UNIVERSITY OF SZEGED

Ph.D. Thesis

Surveying the human population: errors and their corrections

Blanka Szeitl
Supervisor: Tamás Rudas

Doctoral School of Mathematics and Computer Science
Bolyai Institute, University of Szeged

2024

Köszönetnyilvánítás

Köszönöm témavezetőmnek Rudas Tamásnak, hogy az elmúlt évek során folyamatosan támogatta a tanulmányaimat, kutatásaimat és ötleteimet. Köszönöm az eddigi közös munkát és azt a kutatói szemléletet, amit megtanulhattam tőle. Köszönöm Pap Gyulának, hogy fantáziát látott abban, hogy a survey statisztika témakörét a Matematika és Számítástudományok Doktori Iskolában kutassam. Sokat tanultam tőle. Köszönöm a Bolyai Intézetben megismert kollégák segítségét. Külön köszönöm Szűcs Gábor és Nagy-György Judit segítségét az oktatásban és kritikáit a kutatásommal kapcsolatban.

Köszönettel tartozom a Tárki Társadalomkutatási Intézetnek, ahol pályám elején statisztikusként elkezdtem dolgozni és ahol lehetőséget adtak arra, hogy a survey statisztika módszertani kérdéseivel mélyebben is elkezdhessek foglalkozni.

Köszönöm minden korábbi és jelenlegi kollégámnak, akik kérdéseikkel, kritikus meglátásaikkal, vagy éppen pozitív véleményükkel hozzátettek ahhoz, hogy a dolgozatban bemutatott eredmények fejlődjenek.

Köszönöm a családomnak és a barátaimnak a sok támogatást, érdeklődést, motiválást és szeretetet.

Abstract

Most statistical theory assumes that the underlying population is infinite. On the contrary, survey sampling theory is built on a foundation of sampling from a finite population. Additionally, in the case of human population samples, the finite population consists of individuals, involving human factors into the experiment. Human nature and behavior cause difficulties in implementing the mathematical theory of survey sampling and influence the quality of the data in various ways. The quality of the survey data has clearly eroded in the last decade, which motivates re-assessment of the current methodology. The thesis contributes to the existing literature by summarizing the mathematical foundation of human population samples with a focus on data quality. It proposes novel approaches to improve the quality of estimates on aspects (1) when sampled individuals decide not to be observed, which results in uncertainty about the sample composition, and (2) when they answer inconsistently, which leads to uncertainty in the measurement. Regarding sample composition, the thesis introduces a new sample allocation method that takes into account expected response rates (ERRs). For the evaluation of the new method, the main theoretical tool is asymptotic calculations using the δ-method. Within a stratified sample design, ERR allocation leads to lower variances than a traditional allocation method, not only when response rates are correctly specified but under a wide range of conditions. The new sampling method is illustrated with simulations using various combinations of specific population parameters. The thesis also presents a new perspective on the evaluation of survey quality through replication surveys, which is an alternative to the conventional assumption of underlying fixed true population values. Our finding is that the uncertainty of the measurement is highly relevant because respondents are generally inconsistent in their answers. The inconsistency of the answer is presented as an example of regression to the mean (RTM). Although RTM is well known for multivariate normal distributions, the results show that this phenomenon is also relevant in the case of ordinal-scale variables. The general conclusions indicate that the decline in the quality of the survey can be primarily attributed to the uncertainty in the measurement, rather than the uncertainty in the composition of the sample.

Contents

Chapter 1

Introduction

The thesis discusses human population surveys within the framework of finite population samples and presents approaches to improve the quality of estimates. It makes contributions to the fields of mathematical statistics and survey methodology by summarizing the mathematical foundations of human population surveys and proposing new methods for handling the two most pressing aspects: uncertainty in the sample composition and uncertainty in the measurement. The thesis presents theoretical approaches, an evaluation of a specific survey experiment, and simulations.

Survey statistics is an applied field of mathematical statistics, where data are obtained from questionnaire (survey) data collections. Within the field of survey statistics, human population surveys are data collections where the observed units are individuals. Sample surveys from human populations play a crucial role in providing a large portion of quantitative data about our economy and society. National statistical agencies frequently release estimates for various indicators, including unemployment rates, poverty rates, crop production, retail sales, and median family income. Although some statistics may be derived from complete enumerations (censuses), the majority are derived from samples taken from the relevant population [1]. In recent times, there has been a decrease in the precision of survey estimates. This is particularly evident in election forecasts, as they are one of the rare instances where the previously estimated population parameter becomes known. The prominent role of human population surveys and the quality concerns of data have made the statistical evaluation and development of survey-type data collections increasingly relevant. The thesis is a pioneering one as it aims to summarize the mathematical foundations of human population surveys and propose new approaches to improve the quality of estimates derived from human samples and responses. The thesis presents both theoretical and empirical findings of the author.

Chapter 2 presents the background of human population surveys. We introduce the history of the field and highlight how this method developed within mathematical statistics, as well as show the relevance and magnitude of sample surveys. This chapter outlines the current challenges of human population surveys and presents the main motivation to analyze uncertainties in sample composition and measurement.

In Chapter 3 the concept of human population surveys is presented. We introduce the mathematical foundations with the relevant sampling methods and we introduce how sample survey theory can be dealt with within general statistical

theory. Since in human population surveys, the observed units are individuals, the experiment results in errors and biases that are difficult to manage. Quality control in survey sampling, in general, is presented based on the Total Survey Error framework. Errors and biases are also categorized based on mathematical and non-mathematical factors.

Chapter 4 focuses on a new allocation mechanism to handle errors in sample composition. The mechanism takes into account the expected response rates (ERRs). The relative performance of the ERR allocation is assessed by comparing the variances in the resulting estimates. Asymptotic variances are calculated using the δ method and then initially compared by assuming correctly specified response rates. Here, the assumed response rates are subject to random fluctuations, which are then corrected using post-stratification. Variance comparison is made in terms of misspecified response rates and the results of an extensive evaluation using various combinations of specific population parameters are presented.

In Chapter 5 we introduce a new scheme that investigates nonresponse and measurement uncertainties through replication surveys. We present how the general concept of assessing errors and biases can be reconsidered when comparing results with previous surveys. Our approach defines uncertainties regarding sample composition and measurement and decomposes total differences in theory as well as based on a case study. The main results are that the total uncertainty can be decomposed exclusively into nonresponse and measurement uncertainties. Measurement uncertainty is more relevant than nonresponse uncertainty. The respondents are generally inconsistent in their responses at the individual level, which implies uncertainties in the measurements.

In Chapter 6 the results and the potential directions for future studies are summarized.

Chapter 2

Background

2.1 History and relevance

This chapter provides an overview of the most relevant milestones in the history of survey sampling and the debates that have shaped its position in mathematical statistics. We also present the current magnitude and relevance of human population surveys.

The aims of survey statistics are the same as those of classical statistics: to design experiments and analyze the data from them; to estimate unknown quantities from measurements; to test hypotheses; to model random processes and to predict future trends [1]. By definition, the term "survey" covers any activity that collects or acquires statistical data: included are censuses, sample surveys, the collection of data from administrative records, and derived statistical activities ([2], page 7). Complete enumerations (censuses) serve generally as reference points; sample surveys are most often the basis for survey statistics, where the units of observation are persons, households, or organizations. A sample survey is a survey carried out using a sampling method, that is, in which a portion only, not the entire population, is surveyed [3]. The theoretical basis of survey statistics is provided by theorems of mathematical statistics at the level of sampling, data collection, data adjustments, estimation, inference, and analysis. Survey statistics can therefore be considered an applied field of mathematical statistics [4].

Survey sampling theory is a major part of the development of statistics, although its use was only accepted in the 1920s [5]. This theory has sparked many debates in mathematical statistics and highlighted the importance of the evaluation of techniques. Sampling theory has been the subject of multiple approaches, including design-based, model-based, model-assisted, predictive, and Bayesian. A variety of authors have suggested a timeline of breakthroughs in survey theory that reflect the major debates within its evolution ([6, 7, 8, 9, 5]. We present five milestones in the history of survey theory based on the examples of Särndal ([10]) and Tillé ([5]).

The first milestone is the first sample-based experiment: in 1783, Pierre Simon de Laplace presented a method to calculate the population size from birth records based only on a random sample of regions. He proposed taking a sample of regions and calculating the ratio of inhabitants to births and then multiplying it by the total number of births. Laplace even suggested estimating the potential error by referencing the central limit theorem [5]. Even though no questionnaire (survey) was used in the experiment of Laplace, it is considered the first estimation method

based on a sample of the human population. In 1895, Anders Nicolai Kiær, the Director of the Central Statistical Office of Norway, conducted the first sample survey using questionnaires, which can be considered the second milestone. He chose a sample of cities and municipalities and then within each of these he selected certain individuals based on the first letter of their surnames. He employed a two-stage design, but the selection of the units was not random. Kiær advocated for the use of partial data if it was created using a "representative method" [5]. Kiær's notion of representativeness is associated with the quota method, which results in a regulated sample structure based on particular demographic criteria, but it does not qualify as a random sample. His address was followed by a passionate discussion and the start of a statistical discourse on survey techniques [5]. Thirty years after Kiær's survey, and as the result of a series of debates (with Cochran, Royall, Hajek, Neyman, and Godambe) the idea of sampling was officially accepted and the validity of random methods was demonstrated through a rigorous mathematical argument in the International Statistical Institute (ISI) Congress in Rome in 1925 (third milestone). This acceptance of the use of partial data, and especially the recommendation to use random designs, resulted in a rapid mathematization of the theory. Jerzy Neyman was instrumental in developing a large portion of the foundations of the probabilistic theory of sampling for simple, stratified, and cluster designs. He also determined the optimal allocation of a stratified design. His work can be considered the fourth milestone in the history of survey sampling, leading to its formation as a coherent mathematical theory. The professional discourse on how sample survey theory should be incorporated into general statistical theory began in the 1950s [10] and can be considered the fifth milestone. The discourse was then developed as follows: Let a finite population consist of N individuals which are labelled by integers $i = 1, \ldots, N$. Each individual i has a variate value x_i $(i = 1, \ldots, N)$ associated with it [11]. The variate values x_i $(i = 1, \ldots, N)$ are unknown. Hence, the estimation of the population mean is

$$\bar{x}_N = \sum_{1}^{N} x_i/N,$$

let s, represents a subset of n individuals selected through a random sampling procedure (e.g. simple random sampling) without replacement. The observed values of the variates, denoted as x_i, are recorded for each individual i within the sample s. The sample mean, is then calculated.

$$\bar{x}_s = \sum_{i \in s} x_i/n,$$

and this can be considered the unbiased estimator of the population mean. The main statement in this debate was that this is true if the individual labels i are "ignored" [11]. However, in order to perform statistical random sampling, it is necessary to assign labels to the individuals in the population. This requirement of having individual labels means that unbiased estimation is generally not possible, especially when dealing with a survey population that consists of a fixed number of individuals [11].

Based on this contradiction two positions have emerged: (1) general statistical theory should be extended with survey sampling theory with a new model and corresponding formal criteria, or (2) survey sampling theory should be interpreted in

a specific way with which it fits into general statistical theory [10]. This debate continued throughout the '70s and '80s. Nowadays, survey sampling is mainly dealt along the lines of (2), with no formal guidelines on the subject.

Survey statistics are widely used in the social, political, economic, and medical sciences, and the results are usually put into practice. Only a small number of decisions in these fields are based on complete enumerations or administrative data. The majority of these decisions are based on the outcomes of sample surveys conducted at the national, European, and global levels. Survey research is produced by the National Statistical Institutes, academic research networks, and political or market actors. The Eurostat organization brings together the National Statistical Institutes of the European Union. It is responsible for coordinating statistical activities across the Union. According to Eurostat, the National Statistical Institutes conduct an average of 200 surveys annually at the national level [12]. The manifold applications of survey data at the European level are evident: Eurostat, the OECD, and the World Bank all collect, organize, and publish survey data which serves as the foundation for the most relevant international economic and social indicators (e.g. GDP, GINI) in Europe.

