
16-49 years	35	34	27	34	16
16-64 years	35	34	27	34	16
20-24 years	20	19	12	20	7
25-29 years	15	20	14	19	5
25-54 years	35	34	26	34	16
25-64 years	35	34	27	34	16
25-74 years	35	34	27	34	17
25 years and over	35	34	27	34	17
30-34 years	23	19	13	22	3
35-39 years	18	22	16	22	7
40-44 years	25	20	19	23	8
40 years and over	35	33	27	34	17
45-49 years	23	24	20	26	11
50-54 years	23	28	16	24	11
50 years and over	34	32	27	34	17
55-59 years	29	21	21	29	9
55-64 years	31	30	23	31	12
60-64 years	24	25	15	27	10
60 years and over	33	31	27	33	16
65-69 years	30	18	13	27	10
65-74 years	31	26	24	29	13
65 years and over	33	31	27	32	16
70-74 years	25	18	19	22	7
70 years and over	33	29	26	31	15
75-79 years	24	15	13	14	10
80 years and over	20	15	15	20	10

```
. /* Subpopulation analysis of variable "srh" */
. svy, dots(10) subpop(if sex==1 & age>=80) dof(51) :   ///
>     proportion srh, percent cformat(%5.1f)
(running proportion on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Percent estimation          Number of obs    =      8,774
                                    Population size  = 50,168,637
                                    Subpop. no. obs  =        345
                                    Subpop. size     =  2,606,152
                                    Replications     =        152
                                    Design df        =         51
```

	Percent	BRR * std. err.	Logit [95% conf.	interval]
srh				
Bad	7.8	1.6	5.2	11.6
Poor	26.7	2.7	21.7	32.4
Satisfactory	49.7	3.5	42.8	56.6
Good	14.4	2.6	10.0	20.4
Excellent	1.4	0.9	0.4	5.0

Clearly, the subpopulations considered in Tables 4.3 and 4.4 are only a subset of all possible ones. In case the user needs to analyze other subpopulations, we suggest that the approximate number of design degrees of freedom for each subpopulation of interest be determined by applying the following rules of thumb, derived from the available data:

- When the subpopulation spans the whole of Italy:

$$df = \begin{cases} \lfloor n \times (0.12) \rceil, & \text{if } n < 1{,}000 \\ 120, & \text{if } 1{,}000 \leq n < 1{,}500 \\ 130, & \text{if } 1{,}500 \leq n < 2{,}000 \\ 140, & \text{if } n \geq 2{,}000 \end{cases}$$

- When the subpopulation is limited to the North West, the North East, or the South:

$$df = \begin{cases} \lfloor n \times (0.15) \rceil, & \text{if } n < 200 \\ 30, & \text{if } n \geq 200 \end{cases}$$

- When the subpopulation is limited to the Center:

$$df = \begin{cases} \lfloor n \times (0.125) \rceil, & \text{if } n < 200 \\ 25, & \text{if } n \geq 200 \end{cases}$$

- When the subpopulation is limited to the Islands:

$$df = \begin{cases} \lfloor n \times (0.075) \rceil, & \text{if } n < 200 \\ 15, & \text{if } n \geq 200 \end{cases}$$

where $\lfloor \cdot \rceil$ denotes the nearest integer function; and n denotes the size of the subpopulation of interest.

4.4. *Bivariate Analysis*

The purpose of *bivariate analysis* is to describe the relationship between a variable of interest Y and a covariate X, so as to assess whether and how the distribution of Y, or a summary measure of it, varies with the values of X – that is, whether and how there is an *association* between the two variables.

Let us begin with the simplest case, that of the relationship between two *dichotomous variables*. Specifically, let us investigate whether, and to what extent, the probability of suffering from insomnia varies between men and women. One hundred and three respondents did not answer the question on insomnia, so that the number of valid cases for the analysis is 8,675.

The standard Stata command for design-based analysis of relationships between pairs of qualitative variables is `tabulate`, prefixed by `svy`:

```
/* Cross-tabulation of variables "sex" and "insomnia" */
svy : tabulate sex insomnia, row percent ci format(%5.1f)
```

```
. /* Cross-tabulation of variables "sex" and "insomnia" */
. svy : tabulate sex insomnia, row percent ci format(%5.1f)
(running tabulate on estimation sample)

BRR *: for rows

                                                Number of obs    =      8,675
                                                Population size = 49,662,295
                                                Replications     =        152
                                                Design df        =        150
```

Sex	Suffered from insomnia in past 4 weeks: No	Yes	Total
Male	66.2 [63.7,68.5]	33.8 [31.5,36.3]	100.0
Female	55.0 [52.8,57.2]	45.0 [42.8,47.2]	100.0
Total	60.4 [58.4,62.3]	39.6 [37.7,41.6]	100.0

```
Key: Row percentage
     [95% confidence interval for row percentage]
```

```
  Pearson:
    Uncorrected   chi2(1)         =  112.4968
    Design-based  F(1, 150)       =   87.8666     P = 0.0000
```

Data clearly show that women are more likely to suffer from sleep disorders than men. The results of two tests of independence based on Pearson's statistic for two-way tables are also reported. The `Uncorrected` test uses the ordinary χ^2 statistic, while the `Design-based` test uses the statistic adjusted for survey design with the second-order correction of Rao and Scott (1984). Both tests reject the null hypothesis of independence between `insomnia` and `sex`.

The two conditional percent distributions of variable `insomnia` by categories of variable `sex` can also be obtained using command `proportion`:

```
/* Conditional distribution of "insomnia" by "sex" */
svy, dots(10) : proportion insomnia, over(sex) percent cformat(%5.1f)
```

```
. /* Conditional distribution of "insomnia" by "sex" */
. svy, dots(10) : proportion insomnia, over(sex) percent cformat(%5.1f)
(running proportion on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Percent estimation          Number of obs    =      8,675
                                    Population size  = 49,662,295
                                    Replications     =        152
                                    Design df        =        150
```

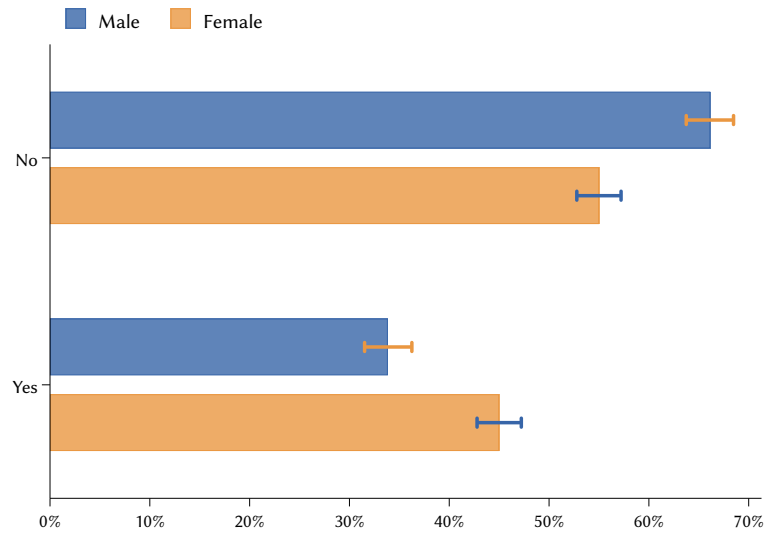
	Percent	BRR * std. err.	Logit [95% conf.	interval]
insomnia@sex				
No Male	66.2	1.2	63.7	68.5
No Female	55.0	1.1	52.8	57.2
Yes Male	33.8	1.2	31.5	36.3
Yes Female	45.0	1.1	42.8	47.2

Once again, however, the best single tool for design-based bivariate analysis is arguably the user-written command `dstat`, which offers many options for analysis, including graphical ones. For example, here is how the conditional distribution of `insomnia` by `sex` can be represented graphically with `dstat`:

```
/* Conditional distribution of "insomnia" by "sex" */
dstat proportion insomnia, over(sex) percent          ///
    vce(svy, dots(10)) graph(merge                    ///
        p1(color("55 101 168")                        ///
            ciopts(color("234 151 65") lwidth(*4))    ///
        )                                             ///
```

```
        p2(color("234 151 65")                          ///
            ciopts(color("55 101 168") lwidth(*4))      ///
        )                                               ///
        xlabel( 0  "0%" 10 "10%" 20 "20%" 30 "30%"      ///
            40 "40%" 50 "50%" 60 "60%" 70 "70%"         ///
        )                                               ///
    )
```

```
. /* Conditional distribution of "insomnia" by "sex" */
. dstat proportion insomnia, over(sex) percent          ///
>     vce(svy, dots(10)) graph(merge                    ///
>         p1(color("55 101 168")                        ///
>             ciopts(color("234 151 65") lwidth(*4))    ///
>         )                                             ///
>         p2(color("234 151 65")                        ///
>             ciopts(color("55 101 168") lwidth(*4))    ///
>         )                                             ///
>         xlabel( 0  "0%" 10 "10%" 20 "20%" 30 "30%"    ///
>             40 "40%" 50 "50%" 60 "60%" 70 "70%"       ///
>         )                                             ///
>     )
(running dstat_svyr to obtain evaluation grid)
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Percent                     Number of obs    =      8,675
                                    Population size  = 49,662,295
                                    Replications     =        152
                                    Design df        =        150

(coefficients table suppressed)
```



The following lines of code, in turn, show how to use `dstat` to estimate the *strength of the association* between `insomnia` and `sex` through five summary measures suitable for the case when both variables X and Y are dichotomous: the difference between the probability of suffering from insomnia for females and the equivalent probability for males, the ratio between these two probabilities, the odds ratio, Cramér's V (Cramér 1946), and the uncertainty coefficient (Agresti 2013):

```
/* Measures of association between "insomnia" and "sex" */
dstat summarize (b rr or cramersv ucl) insomnia, by(sex)   ///
    vce(svy, dots(10))
```

```
. /* Measures of association between "insomnia" and "sex" */
. dstat summarize (b rr or cramersv ucl) insomnia, by(sex)   ///
>     vce(svy, dots(10))
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Summary statistics          Number of obs    =      8,675
                                    Population size  = 49,662,295
                                    Replications     =        152
                                    Design df        =        150
                                    By variable      =        sex

------------------------------------------------------------------------------
             |                BRR *
    insomnia | Coefficient  std. err.      [95% conf. interval]
-------------+----------------------------------------------------------------
           b |   .1114711   .0116785       .0883955    .1345467
          rr |   1.329346   .0421772       1.246008    1.412684
          or |   1.598738   .0805319       1.439615    1.757862
    cramersv |   .1138768   .0119676       .0902301    .1375236
         ucl |   .0096931   .0020655       .0056118    .0137743
------------------------------------------------------------------------------
```

where `b` denotes the probability difference; `rr` denotes the probability ratio; `or` denotes the odds ratio; `cramersv` denotes Cramér's V; and `ucl` denotes the uncertainty coefficient.

With only minor adjustments, the procedure adopted for the analysis of 2×2 contingency tables can also be applied to the analysis of $2 \times C$, $R \times 2$, and $R \times C$ cross-tabulations. Let us examine, for example, the relationship between a dichotomous variable X (*Sex*) and an ordinal variable Y with $C = 5$ categories (*Self-reported health*). First, we calculate the conditional percent distribution of Y given X using command `tabulate`:

```
/* Cross-tabulation of variables "sex" and "srh" */
svy : tabulate sex srh, row percent ci format(%5.1f) nomarginals
```

```
. /* Cross-tabulation of variables "sex" and "srh" */
. svy : tabulate sex srh, row percent ci format(%5.1f) nomarginals
(running tabulate on estimation sample)

BRR *: for rows
                                        Number of obs    =      8,732
                                        Population size  = 49,983,951
                                        Replications     =        152
                                        Design df        =        150
```

Sex	Self-reported health Bad	Poor	Satisfac	Good	Excellen
Male	3.0 [2.1,4.1]	5.0 [4.0,6.3]	20.1 [18.6,21.6]	54.4 [52.1,56.7]	17.5 [15.9,19.3]
Female	1.3 [1.0,1.8]	6.4 [5.6,7.2]	23.7 [22.2,25.1]	55.3 [53.2,57.3]	13.4 [12.0,14.8]

```
Key: Row percentage
     [95% confidence interval for row percentage]

  Pearson:
    Uncorrected   chi2(4)         =   71.3560
    Design-based  F(3.53, 529.91) =   11.7839     P = 0.0000
```

As can be seen, the design-based Pearson test rejects the null hypothesis of independence between variables `rsh` and `sex`. Comparison of sex-specific conditional distributions reveals that women and men are broadly similar in terms of self-reported health, except that the former report both excellent and bad health somewhat less often than the latter and, at the same time, are more likely to report a satisfactory health status. This observation suggests that women's distribution is slightly less variable than men's; as evidence of the above, we can use `dstat` to estimate the difference between the two distributions in terms of Shannon entropy, finding that it is moderately negative:

```
/* Variability of "srh" : Sex differences */
dstat summarize (entropy) srh, over(sex, contrast) vce(svy, dots(10))
```

```
. /* Variability of "srh" : Sex differences */
. dstat summarize (entropy) srh, over(sex, contrast) vce(svy, dots(10))
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Difference in entropy                 Number of obs     =      8,732
                                              Population size   = 49,983,951
                                              Replications      =        152
                                              Design df         =        150
                                              Contrast          =      0.sex
```

srh	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
sex						
Female	-.0424896	.0198941	-2.14	0.034	-.0817985	-.0031807

As in the previous analysis, we can express the overall difference between women and men in terms of self-reported health by having `dstat` compute some summary measures of association applicable to situations – such as the one at hand – in which variable X, variable Y, or both are polytomous: Cramér's V and the uncertainty coefficient (see above), the generalized dissimilarity index (Reardon and Firebaugh 2002), Kendall's τ_a (Kendall 1938), Goodman and Kruskal's γ (Goodman and Kruskal 1954), and Somers' d_{YX} (Somers 1962). The first three measures are suitable for describing the strength of the association between pairs of qualitative variables of any type. The other three, on the other hand, require that both variables be ordinal, or that one be ordinal and the other dichotomous; as such, these measures also indicate the *sign of the association* between the two variables. Here is the relevant Stata code and output:

```
/* Measures of association between "srh" and "sex" */
dstat summarize (cramersv ucl dissim taua gamma somersd) srh,   ///
     by(sex) vce(svy, dots(10))
```

```
. /* Measures of association between "srh" and "sex" */
. dstat summarize (cramersv ucl dissim taua gamma somersd) srh,   ///
>      by(sex) vce(svy, dots(10))
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Summary statistics          Number of obs    =      8,732
                                    Population size = 49,983,951
                                    Replications     =        152
                                    Design df        =        150
                                    By variable      =        sex
```

srh	Coefficient	BRR * std. err.	[95% conf.	interval]
cramersv	.0903979	.0125152	.0656691	.1151267
ucl	.0034422	.000979	.0015078	.0053766
dissim	.0576096	.0086366	.0405444	.0746747
taua	-.026041	.0051109	-.0361397	-.0159423
gamma	-.0834563	.0163814	-.1158243	-.0510882
somersd	-.0521422	.0102333	-.0723623	-.0319221

where `cramersv` denotes Cramér's V; `ucl` denotes the uncertainty coefficient; `dissim` denotes the generalized dissimilarity index; `taua` denotes Kendall's τ_a; `gamma` denotes Goodman and Kruskal's γ; and `somersd` denotes Somers' d_{YX}. Taken together, these measures indicate that (a) there is an association between variables `srh` and `sex`; and (b) this association is relatively weak and of negative sign – on average, women tend to report slightly worse health status than men.

Using results left behind by `dstat`, design-based estimation and inference can also be conducted for quantities not directly supported by the command. For illustration, suppose we want to compute the generalized odds ratio α proposed by Agresti (1980) to describe the association between pairs of variables of ordinal-ordinal or dichotomous-ordinal type. Agresti shows that α is a monotonic transformation of Goodman and Kruskal's γ, namely $\alpha = (1 + \gamma)/(1 - \gamma)$. Now, the previous run of `dstat` saved the design-based point estimates and variance-covariance matrix of all requested measures of association, including γ. Knowing that the Stata internal name for γ is `_b[gamma]`, we can obtain design-based estimates of the generalized odds ratio α and the associated 95% Wald confidence interval using the official commands `nlcom` and `lincom` as follows:

```
/* Association between "srh" and "sex" : Generalized odds ratio */
nlcom ln_alpha : ln((1 + _b[gamma]) / (1 - _b[gamma])), df(150) post
lincom ln_alpha, eform
```

```
. /* Association between "srh" and "sex" : Generalized odds ratio */
. nlcom ln_alpha : ln((1 + _b[gamma]) / (1 - _b[gamma])), df(150) post
    ln_alpha: ln((1 + _b[gamma]) / (1 - _b[gamma]))
```

srh	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ln_alpha	-.1673017	.0329925	-5.07	0.000	-.2324918	-.1021115

```
. lincom ln_alpha, eform
 ( 1)  ln_alpha = 0
```

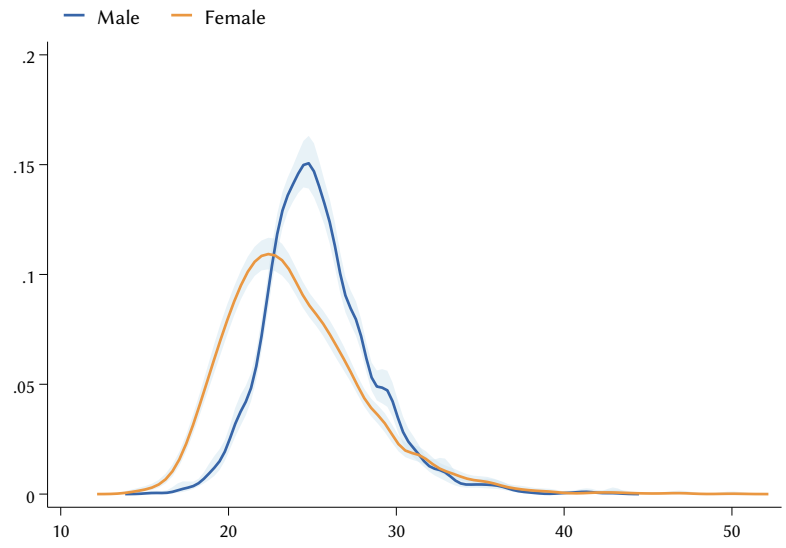
srh	exp(b)	Std. err.	t	P>|t|	[95% conf.	interval]
(1)	.8459444	.0279099	-5.07	0.000	.7925562	.9029288

Let us now look at the relationship between a dichotomous variable X (*Sex*) and a quantitative variable Y (*Body mass index*), with the aim of finding out whether and how the distribution of height-normalized body weight varies

between men and women. We start with a graphical representation of the sex-specific *probability density functions* of variable `bmi`:

```
/* Probability density function of "bmi" by "sex" */
dstat density bmi, over(sex) vce(svy, dots(10)) graph(merge   ///
        p1(color("55 101 168") lwidth(*3))                    ///
        p2(color("234 151 65") lwidth(*3))                    ///
    )
```

```
. /* Probability density function of "bmi" by "sex" */
. dstat density bmi, over(sex) vce(svy, dots(10)) graph(merge   ///
>         p1(color("55 101 168") lwidth(*3))                    ///
>         p2(color("234 151 65") lwidth(*3))                    ///
>     )
(running dstat_svyr to obtain evaluation grid and bandwidth)
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Density                   Number of obs    =      7,311
                                  Population size = 41,317,101
                                  Replications     =        152
                                  Design df        =        150
                                  Kernel           =   gaussian
                                  Bandwidth        =  e(bwidth)

(coefficients table suppressed)
```



The graph shows that the distribution of BMI among men is shifted to the right relative to that of women, although any sex difference disappears at higher BMI levels. This description is corroborated by the following set of percentiles, calculated separately for males and females:

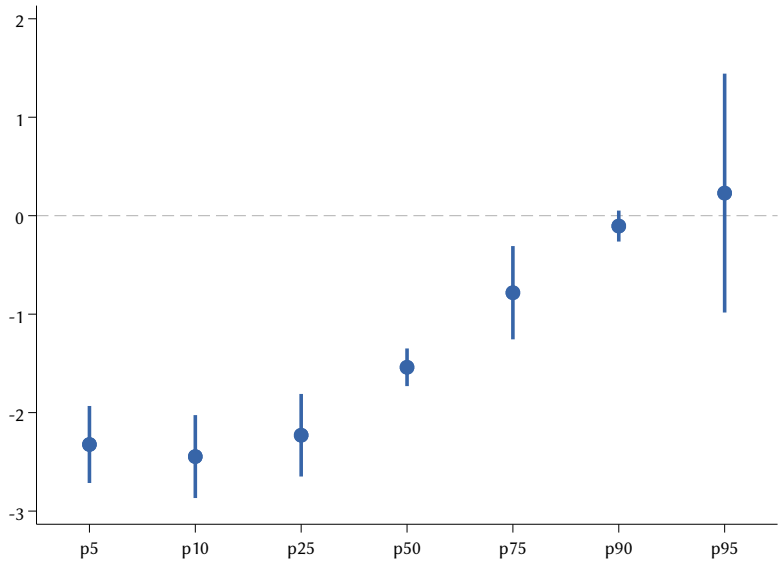
```
/* Select percentiles of "bmi" distribution by "sex" */
dstat summarize (p5 p10 p25 p50 p75 p90 p95) bmi, over(sex)   ///
    vce(svy, dots(10)) cformat(%5.2f)
```

```
. /* Select percentiles of "bmi" distribution by "sex" */
. dstat summarize (p5 p10 p25 p50 p75 p90 p95) bmi, over(sex)   ///
>     vce(svy, dots(10)) cformat(%5.2f)
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Summary statistics          Number of obs    =      7,311
                                    Population size = 41,317,101
                                    Replications     =        152
                                    Design df        =        150

           0: sex = Male
           1: sex = Female

------------------------------------------------------------------------------
             |                    BRR *
         bmi |  Coefficient  std. err.     [95% conf. interval]
-------------+----------------------------------------------------------------
0            |
          p5 |       20.75       0.12        20.52       20.98
         p10 |       21.91       0.15        21.62       22.21
         p25 |       23.32       0.10        23.12       23.53
         p50 |       24.98       0.09        24.80       25.16
         p75 |       27.04       0.13        26.78       27.30
         p90 |       29.40       0.02        29.36       29.44
         p95 |       31.02       0.21        30.60       31.44
-------------+----------------------------------------------------------------
1            |
          p5 |       18.42       0.16        18.11       18.74
         p10 |       19.47       0.16        19.14       19.79
         p25 |       21.09       0.18        20.74       21.45
         p50 |       23.44       0.04        23.37       23.51
         p75 |       26.26       0.20        25.86       26.66
         p90 |       29.30       0.08        29.14       29.45
         p95 |       31.25       0.55        30.16       32.34
------------------------------------------------------------------------------
```

By explicitly calculating the percentile differences between women and men, along with their respective design-based 95% Wald confidence intervals, the findings become even clearer:

```
/* Select percentiles of "bmi" distribution : Sex differences */
dstat summarize (p5 p10 p25 p50 p75 p90 p95) bmi,          ///
    over(sex, contrast) vce(svy, dots(10)) graph(          ///
        vertical yline(0, lpattern(dash) lcolor(gs11))     ///
        p1(msymbol(O) mcolor("55 101 168") msize(*1.5)     ///
            ciopts(color("55 101 168") lwidth(*4))         ///
        )                                                  ///
    )
```

```
. /* Select percentiles of "bmi" distribution : Sex differences */
. dstat summarize (p5 p10 p25 p50 p75 p90 p95) bmi,        ///
>     over(sex, contrast) vce(svy, dots(10)) graph(        ///
>         vertical yline(0, lpattern(dash) lcolor(gs11))   ///
>         p1(msymbol(O) mcolor("55 101 168") msize(*1.5)   ///
>             ciopts(color("55 101 168") lwidth(*4))       ///
>         )                                                ///
>     )
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Difference in summary statistics        Number of obs     =      7,311
                                                Population size   = 41,317,101
                                                Replications      =        152
                                                Design df         =        150
                                                Contrast          =      0.sex

(coefficients table suppressed)
```



To express the overall difference between women and men in terms of BMI, we can ask `dstat` to compute some summary measures of association applicable to situations – such as the one at hand – in which variable X is dichotomous and variable Y is quantitative: the mean difference; Cohen's d (Cohen 1988); the point biserial correlation coefficient r_{pb}, mathematically equivalent to the Pearson's product-moment correlation coefficient r_p (Kraemer 2006); the rank biserial correlation coefficient r_{rb}, mathematically equivalent to the Spearman's rank correlation coefficient r_s (Kraemer 2006); and Somers' d_{YX} (Somers 1962). Here is the relevant Stata code and output:

```
/* Measures of association between "bmi" and "sex" */
dstat summarize (b cohend corr spearman somersd) bmi,   ///
     by(sex) vce(svy, dots(10))
```

```
. /* Measures of association between "bmi" and "sex" */
. dstat summarize (b cohend corr spearman somersd) bmi,   ///
>      by(sex) vce(svy, dots(10))
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Summary statistics          Number of obs    =      7,311
                                    Population size = 41,317,101
                                    Replications     =        152
                                    Design df        =        150
                                    By variable      =        sex

------------------------------------------------------------------------------
             |                    BRR *
         bmi |  Coefficient  std. err.      [95% conf. interval]
-------------+----------------------------------------------------------------
           b |   -1.380588    .126866     -1.631264   -1.129913
      cohend |   -.3739251    .0366396    -.4463215   -.3015286
        corr |   -.1837714    .0174118    -.2181755   -.1493672
    spearman |   -.2306121    .0166296    -.2634705   -.1977537
     somersd |   -.2663322    .0192068    -.3042829   -.2283815
------------------------------------------------------------------------------
```

where b denotes the mean difference; cohend denotes Cohen's d; corr denotes the Pearson's product-moment correlation coefficient r_p; spearman denotes the Spearman's rank correlation coefficient r_s; and somersd denotes Somers' d_{YX}. Overall, these measures support our previous findings: (a) there is a significant association between variables bmi and sex; and (b) this association is negative: on average, women's BMI is lower than men's.