In Hungary, the entire population of roughly 9 million people is only surveyed in the form of a census every 10 years. In addition, however, it conducts on average 150 independent survey data collections with a total of 1.2 million questionnaires per year [13]. On average, 60% of these surveys are annual or less frequent, while the rest are more frequent (biannual or monthly; [13]). Academic research networks such as European Social Survey (ESS), International Social Survey Program (ISSP), European Value Survey (EVS), World Value Survey (WVS), and Survey on Income and Living Conditions (EU-SILC) provide the basis for numerous political and social measures, and thus for international comparisons. Another relevant segment of sample surveys is research on market and public opinion. Although most of these are conducted at the national level, there are also many examples of international or cross-country data collection. ESOMAR, the European Society for Opinion and Marketing Research, provides regular updates on the market and the opinion research industry, both in Europe and around the world [14]. The ESOMAR report for 2020 has concluded that, despite the global economy's 3.3% decrease in real GDP, the market research and public opinion polling industry was able to sustain its prior output, estimated to be worth nearly $90 billion worldwide [14]. Market research is used to guide business decisions that can have an impact on quality of life, while opinion polls are used mainly in politics. Political campaigns and in some cases policy makers rely on the results of public opinion polls. Although there is a lack of comprehensive data regarding the exact size of the political polling industry, it is evident that this sector is expanding. While in 2000, 29 polling agencies were active in the US, this number increased to 69 by 2022 [15]. Another example is that the number of British polls published exploded from appr. 150 polls in 2010 to appr. 500 polls in 2015 [16].

The foundations of survey statistics are based on mathematical statistics, which is shaped by influential figures in the field of mathematical statistics, who emphasize its significance and position within the discipline. Human population surveys are not only relevant within the field of mathematical statistics: they are carried out in

huge quantities and have a crucial role in informing decisions that affect economic and political change and our daily lives. Thus, evaluations and a revision of the methodology are crucial.

2.2 Current challenges

This chapter aims to outline the existing challenges related to the quality of human population surveys and highlights the motivation behind the particular focus of this thesis.

In the context of questionnaire surveys, individuals, or households are observed, which introduce human elements into the experiment. This means that the outcome of the experiment is not solely determined by the sampling procedure, but also influenced by factors such as whether the selected individuals are actually contacted, their willingness to respond to the survey, and to provide certain information, as well as the accuracy of the information they provide. These factors introduce both random errors and specific biases in the estimation of population parameters, and the combination of these errors and biases has an impact on the precision of the collected data.

In recent times, there has been a decrease in the accuracy of survey estimates. This is especially clear in election forecasts, as they are one of the few cases where the population parameter that was previously estimated becomes known. There are reputable cases of failed election forecasts wordwide, just to highlight a few most recent examples: the 2002 Hungarian parliamentary election, the "Brexit" referendum in 2016 in Great Britain, and presidental elections in 2016 and 2020 in the United States (US). In the US, the American Association of Public Opinion Research (AAPOR) regularly evaluates opinion polls in each presidential election year and maintains a task force to examine the performance of opinion polls. The most recent assessment found that the 2020 polls featured polling error of an unusual magnitude: It was the highest in 40 years for the national popular vote (Figure 2.1; [17]).

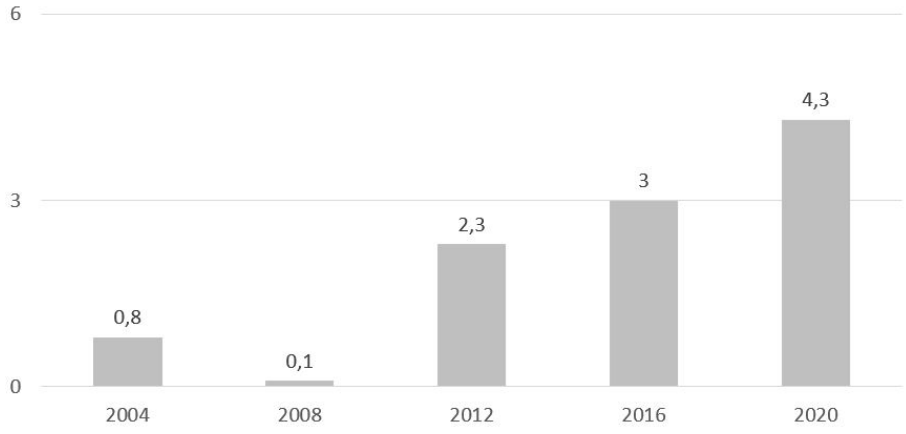


Figure 2.1: Polling error over time in the US National Presidential Polls

Source: Own editing based on AAPOR 2020 report pp.15.

Regarding the factors that can explain the polling error, AAPOR concluded that at least some of the polling error in 2020 was linked to discrepancies in sample composition. This was caused by unit-nonresponse [17].

Unit-nonresponse refers to the discrepancy between the sample and the set of respondents, which is successfully observed [18]. It arises when individuals either cannot be contacted or choose not to participate in the survey, thus it can be related to the response rate. There is a consensus among survey researchers ([19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 19]) that response rates are declining (i.e. nonresponse rates are increasing). Based on surveys conducted by the US Census Bureau, the extent of this decline is of a significant magnitude: the initial response rate during the 1990s was at 85%, however, by 2015 the nonresponse rate had reached nearly 35% [29].

In the ESS data collection a similar pattern can be observed. Figures 2.2 and 2.3 represent the response rate trends of nearly 40 European countries participating in the ESS survey. The countries are divided into two groups: Figure 2.2 includes countries with an average initial response rate in the first round (2002), and Figure 2.3 includes countries with an initial response rate higher than the average in the first round (2002)[1]. It is evident that a significant downward trend can be observed for most countries.

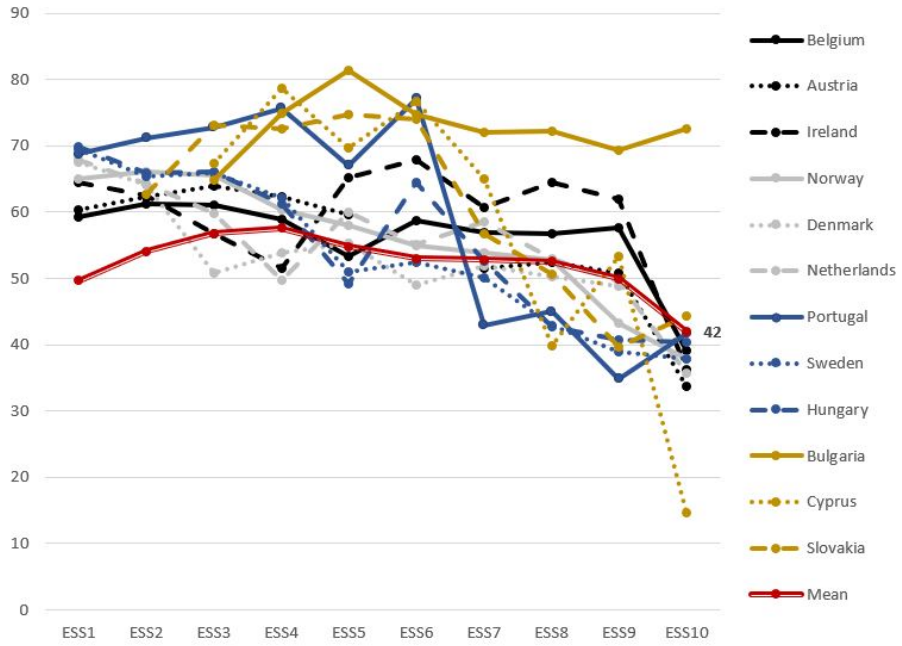


Figure 2.2: Trends in response rates in the ESS between 2002-2020 in the group of countries with average initial response rate in ESS1.

Source: The data presented in the figure represents a collection of technical details about the ESS from various countries that was gathered by the author.

[1]Data from countries with an initial response rate lower than the average are not presented, because those countries had problems with their data collection process and typically also missed some rounds later.

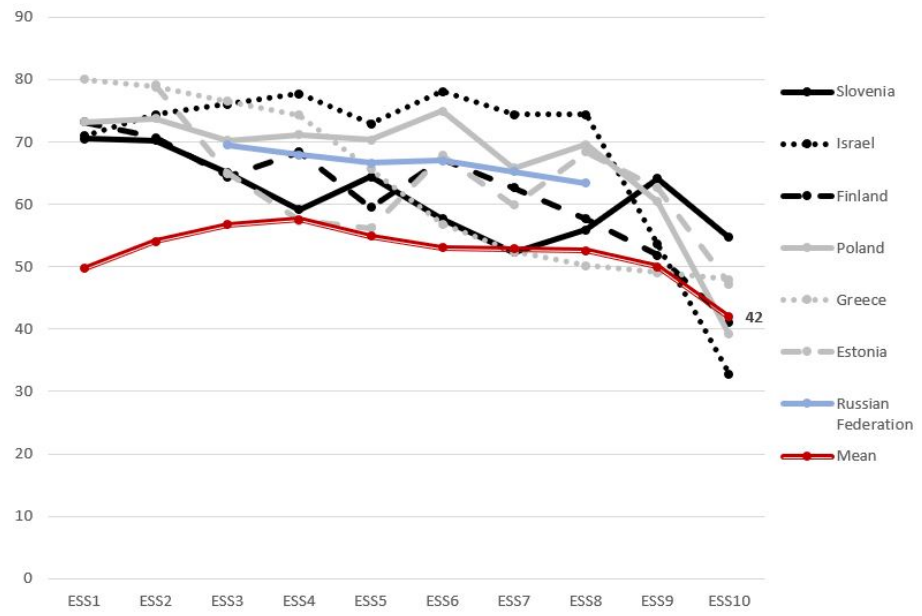


Figure 2.3: Trends in response rates in the ESS between 2002-2020 in the group of countries with a higher than average initial response rate in ESS1.

Source: The data presented in the figure represents a collection of technical details about the ESS from various countries that was gathered by the author.

Looking specifically at Hungary, nonresponse rates have also been increasing over the past 20 years: in 1996, 16% of the selected households refused to answer an optional survey block during the microcensus [31]; by 2016, another survey conducted by the Hungarian Central Statistical Office had a response rate of only 59% [32]. In the most recent methodological study in 2019 a response rate of 40% has been observed [33].

Unit-nonresponse is not random and appears to be linked to certain sociodemographic characteristics of sample elements. Research has indicated that labor-force participation, life stage, socioeconomic status, health, and gender may all play a role [34, 35]. Goyder [36] found that it is more difficult to contact individuals who are single, employed, living in an apartment in a large city, or of a higher socioeconomic status, while it is easier to contact the elderly or larger families. This is corroborated by Campanelli and colleagues [37]. Groves [18] suggests that families with young children and elderly members are more likely to be present in the home, making them easier to contact. Additionally, the size of the household may also be a factor in contactability; the more people in the household, the more likely it is that someone will be available to answer the phone or the door [38]. Additionally, Tucker and Lepkowski [39] observed that the rise in female labor-market involvement could have a negative impact on the ability to make contact [40].

Unit-nonresponse therefore affects the quality of the estimates through the composition of the sample. In addition to this source of discrepancies, there is another important factor which can be linked to the inaccuracy of estimates: the validity of the answers to a given question.

Measurement error is the discrepancy between the ideal measurement and the actual responses obtained [18]. Measurement errors are even present in censuses, where theoretically the entire population is measured. It encompasses various factors, in-

cluding interviewer effects, systematic errors, and random errors [41]. Measurement error occurs when the recorded or observed value deviates from the true value of the variable [42]. Several factors contribute to this disparity, such as the unclear or misleading phrasing of the questions and the context of the preceding questions [43]. Other important factors include changes in the mental state of the respondent, inconsistencies in their answers, social desirability, and the concealment of the true answer[2]. For example, respondents may provide an answer that aligns with the perceived norm rather than their actual opinion due to socially desirable behavior or yea-saying [44].

A simple model indicating the relationship between observed and latent variables of interest can be found in Figure 2.4. We investigate the relationship between two latent variables (two opinions in the population), denoted with f_1 and f_2 in Figure 2.4. In the case of human population surveys, we estimate the relationship based on answers to survey questions, denoted by y_1 and y_2 in Figure 2.4. The relationship between f_1 and y_1 and between f_2 and y_2 will not be perfect due to measurement errors (e_1 and e_2). The standardized effect of the variable of interest f_i on y_i is called the quality coefficient (q_i). The two correlations will be identical only when both measures have a perfect quality (equal to 1), meaning that there are no measurement errors. Unfortunately, this scenario is highly unlikely to happen [45].

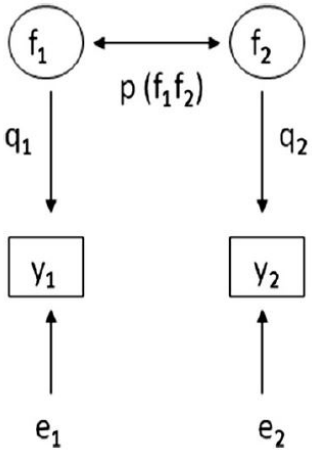


Figure 2.4: The relationship between observed and latent variables of interest

Source: Saris & Revilla (2016): Correction for Measurement Errors in Survey Research: Necessary and Possible pp. 1008.

Alwin ([42]) suggested that around 50% of the variance in the observed survey variables may stem from inaccuracies. This underscores a notable discrepancy between the intended variable under study and the variable actually reflected in the survey question [45].

In summary, the precision of survey estimates is affected by sample composition and measurement.

[2]In the special case of political opinion polling these phenomena are referred to as: "late swing", when voters change their minds between when they were polled and the time they cast their vote; "shy voters" are people not telling the truth; "social desirability" is when supporters of controversial parties or opinions are most likely to conceal their views [16]. But the special case of political opinion polling is not discussed further in the thesis.

As the response rate decreases, conducting survey research becomes more difficult, requiring a larger number of individuals to be contacted to obtain the desired sample size. Consequently, data collection becomes slower and more expensive. The observed sociodemographic patterns in nonresponses introduce bias into the estimates, undermine the implementation of the initial sampling design, and raise concerns about the reliability of the resulting sample.

Measurement error is another type of difficulty, which is present even in the case of complete enumerations. If we do not consider measurement errors caused by human nature and behavior, our findings will be flawed even if we are precise in the sampling stage and obtain a random and representative sample of the population.

This thesis addresses the phenomenon of nonresponse by presenting a new procedure to handle nonresponse during the sampling stage: Allocation of samples that consider expected response rates (ERR). This is discussed in Chapter 4.

Regarding measurement error, we demonstrate how replication surveys can be employed to evaluate measurement errors without relying on the notion of true population parameters and ideal measurements. Additionally, we investigate the presence of the Regression to the Mean phenomenon in the process of answering questions. The new method of replication surveys and the findings on measurement error are presented in Chapter 5. Measurement error is also affected by cultural factors, which makes it particularly important for international comparisons. For the same reason, there is an urgent need to investigate the effects on the response process at the national level. In this respect, the thesis represents added value in terms of understanding the trends observed particularly in Hungary.

Chapter 3

Surveying the human population

Survey sampling theory and the classical sampling theory of mathematical statistics have two major differences. First, unlike the classical sampling theory, the populations in survey sampling are finite. Secondly, the elements being observed are individuals who, by their actions and personal decisions, impact the data quality by hindering the accurate completion of the procedural steps. This chapter introduces the theory of surveying the human population focusing on these two differences.

Chapter 3.1 presents the mathematical foundations of human population samples in terms of sampling and inference. This chapter covers the theoretical basics and shows how survey sampling and inference are carried out in theory.

Chapter 3.2 delves into factors that can influence survey research results and distinguishes mathematical and non-mathematical factors in the total survey error framework. This chapter places particular emphasis on nonresponse and measurement inaccuracies, which are the main focus of the thesis findings.

3.1 Mathematical foundations

We start by describing the general framework for sampling from a finite population with the relevant notation and definitions (Chapter 3.1.1). Chapter 3.1.2 then introduces the different sampling methods in human population surveys reflecting on practical applications of international surveys in Europe. In Chapter 3.1.3 the main theories of finite population inference are presented, covering the concept of superpopulation and the central limit theorem.

3.1.1 Sampling from finite populations

In the general framework of finite population samples, consider a finite set of elements identified by integers $U = \{1, 2, ..., N\}$. This set of identifiers is referred to as the population list. The set U is discrete and consists of a finite number of elements. Each element j_{th} in the list is associated with a vector of characteristics denoted by y_j. In survey sampling applications, y_j is assumed to be real valued. The entire set of N vectors is denoted by $\mathcal{F}$ and is referred to as a finite population or a finite universe. A sample is a subset of elements. In statistical sampling, the goal is to select samples using probability rules in order to establish the probability characteristics of the set of samples defined by the selection rules. The terms "random samples" and "probability samples" are used to refer to samples selected using prob-

ability rules. In this case, the statistic used to derive an estimate from the sample is predetermined, and the method ensures that only one estimate is obtained from a specific sample using that statistic [46]. Let $\mathcal{A}$ denote the set of possible samples under a specific probability procedure, and let A represent the set of indices of U that are included in the sample. In the following, the population from which the samples are selected is assumed to be fixed. Let $P[A = s]$ represent the probability of selecting s (where $s \in A$). A sampling design is a function $p(\cdot)$ that maps s to the interval $[0, 1]$, such that $p(s) = P[A = s]$ for any $s \in \mathcal{A}$. A set of samples of primary importance is the set of all possible samples that contain a fixed number of distinct units. Denote the fixed size by n. A probability sampling method for fixed-size samples of n assigns a probability to each possible sample. The inclusion probability for element k is the sum of the sample probabilities for all samples that contain element k,

$$\pi_k = P(k \in A) = \sum_{s \in A_{(k)}} p(s),$$

where $A_{(k)}$ is the set of samples that contain element k.

Indicator variables are defined in the context of probability sampling schemes to determine which elements are part of the sample. To denote the indicator variable for element k, we use the notation I_k. Thus,

$$\begin{aligned} I_k &= 1 \text{ if element } k \text{ is in the sample} \\ &= 0 \text{ otherwise.} \end{aligned} \tag{3.1}$$

Consider the vector $\boldsymbol{d} = (I_1, I_2, ..., I_N)$, which represents a set of random variables. The probability distribution of $\boldsymbol{d}$ plays a crucial role in determining the probabilistic characteristics of functions derived from the sample. The sampling design defines the probability structure of $\boldsymbol{d}$, with the inclusion probability of element k being equal to the expected value of I_k.

$$\pi_k = E\{I_k\}.$$

The joint inclusion probability, denoted by π_{kl}, for elements k and l ($k \neq l$) is the sum of sample probabilities for all samples that contain both elements k and l. In terms of the indicator variables, the joint inclusion probability for elements k and l is

$$\pi_{kl} = E\{I_k I_l\}.$$

The number of units in a particular sample is

$$n = \sum_{k=1}^{N} I_k$$

and because each I_k is a random variable with expected value π_k, the expected sample size is

$$E\{n\} = \sum_{k=1}^{N} E\{I_k\} = \sum_{k=1}^{N} \pi_k$$

The variance of the sample size is

$$V\{n\} = \left\{\sum_{k=1}^{N} I_k\right\} = \sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl} - \pi_k\pi_l)$$
$$= \sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl} - \left(\sum_{k=1}^{N}\pi_k\right)^2,$$

In the equation, π_{kk} is equal to π_k. When the variance $V\{n\}$ is equal to zero, it indicates that the design is a fixed sample size or fixed-size design [1].

3.1.2 Sampling methods in surveys

This chapter delves into the sampling methods applicable in human population surveys. It begins by discussing direct element sampling approaches, followed by an overview of more complex methods. According to Särndal [10], two fundamental features of direct element sampling are highlighted: (1) the requirement of a sampling frame that lists all population elements and (2) the sampling units being the population elements directly.

The subsequent direct element sampling strategies are applicable in surveys of human populations: Bernoulli sampling (Chapter 3.1.2.1), simple random sampling (Chapter 3.1.2.2), systematic sampling (Chapter 3.1.2.3), Poisson sampling (Chapter 3.1.2.4), and probability proportional to size sampling (Chapter 3.1.2.5). In contrast to direct element designs, complex sampling schemes often involve the selection of population elements in multiple steps, such as choosing other units like a geographic unit. Illustrations of these are outlined in this chapter as stratified sampling (Chapter 3.1.2.6), and multistage sampling (Chapter 3.1.2.7).

Since human population surveys are an applied field of mathematical statistics, the practical feasibility of a sampling design is a key issue. Bernoulli and Poisson sampling techniques lead to random fluctuations in sample sizes and are not commonly used in survey applications. Despite this, they are valuable for illustrating basic principles in survey sampling and can be used as frameworks for comprehending nonresponse mechanisms, which will be later addressed in the survey quality context. In contrast, systematic sampling, proportional probability sampling by size, stratified sampling, and multistage sampling are widely used in survey practice [10]. When applicable, examples are used to illustrate these sampling methods.

3.1.2.1 Bernoulli sampling

Bernoulli sampling is considered the simplest design [10]. In this method, the sample membership indicators $I_1, \ldots, I_N$ are random variables that are independent and identically distributed. Assuming π is a constant with $0 < \pi < 1$, each I_k follows a Bernoulli distribution with the same parameters [10].

$$P(I_k = 1) = \pi; \quad P(I_k = 0) = 1 - \pi$$

If n_s represents the (random) sample size, the sampling scheme is described by

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s} \tag{3.2}$$

The inclusion probabilities are $\pi_k = \pi$ for all k and $\pi_{kl} = \pi^2$ for all $k \neq l$.
In order to choose a Bernoulli sample, the straightforward list-sequential method[1] can be employed. In Bernoulli sampling, the size of the sample, n_s, is a random variable that follows a binomial distribution, with its mean and variance specified as

$$E(n_s) = N\pi; \quad V(n_s) = N\pi(1-\pi)$$

The variance of the sample size n_s, can be assessed by an interval [10]. Using the normal distribution to approximate the binomial, we have n_s obtained within the limits

$$N\pi \pm z_{1-(\alpha/2)}[N\pi(1-\pi)]^{1/2}$$

with a probability of roughly $1-\alpha$, where constant $z_{1-(\alpha/2)}$ is exceeded with probability $\alpha/2$ by the unit normal random variable. For example, in the case of a survey of the Hungarian adult population, where let $N = 8,000,000$ and $\pi = 0.0005$, the 99% interval is

$$4000 \pm 2.58(3,998)^{1/2} = 4,000 \pm 163$$

so the probable variation in this case is roughtly 4% of the expected sample size. In surveys, a limitation of Bernoulli sampling may arise from the inability to determine in advance the precise size of the chosen sample. Moreover, the variability in sample sizes tends to increase the variance of the π estimators [10]. In the Bernoulli design, the π estimator for the population total $t = \sum_U y_k$ of a given characteristics y (as defined in Chapter 3.1.1) is expressed as

$$\hat{t}_\pi = \frac{1}{\pi}\sum_s y_k \tag{3.3}$$

The unbiased variance estimator is given by

$$\widehat{V}(\hat{t}_\pi) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_s y_k^2 \tag{3.4}$$

3.1.2.2 Simple random sampling

The Bernoulli sampling belongs to the category of designs that may be termed equal probability sampling designs. The common feature of such designs is that the first-order inclusion probabilities are all equal, that is, $\pi_k = \text{constant}$ $(k = 1, \ldots, N)$. We now consider an additional equal probability sampling design, the simple random sampling. This sampling design is often taken as a point of reference when discussing

[1] A sequential list scheme entails carrying out a series of random experiments by progressing through the list of elements, without the requirement to reach the end, and conducting an experiment for each element that results in either selecting or not selecting the element [10].

alternative designs [1]. Under the simple random sampling design every sample s of the fixed size n recieves the same probability of being selected. That is

$$p(s) = \begin{cases} 1/\binom{N}{n}, & \text{if } s \text{ is of size } n \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

The inclusion probabilities are

$$\pi_k = \frac{n}{N} \quad k = 1, \ldots, N$$

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)} \quad k \neq l = 1, \ldots, N$$

We call $f = n/N$ the sampling fraction. The simple random sampling design may be carried out with the following list-sequential scheme [10].
Let $\epsilon_1, \epsilon_2, \ldots,$ be independent random numbers drawn from the Unif$(0, 1)$ distribution. If $\epsilon_1 < f$, the element $k = 1$ is selected, otherwise not. For subsequent elements, $k = 2, 3, \ldots,$ let n_k be the number of elements selected among the first $k - 1$ elements in the population list. If

$$\epsilon_k < \frac{n - n_k}{N - k + 1}$$

the element k is selected, otherwise not. The procedure terminates when $n_k = n$. Population lists (registres) are usually only accessible to official bodies. These bodies, in turn, select individuals for a number of data collection purposes. When selecting respondents, it is important to minimize overlap in the sample of different data collections. Non-overlapping is preferred when multiple surveys of the same population need to be conducted quickly. Having little to no overlap helps reduce the burden on respondents and may lead to higher response rates by avoiding excessive approaches. Samples drawn without overlap are called negatively coordinated [10].
The following implementation of the simple random sampling design has the advantage that it permits the simultaneous selection of several non-overlapping simple random samples [10]. This scheme proposes N independent Unif$(0, 1)$ random numbers $\epsilon_1, \ldots, \epsilon_k, \ldots, \epsilon_N$ are first drawn, where ϵ_k is tied to the kth element. These numbers are then ordered according to size;

$$\epsilon_{k_1} < \epsilon_{k_2} < \cdots < \epsilon_{k_N}$$

The notation indicates that the ith smallest among the N ϵ-values corresponds to the element $k_i, i = 1, \ldots, N$. Subsequently, the n smallest ϵ-values $\{k_{n+1}, \ldots, k_{2n}\}$ form a second independent random sample, distinct from the first one, and so forth. Any set of n predetermined positions in the ordered sequence defines an independent random sample. More details on the technique of negatively coordinated samples can be found in [10, 47, 48].
Under the simple random sample the π estimator of the population total $t = \sum_U y_k$ can be written

$$\hat{t}_\pi = N\bar{y}_s = \frac{1}{f}\sum_s y_k \tag{3.6}$$

where $f = \frac{n}{N}$ is the sampling fraction.