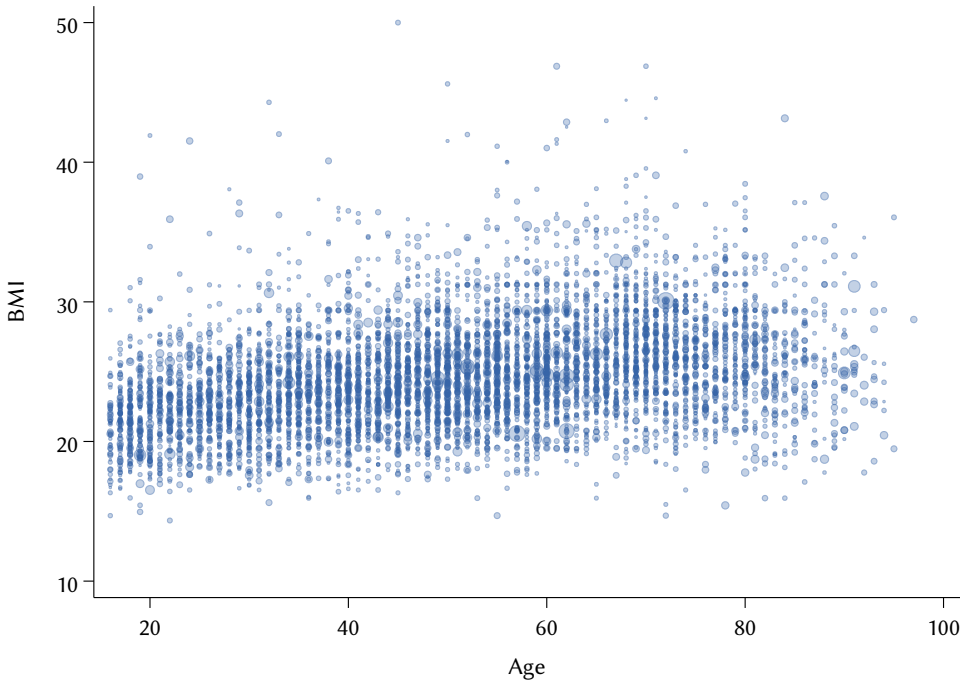
We now consider the relationship between a pair of variables that are both quantitative. Specifically, we examine whether and how the distribution of body mass index (Y) varies with age (X). The typical starting point for such an analysis is the *scatterplot*, which, by showing how subjects are distributed in the two-dimensional space defined by Y and X, provides an easy view of the shape of the relationship between the two variables. Complex survey data, however, "have two features that can make a simple scatterplot less useful. One feature, reflected in the sample weights, is that individuals in the sample represent differing numbers of individuals in the population. A second feature is that the sample sizes can be large. Scatterplots that ignore these features can be misleading or hard to interpret." (Korn and Graubard 1999).

A variant of the scatterplot better suited for representing complex survey data is the *bubble plot*, in which each point on the plot has a size proportional

to its sample weight (Korn and Graubard 1999). In Stata, this plot can be drawn as follows:

```
/* Bubble plot of "bmi" and "age" */
graph twoway scatter bmi age [pw = fiw], msymbol(o) msize(*0.2)   ///
    mcolor("55 101 168%30") ytitle("BMI")
```

```
. /* Bubble plot of "bmi" and "age" */
. graph twoway scatter bmi age [pw = fiw], msymbol(o) msize(*0.2)   ///
>     mcolor("55 101 168%30") ytitle("BMI")
```



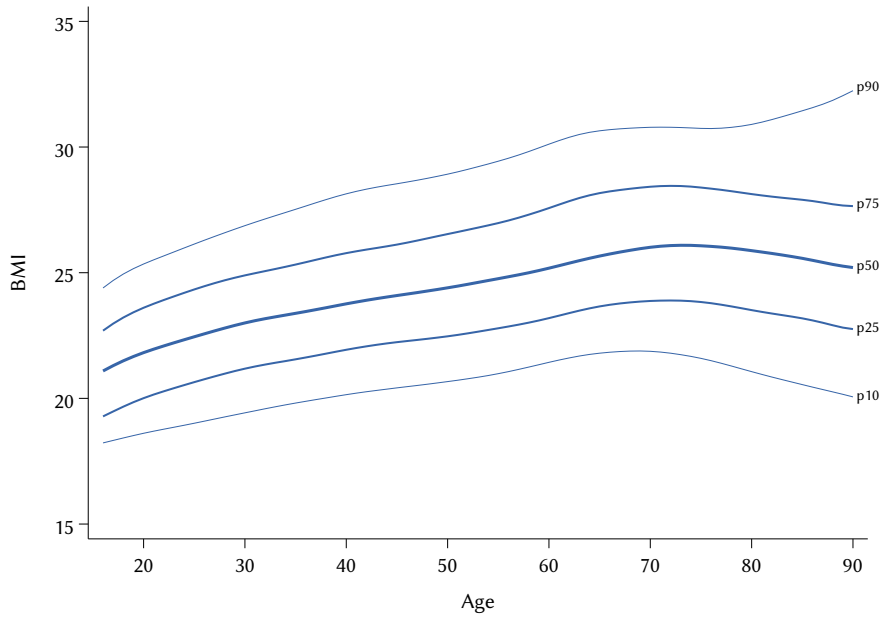
The bubble plot suggests that, for individuals up to age 70, as age increases, BMI also tends to increase. Beyond that age, the association between the two variables seems to vanish or even change sign. The picture provided by the graph, however, is far from clear.

To get a more intelligible graphical representation of the association of interest, we can calculate and plot smoothed trend lines for a selection of percentiles of the conditional distribution of bmi given age (Lohr 2022). Here is how to do it in Stata:

```
/* Smoothed trend lines for percentiles of "bmi", by "age" */
preserve
keep if inrange(age,16,90)
```

```
collapse (p10) pc10=bmi (p25) pc25=bmi (p50) pc50=bmi   ///
    (p75) pc75=bmi (p90) pc90=bmi [pw = fiw], by(age)
graph twoway                                                ///
    (lowess pc10 age, bw(0.4) lcolor("55 101 168") lwidth(*1))   ///
    (lowess pc25 age, bw(0.4) lcolor("55 101 168") lwidth(*2))   ///
    (lowess pc50 age, bw(0.4) lcolor("55 101 168") lwidth(*3))   ///
    (lowess pc75 age, bw(0.4) lcolor("55 101 168") lwidth(*2))   ///
    (lowess pc90 age, bw(0.4) lcolor("55 101 168") lwidth(*1))   ///
    ,                                                       ///
    ytitle("BMI") ylabel(15(5)35)                           ///
    xtitle("Age") xlabel(20(10)90)                          ///
    text(20.2 90.3 "p10", placement(right) size(*0.8))      ///
    text(22.9 90.3 "p25", placement(right) size(*0.8))      ///
    text(25.4 90.3 "p50", placement(right) size(*0.8))      ///
    text(27.85 90.3 "p75", placement(right) size(*0.8))     ///
    text(32.5 90.3 "p90", placement(right) size(*0.8))      ///
    legend(off)
restore
```

```
. /* Smoothed trend lines for percentiles of "bmi", by "age" */
. preserve

. keep if inrange(age,16,90)
(50 observations deleted)

. collapse (p10) pc10=bmi (p25) pc25=bmi (p50) pc50=bmi   ///
>     (p75) pc75=bmi (p90) pc90=bmi [pw = fiw], by(age)

. graph twoway                                                ///
>     (lowess pc10 age, bw(0.4) lcolor("55 101 168") lwidth(*1))   ///
>     (lowess pc25 age, bw(0.4) lcolor("55 101 168") lwidth(*2))   ///
>     (lowess pc50 age, bw(0.4) lcolor("55 101 168") lwidth(*3))   ///
>     (lowess pc75 age, bw(0.4) lcolor("55 101 168") lwidth(*2))   ///
>     (lowess pc90 age, bw(0.4) lcolor("55 101 168") lwidth(*1))   ///
>     ,                                                       ///
>     ytitle("BMI") ylabel(15(5)35)                           ///
>     xtitle("Age") xlabel(20(10)90)                          ///
>     text(20.2 90.3 "p10", placement(right) size(*0.8))      ///
>     text(22.9 90.3 "p25", placement(right) size(*0.8))      ///
>     text(25.4 90.3 "p50", placement(right) size(*0.8))      ///
>     text(27.85 90.3 "p75", placement(right) size(*0.8))     ///
>     text(32.5 90.3 "p90", placement(right) size(*0.8))      ///
>     legend(off)

. restore
```



Overall, this graph corroborates our previous observation: the association between BMI and age follows a nonlinear and non-monotonic trend, being positive up to age 70 and negative after that age.[4]

The nonlinear and non-monotonic nature of the association between variables `bmi` and age implies that its strength cannot be adequately expressed by such standard summary measures as the Pearson's product-moment correlation coefficient r_p and the Spearman's rank correlation coefficient r_s. A viable alternative is to discretize the two quantitative variables of interest and calculate the *uncertainty coefficient*, i.e., the share of entropy (variability) of the distribution of Y accounted for by X. Following Hacine-Gharbi and Ravier (2018), we bin both `bmi` and age in 15 categories of approximately equal size, and then use `dstat` for design-based estimation and inference of the uncertainty coefficient. Here is the relevant Stata code and output:

```
/* Strength of the association between "bmi" and "age" */
tempvar BMI AGE
xtile `BMI' = bmi [pw = fiw], nquantiles(15)
xtile `AGE' = age [pw = fiw], nquantiles(15)
dstat summarize (ucl) `BMI', by(`AGE') vce(svy, dots(10))
```

```
. /* Strength of the association between "bmi" and "age" */
. tempvar BMI AGE
```

4 To ensure an adequate number of observations for each age, the analysis was limited to respondents up to 90 years old.

```
. xtile `BMI´ = bmi [pw = fiw], nquantiles(15)
. xtile `AGE´ = age [pw = fiw], nquantiles(15)
. dstat summarize (ucl) `BMI´, by(`AGE´) vce(svy, dots(10))
(running dstat_svyr on estimation sample)
BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........
Survey: Summary statistics          Number of obs    =      7,311
                                    Population size = 41,317,101
                                    Replications     =        152
                                    Design df        =        150
                                    By variable      =   __000002
```

__000001	Coefficient	BRR * std. err.	[95% conf.	interval]
ucl	.037688	.0040338	.0297175	.0456584

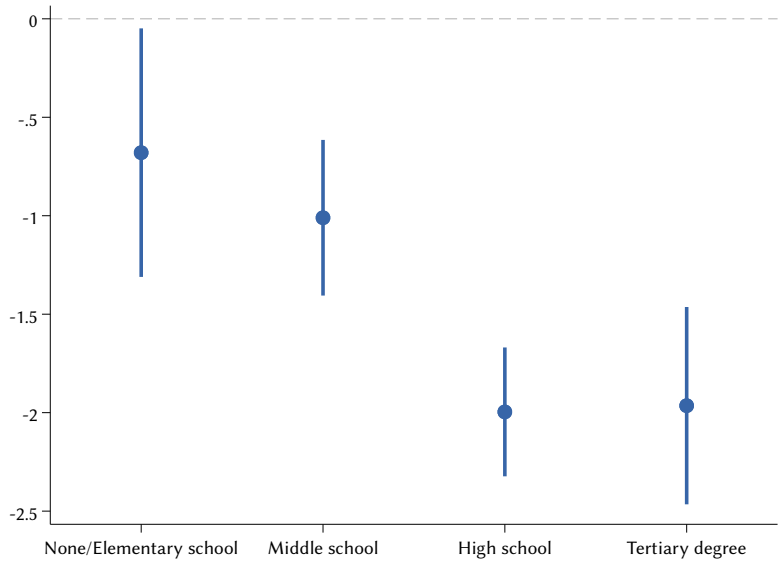
Based on the output, we can conclude that there is a significant association between (discretized) body mass index and age. Knowing the value of the latter variable, in fact, allows us to reduce by 3.8% (95% Wald confidence interval: 3–4.6) the entropy of the distribution of variable `bmi`.

To conclude our overview of bivariate analysis, it is worth pointing out that design-based subpopulation analysis can also be conducted when the quantity of interest is a measure of association between two variables. Suppose, for example, that we are interested in investigating whether the difference between women and men in terms of BMI that we observed earlier varies across levels of education. To this end, we first determine the number of design degrees of freedom most appropriate for the analysis. Looking at the frequency distribution of variable *Educational degree* in our sample, we can see that the lowest frequency, associated with category "Tertiary degree", is 1,014. Applying the rules of thumb specified in Section 4.3, such a frequency corresponds to 120 degrees of freedom. With this information, we can carry out the subpopulation analysis of interest using command `dstat` as follows:

```
/* Association between "bmi" and "sex", by "educ" */
dstat summarize (b) bmi, by(sex) over(educ) table          ///
    vce(svy, dots(10) dof(120)) graph(vertical             ///
        yline(0, lpattern(dash) lcolor(gs11))              ///
        p1(msymbol(O) mcolor("55 101 168") msize(*1.5)     ///
            ciopts(color("55 101 168") lwidth(*4))         ///
        )                                                  ///
    )
```

```
. /* Association between "bmi" and "sex", by "educ" */
. dstat summarize (b) bmi, by(sex) over(educ) table          ///
>     vce(svy, dots(10) dof(120)) graph(vertical             ///
>         yline(0, lpattern(dash) lcolor(gs11))              ///
>         p1(msymbol(O) mcolor("55 101 168") msize(*1.5)     ///
>             ciopts(color("55 101 168") lwidth(*4))         ///
>         )                                                  ///
>     )
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: b                          Number of obs    =      7,311
                                   Population size = 41,317,101
                                   Replications     =        152
                                   Design df        =        120
                                   By variable      =        sex
```

bmi	Coefficient	BRR * std. err.	[95% conf.	interval]
educ				
None/Elementary school	-.6797125	.3187177	-1.310751	-.0486738
Middle school	-1.010203	.1995714	-1.40534	-.6150653
High school	-1.996131	.1652202	-2.323255	-1.669007
Tertiary degree	-1.964464	.2530231	-2.465432	-1.463496



The analysis clearly shows that the association between BMI and sex varies significantly by educational degree. Specifically, the mean difference between women and men in terms of body mass index is substantially smaller among individuals who have not gone beyond the middle school diploma than among those who have a high school or university degree.

4.5. *Multiple Regression Analysis*

The general purpose of *multiple regression analysis* is to describe whether and how the distribution of a given variable of interest Y, or a summary measure of it, varies among subpopulations jointly defined by two or more (i.e., multiple) covariates. As such, multiple regression analysis can be regarded as an extension of bivariate analysis in which the covariate X represents not a single variable, but the *combination* of several variables.

To illustrate this point, suppose we are interested in describing whether and how the mean body mass index varies with sex and age. One possible way to approach this task is to carry out two *separate* bivariate analyses: the first focusing on the relationship between BMI and sex, the other examining the relationship between BMI and age. This is precisely what we did in the previous section. Alternatively, one could perform a *single* multiple regression analysis investigating the relationship between BMI and a "super covariate" X formed by the combination of sex and age. To the extent that these two variables are correlated and/or interact with each other, the two approaches – repeated bivariate analysis or single multiple regression analysis – will yield different results.

Multiple regression analysis can be implemented in several ways. In the simplest one, which we will call the *cross-classification* approach, the super covariate X amounts to a single qualitative variable formed by fully crossing the original covariates of interest. The analysis, then, consists of calculating the distribution of variable Y (or a summary measure of it) within each category of X.

In terms of our example, the cross-classification approach goes like this. First, we generate a qualitative variable X formed by as many categories as there are possible combinations of sex and age; if we group age into seven 10-year intervals (leaving the last interval open-ended), X will thus comprise $2 \times 7 = 14$ categories. In Stata:

```
/* Generate and display super covariate X ("sex" by "agegroup") */
egen X = group(sex agegroup), label
tabulate X
```

```
. /* Generate and display super covariate X ("sex" by "agegroup") */
. egen X = group(sex agegroup), label

. tabulate X

    group(sex agegroup) |      Freq.     Percent        Cum.
------------------------+-----------------------------------
       Male 16-24 years |        442        5.04        5.04
       Male 25-34 years |        521        5.94       10.97
```

Male 35-44 years	594	6.77	17.74
Male 45-54 years	717	8.17	25.91
Male 55-64 years	656	7.47	33.38
Male 65-74 years	628	7.15	40.53
Male 75 years and over	452	5.15	45.68
Female 16-24 years	415	4.73	50.41
Female 25-34 years	598	6.81	57.22
Female 35-44 years	713	8.12	65.35
Female 45-54 years	878	10.00	75.35
Female 55-64 years	802	9.14	84.48
Female 65-74 years	749	8.53	93.02
Female 75 years and over	613	6.98	100.00
Total	8,778	100.00	

Next, we use `dstat` for design-based estimation and inference of the mean values of BMI within each category of X:

```
/* Mean "bmi", by super covariate X */
preserve
dstat summarize (mean) bmi, over(X) cformat(%5.2f) vce(svy, dots(10))
matrix B = e(b)'
matrix CI = e(ci)'
svmat B
svmat CI
egen g_sex = seq() in 1/14, from(0) to(1) block(7)
egen g_age = seq() in 1/14, from(1) to(7) block(1)
label define l_age 1 "16-24 yrs", modify
label define l_age 2 "25-34 yrs", modify
label define l_age 3 "35-44 yrs", modify
label define l_age 4 "45-54 yrs", modify
label define l_age 5 "55-64 yrs", modify
label define l_age 6 "65-74 yrs", modify
label define l_age 7 "75+ yrs", modify
label values g_age l_age
graph twoway                                                    ///
    (rspike CI1 CI2 g_age if g_sex==0, lcolor("55 101 168")     ///
     lwidth(*5)                                                 ///
    )                                                           ///
    (scatter B g_age if g_sex==0, msymbol(O) msize(*1.5)        ///
     mcolor("55 101 168")                                       ///
    )                                                           ///
    (rspike CI1 CI2 g_age if g_sex==1, lcolor("234 151 65")     ///
     lwidth(*5)                                                 ///
    )                                                           ///
    (scatter B g_age if g_sex==1, msymbol(O) msize(*1.5)        ///
     mcolor("234 151 65")                                       ///
    )                                                           ///
    ,                                                           ///
    ytitle("BMI") xtitle("Age") xlabel(1(1)7, valuelabel)       ///
```

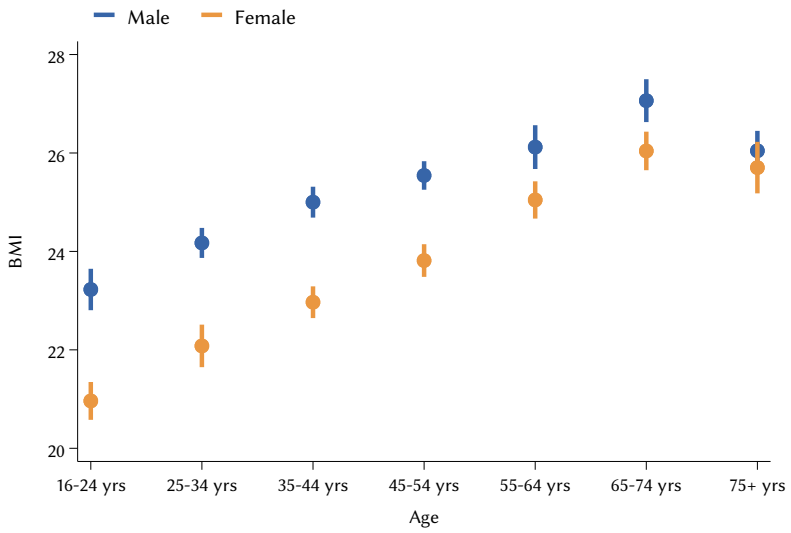
```
    legend(order(1 "Male" 3 "Female"))
restore
drop X
```

```
. /* Mean "bmi", by super covariate X */
. preserve

. dstat summarize (mean) bmi, over(X) cformat(%5.2f) vce(svy, dots(10))
(running dstat_svyr on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: mean                        Number of obs    =      7,311
                                    Population size = 41,317,101
                                    Replications     =        152
                                    Design df        =        150
```

bmi	BRR * Coefficient	std. err.	[95% conf.	interval]
X				
Male 16-24 years	23.23	0.21	22.81	23.65
Male 25-34 years	24.17	0.15	23.87	24.48
Male 35-44 years	25.00	0.16	24.69	25.31
Male 45-54 years	25.54	0.15	25.25	25.83
Male 55-64 years	26.12	0.23	25.67	26.56
Male 65-74 years	27.06	0.22	26.63	27.50
Male 75 years and over	26.05	0.20	25.64	26.45
Female 16-24 years	20.96	0.19	20.58	21.35
Female 25-34 years	22.08	0.22	21.65	22.51
Female 35-44 years	22.97	0.16	22.65	23.29
Female 45-54 years	23.81	0.17	23.48	24.15
Female 55-64 years	25.05	0.19	24.67	25.42
Female 65-74 years	26.04	0.20	25.65	26.43
Female 75 years and over	25.70	0.26	25.18	26.23

```
. matrix B = e(b)'

. matrix CI = e(ci)'

. svmat B

. svmat CI

. egen g_sex = seq() in 1/14, from(0) to(1) block(7)
(8,764 missing values generated)

. egen g_age = seq() in 1/14, from(1) to(7) block(1)
(8,764 missing values generated)

. label define l_age 1 "16-24 yrs", modify

. label define l_age 2 "25-34 yrs", modify

. label define l_age 3 "35-44 yrs", modify

. label define l_age 4 "45-54 yrs", modify

. label define l_age 5 "55-64 yrs", modify

. label define l_age 6 "65-74 yrs", modify

. label define l_age 7 "75+ yrs", modify
```

```
. label values g_age l_age
. graph twoway                                                          ///
>     (rspike CI1 CI2 g_age if g_sex==0, lcolor("55 101 168")           ///
>      lwidth(*5)                                                       ///
>     )                                                                 ///
>     (scatter B g_age if g_sex==0, msymbol(O) msize(*1.5)              ///
>      mcolor("55 101 168")                                             ///
>     )                                                                 ///
>     (rspike CI1 CI2 g_age if g_sex==1, lcolor("234 151 65")           ///
>      lwidth(*5)                                                       ///
>     )                                                                 ///
>     (scatter B g_age if g_sex==1, msymbol(O) msize(*1.5)              ///
>      mcolor("234 151 65")                                             ///
>     )                                                                 ///
>     ,                                                                 ///
>     ytitle("BMI") xtitle("Age") xlabel(1(1)7, valuelabel)             ///
>     legend(order(1 "Male" 3 "Female"))
. restore
. drop X
```



On the one hand, our multiple regression analysis simply confirms the previous bivariate results (see Section 4.4): (a) on average, men's BMI is higher than women's; and (b) BMI tends to increase with age, but decreases among those over 75. On the other hand, however, the analysis reveals a finding that the two separate bivariate analyses could not capture: the difference between men and women gradually decreases with increasing age, until it disappears from age 75 onward.