Although simple random sampling follows simple logic at a theoretical level, it is very difficult to implement in practice.

One contributing factor to the difficulty lies in acquiring population lists: not all countries possess a comprehensive population register, and even if one exists, individual lists may be limited or unavailable. In Hungary, the Ministry of Interior manages the population register, granting access to only a basic random sample of individuals. Additionally, when populations are geographically spread out, a simple random sample will also be geographically spread out, resulting in increased time and costs for data collection (due to the need for more extensive travel to reach each respondent). According to the 2011 census, Hungary comprises 3155 settlements with an average adult population of 2800 individuals. Selecting a sample of 4000 individuals would involve individuals from 7-800 settlements, making data collection significantly challenging. Therefore, general human population surveys with simple random sampling are rarely utilized. Instead, simple random sampling methods are typically integrated into multistage sampling procedures, which will be discussed in Chapter 3.1.2.6.

3.1.2.3 Systematic sampling

In contrast to simple random sampling, systematic sampling in surveys of human populations involves a series of procedures that offer various practical benefits, particularly due to its straightforward implementation [10]. Our focus here is on the basic form of systematic sampling. The initial element is selected randomly and with equal probability from the first a elements on the population list. The value of the positive integer a is predetermined and is known as the sampling interval. Subsequent selections do not require additional random draws. The remaining sample is determined by systematically selecting every ath element thereafter until the end of the list. Consequently, there are only a possible samples, each with an equal probability of selection, namely q/a. The simplicity of a single random draw is a significant advantage. For example, it is straightforward for an interviewer to choose a systematic sample while in the field, such as selecting respondents from a list of addresses or choosing respondents among customers or public transportation users.

For the definition of systematic sampling, let a be the fixed sampling interval and let n be the integer part of N/a, where N is the population size. Then the

$$N = na + c$$

where the integer c satisfies $0 \leq c < a$. if $c = 0$, a sample size n will be drawn by the procedure that we now present. If $c < 0$, the sample size is going to be either n or $n + 1$.
The selection, which can be seen as a list sequential, is as follows:

(i) Select with equal probability $1/a$ a random integer, say r, between 1 and a (inclusively).

(ii) The selected sample is composed as

$$s = \{k : k = r + (j-1)a \leq N; j = 1, 2, \ldots, n_s\} = s_r \tag{3.7}$$

where the sample size n_s is either $n+1$ (when $r \geq c$) or n (when $c < r, \leq a$). the integer r is called the random start.

Under the systematic sampling design, with the sampling interval a, the π estimator of the population total $t = \sum_U y_k$ is

$$\hat{t}_\pi = at_s$$

where $t_s = \sum_s y_k$ is the sample total of y, and s is a member of the set of possible samples $\{s_1, \ldots, s_2, \ldots, s_a\}$, s_r being defined 3.7.

3.1.2.4 Poisson sampling

Bernoulli sampling, simple random sampling, and systematic sampling are designs with equal probabilities, meaning that all π_k are the same in these designs. Having equal probabilities results in straightforward estimators, but this is not a common feature in survey sampling. In practice, most designs involve unequal probabilities, as they tend to be more effective [10].

An example of an unequal probability sampling design is Poisson sampling, which is an extension of Bernoulli sampling [49]. In Poisson sampling, a specific positive inclusion probability π_k is assigned to each element, where $k = 1, \ldots, N$. This design is characterized by the independence of the sample membership indicators I_k, with I_k following a distribution as described below.

$$P(I_k = 1) = \pi_k, \quad P(I_k = 0) = 1 - \pi_k$$

$k = 1, \ldots, N$. The Poisson sampling design is such that the sample s has the probability

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U - s} (1 - \pi_k) \tag{3.8}$$

where $s \in \mathcal{l}$, the set of all 2^N subsets of U. Due to independence, $\pi_{kl} = \pi_k \pi_l$ for any $k \neq l$. Because the π_k can be specified in a variety of ways, Poisson sampling corresponds to a whole class of designs [10].

Given a set of inclusion probabilities $\pi_1, \ldots, \pi_N$, the Poisson design has a simple list-sequential implementation [1]. Let $\epsilon_1, \ldots, \epsilon_N$ be independent random numbers drawn from the Unif$(0, 1)$ distribution. If $\epsilon_k < \pi_k$, the element k is selected, otherwise not; $k =, 1, \ldots, N$.

In Poisson sampling, the sample size n_s is random, with mean

$$E(n_s) = \sum_U \pi_k \tag{3.9}$$

and variance

$$V(n_s) = \sum_U \pi_k (1 - \pi_k) \tag{3.10}$$

Under the Poisson sampling design, the π estimator of the population total $t = \sum_U y_k$ is given by

$$\hat{t}_n = \sum_s \bar{y}_k = \sum_s y_k / \pi_k \tag{3.11}$$

Since $n = \sum_U \pi_k$, we get

$$\pi_k = ny_k / \sum_U y_k \tag{3.12}$$

$k = 1, \ldots, N$, assuming also that $y_k \leq \sum_U y_k/n$ for al k. The values of y_k are currently unknown; however, in certain surveys, we may have information on one or more additional variables, which are variables with known values for the entire population. Let us assume that $x_1, \ldots, x_N$ represents known positive values of an auxiliary variable x. It is also plausible to hypothesize that y is roughly proportional to x. In such instances, we can consider π_k to be proportional to the known x_k. Specifically, for $k = 1, \ldots, N$, assuming $x_k \leq \sum_U x_k/n$ for all k. If $x_k > \sum_U x_k/n$, we should set $\pi_k = 1$. When the ratio y_k/x_k remains relatively constant, the resulting estimator π will exhibit low variance. The inclusion probabilities determined by 3.12 are called proportional to size. The value x_k is viewed as a measure of size for the kth element. Common size measures include total assets or the number of employees for a population of business firms, total territory for a population of farms, etc. [10].

Poisson sampling has the same drawback as Bernoulli sampling, that is, a random sample size [1].

3.1.2.5 Probability proportional-to-size sampling

The examination of Poisson sampling demonstrated that when the variable under study y is closely related to a known auxiliary variable x that is positive, there is a valid reason to opt to select elements with a probability that is proportional to x. For example, x could serve as a rough cost-effective indicator of an initial assessment of the variable being studied [10]. Sampling based on probability proportional to size is beneficial not only for Poisson sampling, but also for various other sampling methodologies [1]. Consider the π estimator in a fixed-size without-replacement design

$$\hat{t}_\pi = \sum_s y_k/\pi_k \tag{3.13}$$

Assume that it is feasible to create a fixed-size design without replacement and a sampling method to execute this design in such a way that

$$y_k/\pi_k = c, \quad k = 1, \ldots, N \tag{3.14}$$

where c is a constant. Then, for any sample s, we would have the following.

$$\hat{t}_\pi = nc$$

where n represents a fixed size of s. As $\hat{t}_\pi$ does not exhibit variability, its variance would be zero. It is evident that a design (along with a corresponding sampling scheme) that satisfies 3.14 cannot be identified as it requires prior knowledge of all y_k. However, if the auxiliary variable x is understood to be roughly proportional to y, then selecting π_k in proportion to the known value x_k will result in approximately constant ratios y_k/π_k. Consequently, the variance of the estimator π will be minimized [10].

3.1.2.6 Stratified samples

In stratified sampling, the population is divided into distinct subgroups known as strata, with a probability sample taken from each stratum independently. This method is highly effective and versatile, making it a commonly utilized approach. According to Särndal [10], three key factors contribute to the popularity of stratified sampling.

(i) If estimates are needed for particular subgroups (geographic units, demographic groups), each subgroup can be considered as a distinct stratum. If the membership of the subgroup is defined in the framework, a suitable probability sample can be chosen from each stratum.

(ii) In a survey, some subgroups may experience higher rates of nonresponse and measurement issues compared to others. The amount of additional information available can also vary greatly. These factors indicate that it may be beneficial to tailor the sampling method to each subgroup to improve the accuracy of the estimation. Therefore, it may be prudent to treat each subgroup as a distinct stratum.

(iii) Due to administrative purposes, the survey organization might have segmented its area into multiple geographic districts, each with a field office. In this scenario, it is common practice to consider each district as a stratum.

The objective is to select an efficient, yet practical, stratified sample. For this, we need to define the stratification variable, which is the characteristics used to subdivide the population into strata (this can be the type of settlement, region, county, or demographic variables such as gender, age, level of education), and the number of strata, which depends on the categories of the stratification variable and n_s. In stratified sampling design, the sampling design and sample size must be specified in each stratum (often the same type of sampling design is applied in all strata. An estimator must be specified for each stratum. Often, this choice is also made uniformly for all strata [10]).
By stratification of a finite population $U = \{1, \ldots, k, \ldots, N\}$ we mean a partition of U into H subpopulations, called strata and denoted $U_1, \ldots, U_h, \ldots, U_H$, where $U_h = \{k : k\}$ belongs to the stratum h [1]. By stratified sampling, we mean that a probability sample s_h is selected from U_h according to the design $p_h(\cdot)$ $(h = 1, \ldots, H)$ and that the selection in one stratum is independent of the selections in all other strata. The resulting total sample, denoted a, will therefore be composed as

$$s = s_1 \cup s_2 \cup s_3 \cup \cdots \cup s_H$$

and, because of the independence,

$$p(s) = p_1(s_1)p_2(s_2)\cdots p_H(s_H)$$

The number of elements within the stratum h, referred to as the size of the stratum h, is represented by N_h, which is considered to be a known value. As the stratum is a subset of U,

$$N = \sum_{h=1}^{H} N_h.$$

The population total can be decomposed as

$$t = \sum_{U} y_k = \sum_{h=1}^{H} t_h = \sum_{h=1}^{H} N_h \bar{y}_{U_h} \tag{3.15}$$

where $t_h = \sum_{U_h} y_k$ represents the total within the stratum, and $\bar{y}_{U_h}$ is the mean within the stratum.
In stratified sampling, the π estimator of the population total $t = \sum_U y_k$ can be written as

$$\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi} \tag{3.16}$$

where $\hat{t}_{h\pi}$ is the π estimator of $t_h = \sum_{U_h} y_k$.
If simple random sampling is applied in all strata, with a fixed sample size of n_h in stratum h, as a consequence of 3.15 and 3.5, we then have the π estimator of the population total $t = \sum_U y_k$ is

$$\hat{t}_\pi = \sum_{h=1}^{H} N_h \bar{y}_{s_h} \tag{3.17}$$

where $\bar{y}_{s_h} = \sum_{s_h} y_k / n_h$.

3.1.2.7 Multistage sampling

The sampling designs outlined in Chapter 3.1.2.1-3.1.2.6 are based on the assumption that direct sampling of elements is feasible, meaning that population elements can serve as sampling units in a single sampling stage. However, in numerous surveys, direct element sampling is not employed mainly due to the following reasons [10, 50]:

1. There is no sampling frame available that can identify every single element of the population, and the cost of creating such a frame is too high.

2. The population elements are distributed across a large area, making direct element sampling lead to a widely dispersed sample as mentioned in the case of simple random sampling in Chapter 3.1.2.2. This would make the cost of fieldwork unaffordable due to the significant travel expenses associated with personal interviews. Moreover, effective supervision of fieldwork might be challenging, potentially leading to high nonresponse rates and significant measurement inaccuracies.

A variety of sampling designs are available for surveys in which the elements sampled are not exclusively individuals. These range from cluster sampling to highly complex multistage sampling designs that use unequal probability sampling at various stages of selection. The simplest version of multistage sampling is two-stage sampling. In the case of a two-stage sample design, the sample of elements is obtained as a result of two stages of sampling:

1. The population units are initially classified into separate subgroups known as primary sampling units (PSUs). A random sample of PSUs is selected, marking the first sampling stage.

2. The second-phase sampling units (SSUs) can be individuals or clusters (groups of individuals). A probability-based sample of SSUs is chosen from each PSU in the first-stage sample, which forms the second-stage sampling procedure. When SSUs are clusters, all individuals within the selected SSUs are surveyed.

Multistage sampling often consists of three or more stages of sampling. There is a hierarchy of sampling units: primary sampling units, secondary sampling units within PSUs, tertiary sampling units with SSUs, etc. The sampling units in the last stage sampling are called ultimate sampling units, and those in the next to the last stage are called penultimate sampling units [10]. Despite their complexity, sampling designs in three or more stages are often used in human population surveys. The objective in this section is to present a general result for sampling in r stages, where $r \geq 2$.

First, he population is divided into primary sampling units (PSUs), $U_1, \ldots, U_i, \ldots, U_N$. Their sizes are often unknown before sampling begins. The set of PSUs will be represented symbolically as

$$U_1 = \{1, \ldots, i, \ldots, N_1\}$$

Let $s_1, p_1(\cdot), \pi_{1i}, \pi_{1ij}, t_i, \hat{t}_i$ denote the same as in Chapter 3.1.2.1-3.1.2.6.
We do not need detailed notation for the subsequent $r-1$ stages of sampling. However, we assume that we construct an estimator $\hat{t}_i$ of the total PSU t_i, such that $\hat{t}_i$ is unbiased with respect to the final $r-1$ stages of selection, that is,

$$E(\hat{t}_i|s_1) = t_i$$

for all i. The ultimate stage sampling units are not necessary elements; they can also be clusters of elements.

Let $V_i = V(\hat{t}_i|s_i)$ be the variance of $\hat{t}_i$ due to the last $r-1$ stages of selection, and let $\hat{V}_i$ be an unbiased estimator of V_i given s_i, that is, $E(\hat{V}_i|s_i) = t_i$.

We assume invariance and independence of the sampling stages subsequent to the first stage. Whenever a certain PSU is selected, subsampling of that PSU follows an invariant rule, and subsampling of one PSU is independent of subsampling of all other PSUs. It is easy to show that an unbiased estimator of the population total t is given by

$$\hat{t} = \sum_{s_1} \hat{t}_i/\pi_{1i} \tag{3.18}$$

where $E(\hat{t}_i|s_1) = t_i$.

Multistage sampling is the most commonly used sampling design in human population surveys. To illustrate these concepts, we present the European Social Survey (ESS) as an example that incorporates all these factors into the sampling design. Since its establishment in 2001, the ESS has been conducting cross-national surveys across Europe[2]. These surveys involve face-to-face interviews with newly selected

[2]More about the ESS and their methodology can be found here: `https://www.europeansocialsurvey.org/about-ess`

cross-sectional samples every two years. It aims to harmonize data collection efforts in 40 countries and ensure comparability over time across the 10 waves of data collection conducted so far.
Multi-stage sample designs are proposed to participating countries to enhance cost-efficiency by clustering the sample within small geographical areas, where each cluster is assigned to one interviewer, or to accommodate constraints imposed by the available sampling frames [50]. Some examples of multi-stage sample designs in the participating countries include:

(i) 2-stage. First stage small geographical areas; second stage persons (population register)

(ii) 3-stage. First stage small geographical areas; second stage dwellings; third stage persons (address list or area sampling)

(iii) 4-stage. First stage small geographical areas; second stage addresses; third stage dwellings; fourth stage persons (address list or area sampling)

Hungary uses a 2-stage sampling design: the first stage consists of small geographical areas (settlements); second stage consists of persons. There are two sampling domains in the first stage of the sampling design. The first domain consists of Budapest and the 141 largest settlements in Hungary, while the second domain includes all other settlements. The allocation of samples to each domain is based on the population of individuals aged 15 and above [50].

In the first domain, a one-stage sampling design is employed, which involves stratified sampling using the settlements as strata. Within each stratum, individuals are selected using a simple random sample. The sample size in the first sampling domain is 2300 individuals [50].

A two-stage sampling design is employed in the second sampling domain. In the first stage, settlements are chosen through a stratified sampling method, with the strata representing 7 geographical regions. The sample size allocation to each stratum is proportional to the size of the target population (15+ population) within that stratum. Within each stratum, settlements are sampled based on the probability proportional to the size of their target population (15+ population). The selected settlements (PSUs), for a given round of the ESS can be found in Figure 3.1.

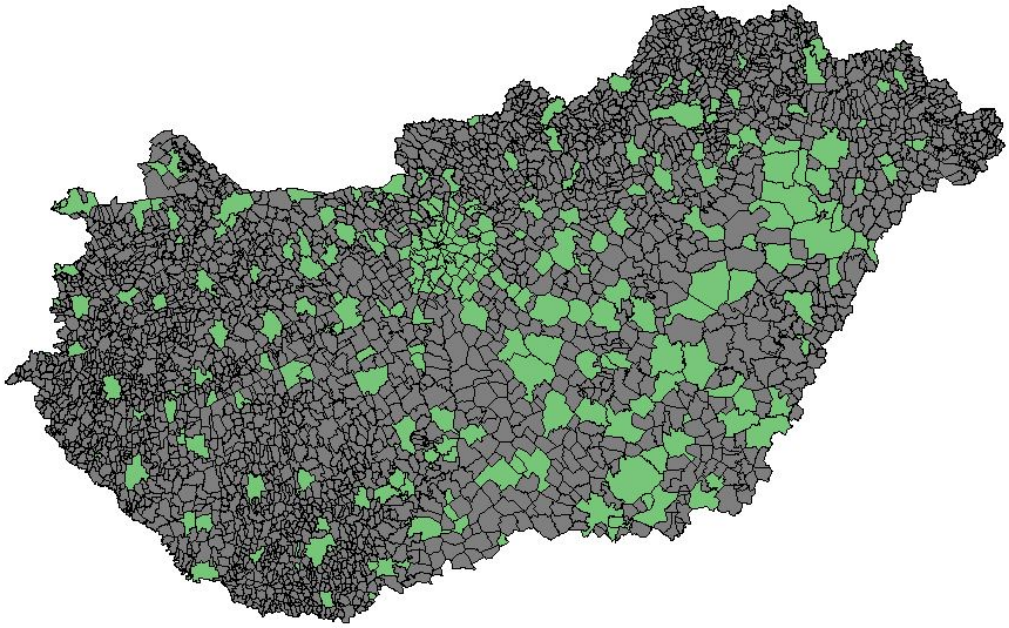


Figure 3.1: Selected settlements (PSUs) during first-stage sampling in Hungary

Source: Own figure.

In the second stage, individuals are selected from each chosen settlement using a simple random sampling technique. In the case of the ESS, individuals are selected from the population register, but there are several other surveys, in the case of which addresses are selected [50]. In these cases, the sampling design is supplemented with an additional stage, which is the within household selection. Within-household selection is a separate field in survey sampling, and not discussed further in this thesis.

The ESS field example in this chapter contains results from the following published paper of the author:
Messing, V., Ságvári, B., Szeitl, B. (2022): Is "push-to-web" an alternative to face-to-face survey?: Experiences from a "push-to-web" hybrid survey in Hungary. (In Hungarian) *Statisztikai Szemle (0039-0690):* 100/3 pp 213-233

3.1.3 Finite population inference

Like other inference procedures, survey statistics relies heavily on limit theorems. Central limit theorems (CLTs) assume that observations are drawn from an infinite population [51], which poses a limitation in the context of human population surveys, where the population is finite. Consequently, the general consideration of CLTs in human population surveys is limited. To overcome this limitation in the case of survey sampling, the concept of superpopulation is applied, whereby CLTs can be interpreted for finite populations. This concept involves the creation of an artificial infinite sequence of finite populations, which allows for the fulfillment of convergence properties. Chapter 3.1.3.1 outlines the superpopulation concept, while Chapter 3.1.3.2 explores the general form of the finite population CLTs based on the artificial sequences of finite populations.

3.1.3.1 The super-population concept

As stated in Chapter 3.1.1, the general assumption in survey sampling is that N is considered to be fixed. This assumption may cause difficulties in making inferences from human population surveys in two aspects: regarding (1) the theoretical scope of conclusions; and (2) the adoptability of limit theories. In this chapter, these aspects are discussed.

In reality, the population undergoes constant changes over time as individuals are born, die, move in, and move out. This means that the conclusions drawn from a survey experiment can be directly generalized only to the state of the population at the given time of sampling (i.e. to Hungary at a given moment) and not to the population in general (i.e. the Hungarian population in the long run). The need to be able to separate these two different targets motivated the development of a more general population concept, called superpopulation [52].

In the concept of superpopulation, there exists an artificial general population, consisting of an infinite number of finite populations [53]. This infinite sequence of finite populations represents a population in the long run. The finite population (N), from which the sample n is directly drawn, is regarded a random sample of the infinite population, thus different sets of N subjects can arise from the infinite superpopulation. With this concept, the superpopulation is postulated to provide an abstract representation of a broader entity from which the population values are generated. It follows that sampling theories can be classified based on the targets of inference. If the target parameters are parameters of the finite population, the sampling procedure is a random sampling process from a fixed finite population, and if the target parameters are parameters of infinite superpopulation, the sampling procedure can be regarded as a random two-step sampling process from an infinite population. The concept of superpopulation was initially introduced in the context of discussing estimators and analyses that utilize repeated survey and census data. Based on this broader concept, a census is only a larger sample, and complete enumerations are unfeasible [54, 53].

In the superpopulation concept, the superpopulation is linked to the potentially observable finite population by defining superpopulation models. These models are most commonly regression models that consider the underlying social and economic cause systems. In addition to general regression models, numerous more complex models have been discussed in the literature [55, 56]. However, we do not discuss superpopulation models in the thesis.

The adoptability of limit theories is another aspect that poses challenges due to the assumed fixed value of N. Just as the superpopulation concept influences the theoretical scope of conclusions, sequences of finite populations are also considered in the context of finite population CLTs.

For formulating superpopulations in this aspect, the sequences are composed of finite populations and their corresponding probability samples. Each element within the finite population sequence is identified by a set of indices. It is commonly assumed that the Nth finite population consists of N elements [1]. Thus, the set of indices for the Nth population is

$$U_N = \{1, 2, ..., N\},$$

where $N = 1, 2, ...$ Associated with the ith element of the Nth population is a column vector of characteristics, denoted by $\mathbf{Y}_{iN}$. Let

$$\mathcal{F}_N = (\mathbf{Y}_{1N}, \mathbf{Y}_{2N}, ..., \mathbf{Y}_{NN})$$

be the set of vectors for the Nth finite population. The set $\mathcal{F}_N$ is often called simply the Nth finite population, or the Nth finite universe. Two types of $\{\mathcal{F}_N\}$ can be specified. In one, the set $\mathcal{F}_N$ is a set of fixed vectors from fixed sequence. In the other, the vectors $\mathbf{Y}_{iN}$, $i = 1, 2, ..., N$ are random variables. For example, the $\{\mathbf{Y}_{iN}\}$, $i = 1, 2, ..., N$, might be the first N element of the sequence $\{Y_i\}$ of *iid* random variables with distribution function $F(Y)$.

Recall that a sample is defined as a subset of population indices, where A_N represents the indices in the sample taken from the Nth finite population. The sample size, denoted as n_N, refers to the number of unique indices in the sample. It is assumed that samples are chosen based on the probability rule $p_N(A)$. A fully specified sequence includes details about the structure of the finite populations and the sampling probabilities. For instance, it may assume that the finite population consists of N independent and identically distributed (iid) random variables with specified probabilities, and that the samples are simple random samples without replacement of size n_N selected from the N population elements [1]. In this scenario, a simple random sample of size n_N taken from the finite universe is a set of iid random variables with a common distribution function $F_y(y)$.

Theorem 3.1.3.1. *(Fuller [2009]: Sampling Statistics Chapter 1.3., pp. 34.) Suppose that $Y_1, Y_2, ..., Y_N$ are iid with distribution function $F(Y)$ and corresponding characteristic function $\varphi(t) = E\{e^{ity}\}$. Let $d = (I_1, I_2, ..., I_N)'$ be a random vector with each component supported $\{0, 1\}$ representing if the element is selected to the sample or not. Assume that $\mathbf{d}$ is independent of $(Y_1, Y_2, ...Y_n)'$. Let $U = \{1, 2, ..., N\}$ and define $A = \{k \in U : I_k = 1\}$. If A is nonempty, the random variables $(Y_k, k \in A)|\mathbf{d}$ are iid with characteristic function $\varphi(t)$.*

The theorem relies on that the probability rule defining membership in the sample, the probability function for d, is independent of $(y_1, y_2, ..., y_N)$. Then it follows that given d with component support on $\{0, 1\}$, the sets $(y_k, k \in A\}$ and $\{y_k, k \notin A\}$ are sets of random variables n and $N - n$ *iid* with distribution function $F_y(y)$.

3.1.3.2 Central limit theorem for finite populations

As summarized by Li [51], Erdös and Rényi [57] and Hájek [58] separately established finite population central limit theorems (CLTs) for simple random sampling from finite populations. In the following, the general CLT for finite populations is outlined based on Hajek [58] and Li [51].