An alternative method of multiple regression analysis is the *modeling* approach, according to which the estimand of interest – i.e., the distribution

of Y (or a summary measure of it) given X – should be viewed as the sum of several components: (a) a *baseline value*; (b) the *main effects* of the constituent covariates of X; and (c) the *interaction (joint) effects* of those covariates.[5] The modeling approach requires that the values of all these components be first calculated, and then appropriately combined to generate the estimates of the quantities of interest.

Let us see how this works in our example. First, we can express symbolically the (conditional) mean values of BMI given X as follows:

$$E(\texttt{BMI}|X) = b_0 + \texttt{sex} + \texttt{agegroup} + \texttt{sex} \times \texttt{agegroup}$$

where $E(\texttt{BMI}|X)$ denotes the mean values of BMI given X; b_0 denotes the baseline value; sex denotes the main effect of variable *Sex*; agegroup denotes the main effect of variable *Age group*; and sex × agegroup denotes the interaction (joint) effect of the two variables. This equation represents a *linear regression model*, whose components can be estimated in Stata using command `regress` prefixed by svy:

```
/* Linear regression of "bmi" on "sex" and "agegroup" (full model) */
svy, dots(10) : regress bmi i.sex i.agegroup i.sex#i.agegroup,   ///
    cformat(%6.3f) noci
```

```
. /* Linear regression of "bmi" on "sex" and "agegroup" (full model) */
. svy, dots(10) : regress bmi i.sex i.agegroup i.sex#i.agegroup,   ///
>     cformat(%6.3f) noci
(running regress on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Linear regression                       Number of obs   =      7,311
                                                Population size = 41,317,101
                                                Replications    =        152
                                                Design df       =        150
                                                F(13, 138)      =      75.53
                                                Prob > F        =     0.0000
                                                R-squared       =     0.1725

------------------------------------------------------------------------------
             |             BRR *
         bmi |  Coefficient  std. err.      t    P>|t|
-------------+----------------------------------------------------------------
         sex |
      Female |     -2.264      0.295    -7.67   0.000
             |
    agegroup |
 25-34 years |      0.946      0.214     4.42   0.000
------------------------------------------------------------------------------
```

5 It is important to stress that the term "effect" is used here in a strictly mathematical sense, without any causal implication whatsoever.

35-44 years	1.774	0.223	7.96	0.000
45-54 years	2.317	0.227	10.22	0.000
55-64 years	2.893	0.324	8.93	0.000
65-74 years	3.837	0.278	13.78	0.000
75 years and over	2.819	0.317	8.90	0.000
sex#agegroup				
Female#25-34 years	0.171	0.362	0.47	0.637
Female#35-44 years	0.231	0.340	0.68	0.498
Female#45-54 years	0.535	0.296	1.81	0.072
Female#55-64 years	1.190	0.398	2.99	0.003
Female#65-74 years	1.242	0.424	2.93	0.004
Female#75 years and over	1.922	0.451	4.27	0.000
_cons	23.226	0.213	109.08	0.000

In the output generated by Stata, _cons represents the baseline value; the coefficient under the heading sex represents the main effect of variable *Sex*; the coefficients under the heading agegroup represent the main effect of variable *Age group*; and the coefficients under the heading sex#agegroup represent the interaction effect of the two variables.

By appropriately combining the values reported in the table above, it is possible to obtain the design-based estimates of the quantities of interest, i.e., the 14 mean values of BMI given X and the associated confidence intervals. For this purpose, we can use the Stata command `predictnl` as follows:

```
/* Predicted mean values of "bmi", by "sex" and "agegroup" */
tempvar MU LB UB
predictnl `MU' = predict() if e(sample), ci(`LB' `UB') df(150)
label variable `LB' "95% lower bound"
label variable `UB' "95% upper bound"
table (agegroup) () (sex), stat(mean `MU' `LB' `UB') nototal   ///
    nformat(%5.2f)
```

```
. /* Predicted mean values of "bmi", by "sex" and "agegroup" */
. tempvar MU LB UB
. predictnl `MU' = predict() if e(sample), ci(`LB' `UB') df(150)
(1,467 missing values generated)
note: confidence intervals calculated using t(150) critical values.
. label variable `LB' "95% lower bound"
. label variable `UB' "95% upper bound"
. table (agegroup) () (sex), stat(mean `MU' `LB' `UB') nototal   ///
>     nformat(%5.2f)
Sex = Male
```

	Prediction	95% lower bound	95% upper bound
Age group			

16-24 years	23.23	22.81	23.65
25-34 years	24.17	23.87	24.48
35-44 years	25.00	24.69	25.31
45-54 years	25.54	25.25	25.83
55-64 years	26.12	25.67	26.56
65-74 years	27.06	26.63	27.50
75 years and over	26.05	25.64	26.45

Sex = Female

	Prediction	95% lower bound	95% upper bound
Age group			
16-24 years	20.96	20.58	21.35
25-34 years	22.08	21.65	22.51
35-44 years	22.97	22.65	23.29
45-54 years	23.81	23.48	24.15
55-64 years	25.05	24.67	25.42
65-74 years	26.04	25.65	26.43
75 years and over	25.70	25.18	26.23

It is worth noting that, in this example, the design-based estimates of the quantities of interest obtained by the modeling approach exactly match those generated by the cross-classification approach. This is because, in general, the cross-classification approach is implicitly equivalent to a *saturated regression model*, i.e., a regression model that, in addition to the indispensable baseline value and main effects of the chosen covariates, includes *all* possible interaction effects among such covariates. Now, the linear regression model used in our example is precisely a saturated model because, apart from the baseline value and the main effects of sex and age group, it includes the interaction effect between these two variables – the only one possible given the scope of the analysis.[6]

Its equivalence to the saturated model, together with the practical difficulties that arise as the number of covariates included in the analysis increases, makes the cross-classification approach unsuitable for most situations. Modeling, on the other hand, is very flexible and can be applied to any regression analysis. The strengths of this approach are particularly evident in analyses involving many covariates, where most interaction effects (especially those of a higher order) make a negligible contribution to determining the estimands

6 Different definitions of a saturated model exist in the literature (cf. Bellocco and Algeri 2013). According to the one adopted here, we call *saturated* any regression model that includes as many parameters as there are distinct covariate patterns. In our example, the analysis involves 14 distinct covariate patterns, which is exactly the number of parameters in the chosen linear regression model: one for the baseline value; one for the main effect of sex; six for the main effect of age group; and six for the interaction effect between the two variables.

of interest. In these cases, the modeling approach allows the analyst to specify a regression model that omits the irrelevant interaction effects, thereby obtaining more efficient (and, possibly, less biased) estimates of the quantities of interest. Thus, modeling can easily be considered as the preferred method for multiple regression analysis.

When using the modeling approach, however, care should be taken to specify a regression model that *fits* the observed data well, i.e., that generates estimates of the quantities of interest consistent with the corresponding observed values. In general, a regression model fits the observed data well when three conditions are met: (a) the main effects of quantitative covariates are specified with the correct functional form; (b) all relevant interaction effects are included in the model; and (c) all irrelevant interaction effects are excluded from the model.

Currently, there are not many tools for assessing the *goodness of fit* of regression models in the analysis of complex survey data. In Stata it is possible to use the `linktest` command, in combination with one or more adjusted Wald tests. The former implements a test "that, conditional on the [model] specification, the independent variables are specified incorrectly" (StataCorp 2021a, p. 1286). In this test, variable Y is regressed on the linear prediction of the model and its square: if the model is specified correctly – i.e., fits the observed data well – then the squared linear prediction will have no explanatory power.

To see how this works in practice, let us resume our example. The regression model we estimated is, as we have noted, a saturated model, so it fits the observed data perfectly by definition. We can ask, however, whether the interaction effect of sex and age group is actually relevant or can instead be excluded from the model as insignificant. To answer this question, we re-estimate the model without interaction and run command `linktest` immediately afterwards:

```
/* Linear regression of "bmi" on "sex" and "agegroup" (main effects) */
svy, dots(10) : regress bmi i.sex i.agegroup, cformat(%6.3f) noci
linktest
```

```
. /* Linear regression of "bmi" on "sex" and "agegroup" (main effects) */
. svy, dots(10) : regress bmi i.sex i.agegroup, cformat(%6.3f) noci
(running regress on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Linear regression                       Number of obs    =      7,311
                                                Population size  = 41,317,101
                                                Replications     =        152
                                                Design df        =        150
```

```
                                                F(7, 144)       =     128.14
                                                Prob > F        =     0.0000
                                                R-squared       =     0.1658
```

bmi	Coefficient	BRR * std. err.	t	P>|t|
sex				
Female	-1.502	0.122	-12.27	0.000
agegroup				
25-34 years	1.031	0.180	5.71	0.000
35-44 years	1.884	0.170	11.08	0.000
45-54 years	2.571	0.176	14.62	0.000
55-64 years	3.475	0.210	16.51	0.000
65-74 years	4.455	0.196	22.74	0.000
75 years and over	3.881	0.256	15.18	0.000
_cons	22.858	0.159	143.52	0.000

```
. linktest
(running regress on estimation sample)
BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..
Survey: Linear regression                       Number of obs   =      7,311
                                                Population size = 41,317,101
                                                Replications    =        152
                                                Design df       =        150
                                                F(2, 149)       =     432.10
                                                Prob > F        =     0.0000
                                                R-squared       =     0.1687
```

bmi	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
_hat	4.629035	.9058621	5.11	0.000	2.839138	6.418933
_hatsq	-.074157	.018528	-4.00	0.000	-.1107666	-.0375474
_cons	-44.22219	11.0385	-4.01	0.000	-66.03322	-22.41115

As we can see, the squared linear prediction (referred to as _hatsq in the output) is statistically significant, so we can conclude that the model without interaction does not fit the observed data well.

The relevance of the interaction between sex and age group can be corroborated by re-estimating the full (saturated) model and performing an adjusted Wald test (Korn and Graubard 1999) on the interaction itself:

```
/* Linear regression of "bmi" on "sex" and "agegroup" (full model) */
svy, dots(10) : regress bmi i.sex i.agegroup i.sex#i.agegroup,   ///
    cformat(%6.3f) noci
testparm i.sex#i.agegroup
```

```
. /* Linear regression of "bmi" on "sex" and "agegroup" (full model) */
. svy, dots(10) : regress bmi i.sex i.agegroup i.sex#i.agegroup,   ///
>     cformat(%6.3f) noci
(running regress on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Survey: Linear regression                       Number of obs    =      7,311
                                                Population size  = 41,317,101
                                                Replications     =        152
                                                Design df        =        150
                                                F(13, 138)       =      75.53
                                                Prob > F         =     0.0000
                                                R-squared        =     0.1725

------------------------------------------------------------------------------
                             |                BRR *
                         bmi |  Coefficient  std. err.      t    P>|t|
-----------------------------+------------------------------------------------
                         sex |
                      Female |     -2.264      0.295    -7.67   0.000
                             |
                    agegroup |
                  25-34 years |      0.946      0.214     4.42   0.000
                  35-44 years |      1.774      0.223     7.96   0.000
                  45-54 years |      2.317      0.227    10.22   0.000
                  55-64 years |      2.893      0.324     8.93   0.000
                  65-74 years |      3.837      0.278    13.78   0.000
            75 years and over |      2.819      0.317     8.90   0.000
                             |
                 sex#agegroup |
          Female#25-34 years |      0.171      0.362     0.47   0.637
          Female#35-44 years |      0.231      0.340     0.68   0.498
          Female#45-54 years |      0.535      0.296     1.81   0.072
          Female#55-64 years |      1.190      0.398     2.99   0.003
          Female#65-74 years |      1.242      0.424     2.93   0.004
    Female#75 years and over |      1.922      0.451     4.27   0.000
                             |
                       _cons |     23.226      0.213   109.08   0.000
------------------------------------------------------------------------------

. testparm i.sex#i.agegroup

Adjusted Wald test

 ( 1)  1.sex#2.agegroup = 0
 ( 2)  1.sex#3.agegroup = 0
 ( 3)  1.sex#4.agegroup = 0
 ( 4)  1.sex#5.agegroup = 0
 ( 5)  1.sex#6.agegroup = 0
 ( 6)  1.sex#7.agegroup = 0

       F(  6,   145) =    4.83
            Prob > F =    0.0002
```

The test clearly confirms the significance of the interaction effect of sex and age group, thereby supporting the choice of the saturated model.

If the chosen regression model fits the observed data well, one can correctly estimate its *predictive power*, i.e., the extent to which the main and interaction effects included in the model "explain" the observed variation in variable Y. For the linear regression model, a popular measure of predictive power is the *coefficient of determination* R^2, for which exists a design-based version (Korn and Graubard 1999). As we can see from the most recent output, for our saturated model the design-based R^2 (referred to as `R-squared` in the output) equals 0.1725, meaning that the main effects of sex and age group, along with their interaction, account for just over 17% of the observed variation in variable `bmi`. To quantify the uncertainty around this value, we can estimate the corresponding design-based 95% Wald confidence interval as follows:

```
/* Predictive power of the full model (point estimate and 95% CI) */
svy brr R2 = e(r2), dots(10) cformat(%6.4f) : regress bmi i.sex   ///
    i.agegroup i.sex#i.agegroup if (bmi < .)
```

```
. /* Predictive power of the full model (point estimate and 95% CI) */
. svy brr R2 = e(r2), dots(10) cformat(%6.4f) : regress bmi i.sex   ///
>     i.agegroup i.sex#i.agegroup if (bmi < .)
(running regress on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

Linear regression                               Number of obs   =      7,311
                                                Population size = 41,317,101
                                                Replications    =        152
                                                Design df       =        150

      Command: regress bmi i.sex i.agegroup i.sex#i.agegroup if (bmi < .)
           R2: e(r2)

------------------------------------------------------------------------------
             |                 BRR *
             | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          R2 |      0.1725     0.0118    14.66   0.000       0.1493       0.1958
------------------------------------------------------------------------------
```

According to the estimated confidence interval, the predictive power of the chosen regression model – as measured by R^2 – is likely to take a value between 14.9% and 19.6% in the target population.

It may sometimes be of interest to estimate the relative contribution of each effect to the predictive power of a given regression model. One of the most popular methods in this respect is *dominance analysis*, which uses an iterative procedure to decompose the value of the chosen measure of predictive power

into a set of additive components, one for each effect in the model of interest (Budescu 1993).

In Stata, design-based dominance analysis can be performed using the user-written command `domin` (Luchman 2021). In the absence of interaction effects, the use of `domin` is simple and straightforward. On the other hand, if the regression model of interest includes interactions, as in our example, a more elaborate procedure must be followed. Specifically, we first use `domin` to calculate the contribution of the main effects of sex and age group to the predictive power of a model from which the interaction between the two variables has been excluded. We then calculate the contribution of the interaction effect as the difference between the predictive power of the full (saturated) model and the predictive power of the model without the interaction. Here is the relevant Stata code:

```
/* Dominance analysis of the full linear regression model */
capture program drop dominance

program dominance, rclass
version 17.0
syntax anything [if] [iw pw]
if "`weight'" != "" {
    local wgtexp "[`weight' `exp']"
}
tempname R2TOT R2SEX R2AGE R2INT
quietly {
    regress bmi i.sex i.agegroup i.sex#i.agegroup `wgtexp'
    scalar `R2TOT' = e(r2)
    domin bmi `wgtexp', reg(regress) fitstat(e(r2))   ///
        sets( (i.sex) (i.agegroup) )
    scalar `R2SEX' = el(e(b),1,1)
    scalar `R2AGE' = el(e(b),1,2)
    scalar `R2INT' = `R2TOT' - `R2SEX' - `R2AGE'
}
return scalar r2tot = `R2TOT'
return scalar r2sex = `R2SEX'
return scalar r2age = `R2AGE'
return scalar r2int = `R2INT'
end

svy brr (Effects : sex = r(r2sex) agegroup = r(r2age)   ///
         sex_by_agegroup = r(r2int))                    ///
        (Total : Model = r(r2tot))                      ///
        , dots(10) cformat(%6.4f) :                     ///
        dominance analysis if (bmi < .)
```

```
. /* Dominance analysis of the full linear regression model */
. capture program drop dominance

.
. program dominance, rclass
  1. version 17.0
  2. syntax anything [if] [iw pw]
  3. if "`weight'" != "" {
  4.     local wgtexp "[`weight' `exp']"
  5. }
  6. tempname R2TOT R2SEX R2AGE R2INT
  7. quietly {
  8.     regress bmi i.sex i.agegroup i.sex#i.agegroup `wgtexp'
  9.     scalar `R2TOT' = e(r2)
 10.     domin bmi `wgtexp', reg(regress) fitstat(e(r2))   ///
>         sets( (i.sex) (i.agegroup) )
 11.     scalar `R2SEX' = el(e(b),1,1)
 12.     scalar `R2AGE' = el(e(b),1,2)
 13.     scalar `R2INT' = `R2TOT' - `R2SEX' - `R2AGE'
 14. }
 15. return scalar r2tot = `R2TOT'
 16. return scalar r2sex = `R2SEX'
 17. return scalar r2age = `R2AGE'
 18. return scalar r2int = `R2INT'
 19. end

.
. svy brr (Effects : sex = r(r2sex) agegroup = r(r2age)   ///
>          sex_by_agegroup = r(r2int))                    ///
>         (Total : Model = r(r2tot))                      ///
>         , dots(10) cformat(%6.4f) :                     ///
>         dominance analysis if (bmi < .)
(running dominance on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

BRR results                                     Number of obs   =      7,311
                                                Population size = 41,317,101
                                                Replications    =        152
                                                Design df       =        150

             Command: dominance analysis if (bmi < .)
         [Effects]sex: r(r2sex)
     [Effects]agegroup: r(r2age)
[Effects]sex_by_ageg~p: r(r2int)
         [Total]Model: r(r2tot)
```

	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
Effects						
sex	0.0368	0.0065	5.65	0.000	0.0239	0.0496
agegroup	0.1291	0.0101	12.73	0.000	0.1090	0.1491
sex_by_agegroup	0.0067	0.0025	2.63	0.009	0.0017	0.0117
Total						
Model	0.1725	0.0118	14.66	0.000	0.1493	0.1958

As can be seen, the dominance analysis shows that most of the predictive power of the regression model under consideration can be attributed to the main effect of age group: out of a total R^2 of 0.1725, the contribution of this effect is 0.1291 (75% in relative terms). This is followed at some distance by the contribution of the main effect of sex, which amounts to 0.0368 (21% in relative terms), and the contribution of the interaction effect of sex and age group, equal to 0.0067 (4%).

Multiple regression analysis can be performed on any type of Y variable. For illustration, let us now consider the case of a dichotomous Y. Specifically, suppose we are interested in describing whether and how, among individuals aged 25-64, the probability of reporting low levels of overall life satisfaction varies with three covariates: region of residence, educational degree and employment status. Thirty-one respondents did not answer the question on life satisfaction, so the number of valid cases for analysis in the age group of interest is 5,459.