Consider a finite population $U_N = \{y_{N1}, y_{N2}, \ldots, y_{NN}\}$ with N units. The sample is a subset of U_N represented by the vector of inclusion indicators $(I_1, \ldots, I_N) \in \{0, 1\}^N$, where $I_i = 1$ if the sample contains unit i, and $I_i = 0$ otherwise. In simple random sampling, the probability that the inclusion vector $(I_1, \ldots, I_N)$ takes a particular value $(z_1, \ldots, z_N)$ is $n!(N - n)!/N!$, where $\sum_{i=1}^N z_i = n$ and $\sum_{i=1}^N (1 - z_i) = N - n$. The sample average $\bar{y}_s = \sum_{i=1}^N I_i y_{Ni}/n$ is a simple estimator of the population mean. In the formula of $\bar{y}_s$, the randomness comes from $(I_1, \ldots, I_N)$, and all

y_{Ni}'s are fixed population quantities. Because of this feature, it is straightforward to show that $\bar{y}_s$ has mean $\bar{y}_N$ and variance

$$var(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right)\nu_N, \tag{3.19}$$

depending on the finite population variance of U_N (see [59]):

$$\nu_N = \frac{1}{N-1}\sum_{i=1}^{N}(y_{Ni} - \bar{y}_N)^2. \tag{3.20}$$

In order to perform statistical inference on $\bar{y}_N$ using $\bar{y}_s$, it is necessary to describe the sampling distribution of $\bar{y}_s$ resulting from simple random sampling.

The finite population asymptotic scheme involves placing U_N within a theoretical infinite sequence of finite populations of increasing sizes, in line with the superpopulation concept discussed in Chapter 3.1.3.1. The asymptotic distribution of any sample statistic is its distribution as the sample size approaches infinity along this theoretical infinite sequence [51]. Similar to the classical Lindeberg–Feller Central Limit Theorem (CLT) [58], the asymptotic properties of $\bar{y}_s$ are heavily influenced by the maximum squared deviation of the $\bar{y}_{Ni}$'s from the population mean $\bar{y}_N$:

$$m_N = \max_{1\leq i\leq N}(y_{Ni} - \bar{y}_N)^2 \tag{3.21}$$

Hájek [58] asserts that, subject to certain regularity conditions on the sequence of finite populations $\{U_N\}_{N=1}^{\infty}$ and the sample sizes of simple random samples, the sample mean converges asymptotically to a normal distribution.

Theorem 3.1.3.2. *Li [51] Let $\bar{y}_s$ denote the mean of a simple random sample of size n taken from a finite population $U_N = \{y_{N1}, y_{N2}, \ldots, y_{NN}\}$. As $N \to \infty$, if*

$$\frac{1}{\min(n, N-n)} \cdot \frac{m_N}{\nu_N} \to 0 \tag{3.22}$$

then $(\bar{y}_s - \bar{y}_N)/\sqrt{var(\bar{y}_s)} \xrightarrow{d} \mathcal{N}(0,1)$, where ν_N is defined in (3.20) and m_N is defined in (3.21).

3.2 Quality control

In human population surveys, there are several stages between the sampling design $p(\cdot)$ and the estimation of a particular population parameter Y, which can lead to errors and biases. This chapter systematically introduces all errors and biases in the research process.

The total survey error (TSE) framework is the most comprehensive method of understanding the sources of errors and biases in survey estimates, classifying them based on the representation and measurement components [60]. Chapter 3.2.1 introduces the TSE framework, which is reconsidered by categorizing the sources of errors and biases into mathematical and non-mathematical factors in Chapter 3.2.2.

3.2.1 The total survey error framework

The basic concept of the TSE refers to all errors that arise in the design, collection, processing, and analysis of survey data [18]. In this context, a survey error is defined as the deviation of a survey estimate from its underlying true value [60]. The TSE framework aims to consider two distinct sets of inferential procedures in surveys. The first set involves steps taken from an estimation derived from a group of respondents to the target population, focusing on aspects like coverage, sampling, nonresponse, and adjustment error characteristics of statistics based on samples. The second set involves steps taken from a single respondent's answer to a question, related to the measurement that has been the subject of psychometric research on measurement error [18]. Figure 3.2 illustrates the various phases involved in collecting human population survey data, along with the potential errors and biases that could arise at each transition between the steps. In the following, we outline these errors and biases in the representation and measurement components of the TSE based on [18].

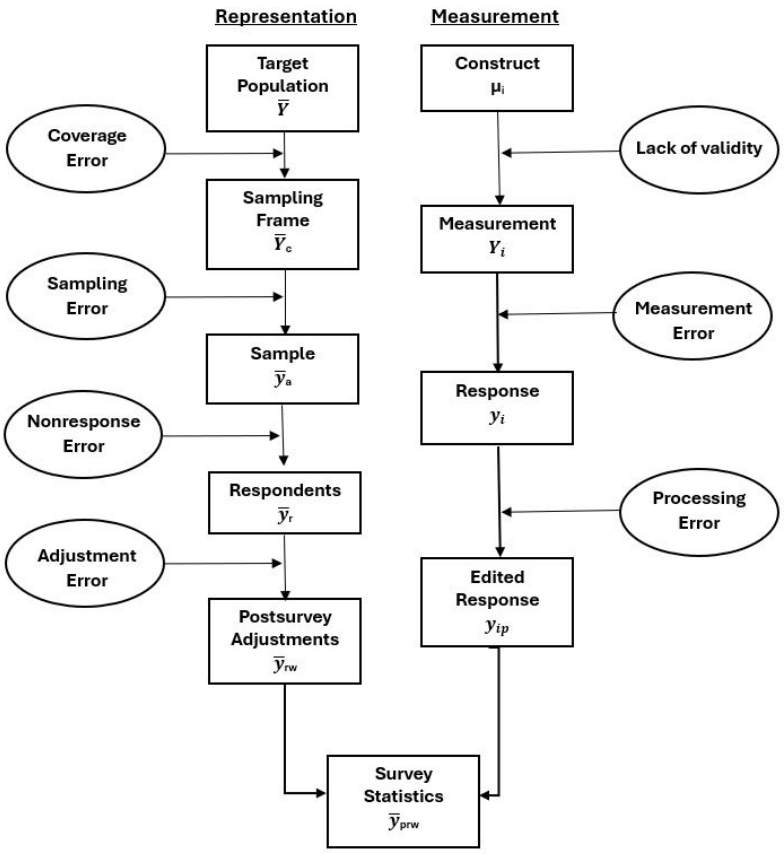


Figure 3.2: Total Survey Error Framework.

Source: Own editing based on Groves (2009), pp. 48. [18]

Coverage error constitutes the initial element of the representation aspect within the TSE framework (on the left side of Figure 3.2). Coverage error represents the non-observational gap between the target population and the sampling frame. Coverage error is always present, as there is no perfect sampling frame (e.g. population register) that totally matches the population. The bias of coverage can be expressed

as:

$$\bar{Y}_C - \bar{Y} = \frac{D}{N}(\bar{Y}_C - \bar{Y}_D),$$

where $\bar{Y}$ denotes the mean value of the target population, while $\bar{Y}_C$ represents the mean value of the population available in the sampling frame. $\bar{Y}_D$ denotes the mean value of the target population that is not included in the sampling frame. N still refers to the total number of individuals in the target population, whereas C represents the total number of individuals covered by the sampling frame and D denotes the number of individuals not covered by the sampling frame. Consequently, the error in the mean value caused by undercoverage[3] can be computed by multiplying the noncoverage rate $\left(\frac{D}{N}\right)$ by the difference between the mean values of the covered and noncovered individuals in the target population. Coverage error may be perceived either as a feature of the population or as a feature of the research setting.

When sampling from populations that have a comprehensive and reliable population register, efforts can be made to reduce this error. However, in situations where the population register of a country is either unreliable or unavailable, coverage error becomes relevant. Addressing coverage error is only possible when there are available data concerning D and $\bar{Y}_D$, which is usually not the case.

Sampling error refers to the difference between the sampling frame and the selected sample [18]. There are two types of sampling error: sampling bias and sampling variance. Sampling bias occurs when certain members of the sampling frame have a reduced chance of being selected. Sampling variance arises because, given the sample design, different realizations of the sample (s) can be drawn. Each of these samples produces a different sample average, which forms the sampling distribution of the average. The dispersion of this distribution measures the sampling variance. The magnitude of the error due to sampling depends on four key principles in the sampling design ($p(\cdot)$): (1) whether the sampling design is a probability sampling process; (2) whether the sample is stratified; (3) whether element or cluster samples are used; and (4) the sample size. Sampling bias is influenced by how probabilities of selection are assigned to different frame elements (π_i). Sampling variance measures the variability of the $\bar{y}_s$ values in all sample realizations:

$$\frac{\sum_{s=1}^{S}\left(\bar{y}_s - \bar{Y}_C\right)^2}{S}$$

When sampling variance is high, the sample means are very unstable, thus the sampling error is high.

Nonresponse error refers to the discrepancy between the sample and the pool of respondents due to non-participation. Despite efforts made in the field, not all individuals in a survey are successfully measured. Let $\bar{y}_s$ represent the mean of the specific sample selected, $\bar{y}_m$ represent the mean of the respondents within the sth sample, $\bar{y}_v$ represent the mean of the nonrespondents within the sth sample, n_s represent the total number of sample members in the sth sample, m_s represent the total number of respondents in the sth sample, and v_s represent the total number of nonrespondents in the sth sample. Similar to coverage bias, nonresponse bias can be expressed as follows:

[3]In sample surveys, it is more typical to encounter undercoverage than overcoverage. Nevertheless, the same reasoning can be applied to situations of overcoverage.

$$\overline{y}_m - \overline{y}_s = \frac{v_s}{n_s}(\overline{y}_m - \overline{y}_s)$$

Therefore, the bias caused by nonresponse for the sample mean can be calculated by multiplying the nonresponse rate with the difference between the mean of the respondents and the mean of the nonrespondents. Nonresponse can occur when one or more variables in a questionnaire do not receive any response. We distinguish unit nonresponse, which happens when there is a complete absence of information for a particular unit, and item nonresponse, which occurs when there is a lack of information for a unit but only for certain variables [5].

- Unit nonresponse can happen due to various reasons. The main theories and current trends are outlined in Chapter 2.2. The most important reasons are: the potential respondent declining to answer or being unable to provide a response, inability to reach the potential respondent, loss or destruction of the questionnaire, or abandonment of the questionnaire at the beginning of the survey.

- Item nonresponse may occur when individuals refuse to answer specific questions, do not understand the questions or responses, abandon the survey before completion, or when certain parts of the questionnaire are invalidated due to inconsistencies in the collected data. The thesis does not delve into item nonresponse any further, but the author addresses this issue in another study [61].

Postsurvey adjustment is a data correction technique, which may also contribute errors. It uses information about the target or frame population, or the response rate on the sample. Generally, adjustments assign greater weight to sample elements (w_i) that are underrepresented in the final data set. An adjusted sample mean is calculated by:

$$\bar{y}_{rw} = \frac{\sum_{i=1}^{r} w_i y_{si}}{\sum_{i=1}^{r} w_i}$$

The error related to the adjusted mean is $(\bar{y}_{rw} - \bar{Y})$.

The second component of the TSE frameworks (right side of Figure 3.2) consists of errors regarding the measurement at the individual level. To define these errors, let us denote:

- μ_i is the true value of a construct for the ith person in the population, $i = 1, 2, \ldots, N$
- Y_i is the value of a measurement for the ith person in the sample, $i = 1, 2, \ldots, n$
- y_i is the value of the response to a given question, $i = 1, 2, \ldots, n$
- y_{ip} is the value of the response after the data processing steps, $i = 1, 2, \ldots, n$.

Lack of validity refers to the discrepancy between constructs and measures in an observational context. Each measurement of the ith individual represents just one of the numerous possible measurements (trials) that could be performed. When a

particular measure of Y is administered, the outcome is not μ_i, but rather the true value plus an error associated with the specific trial ($Y_{it} = \mu_i + \epsilon_{it}$). Validity ($q$) is determined by the correlation between the measurement, Y_i, and the true value, μ_i, between all potential trials and individuals.

$$q = \frac{E_{it}\left[(Y_{it} - \overline{Y})(\mu_i - \mu)\right]}{\sqrt{E_{it}(Y_{it} - \bar{Y})^2}\sqrt{E_{it}(\mu_i - \mu)^2}}$$

When there is a covariance between y and μ, it suggests robust construct validity in the measurement. A dependable indicator of an underlying construct is one that demonstrates a strong correlation with the construct. Evaluating validity is challenging as the latent construct cannot be directly observed.

Measurement error occurs when there is a difference between the actual value of a measurement and the value obtained when measuring a sample unit. In the field of survey measurement theory, a response deviation occurs when $y_i \neq Y_i$. If these deviations in responses consistently point in the same direction in multiple attempts, it may suggest the presence of response bias, thus $E_t(y_{it}) \neq Y_i$. However, unlike measurement bias, response deviations can also result in unreliable responses due to variation. The key difference between response variance and response bias is that the latter is systematic, consistently overestimating or underestimating the quantity being measured, while response variance causes fluctuation in estimated values across attempts [18].
The processing or editing error is the discrepancy between the variable utilized in the estimation and the one provided by the respondent, indicated as $(y_{ip} - y_i)$.

The errors in the TSE framework are classified based on representation (sample) and measurement (answers). This aligns with how we considered the reasons for inaccurate estimates in Chapter 2: the issue could stem from either sample-related concerns, which become more critical as response rates decline, or from the measurement process itself, specifically the response, which can be influenced by various psychological and social factors. In Chapter 3.2.2, the errors are classified into two groups: those that can be rectified using existing mathematical tools (mathematical factors of the total error) and those that cannot be rectified directly with current mathematical tools (non-mathematical factors of the error).

3.2.2 Mathematical and non-mathematical factors of the total error

Certain errors and biases can be accurately detected through mathematical techniques, allowing for estimation of their extent and direct management of their consequences. These can be regarded as mathematical aspects of the inaccuracies. In contrast, there are effcts where the underlying mechanisms are not well-defined, the patterns are only partially understood, making it difficult to address their impacts. These can be seen as non-mathematical aspects. Figure 3.3 illustrates the various types of inaccuracies identified in human population surveys, categorizing them into mathematical and non-mathematical components.

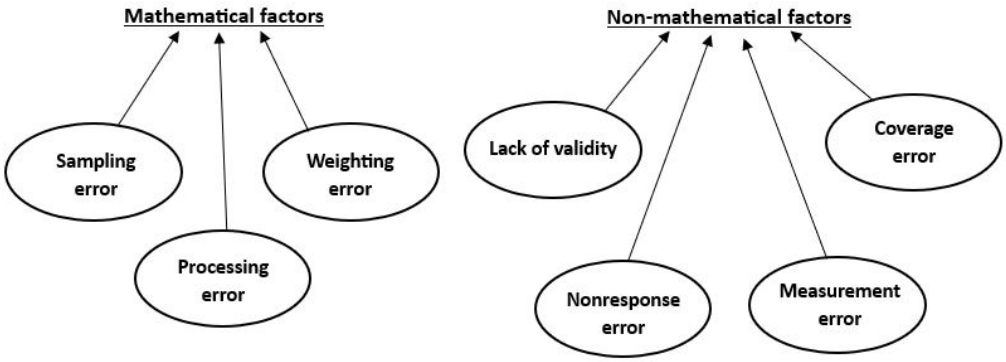


Figure 3.3: The element of the TSE framework in the group of mathematical and non-mathematical factors

Soruce: Own editing.

Mathematical factors include errors that can be managed by sampling design (sampling error), accurate data processing procedures (processing error), and suitable post-survey weights (weighting error). The extent of these errors can be quantified in an exact way using the formulae presented in Chapter 3.2.1.

On the contrary, non-mathematical elements encompass inaccuracies that are impacted by phenomena that are not directly observable. Lack of validity, non-response error, measurement error, and coverage error fall under the umbrella of non-mathematical factors. Lack of validity is dependent on the researcher's ability to adequately capture the phenomena of the studied population, while the true population phenomena remain unknown. Nonresponse and measurement errors are primarily influenced by the behaviors of respondents: nonresponse error is related to participant behavior in survey participation, while measurement error pertains to the accuracy of responses to specific questions. The responses of those who decline to answer will always remain unknown, making it impossible to quantify the extent of the error. Similarly, the true responses of those who provide different answers will never be known, making it challenging to quantify the degree of error. Coverage error is influenced by the research infrastructure and the specific population being studied. For instance, the population not included in a population register is unknown due to lack of data.

Errors within the category of non-mathematical factors are inevitable, yet their underlying mechanisms remain elusive.

As discussed in Chapter 2.2, these specific issues often lead to inaccurate predictions as they are more challenging to manage compared to mathematical factors. The thesis will now delve into two errors within the realm of non-mathematical factors: nonresponse error, which impacts the composition of the sample, and measurement error, which affects the accuracy of responses.

Chapter 4

A new sample allocation method

As mentioned in Chapter 3.2, the issue of bias caused by nonresponse is commonly addressed through post-stratification. Although this approach is generally effective when there is enough supplementary data, it increases variance in cases where certain cells have very limited or no observations. This chapter introduces a novel method that addresses the issue of nonresponse bias and reduces variance by allocating samples based on expected response rates. The approach is specifically designed for stratified sample designs (Chapter 3.1.2.6), which are commonly used in survey implementations.

We demonstrate that if specific response rates are available for different strata, the new sample allocation that takes these into account has certain advantages over the traditional proportional to size (PS) allocation methods. This is accomplished not only when the response rates are accurately known, but also when they are only approximately defined. In fact, an allocation that considers the expected response rates (ERR) leads to lower variance compared to using a PS allocation. By implementing the new allocation method, it becomes possible to effectively address one of the most critical quality concerns in human population surveys, the nonresponse bias. This method allows better control over sample composition, thereby reducing variance and ensuring an appropriate representation of the population.

After a short introduction, we briefly introduce the traditional PS procedure with post-stratification (Chapter 4.2) before Chapter 4.3 formally presents the method of ERRs allocation. The relative performance of ERRs allocation is assessed by comparing the variances in the resulting estimates in Chapter 4.4. The asymptotic variances are calculated using the d-method in Chapter 4.5.1 and are then initially compared by assuming correctly specified response rates in Chapter 4.5.2. Here, the assumed response rates are subject to random fluctuations, which are then corrected using post-stratification. In Chapter 4.5.3, variance comparison is performed in terms of misspecified response rates, and the results of an extensive assessment using various combinations of specific population parameters are presented.

This chapter is published as:
B. Szeitl & T. Rudas (2022) "Reducing variance with sample allocation based on expected response rates in stratified sample designs" *Journal of Survey Statistics and Methodology* (10), 1107–1120 https://doi.org/10.1093/jssam/smab021

4.1 Introduction

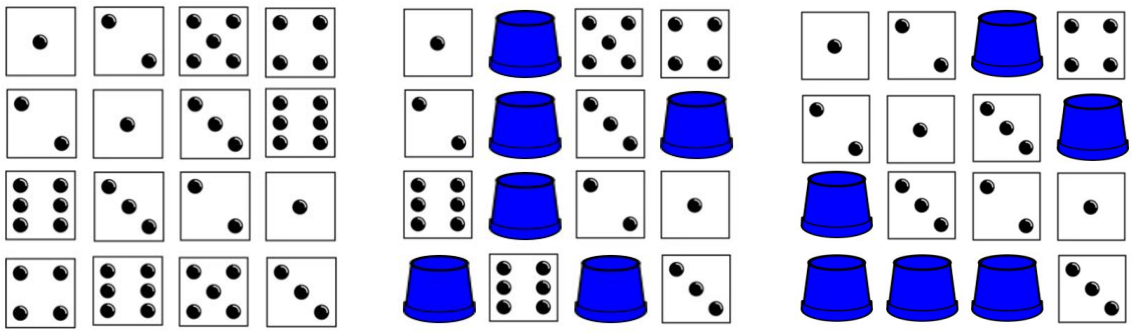
Correcting for nonresponse starts with identifying the nonresponse mechanism. Rubin (1976, [62]) proposed a typology for the different types of nonresponse mechanism:

(i) Nonresponse is called uniform or missing completely at random (MCAR) if it does not depend on either the variable of interest or auxiliary variables. The response probability is then constant for all units in the population.

(ii) Nonresponse is called ignorable or missing at random (MAR) if the response probability depends on auxiliary variables, which are not affected by nonresponse, but do not depend on the variable of interest.

(iii) Nonresponse is called nonignorable or missing not at random (MNAR) if it depends on the variable of interest which is affected by nonresponse. For example, if the nonresponse of an income variable depends on the income itself.

The concept of nonresponse mechanisms is illustrated using the analogy of the traditional dice roll experiments in Figure 4.1. The complete set of true values (i.e. all of the 16 individuals in the sample) is shown in Figure 4.1a. Figure 4.1b demonstrates the MCAR situation, where some results are not visible but are independent of the true values in Figure 4.1a. On the contrary, Figure 4.1c presents the MNAR/-MAR scenario, where the unobserved outcomes depend on the rolled values (which can coinsidered as an auxiliary variable or the variable of interest). In this case, there could be a hidden process that hides larger values, leading to biased estimates. When addressing data quality, the modeling task involves recognizing the nonresponse mechanism associated with the variable being studied.



(a) Total sample (b) MCAR (c) MNAR (MAR)

Figure 4.1: Nonresponse mechanism based on the analogy of the dice roll

Source: Own figure.

Nonresponse can therefore seriously distort the results and even lead to incorrect conclusions. As we discussed in Chapter 2, previous field experiences and the analysis of current survey meta-data indicate that the overall increase in survey nonresponse does not equally apply to different population subgroups [29, 63]. The resulting distortion of sample composition is usually dealt with using post-stratification [18]. It has been found that single-person households, renters and individuals outside of

the labour force are less likely to participate in surveys than members of other social groups [64, 29]. This suggests that giving a larger proportional allocation to these groups may improve the realised sample. To be able to determine the exact proportions during the allocation procedure, estimates from previous surveys are needed. In case of item-nonresponse the expected historical response rates are easy to determine using publicly available survey data. In the case of unit-nonresponse, the contact data (or survey meta-data) are typically not available publicly, but survey organizations can use their own historical data.

As we introduced in Chapter 3.1.2.7, to achieve an optimum balance between data collection costs and estimation efficiency (variance reduction), complex selection methods are typically required for the sampling design of human population surveys.

Samples that are representative according to previously appointed variables may be obtained via a precise allocation of the sample sizes within different strata, if the relevant information is available both for the entire population, e.g., from a census, and also for every individual in a sampling frame, e.g., in a register. Generally, the proportional-to-stratum size (PS) allocation method [65] is used. However, the realised (observed) sample sizes within the strata tend to differ from the planned (allocated) ones.

4.2 Allocation proportional to size

Let N denote the population size and let N_h $(h = 1, 2, \ldots, H)$, be the sizes of the strata relevant to the sampling procedure, with $N = N_1 + ... + N_H$. In a stratified random sample, a simple random sample of n_h elements is taken from each stratum h $(h = 1, 2, ..., H)$, with a total sample size of n elements.

When the survey aims to collect m responses, the response rate which characterizes the population needs to be taken into account in deciding about the attempted sample size. Of course, such decisions should be made based on the true response rate, but it is rarely known. Thus, the ERR, say r, is used which is based on former experience. Then, a total of $n = m/r$ observations are allocated.

In the case of allocation proportional to size (PS), let n_h^{PS} $(h = 1, 2, ..., H)$ denote the subsample size within stratum h. The sampling fraction n_h^{PS}/N_h is specified to be the same for each stratum and thus

$$n_h^{PS} = \frac{1}{r}\frac{N_h}{N}m \qquad h=1,\ldots,H, \tag{4.1}$$

which implies that the overall sampling fraction n/N is the same as the fraction taken from each stratum. The total allocated sample size is then as follows:

$$n^{PS} = m\sum_{h=1}^{H}\frac{N_h}{N}\frac{1}{r} = \frac{m}{r} \tag{4.2}$$

4.3 A new allocation based on ERRs

In the case of ERR allocation, let n_h^{ERR} $(h = 1, 2, ..., H)$ denote the allocated sub-sample size within stratum h. Let r_h $(h = 1, 2, ..., H)$ denote the stratum-specific ERRs, which are also assumed to be population parameters. Clearly,

$$r = \sum_{h=1}^{H} \frac{r_h N_h}{N}.$$

In ERR allocation, the allocated sample size in each stratum n_h^{ERR} is specified using, instead of the population level ERR, the stratum-specific ERRs. The allocated sample size in each stratum is

$$n_h^{ERR} = \frac{1}{r_h} \frac{N_h}{N} m \qquad h=1,...,H. \tag{4.3}$$

Consequently, the total allocated sample size is

$$n^{ERR} = m \sum_{h=1}^{H} \frac{N_h}{N} \frac{1}{r_h}. \tag{4.4}$$

4.4 Estimation procedures

To assess the ERRs and PS allocations, the variances of the estimates obtained will be compared in Chapter 4.5.1 using the δ-method. Here, we describe the estimating procedures.

The main aim is to estimate the proportion of respondents within a given population who would choose a fixed category, e.g., 'yes', of a given close-ended question based on observed samples in terms of both ERRs and PS allocations. In both cases, post-stratification is applied prior to the estimation to appropriately reproduce the relative sizes of the strata in the population [18].