In Stata, we first create the new Y variable `unsatisfied`:

```
/* Create variable "unsatisfied" */
generate unsatisfied = (lifesat < 6) if (lifesat < .)
label variable unsatisfied "Overall life satisfaction < 6"
label define l_unsatisfied 0 "No", modify
label define l_unsatisfied 1 "Yes", modify
label values unsatisfied l_unsatisfied
```

```
. /* Create variable "unsatisfied" */
. generate unsatisfied = (lifesat < 6) if (lifesat < .)
(31 missing values generated)
. label variable unsatisfied "Overall life satisfaction < 6"
. label define l_unsatisfied 0 "No", modify
. label define l_unsatisfied 1 "Yes", modify
. label values unsatisfied l_unsatisfied
```

Here is the percent frequency distribution of the new variable in the age group of interest:

```
/* Percent distribution of variable "unsatisfied" */
svy, subpop(if inrange(age,25,64)) dof(146) :           ///
    tabulate unsatisfied, percent se ci format(%5.1f)
```

```
. /* Percent distribution of variable "unsatisfied" */
. svy, subpop(if inrange(age,25,64)) dof(146) :           ///
>     tabulate unsatisfied, percent se ci format(%5.1f)
(running tabulate on estimation sample)
```

```
                                Number of obs    =      8,758
                                Population size  = 50,032,626
                                Subpop. no. obs  =      5,459
                                Subpop. size     = 31,585,318
                                Replications     =        152
                                Design df        =        146
```

Overall life satisfaction < 6	percentage	se	lb	ub
No	91.0	0.7	89.5	92.4
Yes	9.0	0.7	7.6	10.5
Total	100.0			

```
Key: percentage = Cell percentage
             se = Brr standard error of cell percentage
             lb = Lower 95% confidence bound for cell percentage
             ub = Upper 95% confidence bound for cell percentage
```

The quantities of interest in our analysis are the (conditional) probabilities of reporting low levels of overall life satisfaction given the region of residence, the educational degree and the employment status. To estimate these quantities, we use a *binomial logistic regression model* that expresses the target probabilities as a function of the effects of the three selected covariates. We start with an extensive specification that includes the baseline value, the main effects of the covariates, and all possible two-way interaction effects. In Stata:

```
/* Binomial logistic regression model : Initial specification */
svy, subpop(if inrange(age,25,64)) dof(146) dots(10) :     ///
    logit unsatisfied i.region i.educ i.empstat            ///
    i.region#i.educ i.region#i.empstat i.educ#i.empstat,   ///
    cformat(%6.3f) noci
```

```
. /* Binomial logistic regression model : Initial specification */
. svy, subpop(if inrange(age,25,64)) dof(146) dots(10) :     ///
>     logit unsatisfied i.region i.educ i.empstat            ///
>     i.region#i.educ i.region#i.empstat i.educ#i.empstat,   ///
>     cformat(%6.3f) noci
(running logit on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

note: 1b.region#3.empstat != 0 predicts failure perfectly;
      1b.region#3.empstat omitted and 20 obs not used.
note: 3.region#3.empstat != 0 predicts failure perfectly;
      3.region#3.empstat omitted and 22 obs not used.
note: 5.region#3.empstat != 0 predicts failure perfectly;
      5.region#3.empstat omitted and 5 obs not used.
note: 2.educ#3.empstat != 0 predicts failure perfectly;
```

```
      2.educ#3.empstat omitted and 1 obs not used.
note: 4.region#3.empstat omitted because of collinearity.
note: 1b.educ#3.empstat != 0 predicts failure perfectly;
      1b.educ#3.empstat omitted and 0 obs not used.
note: 4.educ#3.empstat omitted because of collinearity.
Survey: Logistic regression                     Number of obs   =      8,710
                                                Population size = 49,693,185
                                                Subpop. no. obs =      5,411
                                                Subpop. size    = 31,245,877
                                                Replications    =        152
                                                Design df       =        146
                                                F(46, 101)      =       4.86
                                                Prob > F        =     0.0000
```

unsatisfied	Coefficient	BRR * std. err.	t	P>|t|
region				
North-East	-0.774	0.947	-0.82	0.415
Center	0.227	1.094	0.21	0.836
South	0.362	0.777	0.47	0.642
Islands	-2.392	1.012	-2.36	0.019
educ				
Middle school	0.786	0.899	0.87	0.383
High school	-0.050	0.923	-0.05	0.957
Tertiary degree	-0.871	1.132	-0.77	0.443
empstat				
Job seeker	2.956	0.829	3.56	0.000
Student	1.940	1.235	1.57	0.118
Retired	0.361	1.215	0.30	0.767
Homemaker/Other	1.314	0.919	1.43	0.155
region#educ				
North-East#Middle school	-0.255	0.973	-0.26	0.794
North-East#High school	-0.222	1.022	-0.22	0.828
North-East#Tertiary degree	-0.064	1.430	-0.04	0.964
Center#Middle school	-0.571	1.170	-0.49	0.626
Center#High school	-1.044	1.238	-0.84	0.400
Center#Tertiary degree	0.253	1.539	0.16	0.870
South#Middle school	-0.031	0.661	-0.05	0.963
South#High school	-0.096	0.738	-0.13	0.896
South#Tertiary degree	0.250	1.033	0.24	0.809
Islands#Middle school	1.141	0.867	1.32	0.190
Islands#High school	0.999	1.069	0.93	0.352
Islands#Tertiary degree	1.299	1.656	0.78	0.434
region#empstat				
North-West#Student	0.000	(empty)		
North-East#Job seeker	0.779	0.729	1.07	0.287
North-East#Student	0.447	1.883	0.24	0.813
North-East#Retired	2.868	1.227	2.34	0.021
North-East#Homemaker/Other	0.855	0.828	1.03	0.303
Center#Job seeker	-1.264	0.520	-2.43	0.016
Center#Student	0.000	(empty)		
Center#Retired	1.216	1.689	0.72	0.473
Center#Homemaker/Other	0.380	0.838	0.45	0.651

South#Job seeker	-0.635	0.357	-1.78	0.078
South#Student	0.000	(omitted)		
South#Retired	1.256	1.230	1.02	0.309
South#Homemaker/Other	0.433	0.608	0.71	0.478
Islands#Job seeker	0.452	0.793	0.57	0.569
Islands#Student	0.000	(empty)		
Islands#Retired	2.533	1.587	1.60	0.113
Islands#Homemaker/Other	1.061	0.854	1.24	0.216
educ#empstat				
None/Elementary school#Student	0.000	(empty)		
Middle school#Job seeker	-1.384	0.792	-1.75	0.083
Middle school#Student	0.000	(empty)		
Middle school#Retired	-1.604	0.979	-1.64	0.103
Middle school#Homemaker/Other	-1.301	0.771	-1.69	0.094
High school#Job seeker	-1.133	0.772	-1.47	0.144
High school#Student	-2.200	1.591	-1.38	0.169
High school#Retired	-2.287	1.339	-1.71	0.090
High school#Homemaker/Other	-1.579	0.812	-1.95	0.054
Tertiary degree#Job seeker	-1.420	1.022	-1.39	0.167
Tertiary degree#Student	0.000	(omitted)		
Tertiary degree#Retired	0.284	1.779	0.16	0.873
Tertiary degree#Homemaker/Other	-1.341	1.012	-1.32	0.188
_cons	-2.634	0.935	-2.82	0.006

The overall specification test (see above) suggests that this preliminary model fits the observed data well:

```
/* Intial model specification test */
linktest
```

```
. /* Intial model specification test */
. linktest
(running logit on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..

Survey: Logistic regression                     Number of obs    =      8,710
                                                Population size  = 49,693,185
                                                Subpop. no. obs  =      5,411
                                                Subpop. size     = 31,245,877
                                                Replications     =        152
                                                Design df        =        150
                                                F(2, 149)        =      69.48
                                                Prob > F         =     0.0000
```

unsatisfied	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
_hat	1.156245	.2515863	4.60	0.000	.6591342	1.653356

```
       _hatsq |   .0384447   .0599678     0.64   0.522    -.0800461    .1569354
        _cons |   .1234104   .2670881     0.46   0.645    -.4043304    .6511511
------------------------------------------------------------------------------
```

However, the model may be overspecified, i.e., it may include one or more irrelevant interaction effects. Therefore, we perform an adjusted Wald test on each of the three interactions included in the model:

```
/* Adjusted Wald tests of interaction effects */
testparm i.region#i.educ
testparm i.region#i.empstat
testparm i.educ#i.empstat
```

```
. /* Adjusted Wald tests of interaction effects */
. testparm i.region#i.educ

Adjusted Wald test

 ( 1)  [unsatisfied]2.region#2.educ = 0
 ( 2)  [unsatisfied]2.region#3.educ = 0
 ( 3)  [unsatisfied]2.region#4.educ = 0
 ( 4)  [unsatisfied]3.region#2.educ = 0
 ( 5)  [unsatisfied]3.region#3.educ = 0
 ( 6)  [unsatisfied]3.region#4.educ = 0
 ( 7)  [unsatisfied]4.region#2.educ = 0
 ( 8)  [unsatisfied]4.region#3.educ = 0
 ( 9)  [unsatisfied]4.region#4.educ = 0
 (10)  [unsatisfied]5.region#2.educ = 0
 (11)  [unsatisfied]5.region#3.educ = 0
 (12)  [unsatisfied]5.region#4.educ = 0

       F( 12,   135) =    0.64
            Prob > F =    0.8033

. testparm i.region#i.empstat

Adjusted Wald test

 ( 1)  [unsatisfied]2.region#2.empstat = 0
 ( 2)  [unsatisfied]2.region#3.empstat = 0
 ( 3)  [unsatisfied]2.region#4.empstat = 0
 ( 4)  [unsatisfied]2.region#5.empstat = 0
 ( 5)  [unsatisfied]3.region#2.empstat = 0
 ( 6)  [unsatisfied]3.region#4.empstat = 0
 ( 7)  [unsatisfied]3.region#5.empstat = 0
 ( 8)  [unsatisfied]4.region#2.empstat = 0
 ( 9)  [unsatisfied]4.region#4.empstat = 0
 (10)  [unsatisfied]4.region#5.empstat = 0
 (11)  [unsatisfied]5.region#2.empstat = 0
 (12)  [unsatisfied]5.region#4.empstat = 0
 (13)  [unsatisfied]5.region#5.empstat = 0

       F( 13,   134) =    1.71
            Prob > F =    0.0648

. testparm i.educ#i.empstat

Adjusted Wald test

 ( 1)  [unsatisfied]2.educ#2.empstat = 0
 ( 2)  [unsatisfied]2.educ#4.empstat = 0
```

```
 ( 3)  [unsatisfied]2.educ#5.empstat = 0
 ( 4)  [unsatisfied]3.educ#2.empstat = 0
 ( 5)  [unsatisfied]3.educ#3.empstat = 0
 ( 6)  [unsatisfied]3.educ#4.empstat = 0
 ( 7)  [unsatisfied]3.educ#5.empstat = 0
 ( 8)  [unsatisfied]4.educ#2.empstat = 0
 ( 9)  [unsatisfied]4.educ#4.empstat = 0
 (10)  [unsatisfied]4.educ#5.empstat = 0

       F( 10,   137) =    0.87
            Prob > F =    0.5661
```

As can be seen, all three interaction effects are not significant and can therefore be omitted from the regression model:

```
/* Binomial logistic regression model : Final specification */
svy, subpop(if inrange(age,25,64)) dof(146) dots(10) :     ///
    logit unsatisfied i.region i.educ i.empstat,           ///
    cformat(%6.3f) noci baselevels
linktest
```

```
. /* Binomial logistic regression model : Final specification */
. svy, subpop(if inrange(age,25,64)) dof(146) dots(10) :     ///
>     logit unsatisfied i.region i.educ i.empstat,           ///
>     cformat(%6.3f) noci baselevels
(running logit on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Logistic regression                    Number of obs    =      8,758
                                               Population size  = 50,032,626
                                               Subpop. no. obs  =      5,459
                                               Subpop. size     = 31,585,318
                                               Replications     =        152
                                               Design df        =        146
                                               F(11, 136)       =      14.37
                                               Prob > F         =     0.0000

------------------------------------------------------------------------------
             |                 BRR *
  unsatisfied | Coefficient  std. err.      t    P>|t|
-------------+----------------------------------------------------------------
      region |
  North-West |      0.000  (base)
  North-East |     -0.547      0.300    -1.83   0.070
      Center |     -0.438      0.274    -1.60   0.113
       South |      0.326      0.245     1.33   0.186
     Islands |     -0.810      0.398    -2.04   0.043
             |
        educ |
None/Elementary school |      0.000  (base)
 Middle school |     -0.282      0.290    -0.97   0.334
   High school |     -1.225      0.296    -4.14   0.000
Tertiary degree |     -1.556      0.390    -3.99   0.000
             |
      empstat |
```

Employed	0.000	(base)		
Job seeker	1.378	0.167	8.23	0.000
Student	-0.014	0.750	-0.02	0.985
Retired	0.391	0.336	1.16	0.247
Homemaker/Other	0.476	0.196	2.43	0.016
_cons	-1.664	0.328	-5.07	0.000

```
. linktest
(running logit on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..

Survey: Logistic regression                     Number of obs    =      8,758
                                                Population size  = 50,032,626
                                                Subpop. no. obs  =      5,459
                                                Subpop. size     = 31,585,318
                                                Replications     =        152
                                                Design df        =        150
                                                F(2, 149)        =      56.10
                                                Prob > F         =     0.0000
```

unsatisfied	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
_hat	.8114992	.2724867	2.98	0.003	.2730912	1.349907
_hatsq	-.0471518	.07092	-0.66	0.507	-.187283	.0929793
_cons	-.1512053	.287558	-0.53	0.600	-.7193926	.416982

We can now use the results of the final regression model to estimate the conditional probabilities of interest, one for each possible covariate pattern. To do this, we first use the command `collapse` to create a dataset with all possible covariate patterns:

```
/* Create covariate pattern dataset */
collapse unsatisfied, by(region educ empstat)
summarize
```

```
. /* Create covariate pattern dataset */
. collapse unsatisfied, by(region educ empstat)

. summarize
```

Variable	Obs	Mean	Std. dev.	Min	Max
region	97	2.989691	1.425182	1	5
educ	97	2.546392	1.108872	1	4
empstat	97	3	1.443376	1	5
unsatisfied	97	.0991172	.1403503	0	1

Although there are 100 possible covariate patterns (five regions times four educational degrees times five employment conditions), the resulting dataset contains only 97 of them – three of the possible covariate patterns are not represented in the working sample. We generate the three missing covariate patterns using the command `fillin`:

```
/* Complete covariate pattern dataset */
fillin region educ empstat
drop _fillin
summarize
```

```
. /* Complete covariate pattern dataset */
. fillin region educ empstat

. drop _fillin

. summarize
```

Variable	Obs	Mean	Std. dev.	Min	Max
region	100	3	1.421338	1	5
educ	100	2.5	1.123666	1	4
empstat	100	3	1.421338	1	5
unsatisfied	97	.0991172	.1403503	0	1

Finally, we use the command `predict` to estimate the conditional probabilities of interest and the associated design-based 95% logit-transformed confidence intervals:

```
/* Estimate conditional probabilities and 95% CIs */
egen CP = group(region educ empstat), label
predict PI
predict XB, xb
predict SE, stdp
generate LB = XB - invt(146,0.975) * SE
replace LB = invlogit(LB)
generate UB = XB + invt(146,0.975) * SE
replace UB = invlogit(UB)
keep CP PI LB UB
label variable CP "Covariate pattern"
label variable PI "Probability"
label variable LB "95% lower bound"
label variable UB "95% upper bound"
format PI LB UB %5.3f
gsort -PI
describe
```

```
. /* Estimate conditional probabilities and 95% CIs */
. egen CP = group(region educ empstat), label

. predict PI
```

```
(option pr assumed; Pr(unsatisfied))
. predict XB, xb
. predict SE, stdp
. generate LB = XB - invt(146,0.975) * SE
. replace LB = invlogit(LB)
(100 real changes made)
. generate UB = XB + invt(146,0.975) * SE
. replace UB = invlogit(UB)
(100 real changes made)
. keep CP PI LB UB
. label variable CP "Covariate pattern"
. label variable PI "Probability"
. label variable LB "95% lower bound"
. label variable UB "95% upper bound"
. format PI LB UB %5.3f
. gsort -PI
. describe
Contains data
 Observations:           100
    Variables:             4
```

Variable name	Storage type	Display format	Value label	Variable label
CP	float	%49.0g	CP	Covariate pattern
PI	float	%5.3f		Probability
LB	float	%5.3f		95% lower bound
UB	float	%5.3f		95% upper bound

```
Sorted by:
     Note: Dataset has changed since last saved.
```

The resulting dataset contains the design-based estimates of all quantities of interest. For example, here are the estimates for the covariate patterns characterized by the ten highest and ten lowest probabilities of reporting low levels of overall life satisfaction:

```
/* Display select estimates */
clist CP PI LB UB in 1/10, noobs
clist CP PI LB UB in -10/l, noobs
```

```
. /* Display select estimates */
. clist CP PI LB UB in 1/10, noobs
                                              CP     PI     LB     UB
          South None/Elementary school Job seeker  0.510  0.328  0.689
                   South Middle school Job seeker  0.440  0.332  0.554
     North-West None/Elementary school Job seeker  0.429  0.264  0.611
              North-West Middle school Job seeker  0.362  0.267  0.469
         Center None/Elementary school Job seeker  0.327  0.194  0.494
```

```
     North-East None/Elementary school Job seeker  0.303  0.157  0.502
     South None/Elementary school Homemaker/Other  0.297  0.177  0.454
             South None/Elementary school Retired  0.279  0.159  0.443
                  Center Middle school Job seeker  0.268  0.188  0.366
        Islands None/Elementary school Job seeker  0.250  0.119  0.452
```

```
. clist CP PI LB UB in -10/l, noobs
```

```
                                          CP     PI     LB     UB
          Islands Tertiary degree Homemaker/Other  0.028  0.011  0.067
                  Islands Tertiary degree Retired  0.026  0.009  0.073
                  Center Tertiary degree Employed  0.025  0.012  0.051
                   Center Tertiary degree Student  0.025  0.005  0.108
                     Islands High school Employed  0.024  0.012  0.050
                      Islands High school Student  0.024  0.005  0.112
              North-East Tertiary degree Employed  0.023  0.011  0.044
               North-East Tertiary degree Student  0.022  0.005  0.101
                 Islands Tertiary degree Employed  0.017  0.008  0.040
                  Islands Tertiary degree Student  0.017  0.003  0.088
```

As can be seen, at one end of the range we have those living in the mainland regions of the South, poorly educated and unemployed, with an estimated probability of being unsatisfied at 51% (95% confidence interval: 32.8-68.9). At the other end we find people residing in Sicily or Sardinia who have a university degree and are still studying, with an estimated probability of discontent equal to 1.7% (95% confidence interval: 0.3-8.8).

To conclude the example, we use dominance analysis to estimate the relative contribution of each covariate to the predictive power of the chosen regression model, as measured by McFadden's (1974) pseudo R^2:

```
/* Dominance analysis of the regression model */
use "itali.dta", clear
generate unsatisfied = (lifesat < 6) if (lifesat < .)
svy brr _b, subpop(if inrange(age,25,64)) dof(146)      ///
    dots(10) cformat(%6.4f) : domin unsatisfied         ///
    if (unsatisfied < .), reg(logit) fitstat(e(r2_p))   ///
    sets( (i.region) (i.educ) (i.empstat) )
nlcom _b[set1] + _b[set2] + _b[set3], cformat(%6.4f)
```

```
. /* Dominance analysis of the regression model */
. use "itali.dta", clear
```

```
. generate unsatisfied = (lifesat < 6) if (lifesat < .)
(31 missing values generated)
```

```
. svy brr _b, subpop(if inrange(age,25,64)) dof(146)      ///
>     dots(10) cformat(%6.4f) : domin unsatisfied         ///
>     if (unsatisfied < .), reg(logit) fitstat(e(r2_p))   ///
>     sets( (i.region) (i.educ) (i.empstat) )
(running domin on estimation sample)
```

```
BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....
```

```
Dominance analysis                              Number of obs    =      8,747
                                                Population size  = 49,964,026
                                                Subpop. no. obs  =      5,459
                                                Subpop. size     = 31,585,318
                                                Replications     =        152
                                                Design df        =        146
```

unsatisfied	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
set1	0.0246	0.0106	2.33	0.021	0.0037	0.0456
set2	0.0437	0.0102	4.30	0.000	0.0236	0.0637
set3	0.0377	0.0084	4.49	0.000	0.0211	0.0543

```
. nlcom _b[set1] + _b[set2] + _b[set3], cformat(%6.4f)
      _nl_1: _b[set1] + _b[set2] + _b[set3]
```

unsatisfied	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	0.1060	0.0187	5.65	0.000	0.0693	0.1427

The Stata output shows that the predictive power of the model is 0.1060. The largest contribution to this value comes from the main effect of educational degree, and amounts to 0.0437 (41% in relative terms). This is closely followed by the contribution of the main effect of employment status at 0.0377 (36% of the total) and the contribution of the main effect of region of residence at 0.0246 (23%). It should be noted that the contributions of the three covariates to the predictive power of the model appear to be essentially equivalent when estimation uncertainty is taken into account. This equality is confirmed by the following adjusted Wald test:

```
/* Dominance analysis : Test of equality of contributions */
test _b[set1] = _b[set2] = _b[set3]
```

```
. /* Dominance analysis : Test of equality of contributions */
. test _b[set1] = _b[set2] = _b[set3]
Adjusted Wald test
 ( 1)  set1 - set2 = 0
 ( 2)  set1 - set3 = 0
       F(  2,   145) =    1.02
            Prob > F =    0.3644
```

4.6. *Treatment Effect Estimation*

The purpose of *treatment effect estimation* is to evaluate the impact of a given treatment or intervention or exposure on an outcome of interest. In practice, this amounts to estimating the extent to which a change in a treatment variable T causes a change in an outcome variable Y.[7]

Many studies estimate the treatment effects of interest using observational data collected in complex surveys. Often, however, they do so without due consideration of all the elements of the survey design, resulting in biased estimates of both the quantities of interest and the uncertainty surrounding them (Lenis *et al.* 2019; Levin and Sinclair 2018). The implication should be clear: for sample estimates of treatment effects to be generalizable to the entire target population, these estimates must be obtained by applying the rules of design-based estimation and inference.

Although Stata has several commands for treatment effect estimation, most are not tailored to the design-based analysis of complex survey data. A notable exception is the user-written command `kmatch` (Jann 2017), which allows for all elements of complex survey designs and provides many treatment effect estimators, including multivariate-distance matching, propensity-score matching, coarsened exact matching, entropy balance, inverse probability weighting, and regression adjustment (Hainmueller 2012; Iacus *et al.* 2012; Imbens and Rubin 2015; Rosenbaum 2020; Stuart 2010).

In the following, we illustrate the main features of `kmatch` using a toy example, which consists in estimating the effect of pre-school attendance on future chances of obtaining a university degree. Since college studies in Italy take place, on average, between the ages of 19 and 24, we will only look at individuals who are 25 years old or older. Seventy-three respondents in the age group of interest did not answer the question on pre-school attendance, so the number of valid cases for analysis is 7,848.

In our working dataset, the treatment is represented by variable `preschool`:

```
/* Percent distribution of treatment variable "preschool"  */
svy, subpop(if age >= 25) dof(147) :   ///
    tabulate preschool, percent se     ///
    ci format(%5.1f)
```

```
. /* Percent distribution of treatment variable "preschool"  */
. svy, subpop(if age >= 25) dof(147) :   ///
```

7 In the following discussion, the expression "treatment variable" is used as a general term to refer to any variable that represents two or more alternative treatments, interventions, or exposures being compared.

```
>     tabulate preschool, percent se     ///
>     ci format(%5.1f)
(running tabulate on estimation sample)
                                                Number of obs    =      8,705
                                                Population size = 49,769,079
                                                Subpop. no. obs =      7,848
                                                Subpop. size    = 44,839,255
                                                Replications    =        152
                                                Design df       =        147
```

Attended pre-primary school 1+ years	percentage	se	lb	ub
No	38.7	1.4	35.9	41.6
Yes	61.3	1.4	58.4	64.1
Total	100.0			

```
Key: percentage = Cell percentage
             se = Brr standard error of cell percentage
             lb = Lower 95% confidence bound for cell percentage
             ub = Upper 95% confidence bound for cell percentage
```

The outcome variable Y, on the other hand, is generated as follows:

```
/* Generate outcome variable Y */
generate Y = (educ == 4)
label variable Y "University degree"
label define l_Y 0 "No", modify
label define l_Y 1 "Yes", modify
label values Y l_Y

/* Percent distribution of outcome variable Y  */
svy, subpop(if age >= 25) dof(147) :   ///
    tabulate Y if (preschool < .),     ///
    percent se ci format(%5.1f)
```

```
. /* Generate outcome variable Y */
. generate Y = (educ == 4)
. label variable Y "University degree"
. label define l_Y 0 "No", modify
. label define l_Y 1 "Yes", modify
. label values Y l_Y
.
. /* Percent distribution of outcome variable Y  */
. svy, subpop(if age >= 25) dof(147) :   ///
>     tabulate Y if (preschool < .),     ///
>     percent se ci format(%5.1f)
(running tabulate on estimation sample)
                                                Number of obs    =      8,702
```

```
                                        Population size = 49,747,623
                                        Subpop. no. obs =      7,848
                                        Subpop. size    = 44,839,255
                                        Replications    =        152
                                        Design df       =        147
```

Universit y degree	percentage	se	lb	ub
No	83.5	0.2	83.2	83.8
Yes	16.5	0.2	16.2	16.8
Total	100.0			

```
Key: percentage = Cell percentage
             se = Brr standard error of cell percentage
             lb = Lower 95% confidence bound for cell percentage
             ub = Upper 95% confidence bound for cell percentage
```

Finally, we select three control variables: sex, age, and current region of residence – the latter used as a proxy for region of residence during pre-school age. Based on the results of preliminary data inspection, both age and age squared will be included in the analysis.

Since kmatch does not support the svy prefix, prior to starting the analysis we must globally change the survey design settings by entering the appropriate number of design degrees of freedom for the subpopulation of interest (see Table 4.3):

```
/* Adjust number of design degrees of freedom */
svyset, dof(147) noclear
```

```
. /* Adjust number of design degrees of freedom */
. svyset, dof(147) noclear
Sampling weights: fiw
             VCE: brr
             MSE: on
     BRR weights: brr_iw_1 .. brr_iw_152
Fay's adjustment: .5
       Design df: 147
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: <observations>
           FPC 1: <zero>
```

Now we are ready for the analysis. First, we estimate the naive average treatment effect, which is the simple bivariate association between treatment and outcome:

```
/* Estimate the naive average treatment effect */
kmatch ra preschool (Y) if (preschool < .), svy       ///
    subpop(if age >= 25) po nomtable cformat(%6.4f)
```

```
. /* Estimate the naive average treatment effect */
. kmatch ra preschool (Y) if (preschool < .), svy       ///
>     subpop(if age >= 25) po nomtable cformat(%6.4f)
(running kmatch on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..

Regression adjustment                             Number of obs    =      8,702
                                                  Population size  = 49,747,623
                                                  Subpop. no. obs  =      7,848
                                                  Subpop. size     = 44,839,255
                                                  Replications     =        152
                                                  Design df        =        147

Treatment    : preschool = 1

Treatment-effects estimation
```

Y	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
ATE	0.0952	0.0132	7.19	0.000	0.0690	0.1214
Y1	0.2019	0.0054	37.63	0.000	0.1913	0.2125
Y0	0.1067	0.0084	12.67	0.000	0.0900	0.1233

The output returns three relevant pieces of information: (a) the proportion of individuals with pre-school experience who have earned a university degree (referred to as `Y1` in the output); (b) the proportion of individuals without pre-school experience who have earned a university degree (referred to as `Y0` in the output); and (c) the difference between the two proportions (referred to as `ATE` in the output). The latter is the estimate of the naive average treatment effect. As can be seen, at the bivariate level, pre-school attendance appears to have a large effect on the probability of obtaining a university degree, almost doubling it.

The naive average treatment effect is not a plausible estimator of the true treatment effect of interest because there is good reason to believe that the association between treatment and outcome is confounded by one or more covariates. As anticipated, in this toy analysis we will consider only three potential confounders: sex, age (in quadratic form), and region of residence. Let us see what happens when we introduce these covariates into the analysis and estimate the treatment effect of interest by simple regression adjustment:

```
/* ATE, ATT, ATC estimation : Regression adjustment */
kmatch ra preschool (Y = i.sex c.age##c.age i.region)   ///
    if (preschool < .), svy subpop(if age >= 25)        ///
    ate att atc nomtable cformat(%6.4f)
```

```
. /* ATE, ATT, ATC estimation : Regression adjustment */
. kmatch ra preschool (Y = i.sex c.age##c.age i.region)   ///
>     if (preschool < .), svy subpop(if age >= 25)        ///
>     ate att atc nomtable cformat(%6.4f)
(running kmatch on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..

Regression adjustment                           Number of obs    =      8,702
                                                Population size  = 49,747,623
                                                Subpop. no. obs  =      7,848
                                                Subpop. size     = 44,839,255
                                                Replications     =        152
                                                Design df        =        147

Treatment   : preschool = 1
RA equations: Y = i.sex age c.age#c.age i.region _cons

Treatment-effects estimation
```

Y	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
ATE	0.0527	0.0145	3.63	0.000	0.0240	0.0813
ATT	0.0632	0.0168	3.77	0.000	0.0301	0.0963
ATC	0.0359	0.0121	2.97	0.003	0.0121	0.0598

The results of the analysis suggest two main conclusions. First, after adjusting for the selected confounders, the average treatment effect (ATE) is significantly lower than that estimated with the naive approach (0.0527 vs 0.0952). Second, the average treatment effect is larger in treated individuals (ATT = 0.0632) than in untreated individuals (ATC = 0.0359). In essence, this means that the benefit of pre-school for those who did attend is greater than it would have been for those who did not attend, had they actually attended. Such a difference suggests the presence of some residual confounding not controlled for in the analysis.

We now estimate the three variants of the treatment effect of interest using multivariate-distance matching:

```
/* ATE, ATT, ATC estimation : Multivariate-distance matching */
kmatch md preschool i.sex c.age##c.age i.region (Y)   ///
```

```
    if (preschool < .), svy subpop(if age >= 25)      ///
    ate att atc cformat(%6.4f)
```

```
    . /* ATE, ATT, ATC estimation : Multivariate-distance matching */
. kmatch md preschool i.sex c.age##c.age i.region (Y)   ///
>     if (preschool < .), svy subpop(if age >= 25)      ///
>     ate att atc cformat(%6.4f)
(running kmatch on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..

Multivariate-distance kernel matching             Number of obs    =      8,702
                                                  Population size = 49,747,623
                                                  Subpop. no. obs =      7,848
                                                  Subpop. size    = 44,839,255
                                                  Replications    =        152
                                                  Design df       =        147
                                                 Kernel           =       epan

Treatment   : preschool = 1
Metric      : mahalanobis
Covariates  : i.sex age c.age#c.age i.region

Matching statistics
```

	Matched			Controls			Bandwidth
	Yes	No	Total	Used	Unused	Total	
Treated	5008	11	5019	2805	24	2829	.5427061
Untreated	2825	4	2829	5018	1	5019	.9843852
Combined	7833	15	7848	7823	25	7848	

Treatment-effects estimation

Y	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
ATE	0.0556	0.0139	4.01	0.000	0.0282	0.0830
ATT	0.0623	0.0158	3.94	0.000	0.0311	0.0936
ATC	0.0451	0.0124	3.63	0.000	0.0205	0.0696

As can be seen, the estimates of all three effects are very similar to those obtained by regression adjustment. However, the difference between ATT and ATC now appears smaller.

In treatment effect estimation, the aim of matching is to achieve adequate covariate balance, i.e., a high degree of similarity in the distribution of covariates across levels of the treatment variable. To assess covariate balance, `kmatch` provides several postestimation commands. In particular, the command `kmatch summarize` calculates and displays standardized mean differences and variance ratios between treatment groups, before and after matching. For example, here is the Stata code for checking covariate balance after ATE estimation by multivariate-distance matching:

```
/* Covariate balance check after MDM */
kmatch summarize, ate
```

```
. /* Covariate balance check after MDM */
. kmatch summarize, ate
(refitting the model using the generate() option)
```

Means	Raw Treated	Raw Untreated	Raw StdDif	Matched(ATE) Treated	Matched(ATE) Untreated	Matched(ATE) StdDif
1.sex	.5137917	.5365504	-.0455807	.523994	.5241263	-.000265
age	49.99612	61.95894	-.7454937	54.24867	54.69422	-.0277655
c.age#c.age	2739.692	4113.688	-.7420199	3220.006	3270.79	-.0274252
2.region	.2039267	.1805181	.0594237	.1954115	.1960005	-.0014953
3.region	.182249	.2181354	-.0897614	.1966968	.1963756	.0008033
4.region	.2335238	.237381	-.00909	.2355975	.2358194	-.0005229
5.region	.0991162	.1235522	-.0777342	.1063978	.1052945	.0035096

Variances	Raw Treated	Raw Untreated	Raw Ratio	Matched(ATE) Treated	Matched(ATE) Untreated	Matched(ATE) Ratio
1.sex	.2498596	.248752	1.004453	.249474	.2495069	.9998683
age	240.1283	274.8744	.8735927	277.1438	279.432	.9918114
c.age#c.age	2775171	4082389	.6797909	3571589	3601418	.9917173
2.region	.162373	.1479836	1.097236	.1572572	.1576405	.9975683
3.region	.149064	.1706127	.8736984	.1580387	.1578685	1.001078
4.region	.1790261	.1810953	.9885741	.1801272	.1802729	.999192
5.region	.08931	.1083254	.8244608	.0950962	.0942412	1.009073

As can be seen, matching performed very well in this case, almost completely eliminating the differences between treatment groups in terms of the distribution (mean and variance) of the covariates.

Finally, we repeat the estimation exercise using inverse probability weighting by the propensity score:

```
/* ATE, ATT, ATC estimation : Inverse probability weighting */
kmatch ipw preschool i.sex c.age##c.age i.region (Y)   ///
    if (preschool < .), svy subpop(if age >= 25)       ///
    ate att atc cformat(%6.4f) nomtable
```

```
. /* ATE, ATT, ATC estimation : Inverse probability weighting */
. kmatch ipw preschool i.sex c.age##c.age i.region (Y)   ///
>     if (preschool < .), svy subpop(if age >= 25)       ///
>     ate att atc cformat(%6.4f) nomtable
(running kmatch on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..
```

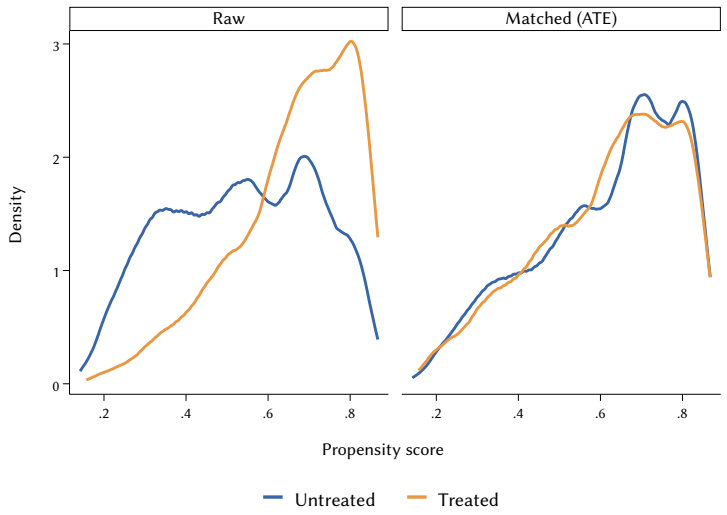
```
Inverse probability weighting                   Number of obs    =      8,702
                                                Population size  = 49,747,623
                                                Subpop. no. obs  =      7,848
                                                Subpop. size     = 44,839,255
                                                Replications     =        152
                                                Design df        =        147

Treatment   : preschool = 1
Covariates  : i.sex age c.age#c.age i.region
PS model    : logit (pr)

Treatment-effects estimation
```

Y	Coefficient	BRR * std. err.	t	P>|t|	[95% conf.	interval]
ATE	0.0523	0.0150	3.50	0.001	0.0227	0.0819
ATT	0.0620	0.0172	3.60	0.000	0.0279	0.0960
ATC	0.0370	0.0127	2.90	0.004	0.0118	0.0622

Again, the estimates of the treatment effect of interest and the overall conclusions remain essentially unchanged. To assess covariate balance, in this case – in addition to `kmatch summarize` – one can use postestimation plots that show the degree of similarity in the distribution of propensity scores across treatment groups, before and after weighting. For example:

```
/* Propensity score balance check after IPW */
kmatch density, ate lwidth(*4 *4) lcolor("55 101 168" "234 151 65")
```

```
. /* Propensity score balance check after IPW */
. kmatch density, ate lwidth(*4 *4) lcolor("55 101 168" "234 151 65")
(refitting the model using the generate() option)
(applying 0-1 boundary correction to density estimation of propensity score)
(bandwidth for propensity score = .05704863)
```



The plots clearly show that the distribution of propensity scores after weighting is essentially the same between the two treatment groups being compared.

4.7. *Event History Analysis*

The purpose of *event history analysis* – also known as survival analysis, duration analysis, or transition analysis – is to describe the occurrence and timing of various kinds of life course events. For simplicity, here we will just consider *single non-repeatable events*, that is, events of a single type (or treated alike) that can occur only once, such as entry into the first job, entry into the first marriage, or birth of the first child (Vermunt and Moors 2005).

When studying such events, each unit in the sample is observed for a fixed period of time $[t_{\text{start}}, t_{\text{end}}]$, called the *observation window*. At start time t_{start}, all units are in the same state, which we call the *origin state*. From this moment on, the units are observed over a period – ending at time t_{end} – during which they are susceptible (at "risk") of transitioning to another predefined state, which we call the *destination state*. If such a transition occurs at some time t_{obs} within the observation window, then we say that the corresponding sample unit has experienced the event of interest at time t_{obs}. On the other hand, if the transition does not occur during the observation period, the corresponding unit is said to be *right-censored* at time t_{end}; in this case, all we can conclude is that the event of interest has not yet occurred by the end of the observation window, and we do not know if and when it will occur in the future. Determining whether the event of interest has occurred during the observation period – and, if so, when – makes it possible to estimate the distribution of the time of occurrence of the event itself, the variation in the rate or risk of event occurrence over the observation period, and how these phenomena vary according to the values of one or more covariates (Blossfeld *et al.* 2019; Cleves *et al.* 2016).

Although event history analysis is a well-developed branch of statistics, its application to complex survey data still lacks theoretical and software development. Stata is no exception, with limited direct support for design-based analysis of event history data. Still, with some tinkering, it is possible to get Stata to perform a wide range of analyses of event history data that properly account for survey design. In this section, we will explore what Stata has to offer in this area.

For illustration, we will use as a running example the analysis of the entry into the first job of women and men born between 1946 and 1995. In the example, age is used as the time scale, so that t_{start} corresponds to the year of

birth, t_{end} is the year of the interview, and t_{obs} is the year in which respondents with work experience entered their first job – that is, they transitioned from the origin state "Never worked" to the destination state "Currently working". To limit the analysis to a reasonable age range for the event of interest, only individuals who started working after age 10 are considered. For the same reason, the maximum follow-up age is set at 50.

To begin, we create a new working data file that contains all the information needed to perform the analyses of interest:

```
/* Extract data of interest from file "household_grid.dta" */
use "household_grid.dta", clear
keep if (W19QUEST == 1)
keep W19CID W19BIRTH_Y W19INTSTR_Y
tempfile JOB
save `JOB', replace

/* Extract data of interest from file "job_history.dta" */
use "job_history.dta", clear
keep if (W19JH001 == 1)
keep if inlist(W19JHEMST,1,2)
tempvar MARK
sort W19CID W19JHSPL
by W19CID : generate `MARK' = (_n == 1)
keep if `MARK'
keep W19CID W19JHSTR_Y
generate status = 1
label variable status "Event occurred during observation period"
label define l_status 0 "No", modify
label define l_status 1 "Yes", modify
label values status l_status

/* Merge files */
merge n:1 W19CID using `JOB'
drop _merge
recode status (. = 0)

/* Generate time variables */
generate y_birth = W19BIRTH_Y
label variable y_birth "Year of birth"
generate y_job1entry = W19JHSTR_Y
label variable y_job1entry "Year of entry into the first job"
generate y_interview = W19INTSTR_Y
label variable y_interview "Year of interview"
generate y_lastobs = cond(status == 1, y_job1entry, y_interview)
label variable y_lastobs "Year of last obs. (event or censoring)"
```

```
/* Adjust status variable */
replace status = 0 if (y_job1entry - y_birth) > 50

/* Generate variable "Birth cohort" */
generate bcohort = W19BIRTH_Y
recode bcohort (1921/1945 = 1) (1946/1955 = 2) (1956/1965 = 3)   ///
    (1966/1980 = 4) (1981/1995 = 5) (1996/2004 = 6)
label variable bcohort "Birth cohort"
label define l_bcohort 1 "1921-1945", modify
label define l_bcohort 2 "1946-1955", modify
label define l_bcohort 3 "1956-1965", modify
label define l_bcohort 4 "1966-1980", modify
label define l_bcohort 5 "1981-1995", modify
label define l_bcohort 6 "1996-2004", modify
label values bcohort l_bcohort

/* Create working data file */
keep W19CID bcohort status y_*
compress
save `JOB', replace
use "itali.dta", clear
merge 1:1 W19CID using `JOB'
drop _merge
drop agegroup empstat-bmi
move bcohort educ

/* svyset data file (with respondent's id) */
svyset W19CID [pw = fiw], vce(brr) brrweight(brr_iw_*)   ///
    fay(0.5) dof(150) mse

/* Save working data file */
compress
save "itali-eha.dta", replace
```

```
. /* Extract data of interest from file "household_grid.dta" */
. use "household_grid.dta", clear

. keep if (W19QUEST == 1)
(2,611 observations deleted)

. keep W19CID W19BIRTH_Y W19INTSTR_Y

. tempfile JOB

. save `JOB´, replace
(file /var/folders/ww/06m0tz_s5_q9_d1hm0cy2_gr0000gn/T//S_08839.000001 not found)
file /var/folders/ww/06m0tz_s5_q9_d1hm0cy2_gr0000gn/T//S_08839.000001 saved as .dta
    format

.
. /* Extract data of interest from file "job_history.dta" */
. use "job_history.dta", clear

. keep if (W19JH001 == 1)
```

```
(2,038 observations deleted)
. keep if inlist(W19JHEMST,1,2)
(9,096 observations deleted)
. tempvar MARK
. sort W19CID W19JHSPL
. by W19CID : generate `MARK´ = (_n == 1)
. keep if `MARK´
(5,778 observations deleted)
. keep W19CID W19JHSTR_Y
. generate status = 1
. label variable status "Event occurred during observation period"
. label define l_status 0 "No", modify
. label define l_status 1 "Yes", modify
. label values status l_status
.
. /* Merge files */
. merge n:1 W19CID using `JOB´
(variable W19CID was str6, now str7 to accommodate using data´s values)
    Result                      Number of obs
    ----------------------------------------
    Not matched                         1,926
        from master                         0  (_merge==1)
        from using                      1,926  (_merge==2)
    Matched                             6,852  (_merge==3)
    ----------------------------------------
. drop _merge
. recode status (. = 0)
(1926 changes made to status)
.
. /* Generate time variables */
. generate y_birth = W19BIRTH_Y
. label variable y_birth "Year of birth"
. generate y_job1entry = W19JHSTR_Y
(1,926 missing values generated)
. label variable y_job1entry "Year of entry into the first job"
. generate y_interview = W19INTSTR_Y
. label variable y_interview "Year of interview"
. generate y_lastobs = cond(status == 1, y_job1entry, y_interview)
. label variable y_lastobs "Year of last obs. (event or censoring)"
.
. /* Generate variable "Birth cohort" */
. generate bcohort = W19BIRTH_Y
. recode bcohort (1921/1945 = 1) (1946/1955 = 2) (1956/1965 = 3)   ///
>     (1966/1980 = 4) (1981/1995 = 5) (1996/2004 = 6)
(8778 changes made to bcohort)
. label variable bcohort "Birth cohort"
```

```
. label define l_bcohort 1 "1921-1945", modify
. label define l_bcohort 2 "1946-1955", modify
. label define l_bcohort 3 "1956-1965", modify
. label define l_bcohort 4 "1966-1980", modify
. label define l_bcohort 5 "1981-1995", modify
. label define l_bcohort 6 "1996-2004", modify
. label values bcohort l_bcohort
.
. /* Create working data file */
. keep W19CID bcohort status y_*
. compress
  variable status was float now byte
  variable y_birth was float now int
  variable y_job1entry was float now int
  variable y_interview was float now int
  variable y_lastobs was float now int
  variable bcohort was float now byte
  variable W19CID was str7 now str6
  (131,670 bytes saved)
. save `JOB', replace
file /var/folders/ww/06m0tz_s5_q9_d1hm0cy2_gr0000gn/T//S_08839.000001 saved as .dta
    format
. use "itali.dta", clear
. merge 1:1 W19CID using `JOB'
    Result                      Number of obs
    ------------------------------------------
    Not matched                             0
    Matched                             8,778  (_merge==3)
    ------------------------------------------
. drop _merge
. drop agegroup empstat-bmi
. move bcohort educ
.
. /* svyset data file (with respondent's id) */
. svyset W19CID [pw = fiw], vce(brr) brrweight(brr_iw_*)   ///
>     fay(0.5) dof(150) mse
Sampling weights: fiw
             VCE: brr
             MSE: on
     BRR weights: brr_iw_1 .. brr_iw_152
Fay's adjustment: .5
       Design df: 150
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: W19CID
           FPC 1: <zero>
.
. /* Save working data file */
. compress
  (0 bytes saved)
```

```
. save "itali-eha.dta", replace
file itali-eha.dta saved
```

In principle, entry into the first job can occur on any day of the year, so the time of occurrence of this event can be regarded as continuous. In our working data file, however, all dates are *interval-censored*, that is, they are grouped into discrete-time intervals – in this case, all of the same length (one year). This means, for example, that if an event is recorded as occurring in 2016, all we know is that the event took place on some day between January 1 and December 31, 2016, but we do not know exactly which day (Cleves *et al.* 2016). Interval-censored event history data can be analyzed in three different ways (Canette 2016): using continuous-time methods, using discrete-time methods, or using dedicated methods. In the following, we will focus on continuous-time methods.

To use continuous-time methods, it is necessary that the dataset in memory be `stset`, so that Stata knows, for each unit in the sample, whether and when the event of interest occurred during the observation period (Cleves *et al.* 2016). In our example:

```
/* Open working dataset */
use "itali-eha.dta", clear

/* stset data in memory */
stset y_lastobs if inrange(y_birth,1946,1995) [pw = fiw],    ///
    id(W19CID) failure(status) origin(time y_birth)          ///
    enter(time y_birth + 10) exit(time y_birth + 50)
```

```
. /* Open working dataset */
. use "itali-eha.dta", clear

.
. /* stset data in memory */
. stset y_lastobs if inrange(y_birth,1946,1995) [pw = fiw],   ///
>     id(W19CID) failure(status) origin(time y_birth)         ///
>     enter(time y_birth + 10) exit(time y_birth + 50)

Survival-time data settings

          ID variable: W19CID
        Failure event: status!=0 & status<.
Observed time interval: (y_lastobs[_n-1], y_lastobs]
     Enter on or after: time y_birth + 10
     Exit on or before: time y_birth + 50
     Time for analysis: (time-origin)
               Origin: time y_birth
               Weight: [pweight=fiw]
      Keep observations
               if exp: inrange(y_birth,1946,1995)

------------------------------------------------------------------------------
      8,778  total observations
```

```
      1,929  ignored at outset because of if exp
         45  observations end on or before enter()
------------------------------------------------------------------------------

      6,804  observations remaining, representing
      6,804  subjects
      5,789  failures in single-failure-per-subject data
     98,781  total analysis time at risk and under observation
                                             At risk from t =         0
                                  Earliest observed entry t =        10
                                       Last observed exit t =        50
```

The Stata output informs us that 1,929 individuals were ignored because they were born before or after the time period considered for the analysis (1946-1995), while another 45 respondents were excluded because they reported having started working before the lower age limit we set (11 years). Therefore, the valid cases for analysis are 6,804. Of these, 5,789 (referred to as `failures` in the output) started their first job during the follow-up period (ages 11 to 50), whereas the remaining 1,015 are right-censored.

It is also worth noting that the command `stset` created five new variables. Variable `_t` represents the *analysis time*, defined as the difference between the year of entry into the first job and the year of birth for respondents with work experience, and as the difference between the year of the interview and the year of birth for respondents without work experience; values of `_t` greater than the maximum follow-up age (50) were truncated at 50. Variable `_d` represents the *observation status* and takes value 1 if the event of interest occurred during the observation period, value 0 if the observation is right-censored. Variable `_origin` represents the *origin time* and, in our example, is just a copy of variable `y_birth` (year of birth). Variable `_t0` represents the *entry time*, which in our example is age 10. Finally, `_st` is a binary variable that indicates which respondents are included in (value 1) or excluded from (value 0) the analysis.

The key variable in event history analysis is the so-called *survival time*, that is, the time spent by each individual in the origin state before transitioning to the destination state. Formally, we can represent survival time as a positive random variable T with a given probability distribution. This distribution can be described in various ways, but there are two functions most commonly used for this purpose: the hazard function and the survival function (Blossfeld *et al.* 2019; Cleves *et al.* 2016).

In our example, the *hazard function* $h(t)$ can be properly interpreted as the conditional probability that the event of interest will occur at age t, given that it has not occurred at earlier ages. In turn, the *survival function* $S(t)$ expresses the probability that the event of interest will occur after age t. The

two functions are related as follows (Allison 1982):

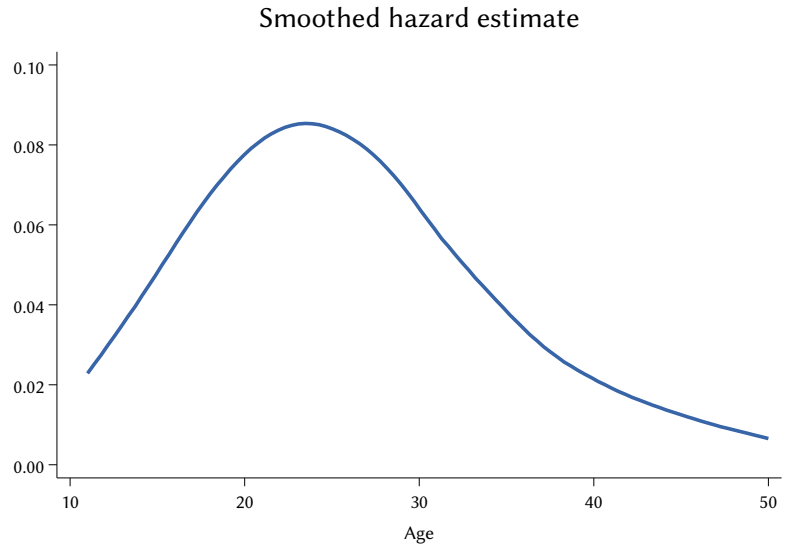
$$S(t) = \prod_{j=11}^{t} \big(1 - h(j)\big) \tag{4.1}$$

Note that the counter j starts at 11, since in our example the observation period is between ages 11 and 50 (see above).

In general, the starting point for any analysis of time-to-event data is what we might call *univariate event history analysis*, which consists of estimating the hazard function and the survival function for the entire target population. In Stata, you can use the command `sts graph` to get a quick graphical representation of the two functions:

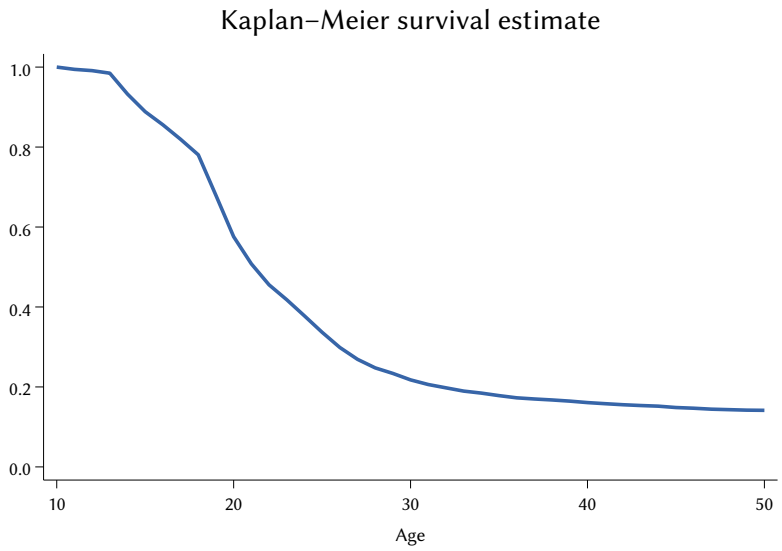
```
/* Graphical representation of the (smoothed) hazard function */
sts graph, hazard noboundary ylabel(0(0.02)0.1, format(%4.2f))   ///
    xtitle("Age") plotopts(lwidth(*4) lcolor("55 101 168"))      ///
    noshow
```

```
. /* Graphical representation of the (smoothed) hazard function */
. sts graph, hazard noboundary ylabel(0(0.02)0.1, format(%4.2f))   ///
>     xtitle("Age") plotopts(lwidth(*4) lcolor("55 101 168"))      ///
>     noshow
```




```
/* Graphical representation of the survival function */
sts graph, survival tmin(10) ylabel(0(0.2)1, format(%4.1f))   ///
    xtitle("Age") plotopts(lwidth(*4) lcolor("55 101 168")    ///
    connect(direct)) noshow
```

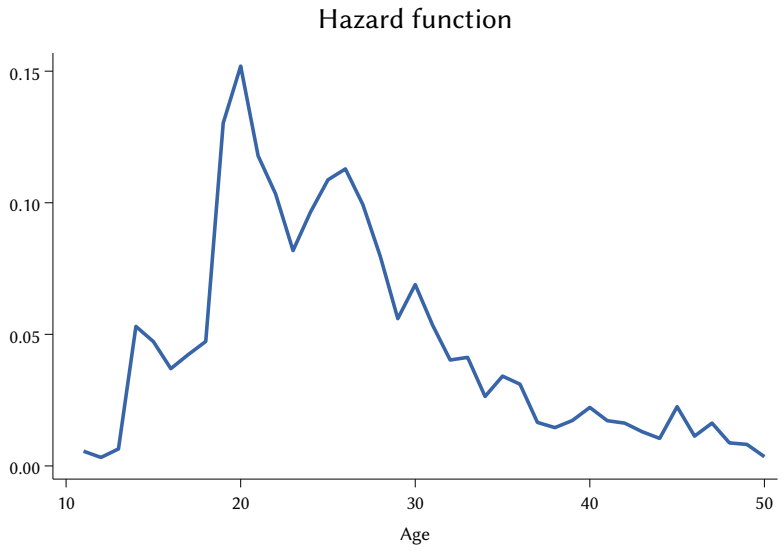
```
. /* Graphical representation of the survival function */
. sts graph, survival tmin(10) ylabel(0(0.2)1, format(%4.1f))   ///
>     xtitle("Age") plotopts(lwidth(*4) lcolor("55 101 168")    ///
>     connect(direct)) noshow
```



Alternatively, one can use the command `sts generate` to create two new variables containing the estimates of $h(t)$ and $S(t)$, and then use these variables to graph the two functions:

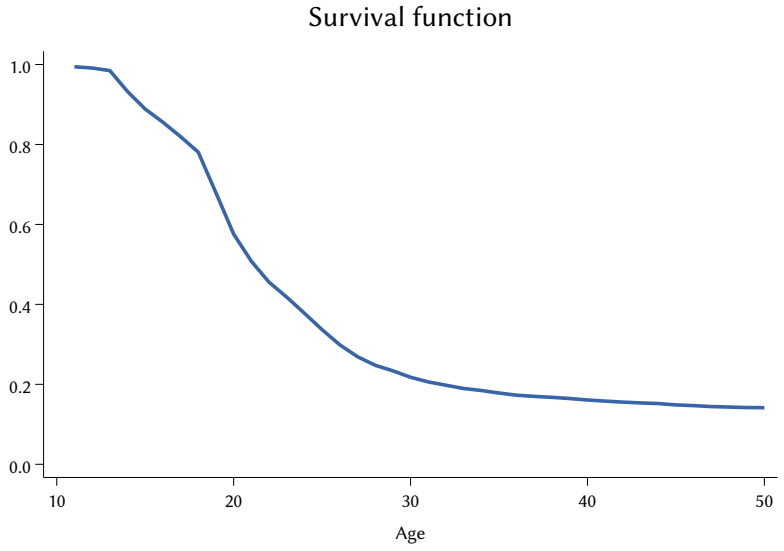
```
/* Calculation and graphical representation of the hazard function */
sts generate h_t = h

graph twoway line h_t _t, sort title("Hazard function")   ///
    ytitle("") ylabel(0(0.05)0.15, format(%4.2f))         ///
    xtitle("Age") lwidth(*4) lcolor("55 101 168")
```

```
. /* Calculation and graphical representation of the hazard function */
. sts generate h_t = h

.
. graph twoway line h_t _t, sort title("Hazard function")   ///
>     ytitle("") ylabel(0(0.05)0.15, format(%4.2f))         ///
>     xtitle("Age") lwidth(*4) lcolor("55 101 168")
```




```
/* Calculation and graphical representation of the survival function */
sts generate S_t = s

graph twoway line S_t _t, sort title("Survival function")   ///
    ytitle("") ylabel(0(0.2)1, format(%4.1f))               ///
    xtitle("Age") lwidth(*4) lcolor("55 101 168")
```

```
. /* Calculation and graphical representation of the survival function */
. sts generate S_t = s

.
. graph twoway line S_t _t, sort title("Survival function")   ///
>     ytitle("") ylabel(0(0.2)1, format(%4.1f))               ///
>     xtitle("Age") lwidth(*4) lcolor("55 101 168")
```



From a design-based perspective, the approach just outlined provides correct point estimates of $h(t)$ and $S(t)$, but it does not allow for a correct estimate of the uncertainty around the point estimates. To obtain such an estimate, we must convert the data in memory into a person-time dataset, fit an appropriate regression model to the latter, and use the resulting parameter estimates to compute the quantities of interest. Here is one way to implement this procedure in Stata:

```
/* Preserve data in memory */
preserve

/* Generate person-year dataset */
stsplit pyear, every(1)
replace pyear = pyear + 1

/* Design-based estimation of piecewise exponential regression model */
svy, dots(10) : streg ibn.pyear, distribution(exponential) nocons

/* Create dataset for estimation of quantities of interest */
collapse _st if _st, by(pyear _t)

/* Compute design-based point estimates (h_t_est) and corresponding
   standard errors (h_t_se) of hazard function */
predictnl h_t_est = predict(hazard), se(h_t_se)

/* Compute lower (h_t_lb) and upper (h_t_ub) limits of design-based
   95% confidence intervals around point estimates of hazard function
   using logit transformation */
generate h_t_lb = logit(h_t_est) -                          ///
    invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))
replace h_t_lb = invlogit(h_t_lb)
generate h_t_ub = logit(h_t_est) +                          ///
    invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))
replace h_t_ub = invlogit(h_t_ub)

/* Compute design-based point estimates (S_t_est) and corresponding
   variances (S_t_var) of survival function */
predictnl S_t_est = exp(sum(ln((1 - predict(hazard))))),   ///
    variance(S_t_var)

/* Compute lower (S_t_lb) and upper (S_t_ub) limits of design-based
   95% confidence intervals around point estimates of survival function
   using log-log transformation as per Heeringa et al. (2017) */
generate S_t_lb = ln(-ln(S_t_est)) +                                ///
    invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)
replace S_t_lb = exp(-exp(S_t_lb))
```

```
generate S_t_ub = ln(-ln(S_t_est)) -                                    ///
    invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)
replace S_t_ub = exp(-exp(S_t_ub))

/* Graphical representation of hazard function with 95% CI */
graph twoway                                                            ///
    (rcap h_t_lb h_t_ub _t, color("55 101 168"))                        ///
    (connected h_t_est _t, lwidth(*3) lcolor("55 101 168")              ///
     msize(*0.5) mlcolor("55 101 168") mfcolor(white))                  ///
    ,                                                                   ///
    title("Hazard function")                                            ///
    subtitle("Design-based point estimates and"                         ///
             "95% confidence intervals")                                ///
    ytitle("") ylabel(0(0.05)0.20, format(%4.2f))                       ///
    xtitle("Age") xlabel(10(5)50) legend(off)                           ///
    name(h_t, replace)

/* Graphical representation of survival function with 95% CI */
graph twoway                                                            ///
    (rcap S_t_lb S_t_ub _t, color("55 101 168"))                        ///
    (connected S_t_est _t, lwidth(*3) lcolor("55 101 168")              ///
     msize(*0.5) mlcolor("55 101 168") mfcolor(white))                  ///
    ,                                                                   ///
    title("Survival function")                                          ///
    subtitle("Design-based point estimates and"                         ///
             "95% confidence intervals")                                ///
    ytitle("") ylabel(0(0.2)1, format(%4.1f))                           ///
    xtitle("Age") xlabel(10(5)50) legend(off)                           ///
    name(S_t, replace)

/* Restore initial data */
restore
```

```
. /* Preserve data in memory */
. preserve

.
. /* Generate person-year dataset */
. stsplit pyear, every(1)
(91,977 observations (episodes) created)

. replace pyear = pyear + 1
(98,781 real changes made)

.
. /* Design-based estimation of piecewise exponential regression model */
. svy, dots(10) : streg ibn.pyear, distribution(exponential) nocons
(running streg on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............
```

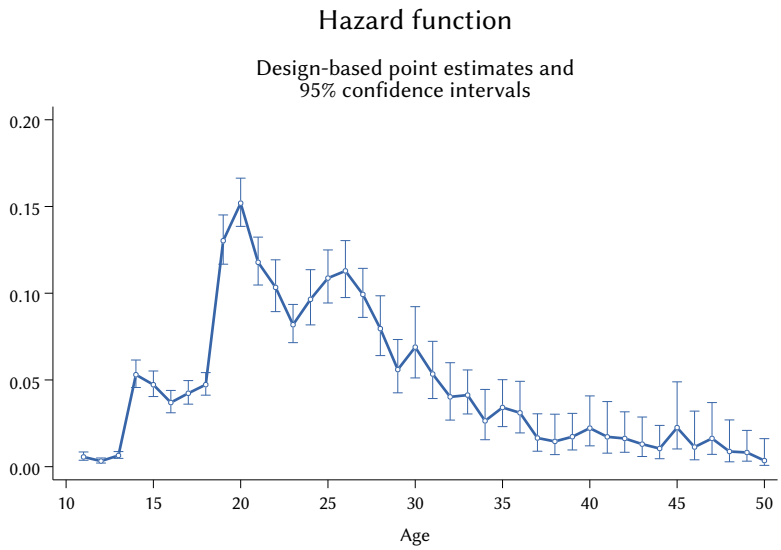
```
Survey: Exponential PH regression                 Number of obs    =     100,755
                                                  Population size = 593,701,831
                                                  Replications     =         152
                                                  Design df        =         150
                                                  F(40, 111)       =      507.45
                                                  Prob > F         =      0.0000
```

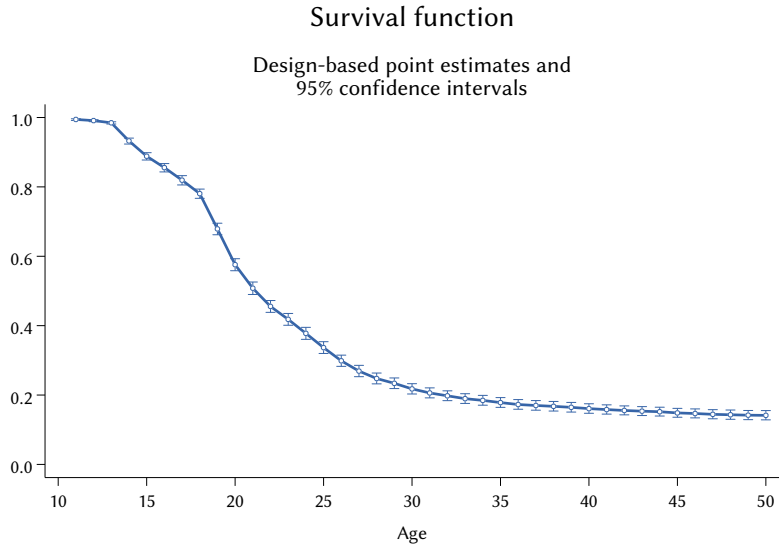
_t	Haz. ratio	BRR * std. err.	t	P>|t|	[95% conf.	interval]
pyear						
11	.0056065	.0011741	-24.75	0.000	.0037068	.00848
12	.0032395	.0007124	-26.07	0.000	.0020978	.0050026
13	.0064593	.0009758	-33.38	0.000	.0047924	.008706
14	.0530156	.0039905	-39.02	0.000	.045689	.061517
15	.0472496	.0037119	-38.85	0.000	.0404561	.0551839
16	.0369719	.0032437	-37.59	0.000	.0310875	.0439701
17	.0423278	.0034216	-39.12	0.000	.0360793	.0496585
18	.0472997	.0032822	-43.97	0.000	.0412394	.0542505
19	.1302792	.0071732	-37.02	0.000	.1168494	.1452526
20	.1519202	.0070272	-40.74	0.000	.1386507	.1664596
21	.1178315	.0069758	-36.12	0.000	.1048235	.1324536
22	.1033709	.007537	-31.13	0.000	.0895016	.1193894
23	.0818448	.0055631	-36.82	0.000	.0715589	.0936092
24	.0964806	.0080258	-28.11	0.000	.081857	.1137166
25	.1087201	.0077208	-31.25	0.000	.0944866	.1250978
26	.1128701	.0082906	-29.70	0.000	.097622	.1304999
27	.0993032	.0071365	-32.14	0.000	.0861575	.1144547
28	.0795818	.0086754	-23.22	0.000	.0641605	.0987097
29	.0559872	.007698	-20.97	0.000	.0426677	.0734646
30	.068916	.010285	-17.92	0.000	.0513162	.0925521
31	.0534194	.0082423	-18.99	0.000	.0393818	.0724607
32	.0402557	.0081833	-15.80	0.000	.0269393	.0601548
33	.0412529	.0063362	-20.76	0.000	.0304545	.0558801
34	.0264036	.0070414	-13.63	0.000	.015589	.0447209
35	.0341247	.0066976	-17.21	0.000	.023155	.0502912
36	.0310608	.0072943	-14.78	0.000	.0195295	.0494008
37	.0165351	.0051482	-13.18	0.000	.0089378	.0305904
38	.014563	.0054239	-11.36	0.000	.0069767	.0303987
39	.0172591	.0050655	-13.83	0.000	.0096641	.0308231
40	.0222293	.0068856	-12.29	0.000	.0120537	.040995
41	.0172033	.0068508	-10.20	0.000	.0078322	.0377869
42	.0162665	.0055074	-12.16	0.000	.0083321	.0317565
43	.012999	.0052246	-10.81	0.000	.005875	.0287616
44	.0104696	.0043712	-10.92	0.000	.0045882	.0238898
45	.0225177	.0089398	-9.56	0.000	.0102764	.0493411
46	.0113316	.0060093	-8.45	0.000	.0039739	.0323122
47	.0162893	.0068114	-9.85	0.000	.0071298	.0372159
48	.0087496	.0050155	-8.27	0.000	.0028189	.0271574
49	.0081819	.0039056	-10.07	0.000	.0031859	.0210125
50	.0035235	.0027257	-7.30	0.000	.0007641	.0162483

```
.
. /* Create dataset for estimation of quantities of interest */
. collapse _st if _st, by(pyear _t)

.
. /* Compute design-based point estimates (h_t_est) and corresponding
```

```
>    standard errors (h_t_se) of hazard function */
. predictnl h_t_est = predict(hazard), se(h_t_se)

.
. /* Compute lower (h_t_lb) and upper (h_t_ub) limits of design-based
>    95% confidence intervals around point estimates of hazard function
>    using logit transformation */
. generate h_t_lb = logit(h_t_est) -                          ///
>     invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))

. replace h_t_lb = invlogit(h_t_lb)
(40 real changes made)

. generate h_t_ub = logit(h_t_est) +                          ///
>     invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))

. replace h_t_ub = invlogit(h_t_ub)
(40 real changes made)

.
. /* Compute design-based point estimates (S_t_est) and corresponding
>    variances (S_t_var) of survival function */
. predictnl S_t_est = exp(sum(ln((1 - predict(hazard))))),   ///
>     variance(S_t_var)

.
. /* Compute lower (S_t_lb) and upper (S_t_ub) limits of design-based
>    95% confidence intervals around point estimates of survival function
>    using log-log transformation as per Heeringa et al. (2017) */
. generate S_t_lb = ln(-ln(S_t_est)) +                                ///
>     invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)

. replace S_t_lb = exp(-exp(S_t_lb))
(40 real changes made)

. generate S_t_ub = ln(-ln(S_t_est)) -                                ///
>     invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)

. replace S_t_ub = exp(-exp(S_t_ub))
(40 real changes made)

.
. /* Graphical representation of hazard function with 95% CI */
. graph twoway                                                  ///
>     (rcap h_t_lb h_t_ub _t, color("55 101 168"))              ///
>     (connected h_t_est _t, lwidth(*3) lcolor("55 101 168")    ///
>      msize(*0.5) mlcolor("55 101 168") mfcolor(white))        ///
>     ,                                                         ///
>     title("Hazard function")                                  ///
>     subtitle("Design-based point estimates and"               ///
>              "95% confidence intervals")                      ///
>     ytitle("") ylabel(0(0.05)0.20, format(%4.2f))             ///
>     xtitle("Age") xlabel(10(5)50) legend(off)                 ///
>     name(h_t, replace)

.
. /* Graphical representation of survival function with 95% CI */
. graph twoway                                                  ///
>     (rcap S_t_lb S_t_ub _t, color("55 101 168"))              ///
>     (connected S_t_est _t, lwidth(*3) lcolor("55 101 168")    ///
>      msize(*0.5) mlcolor("55 101 168") mfcolor(white))        ///
>     ,                                                         ///
>     title("Survival function")                                ///
>     subtitle("Design-based point estimates and"               ///
>              "95% confidence intervals")                      ///
```

```
>     ytitle("") ylabel(0(0.2)1, format(%4.1f))                ///
>     xtitle("Age") xlabel(10(5)50) legend(off)                ///
>     name(S_t, replace)

.
. /* Restore initial data */
. restore
```





Because it has been estimated using a separate parameter for each year, the hazard function shown above appears rather noisy, with frequent changes in direction that are implausible on a substantive level and can easily be attributed to random variation. A smoother representation of the hazard function can

be obtained by modeling time with an appropriate number of restricted cubic splines (Royston and Lambert 2011). For example:

```
/* Preserve data in memory */
preserve

/* Generate person-semester dataset */
stsplit semester, every(0.5)
replace semester = semester + 0.5

/* Generate restricted cubic splines */
rcsgen semester, df(5) orthog gen(age_rcs)

/* Design-based estimation of exponential regression model */
svy, dots(10) : streg age_rcs*, distribution(exponential)

/* Create dataset for estimation of quantities of interest */
collapse _st if _st, by(_t age_rcs*)

/* Compute design-based point estimates (h_t_est) and corresponding
   standard errors (h_t_se) of hazard function */
predictnl h_t_est = predict(hazard), se(h_t_se)

/* Compute lower (h_t_lb) and upper (h_t_ub) limits of design-based
   95% confidence intervals around point estimates of hazard function
   using logit transformation */
generate h_t_lb = logit(h_t_est) -                          ///
    invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))
replace h_t_lb = invlogit(h_t_lb)
generate h_t_ub = logit(h_t_est) +                          ///
    invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))
replace h_t_ub = invlogit(h_t_ub)

/* Compute design-based point estimates (S_t_est) and corresponding
   variances (S_t_var) of survival function */
predictnl S_t_est = exp(sum(ln((1 - predict(hazard))))),   ///
    variance(S_t_var)

/* Compute lower (S_t_lb) and upper (S_t_ub) limits of design-based
   95% confidence intervals around point estimates of survival function
   using log-log transformation as per Heeringa et al. (2017) */
generate S_t_lb = ln(-ln(S_t_est)) +                                ///
    invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)
replace S_t_lb = exp(-exp(S_t_lb))
generate S_t_ub = ln(-ln(S_t_est)) -                                ///
    invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)
replace S_t_ub = exp(-exp(S_t_ub))
```

```
/* Graphical representation of hazard function with 95% CI */
graph twoway                                                    ///
    (rarea h_t_lb h_t_ub _t, color("55 101 168%10"))            ///
    (line h_t_est _t, lwidth(*2) lcolor("55 101 168"))          ///
    ,                                                           ///
    title("Hazard function")                                    ///
    subtitle("Design-based point estimates and"                 ///
             "95% confidence intervals")                        ///
    ytitle("") ylabel(0(0.05)0.20, format(%4.2f))               ///
    xtitle("Age") xlabel(10(5)50) legend(off)                   ///
    name(h_t, replace)

/* Graphical representation of survival function with 95% CI */
graph twoway                                                  ///
    (rarea S_t_lb S_t_ub _t, color("55 101 168%10"))          ///
    (line S_t_est _t, lwidth(*2) lcolor("55 101 168"))        ///
    ,                                                         ///
    title("Survival function")                                ///
    subtitle("Design-based point estimates and"               ///
             "95% confidence intervals")                      ///
    ytitle("") ylabel(0(0.2)1, format(%4.1f))                 ///
    xtitle("Age") xlabel(10(5)50) legend(off)                 ///
    name(S_t, replace)

/* Restore initial data */
restore
```

```
. /* Preserve data in memory */
. preserve

.
. /* Generate person-semester dataset */
. stsplit semester, every(0.5)
(190,758 observations (episodes) created)

. replace semester = semester + 0.5
(197,562 real changes made)

.
. /* Generate restricted cubic splines */
. rcsgen semester, df(5) orthog gen(age_rcs)
Variables age_rcs1 to age_rcs5 were created

.
. /* Design-based estimation of exponential regression model */
. svy, dots(10) : streg age_rcs*, distribution(exponential)
(running streg on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Exponential PH regression               Number of obs    =       199,536
                                                Population size = 1,175,495,316
                                                Replications     =           152
```

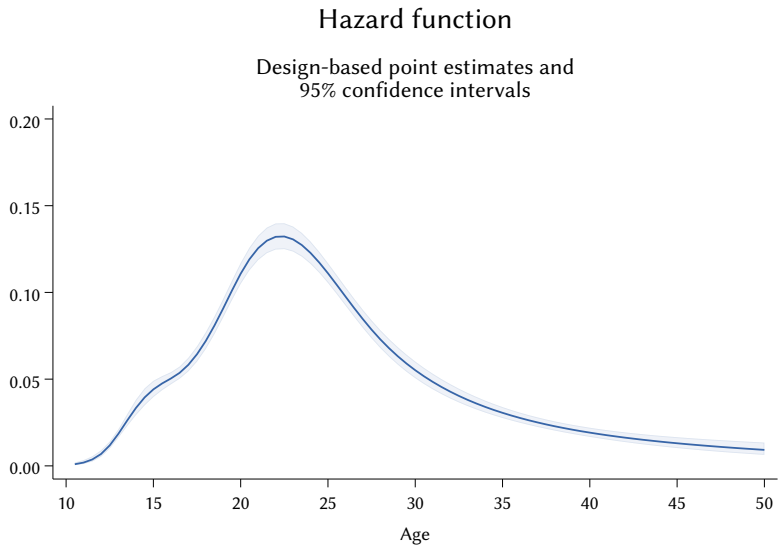
```
                                                Design df       =           150
                                                F(5, 146)       =        291.52
                                                Prob > F        =        0.0000
```

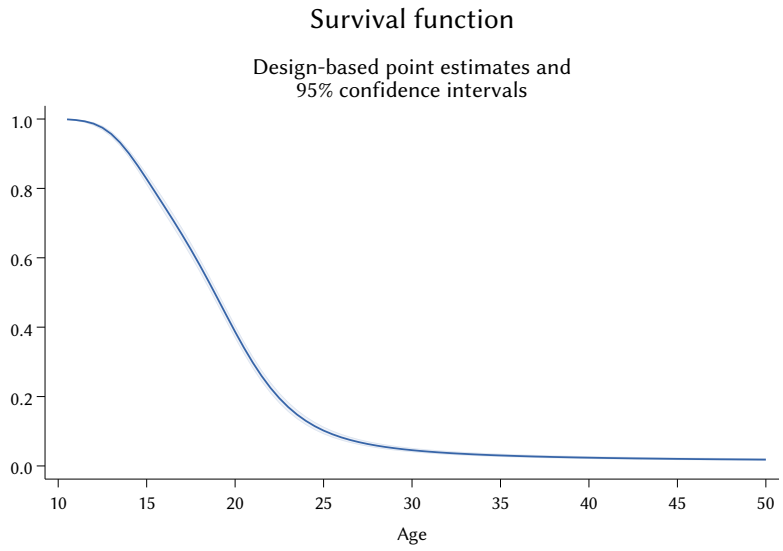
_t	Haz. ratio	BRR * std. err.	t	P>|t|	[95% conf.	interval]
age_rcs1	1.288529	.057419	5.69	0.000	1.179926	1.407128
age_rcs2	2.958742	.1558475	20.59	0.000	2.666285	3.283277
age_rcs3	.5877044	.0401207	-7.79	0.000	.5135437	.6725745
age_rcs4	.9391254	.0250994	-2.35	0.020	.8908181	.9900522
age_rcs5	.8171342	.0139957	-11.79	0.000	.7899427	.8452616
_cons	.032945	.0013717	-81.97	0.000	.0303433	.0357699

Note: **_cons** estimates baseline hazard.

```
.
. /* Create dataset for estimation of quantities of interest */
. collapse _st if _st, by(_t age_rcs*)

.
. /* Compute design-based point estimates (h_t_est) and corresponding
>    standard errors (h_t_se) of hazard function */
. predictnl h_t_est = predict(hazard), se(h_t_se)

.
. /* Compute lower (h_t_lb) and upper (h_t_ub) limits of design-based
>    95% confidence intervals around point estimates of hazard function
>    using logit transformation */
. generate h_t_lb = logit(h_t_est) -                          ///
>     invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))

. replace h_t_lb = invlogit(h_t_lb)
(80 real changes made)

. generate h_t_ub = logit(h_t_est) +                          ///
>     invt(140,0.975) * h_t_se / (h_t_est * (1 - h_t_est))

. replace h_t_ub = invlogit(h_t_ub)
(80 real changes made)

.
. /* Compute design-based point estimates (S_t_est) and corresponding
>    variances (S_t_var) of survival function */
. predictnl S_t_est = exp(sum(ln((1 - predict(hazard))))),   ///
>     variance(S_t_var)

.
. /* Compute lower (S_t_lb) and upper (S_t_ub) limits of design-based
>    95% confidence intervals around point estimates of survival function
>    using log-log transformation as per Heeringa et al. (2017) */
. generate S_t_lb = ln(-ln(S_t_est)) +                              ///
>     invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)

. replace S_t_lb = exp(-exp(S_t_lb))
(80 real changes made)

. generate S_t_ub = ln(-ln(S_t_est)) -                              ///
>     invt(140,0.975) * sqrt(S_t_var / (S_t_est * ln(S_t_est))^2)

. replace S_t_ub = exp(-exp(S_t_ub))
(80 real changes made)

.
. /* Graphical representation of hazard function with 95% CI */
```

```
. graph twoway                                                  ///
>     (rarea h_t_lb h_t_ub _t, color("55 101 168%10"))          ///
>     (line h_t_est _t, lwidth(*2) lcolor("55 101 168"))        ///
>     ,                                                         ///
>     title("Hazard function")                                  ///
>     subtitle("Design-based point estimates and"               ///
>              "95% confidence intervals")                      ///
>     ytitle("") ylabel(0(0.05)0.20, format(%4.2f))             ///
>     xtitle("Age") xlabel(10(5)50) legend(off)                 ///
>     name(h_t, replace)

.
. /* Graphical representation of survival function with 95% CI */
. graph twoway                                                 ///
>     (rarea S_t_lb S_t_ub _t, color("55 101 168%10"))         ///
>     (line S_t_est _t, lwidth(*2) lcolor("55 101 168"))       ///
>     ,                                                        ///
>     title("Survival function")                               ///
>     subtitle("Design-based point estimates and"              ///
>              "95% confidence intervals")                     ///
>     ytitle("") ylabel(0(0.2)1, format(%4.1f))                ///
>     xtitle("Age") xlabel(10(5)50) legend(off)                ///
>     name(S_t, replace)

.
. /* Restore initial data */
. restore
```





Sometimes it is useful to estimate percentiles of survival time. In Stata, the command `stsum` provides correct design-based point estimates of the three quartiles of survival time:

```
/* Compute quartiles of survival time (point estimates) */
stsum, noshow
```

```
    . /* Compute quartiles of survival time (point estimates) */
. stsum, noshow
```

	Time at risk	Incidence rate	Number of subjects	Survival time 25%	50%	75%
Total	581793484.9	.0552336	3.83e+07	19	22	28

Although `stsum` does not directly support design-based inference, it can still be used for this purpose with a little bit of coding:

```
/* Quartiles of survival time (design-based inference) */
capture program drop survperc

program survperc, rclass
version 17.0
syntax anything [if] [iw pw]
if "`weight'" != "" {
    local wgtexp "[`weight' `exp']"
}
quietly {
    streset `wgtexp'
    stsum
```

```
}
return scalar Q1 = r(p25)
return scalar Q2 = r(p50)
return scalar Q3 = r(p75)
quietly {
    streset [pw = fiw]
}
end

svy brr Q1=r(Q1) Median=r(Q2) Q3=r(Q3), dots(10) :   ///
    survperc estimate if _st
```

```
    . /* Quartiles of survival time (design-based inference) */
. capture program drop survperc

. program survperc, rclass
  1. version 17.0
  2. syntax anything [if] [iw pw]
  3. if "`weight'" != "" {
  4.     local wgtexp "[`weight' `exp']"
  5. }
  6. quietly {
  7.     streset `wgtexp'
  8.     stsum
  9. }
 10. return scalar Q1 = r(p25)
 11. return scalar Q2 = r(p50)
 12. return scalar Q3 = r(p75)
 13. quietly {
 14.     streset [pw = fiw]
 15. }
 16. end

. svy brr Q1=r(Q1) Median=r(Q2) Q3=r(Q3), dots(10) :   ///
>     survperc estimate if _st
(running survperc on estimation sample)
BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..............

BRR results                                     Number of obs    =      6,804
                                                Population size  = 38,284,250
                                                Replications     =        152
                                                Design df        =        150

      Command: survperc estimate if _st
           Q1: r(Q1)
       Median: r(Q2)
           Q3: r(Q3)

------------------------------------------------------------------------------
             |                  BRR *
             | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          Q1 |          19          .        .       .            .           .
      Median |          22   .2809757    78.30   0.000     21.44482    22.55518
          Q3 |          28   1.051315    26.63   0.000      25.9227     30.0773
------------------------------------------------------------------------------
```

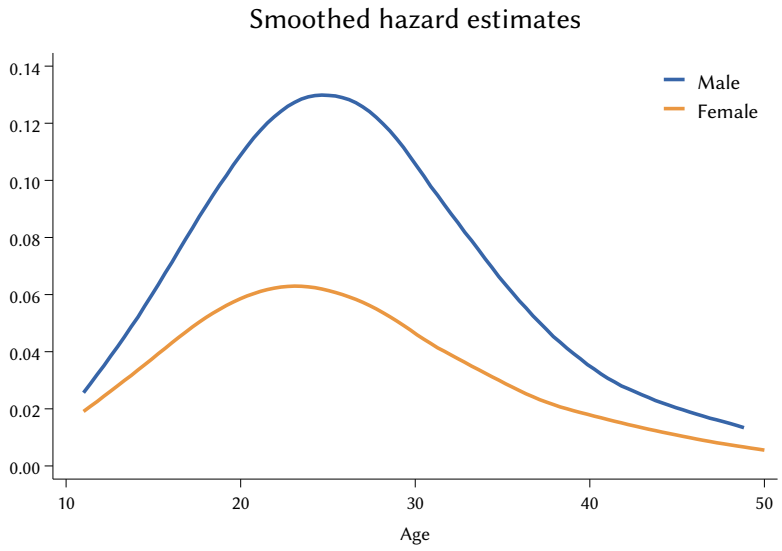
As can be seen, it was not possible to calculate the design-based confidence interval around the first quartile. This is because there is not enough sampling variation in the data to calculate the standard error of the corresponding estimator.

Typically, the description of the hazard and survival functions for the entire target population is followed by more extensive analyses aimed at determining, either from a predictive or from a causal perspective, whether and how the risk and time of occurrence of the event of interest vary with the values of one or more covariates. Here we present only a simple example of this type of analysis, namely one that aims at investigating whether and how the risk and time of entry into the first job vary by sex. The analysis can have three possible outcomes: (a) there is no significant difference between men and women; (b) there is a difference between men and women in terms of the hazard function and this difference, as measured by the hazard ratio, is constant over time (proportional hazards); or (c) there is a difference between men and women in terms of the hazard function and this difference, as measured by the hazard ratio, varies over time (nonproportional hazards).

For a first graphical exploration of the association between entry into the first job and sex, we can use the aforementioned command `sts graph`:

```
/* Smoothed hazard function, by sex */
sts graph, by(sex) hazard noboundary noshow            ///
    ylabel(0(0.02)0.14, format(%4.2f)) xtitle("Age")   ///
    plot1opts(lwidth(*4) lcolor("55 101 168"))         ///
    plot2opts(lwidth(*4) lcolor("234 151 65"))         ///
    legend(order(1 "Male" 2 "Female")                  ///
    cols(1) ring(0) position(2))
```

```
    . /* Smoothed hazard function, by sex */
. sts graph, by(sex) hazard noboundary noshow            ///
>     ylabel(0(0.02)0.14, format(%4.2f)) xtitle("Age")   ///
>     plot1opts(lwidth(*4) lcolor("55 101 168"))         ///
>     plot2opts(lwidth(*4) lcolor("234 151 65"))         ///
>     legend(order(1 "Male" 2 "Female")                  ///
>     cols(1) ring(0) position(2))
```



This graph suggests two preliminary conclusions: first, the hazard function for women is lower than that for men at all ages; second, the difference between women and men, as measured by the female/male hazard ratio, appears to vary over time.

There are several ways to formally test these tentative conclusions. The quickest is to estimate a *Cox regression model* with variable sex as the only covariate. Then, to determine whether the hazard function is on average lower for women than for men, we need only examine the regression coefficient associated with sex, which corresponds to the female/male hazard ratio: if its estimated value is significantly less than one, we can conclude that the hazard function of women is indeed on average lower than that of men. Here is the relevant Stata code and output:

```
/* Design-based Cox regression analysis */
svy, dots(10) : stcox sex if _st
```

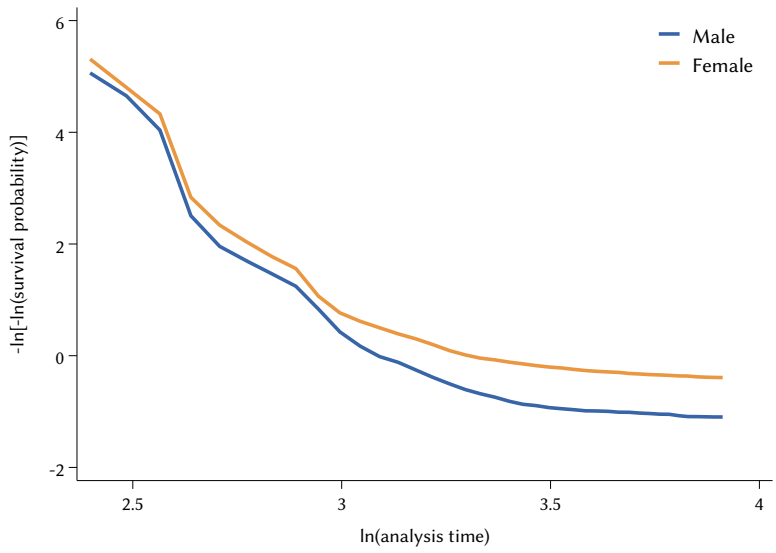
```
    . /* Design-based Cox regression analysis */
. svy, dots(10) : stcox sex if _st
(running stcox on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Cox regression                          Number of obs    =      6,804
                                                Population size  = 38,284,250
                                                Replications     =        152
                                                Design df        =        150
                                                F(1, 150)        =     332.81
                                                Prob > F         =     0.0000
```

_t	Haz. ratio	BRR * std. err.	t	P>|t|	[95% conf.	interval]
sex	.5732749	.0174842	-18.24	0.000	.5397482	.6088841

The output corroborates our first conclusion: on average, the hazard function of women is between 54% and 61% of that of men. To test whether this ratio is constant over the entire period under consideration or rather varies with age, we can use the command `stphplot` as follows:

```
/* Test of proportionality of hazard functions by sex */
stphplot, by(sex) noshow                                    ///
    plot1opts(lwidth(*4) lcolor("55 101 168") msymbol(i))   ///
    plot2opts(lwidth(*4) lcolor("234 151 65") msymbol(i))   ///
    legend(order(1 "Male" 2 "Female") cols(1) ring(0)       ///
    position(2))
```

```
. /* Test of proportionality of hazard functions by sex */
. stphplot, by(sex) noshow                                    ///
>     plot1opts(lwidth(*4) lcolor("55 101 168") msymbol(i))   ///
>     plot2opts(lwidth(*4) lcolor("234 151 65") msymbol(i))   ///
>     legend(order(1 "Male" 2 "Female") cols(1) ring(0)       ///
>     position(2))
.
```



This "log-log" plot is easy to read: the proportional hazards assumption – that is, the hypothesis of stability of the hazard ratio over time – holds if the two curves are parallel. Since this is clearly not the case, our second conclusion

also seems to be supported: the difference between women and men in the (conditional) risk of entry into the first job varies over time.

Note, however, that the "log-log" plot does not take into account the sampling uncertainty associated with the estimates. To draw more robust conclusions about the proportionality or nonproportionality of the hazard functions for women and men, appropriate design-based inference is required. For this purpose, the powerful and versatile command gsem can be used as follows:

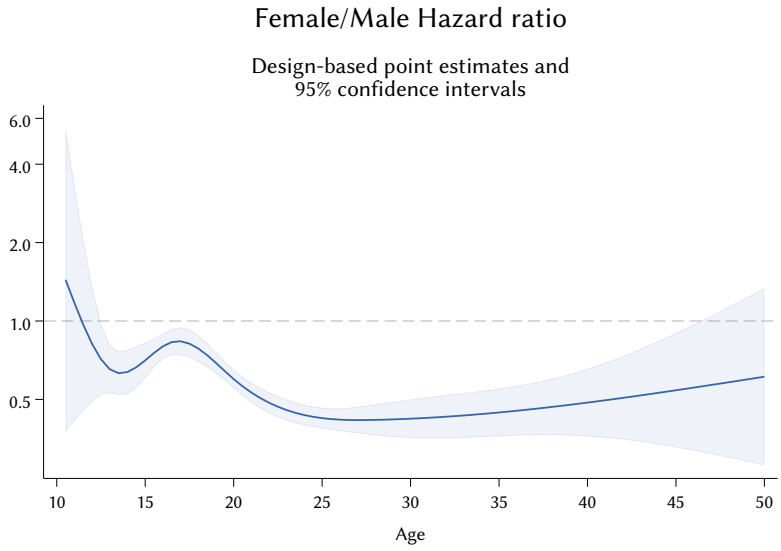
```
/* Preserve data in memory */
preserve

/* Generate person-semester dataset */
stsplit semester, every(0.5)
replace semester = semester + 0.5

/* Generate restricted cubic splines */
rcsgen semester, df(5) orthog gen(age_rcs)

/* Design-based estimation of exponential regression model */
clonevar _tm = _t if (sex == 0)
clonevar _tf = _t if (sex == 1)
svy, dots(10) : gsem (_tm <- age_rcs*) (_tf <- age_rcs*) if _st,   ///
    family(exponential, failure(_d) ltruncated(_t0) ph)

/* Compute design-based point estimates (hr_t_est) and corresponding
   standard errors (hr_t_se) of log-hazard ratio */
predictnl hr_t_est = predict(expression(eta(_tf) - eta(_tm))),   ///
    se(hr_t_se)

/* Simplify dataset */
collapse hr_t_est hr_t_se if _st, by(_t)

/* Compute lower (hr_t_lb) and upper (hr_t_ub) limits of design-based
   95% confidence intervals around point estimates of hazard ratio */
generate hr_t_lb = hr_t_est - invt(140,0.975) * hr_t_se
replace hr_t_lb = exp(hr_t_lb)
generate hr_t_ub = hr_t_est + invt(140,0.975) * hr_t_se
replace hr_t_ub = exp(hr_t_ub)

/* Compute hazard ratio */
replace hr_t_est = exp(hr_t_est)

/* Graphical representation of hazard ratio with 95% CI */
graph twoway                                                    ///
    (rarea hr_t_lb hr_t_ub _t, color("55 101 168%10"))          ///
    (line hr_t_est _t, lwidth(*2) lcolor("55 101 168"))         ///
```

```
    ,                                                                   ///
    title("Female/Male Hazard ratio")                                   ///
    subtitle("Design-based point estimates and"                         ///
             "95% confidence intervals")                                ///
    yscale(log) ytitle("") ylabel(0 0.5 1 2 4 6, format(%3.1f))         ///
    yline(1, lpattern(dash) lcolor(gs11))                               ///
    xtitle("Age") xlabel(10(5)50) legend(off)

/* Restore initial data */
restore
```

```
. /* Preserve data in memory */
. preserve

.
. /* Generate person-semester dataset */
. stsplit semester, every(0.5)
(190,758 observations (episodes) created)

. replace semester = semester + 0.5
(197,562 real changes made)

.
. /* Generate restricted cubic splines */
. rcsgen semester, df(5) orthog gen(age_rcs)
Variables age_rcs1 to age_rcs5 were created

.
. /* Design-based estimation of exponential regression model */
. clonevar _tm = _t if (sex == 0)
(127,840 missing values generated)

. clonevar _tf = _t if (sex == 1)
(73,670 missing values generated)

. svy, dots(10) : gsem (_tm <- age_rcs*) (_tf <- age_rcs*) if _st,   ///
>     family(exponential, failure(_d) ltruncated(_t0) ph)
(running gsem on estimation sample)

BRR replications (152)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..........

Survey: Generalized structural equation model

                                                Number of obs   =       197,562
                                                Population size = 1,163,586,970
                                                Replications    =           152
                                                Design df       =           150

Response: _tm                                   Number of obs   =    71,696
Family:   Exponential                           No. of failures =     2,951
Form:     Proportional hazards                  Time at risk    = 35,848.00
Link:     Log

Response: _tf                                   Number of obs   =   125,866
Family:   Exponential                           No. of failures =     2,838
Form:     Proportional hazards                  Time at risk    = 62,933.00
Link:     Log

------------------------------------------------------------------------------
             |                BRR *
```

	Coefficient	std. err.	t	P>|t|	[95% conf.	interval]
_tm						
age_rcs1	.4200158	.0783285	5.36	0.000	.265246	.5747855
age_rcs2	1.162526	.0727868	15.97	0.000	1.018706	1.306345
age_rcs3	-.53068	.0921493	-5.76	0.000	-.7127583	-.3486016
age_rcs4	.0189924	.0348268	0.55	0.586	-.0498221	.087807
age_rcs5	-.2115933	.0261563	-8.09	0.000	-.2632757	-.1599109
_cons	-3.15592	.0581316	-54.29	0.000	-3.270782	-3.041057
_tf						
age_rcs1	.2104843	.057831	3.64	0.000	.0962157	.324753
age_rcs2	.993498	.0722012	13.76	0.000	.8508353	1.136161
age_rcs3	-.4998676	.0956456	-5.23	0.000	-.6888542	-.310881
age_rcs4	-.1214244	.0344602	-3.52	0.001	-.1895145	-.0533343
age_rcs5	-.147755	.0230763	-6.40	0.000	-.1933516	-.1021584
_cons	-3.602534	.0578516	-62.27	0.000	-3.716843	-3.488224

```
.
. /* Compute design-based point estimates (hr_t_est) and corresponding
>    standard errors (hr_t_se) of log-hazard ratio */
. predictnl hr_t_est = predict(expression(eta(_tf) - eta(_tm))),   ///
>     se(hr_t_se)
(1,974 missing values generated)

.
. /* Simplify dataset */
. collapse hr_t_est hr_t_se if _st, by(_t)

.
. /* Compute lower (hr_t_lb) and upper (hr_t_ub) limits of design-based
>    95% confidence intervals around point estimates of hazard ratio */
. generate hr_t_lb = hr_t_est - invt(140,0.975) * hr_t_se

. replace hr_t_lb = exp(hr_t_lb)
(80 real changes made)

. generate hr_t_ub = hr_t_est + invt(140,0.975) * hr_t_se

. replace hr_t_ub = exp(hr_t_ub)
(80 real changes made)

.
. /* Compute hazard ratio */
. replace hr_t_est = exp(hr_t_est)
(80 real changes made)

.
. /* Graphical representation of hazard ratio with 95% CI */
. graph twoway                                                          ///
>     (rarea hr_t_lb hr_t_ub _t, color("55 101 168%10"))                ///
>     (line hr_t_est _t, lwidth(*2) lcolor("55 101 168"))               ///
>     ,                                                                 ///
>     title("Female/Male Hazard ratio")                                 ///
>     subtitle("Design-based point estimates and"                       ///
>              "95% confidence intervals")                              ///
>     yscale(log) ytitle("") ylabel(0 0.5 1 2 4 6, format(%3.1f))       ///
>     yline(1, lpattern(dash) lcolor(gs11))                             ///
>     xtitle("Age") xlabel(10(5)50) legend(off)

.
. /* Restore initial data */
. restore
```



Our design-based analysis supports the conclusion of nonproportionality: the female/male hazard ratio is not constant, but varies with age. However, the variation is only observed up to the age of 25, after which the hazard ratio appears to be rather stable.

REFERENCES

AAPOR (2016) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, American Association for Public Opinion Research, https://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx (visited on 03/19/2022). [Cit. on pp. 10, 12.]

Agresti, A. (1980) "Generalized odds ratios for ordinal data," *Biometrics* 36(1), pp. 59-67. [Cit. on p. 95.]

– (2013) *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley. [Cit. on pp. 17, 92.]

– (2018) *Statistical Methods for the Social Sciences*. 5th ed. Boston, MA: Pearson. [Cit. on p. 60.]

Allison, P. D. (1982) "Discrete-time methods for the analysis of event histories," *Sociological Methodology* 13(1), pp. 61-98. [Cit. on p. 144.]

Battaglia, M. P. (2008) "EPSEM sample," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 234-235. [Cit. on p. 9.]

Bellocco, R. and Algeri, S. (2013) "Goodness-of-fit tests for categorical data," *The Stata Journal* 13(2), pp. 356-365. [Cit. on p. 111.]

Bergmann, M., Kneip, T., De Luca, G., and Scherpenzeel, A. (2019) *Survey Participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1-7*, Working Paper Series 41-2019, http://www.share-project.org/fileadmin/pdf_documentation/Working_Paper_Series/WP_Series_41_2019_Bergmann_et_al.pdf (visited on 03/20/2022). [Cit. on p. 12.]

Bethlehem, J., Cobben, F., and Schouten, B. (2011) *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: Wiley. [Cit. on pp. 15, 16.]

Biemer, P. P. and Lyberg, L. (2003) *Introduction to Survey Quality*. Hoboken, NJ: Wiley. [Cit. on p. 10.]

Blair, J. and Lacy, M. G. (2000) "Statistics of ordinal variation," *Sociological Methods & Research* 28(3), pp. 251-280. [Cit. on p. 79.]

Blossfeld, H.-P., Rohwer, G., and Schneider, T. (2019) *Event History Analysis with Stata*. 2nd ed. Abingdon: Routledge. [Cit. on pp. 137, 143.]

Budescu, D. V. and Budescu, M. (2012) "How to measure diversity when you must," *Psychological Methods* 17(2), pp. 215-227. [Cit. on p. 78.]

Budescu, D. V. (1993) "Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression," *Psychological Bulletin* 114(3), pp. 542-551. [Cit. on p. 116.]

Canette, I. (2016) *Discrete-time Survival Analysis with Stata*, 2016 Stata Users Group Meeting, Barcelona, https://www.stata.com/meeting/spain16/slides/canette-spain16.pdf (visited on 02/11/2023). [Cit. on p. 142.]

Caughey, D., Berinsky, A. J., Chatfield, S., Hartman, E., Schickler, E., and Sekhon, J. S. (2020) *Target Estimation and Adjustment Weighting for Survey Nonresponse and Sampling Bias*. Cambridge: Cambridge University press. [Cit. on p. 32.]

Chen, T.-C. and Parker, J. (2016) "Subsample weights studies: Alternate methods for constructing BRR weights for NHANES single year samples," in *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 3928-3936. [Cit. on p. 49.]

Cleves, M., Gould, W. W., and Marchenko, Y. V. (2016) *An Introduction to Survival Analysis Using Stata*. Revised 3rd ed. College Station, TX: Stata Press. [Cit. on pp. 137, 142, 143.]

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates. [Cit. on p. 98.]

Copi, I. M., Cohen, C., and Rodych, V. (2019) *Introduction to Logic*. 15th ed. New York, NY: Routledge. [Cit. on p. 41.]

Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press. [Cit. on p. 92.]

Daniel, J. (2012) *Sampling Essentials: Practical Guidelines for Making Sampling Choices*. Thousand Oaks, CA: Sage. [Cit. on p. 15.]

Dean, N. and Pagano, M. (2015) "Evaluating confidence interval methods for binomial proportions in clustered surveys," *Journal of Survey Statistics and Methodology* 3(4), pp. 484-503. [Cit. on p. 76.]

Deming, W. E. and Stephan, F. F. (1940) "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *The Annals of Mathematical Statistics* 11(4), pp. 427-444. [Cit. on p. 32.]

Dippo, C. S., Fay, R. E., and Morganstein, D. H. (1984) "Computing variances from complex samples with replicate weights," in *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 489-494. [Cit. on p. 47.]

Duncan, O. D. and Duncan, B. (1955) "A methodological analysis of segregation indexes," *American Sociological Review* 20(2), pp. 210-217. [Cit. on p. 17.]

Eurostat (2021) *Applying the Degree of Urbanisation: A Methodological Manual to Define Cities, Towns and Rural Areas for International Comparisons*. Luxembourg: Publications Office of the European Union. [Cit. on p. 3.]

Fay, R. E. (1989) "Theory and application of replicate weighting for variance calculations," in *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 212-217. [Cit. on p. 47.]

Flegal, K. M., Kit, B. K., and Graubard, B. I. (2014) "Body mass index categories in observational studies of weight and risk of death," *American Journal of Epidemiology* 180(3), pp. 288-296. [Cit. on p. 82.]

Franco, C., Little, R. J. A., Louis, T. A., and Slud, E. V. (2019) "Comparative study of confidence intervals for proportions in complex sample surveys," *Journal of Survey Statistics and Methodology* 7(3), pp. 334-364. [Cit. on p. 76.]

Gardner, E., Kimpel, T., and Zhao, Y. (2015) *American Community Survey User Guide*, Office of Financial Management, https://ofm.wa.gov/sites/default/files/public/legacy/pop/acs/ofm_acs_user_guide.pdf (visited on 07/18/2022). [Cit. on p. 51.]

Goodman, L. A. and Kruskal, W. H. (1954) "Measures of association for cross classifications," *Journal of the American Statistical Association* 49(268), pp. 732-764. [Cit. on p. 94.]

Groves, R. M. (1989) *Survey Errors and Survey Costs*. Hoboken, NJ: Wiley. [Cit. on p. 42.]

– (2006) "Nonresponse rates and nonresponse bias in household surveys," *Public Opinion Quarterly* 70(5), pp. 646-675. [Cit. on pp. 15, 16.]

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009) *Survey Methodology*. 2nd ed. Hoboken, NJ: Wiley. [Cit. on pp. 1, 3, 10, 42.]

Hacine-Gharbi, A. and Ravier, P. (2018) "A binning formula of bi-histogram for joint entropy estimation using mean square error minimization," *Pattern Recognition Letters* 101, pp. 21-28. [Cit. on p. 102.]

Hainmueller, J. (2012) "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis* 20(1), pp. 25-46. [Cit. on p. 129.]

Hall, J. (2008) "Area probability sample," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 33-36. [Cit. on p. 3.]

Haziza, D. and Beaumont, J.-F. (2017) "Construction of weights in surveys: A review," *Statistical Science* 32(2), pp. 206-226. [Cit. on p. 25.]

Heeringa, S. G., West, B. T., and Berglund, P. A. (2017) *Applied Survey Data Analysis*. 2nd ed. Boca Raton, FL: CRC Press. [Cit. on pp. 25, 44, 45, 51, 54.]

Iacus, S. M., King, G., and Porro, G. (2012) "Causal inference without balance checking: Coarsened exact matching," *Political Analysis* 20(1), pp. 1-24. [Cit. on p. 129.]

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press. [Cit. on p. 129.]

Istat (2019) *Annuario statistico italiano 2019*. Roma: Istituto nazionale di statistica. [Cit. on p. 2.]

Jann, B. (2008) "Multinomial goodness-of-fit: Large-sample tests with survey design correction and exact tests for small samples," *The Stata Journal* 8(2), pp. 147-169. [Cit. on p. 17.]

– (2017) *kmatch: Stata module for multivariate-distance and propensity-score matching, including entropy balancing, inverse probability weighting, (coarsened) exact matching, and regression adjustment*, Statistical Software Components S458346, Boston College Department of Economics, revised 19 Sep 2020, http://ideas.repec.org/c/boc/bocode/s458346.html (visited on 01/27/2023). [Cit. on p. 129.]

– (2020) *dstat: Stata module to compute summary statistics and distribution functions including standard errors and optional covariate balancing*, Statistical Software Components S458874, Boston College Department of Economics, revised 24 Nov 2022, http://ideas.repec.org/c/boc/bocode/s458874.html (visited on 12/04/2022). [Cit. on p. 77.]

– (2021) "Relative distribution analysis in Stata," *The Stata Journal* 21(4), pp. 885-951. [Cit. on pp. 17, 37.]

Jenkins, S. P. (2020) "Comparing distributions of ordinal data," *The Stata Journal* 20(3), pp. 505-531. [Cit. on p. 79.]

– (2021) "Inequality comparisons with ordinal data," *Review of Income and Wealth* 67(3), pp. 547-563. [Cit. on p. 79.]

Judkins, D. R. (1990) "Fay's method for variance estimation," *Journal of Official Statistics* 6(3), pp. 223-239. [Cit. on pp. 47-49.]

Kalton, G. and Flores-Cervantes, I. (2003) "Weighting methods," *Journal of Official Statistics* 19(2), pp. 81-97. [Cit. on p. 25.]

Kalton, G., Kali, J., and Sigman, R. (2014) "Handling frame problems when address-based sampling is used for in-person household surveys," *Journal of Survey Statistics and Methodology* 2(3), pp. 283-304. [Cit. on p. 7.]

Kendall, M. G. (1938) "A new measure of rank correlation," *Biometrika* 30(1/2), pp. 81-93. [Cit. on p. 94.]

Kim, J. K. and Wu, C. (2013) "Sparse and efficient replication variance estimation for complex surveys," *Survey Methodology* 39(1), pp. 91-120. [Cit. on p. 50.]

King, G., Keohane, R. O., and Verba, S. (1994) *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press. [Cit. on p. 41.]

Kish, L. (1965) *Survey Sampling*. New York, NY: John Wiley & Sons. [Cit. on p. 15.]

– (1995) "Methods for design effects," *Journal of Official Statistics* 11(1), pp. 55-77. [Cit. on p. 51.]

Knoke, D., Bohrnstedt, G. W., and Mee, A. P. (2002) *Statistics for Social Data Analysis*. 4th ed. Itasca, IL: F.E. Peacock Publishers. [Cit. on p. 60.]

Kohler, U. and Kreuter, F. (2012) *Data Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press. [Cit. on p. 60.]

Kolenikov, S. (2010) "Resampling variance estimation for complex survey data," *The Stata Journal* 10(2), pp. 165-199. [Cit. on pp. 45, 46.]

Korn, E. L. and Graubard, B. I. (1999) *Analysis of Health Surveys*. New York, NY: Wiley. [Cit. on pp. 49, 76, 99, 100, 113, 115.]

Kraemer, H. C. (2006) "Correlation coefficients in medical research: From product moment correlation to the odds ratio," *Statistical Methods in Medical Research* 15(6), pp. 525-545. [Cit. on p. 98.]

Larsen, M. D. (2008) "Proportional allocation to strata," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 629-630. [Cit. on p. 6.]

Lavrakas, P. J. (ed.) (2008a) *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage. [Cit. on p. 10.]

– (2008b) "Sample replicates," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 780-781. [Cit. on p. 6.]

Lenis, D., Nguyen, T. Q., Dong, N., and Stuart, E. A. (2019) "It's all about balance: Propensity score matching in the context of complex survey data," *Biostatistics* 20(1), pp. 147-163. [Cit. on p. 129.]

Levin, I. and Sinclair, B. (2018) "Causal inference with complex survey designs: Generating population estimates using survey weights," in *The Oxford Handbook of Polling and Survey Methods*, ed. by L. R. Atkeson and R. M. Alvarez, Oxford: Oxford University Press, pp. 299-315. [Cit. on p. 129.]

Levy, P. S. and Lemeshow, S. (2008) *Sampling of Populations: Methods and Applications*. 4th ed. Hoboken, NJ: Wiley. [Cit. on p. 2.]

Lohr, S. L. (2022) *Sampling: Design and Analysis*. 3rd ed. Boca Raton, FL: CRC Press. [Cit. on pp. 1, 2, 16, 41, 45, 46, 50, 59, 100.]

Lucchini, M., Argentin, G., Bussi, D., Consolazio, D., De Santis, G., Gerosa, T., Guidi, G., Negrelli, S., Piazzoni, C., Pisati, M., Respi, C., Riva, E., Sala, E., Scisci, D., and Terraneo, M. (2023) *Quality Profile: Questionnaires, Fieldwork, and Data Preparation*. Milano: Institute for Advanced Study of Social Change. [Cit. on pp. 6, 7, 10, 12, 60.]

Luchman, J. N. (2021) "Determining relative importance in Stata using dominance analysis: domin and domme," *The Stata Journal* 21(2), pp. 510-538. [Cit. on p. 116.]

Luiten, A., Hox, J., and de Leeuw, E. (2020) "Survey nonresponse trends and fieldwork effort in the 21st Century: Results of an international study across countries and surveys," *Journal of Official Statistics* 36(3), pp. 469-487. [Cit. on p. 10.]

Lynn, P. and Borkowska, M. (2018) *Some Indicators of Sample Representativeness and Attrition Bias for BHPS and Understanding Society*, Understanding Society Working Paper Series 2018-01, https://www.understandingsociety.ac.uk/research/publications/524851 (visited on 04/03/2022). [Cit. on p. 17.]

McCarthy, P. J. (1966) *Replication: An Approach to the Analysis of Data from Complex Surveys*, National Center for Health Statistics. [Cit. on p. 46.]

– (1969) "Pseudo-replication: Half samples," *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 37(3), pp. 239-264. [Cit. on p. 46.]

McFadden, D. L. (1974) "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, New York, NY: Academic Press, pp. 104-142. [Cit. on p. 127.]

Mehmetoglu, M. and Jakobsen, G. (2022) *Applied Statistics Using Stata: A Guide for the Social Sciences*. 2nd ed. Thousand Oaks, CA: Sage. [Cit. on p. 60.]

Pedlow, S. (2008) "Variance estimation," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 942-944. [Cit. on p. 45.]

Potter, F. (2008) "List sampling," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 434-436. [Cit. on p. 2.]

Potter, F., Jang, D., Friedman, E., Diaz-Tena, N., and Ghosh, B. (2003) "Comparison of procedures to account for certainty primary sampling units," in *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 3360-3365. [Cit. on p. 49.]

Rao, J. N. K. and Scott, A. J. (1984) "On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data," *The Annals of Statistics* 12(1), pp. 46-60. [Cit. on p. 90.]

Rao, J. N. K. and Shao, J. (1999) "Modified balanced repeated replication for complex survey data," *Biometrika* 86(2), pp. 403-415. [Cit. on pp. 49, 50.]

Reardon, S. F. and Firebaugh, G. (2002) "Measures of multigroup segregation," *Sociological Methodology* 32(1), pp. 33-67. [Cit. on p. 94.]

Rosenbaum, P. R. (2020) "Modern algorithms for matching in observational studies," *Annual Review of Statistics and Its Application* 7(1), pp. 143-176. [Cit. on p. 129.]

Rothbaum, J. and Hokayem, C. (2021) *How did the pandemic affect survey response: Using administrative data to evaluate nonresponse in the 2021 Current Population Survey Annual Social and Economic Supplement*, https://www.census.gov/newsroom/blogs/research-matters/2021/09/pandemic-affect-survey-response.html (visited on 03/16/2022). [Cit. on p. 10.]

Royston, P. and Lambert, P. C. (2011) *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, TX: Stata Press. [Cit. on p. 152.]

Rust, K. (1985) "Variance estimation for complex estimators in sample surveys," *Journal of Official Statistics* 1(4), pp. 381-397. [Cit. on p. 45.]

Rust, K. F. and Rao, J. N. K. (1996) "Variance estimation for complex surveys using replication techniques," *Statistical Methods in Medical Research* 5(3), pp. 283-310. [Cit. on pp. 45, 46, 85.]

Särndal, C.-E. (1978) "Design-based and model-based inference in survey sampling," *Scandinavian Journal of Statistics* 5(1), pp. 27-43. [Cit. on p. 41.]

– (1985) "How survey methodologists communicate," *Journal of Official Statistics* 1(1), pp. 49-63. [Cit. on p. 41.]

Shao, J. (1996) "Resampling methods in sample surveys," *Statistics* 27(3-4), pp. 203-237. [Cit. on p. 46.]

Shapiro, G. M. (2008a) "Sample design," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 776-777. [Cit. on p. 1.]

– (2008b) "Sample size," in *Encyclopedia of Survey Research Methods*, ed. by P. J. Lavrakas, Thousand Oaks, CA: Sage, pp. 781-783. [Cit. on p. 51.]

Somers, R. H. (1962) "A new asymmetric measure of association for ordinal variables," *American Sociological Review* 27(6), pp. 799-811. [Cit. on pp. 94, 98.]

StataCorp (2021a) *Stata 17 Base Reference Manual*. College Station, TX: Stata Press. [Cit. on p. 112.]

– (2021b) *Stata 17 Survey Data Reference Manual*. College Station, TX: Stata Press. [Cit. on p. 75.]

– (2021c) *Stata Statistical Software: Release 17*. College Station, TX: StataCorp LLC. [Cit. on p. 59.]

Statistics Canada (2020) *Guide to the Labour Force Survey 2020*, https://www150.statcan.gc.ca/n1/en/pub/71-543-g/71-543-g2020001-eng.pdf (visited on 07/18/2022). [Cit. on p. 51.]

Stuart, E. A. (2010) "Matching methods for causal inference: A review and a look forward," *Statistical Science* 25(1), pp. 1-21. [Cit. on p. 129.]

U.S. Bureau of Labor Statistics (2022) *Effects of COVID-19 pandemic and response on the Consumer Expenditure Surveys*, https://www.bls.gov/covid19/effects-of-covid-19-pandemic-and-response-on-the-consumer-expenditure-surveys.htm (visited on 03/16/2022). [Cit. on p. 10.]

U.S. Census Bureau (2022) *Statistical Quality Standards*, https://www2.census.gov/about/policies/quality/quality-standards.pdf (visited on 07/18/2022). [Cit. on p. 52.]

Valliant, R. and Dever, J. A. (2018) *Survey Weights: A Step-by-Step Guide to Calculation*. College Station, TX: Stata Press. [Cit. on pp. 26, 45, 47, 48.]

Valliant, R., Dever, J. A., and Kreuter, F. (2018) *Practical Tools for Designing and Weighting Survey Samples*. 2nd ed. Cham: Springer. [Cit. on pp. 6, 7, 25, 26, 51.]

Vermunt, J. K. and Moors, G. (2005) "Event history analysis," in *Encyclopedia of Statistics in Behavioral Science*, ed. by B. S. Everitt and D. C. Howell, Chichester: Wiley, pp. 568-575. [Cit. on p. 137.]

West, B. T., Berglund, P., and Heeringa, S. G. (2008) "A closer examination of subpopulation analysis of complex-sample survey data," *The Stata Journal* 8(4), pp. 520-531. [Cit. on pp. 84, 85.]

West, B. T., Sakshaug, J. W., and Aurelien, G. A. S. (2016) "How big of a problem is analytic error in secondary analyses of survey data," *PLoS One* 11(6), e0158120. [Cit. on p. 59.]

– (2018) "Accounting for complex sampling in survey estimation: A review of current software tools," *Journal of Official Statistics* 34(3), pp. 721-752. [Cit. on p. 59.]

Wolter, K. M. (2007) *Introduction to Variance Estimation*. 2nd ed. New York, NY: Springer. [Cit. on pp. 45-47, 50.]