It is assumed that responding to the survey is probabilistic and occurs in stratum h with probability p_h and is independent from the true answer to the question of interest. It should be noted that the r_h response rates represent the expectation of the researcher based on previous knowledge and that p_h is the true probability of responding. The probability of nonresponse[1] is therefore $1 - p_h$ in each stratum h. Thus, the nonresponse mechanism is MCAR [62]. The probability of a 'yes' response is assumed to be q_h in each stratum h.
Under the previous assumptions, the complete data for each stratum, would be the observation of a variable $\mathbf{Z}_h$ with the following four components:

1. Z_{h1} counts the number of cases when the selected respondent did answer and the answer was 'yes'.

[1]For the present argument, it is irrelevant whether nonresponse applies to the entire survey because of no-contact or refusal or only to the current question.

2. Z_{h2} counts the number of cases when the selected respondent did answer and the answer was 'no';

3. Z_{h3} counts the number of cases when the selected respondent did not answer and the answer would have been 'yes';

4. Z_{h4} counts the number of cases when the selected respondent did not answer and the answer would have been 'no';

Within stratum h, $\mathbf{Z}_h$ has a multinomial distribution with parameters n_h and $\mathbf{q}_h$, where n_h is the allocated sample size for stratum h, which depends on the type of allocation, and under the assumed independence of the true response from whether or not the answer is received,

$$\mathbf{q}_h = (p_h q_h,\, p_h(1-q_h),\, (1-p_h)q_h,\, (1-p_h)(1-q_h)). \tag{4.5}$$

The observed sample size is $o_h = Z_{h1} + Z_{h2}$ in stratum h, and for each observation, a post-stratification weight of

$$\frac{\frac{N_h}{N}\sum_{i=1}^{H} o_i}{o_h} \qquad \text{h=1,...,H.}$$

is applied, which adjusts the fraction of the sample size in stratum h to be equal to the population fraction of stratum h but does not change the total observed sample size. After weight is applied, Z_{h_j} is replaced by

$$\frac{N_h}{N} \cdot \frac{\sum_{i=1}^{H}(Z_{i1}+Z_{i2})}{Z_{h1}+Z_{h2}} Z_{h_j}, \qquad \text{j=1,2,3,4} \qquad \text{h=1,...,H.}$$

As such, the natural estimator for the fraction of 'yes' responses in stratum h is

$$\hat{q}_h = \frac{\frac{N_h}{N} \cdot \frac{\sum_{i=1}^{H}(Z_{i1}+Z_{i2})}{Z_{h1}+Z_{h2}} Z_{h1}}{\frac{N_h}{N} \cdot \frac{\sum_{i=1}^{H}(Z_{i1}+Z_{i2})}{Z_{h1}+Z_{h2}} Z_{h2} + \frac{N_h}{N} \cdot \frac{\sum_{i=1}^{H}(Z_{i1}+Z_{i2})}{Z_{h1}+Z_{h2}} Z_{h1}} = \frac{Z_{h1}}{Z_{h1}+Z_{h2}}, \tag{4.6}$$

which is the relative frequency of 'yes' responses among all responses observed in stratum h. It should be noted that as $\hat{q}_h$ refers to a single stratum, the post-stratification weights are cancelled out because they are identical within each stratum. For the entire sample, the $\mathbf{Z}_h$ variables have a product multinomial distribution. The estimator for the fraction of 'yes' responses in the total sample is

$$\begin{aligned}\hat{q} &= \frac{\sum_{h=1}^{H} \frac{N_h}{N} \cdot \frac{\sum_{i=1}^{H}(Z_{i1}+Z_{i2})}{Z_{h1}+Z_{h2}} Z_{h1}}{\sum_{h=1}^{H} \left(\frac{N_h}{N} \cdot \frac{\sum_{i=1}^{H}(Z_{i1}+Z_{i2})}{Z_{h1}+Z_{h2}} Z_{h2} + \frac{N_h}{N} \cdot \frac{\sum_{i=1}^{H}(Z_{i1}+Z_{i2})}{Z_{h1}+Z_{h2}} Z_{h1} \right)} \\ &= \frac{1}{N}\sum_{h=1}^{H} N_h \frac{Z_{h1}}{Z_{h1}+Z_{h2}}\end{aligned} \tag{4.7}$$

which is the weighted fraction of 'yes' responses among all responses observed in the total sample. Here, post-stratification has the effect of weighting the stratum-specific estimates in terms of their population weights.

4.5 Variance comparison

In this chapter we compare the variances of the estimates derived from the ERRs and PS allocations using the δ-method and present the results in the case of correctly specified and misspecified response rates.

4.5.1 The δ-method

Theorem 4.5.1.1 (Multidimensional δ-method)**.** *Let $X_n, n = 1, 2, ...$ be a sequence of k-dimensional vector-valued random variables such that,*

$$\sqrt{n}(X_n - a) \xrightarrow{d} Y, \tag{4.8}$$

where $a \in \mathbb{R}^k$ and $Y \sim N(0, \Sigma)$. If a function $f : \mathbb{R}^k \to \mathbb{R}^l$ is differentiable at $a \in \mathbb{R}^k$, and D is its $l \times k$ matrix of partial derivatives with $d_{ij} = \frac{\partial f_i(a)}{\partial x_j}$, then

$$\sqrt{n}(f(X_n) - f(a)) \xrightarrow{d} Z, Z \sim N(0, D\Sigma D^T). \tag{4.9}$$

The proof of the Theorem can be found in [66] or [49].
As condition (4.8) holds for the multinomial distribution, theorem 4.5.1.1 may be applied within each stratum. The estimator for the proportion of 'yes' responses in the total population (4.7) in Chapter 4.4, is the weighted fraction of 'yes' responses among all responses observed across all strata.
In one strata, when omitting the index h, the estimation function is $f(Z) = \frac{Z_1}{Z_1+Z_2}$ and the partial derivatives are as follows:

$$\frac{df}{dZ_1} = \frac{Z_2}{(Z_1+Z_2)^2} \qquad \frac{df}{dZ_2} = -\frac{Z_1}{(Z_1+Z_2)^2}$$
$$\frac{df}{dZ_3} = 0 \qquad \frac{df}{dZ_4} = 0$$

The partial derivative vector D with the components evaluated above at the expectations $E(Z_1) = np_1$ and $E(Z_2) = np_2$, is

$$D = \begin{bmatrix} -\frac{np_1}{(np_1+np_2)^2} \\ \\ \frac{np_2}{(np_1+np_2)^2} \\ \\ 0 \\ \\ 0 \end{bmatrix} \tag{4.10}$$

As $\boldsymbol{Z}$ has a multinomial distribution with the probability vector given in (4.5), its covariance matrix is

$$\Sigma = \begin{bmatrix} np_1(1-p_1) & -np_1p_2 & -np_1p_3 & -np_1p_4 \\ -np_2p_1 & np_2(1-p_2) & -np_2p_3 & -np_2p_4 \\ -np_3p_1 & -np_3p_2 & np_3(1-p_3) & -np_3p_4 \\ -np_4p_1 & -np_4p_2 & -np_4p_3 & np_4(1-p_4) \end{bmatrix} \tag{4.11}$$

Then, one has the following results for the asymptotic variances.

Theorem 4.5.1.2 (Variance of the estimates)**.** *Let the population size be N, and let the population be divided into H strata of respective sizes of N_h, $(h = 1, .., H)$. Let m be the intended total sample size, r the ERR in the entire population and r_h the respective ERRs in the strata. The true population proportion of those possessing the characteristics of interest is denoted by q_h, which is the parameter to be estimated in each stratum h. Finally, let p_h be the true response rate in stratum h. Then, the asymptotic variances of the estimates obtained from samples based on PS and ERRs allocations, with post-stratification applied, are as follows.*

$$V^{PS}(\hat{q}) = \frac{1}{Nm}\sum_{h=1}^{H} N_h q_h(1-q_h)\frac{r}{p_h} \tag{4.12}$$

$$V^{ERR}(\hat{q}) = \frac{1}{Nm}\sum_{h=1}^{H} N_h q_h(1-q_h)\frac{r_h}{p_h} \tag{4.13}$$

Proof. As stratified sampling leads to a product multinomial distribution (see, e.g., Rudas, 2018), Theorem 4.5.1.1 is applied for each stratum. Then, the asymptotic variance is obtained as follows:

$$D^T \Sigma D = \frac{(p_h q_h)(p_h(1-q_h))^2 + (p_h q_h)^2(p_h(1-q_h))}{n_h((p_h(1-q_h)+p_h q_h)^4}$$

$$= \frac{p_h^3 q_h - p_h^3 q_h^2}{n_h p_h^4} = \frac{p_h^3 q_h(1-q_h)}{n_h p_h^4} = \frac{q_h(1-q_h)}{n_h p_h}$$

As the allocated stratum-specific sample sizes n_h are different in the PS and ERR allocations, different asymptotic variances will be obtained.
In the case of PS allocation, using (4.1),

$$V_h^{PS}(\hat{q}) = \frac{q_h(1-q_h)}{n_h^{PS} p_h} = \frac{q_h(1-q_h)}{\left(\frac{1}{r}\frac{N_h}{N}m\right)p_h},$$

whereas for the total sample, the following is obtained:

$$V^{PS}(\hat{q}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \hat{V}_h(\hat{q}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{q_h(1-q_h)}{\left(\frac{1}{r}\frac{N_h}{N}m\right)p_h}$$

$$= \frac{1}{Nm}\sum_{h=1}^{H} N_h q_h(1-q_h)\frac{r}{p_h}.$$

The asymptotic variance in stratum h in case of the ERR allocation with (4.3) is

$$V_h^{ERR}(\hat{q}) = \frac{q_h(1-q_h)}{n_h^{ERR} p_h} = \frac{q_h(1-q_h)}{\left(\frac{1}{r_h}\frac{N_h}{N}m\right)p_h}$$

whereas for the total sample, the following is obtained:

$$V^{ERR}(\hat{q}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \hat{V}_h(\hat{q}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{q_h(1-q_h)}{\left(\frac{1}{r_h}\frac{N_h}{N}m\right)p_h}$$

$$= \frac{1}{Nm} \sum_{h=1}^{H} N_h q_h (1 - q_h) \frac{r_h}{p_h}$$

□

In terms of a general comparison of the variances obtained above, the difference of the variances for the ERR and PS allocations, disregarding a positive constant multiplier, may be written as a weighted sum of the quantities

$$\frac{r_h - r}{p_h}, \tag{4.14}$$

with weights equal to

$$N_h q_h (1 - q_h). \tag{4.15}$$

Large negative values and small positive values of (4.14) point to a better performance of the ERR allocation than of the PS allocation. The value of (4.14) is sometimes negative and sometimes positive, as r is a weighted average of the r_h values. Negative values of (4.14) are obtained when r_h is smaller than average and they will be made larger if p_h is small. Positive values of (4.14) are obtained when r_h is greater than average and will be made smaller if p_h is large. Thus, (4.14) may be viewed as a measure of how well the ERRs r_h approximate the true response rates p_h, with large negative and small positive values meaning better approximation.

Consequently, $V^{ERR}(\hat{q}) - V^{PS}(\hat{q})$ may be seen as a weighted average of how well r_h approximates p_h, as measured by (4.14), where the weights are the total variances of the strata, given in (4.15). The better the approximation, in particular in the strata with large total variances, the better the ERR allocation performs relative to the PS allocation.

In the following, more detailed comparisons are given.

4.5.2 Comparison under correctly specified response rates

In this section, we prove that in the case of correctly specified response rates ($r_h = p_h$), the variance of the estimate based on the ERRs allocation is less than or equal to that derived from the PS allocation:

Theorem 4.5.2.1 (Relationships among the variances)**.** *Let $\hat{V}^{PS}(\hat{q})$ be the total variance of the estimates based on a sample drawn via the PS allocation given in (4.12), and let $\hat{V}^{ERR}(\hat{q})$ be the total variance of the estimates based on a sample drawn by the allocation based on different ERRs, as given in (4.13). If the observed response rates are equal to the ERRs, then,*

$$\hat{V}^{ERR}(\hat{q}) \leq \hat{V}^{PS}(\hat{q}) \tag{4.16}$$

Proof. If $r_h = p_h$, the response rates are correctly specified, and then r is also the average ERR among all strata. Because N, N_h and q_h are population parameters, and m is a fixed constant, it is enough to see that

$$\sum_{h=1}^{H} N_h q_h (1 - q_h) \leq \sum_{h=1}^{H} N_h q_h (1 - q_h) \frac{\frac{1}{H}\sum_{j=1}^{H} p_j}{p_h}$$

or

$$\frac{1}{\sum_{h=1}^{H} w_h} \sum_{h=1}^{H} w_h \frac{1}{\frac{1}{H}\sum_{j=1}^{H} p_j} \leq \frac{1}{\sum_{h=1}^{H} w_h} \sum_{h=1}^{H} w_h \frac{1}{p_h}.$$

As the left hand side is the weighted harmonic mean of the values $\frac{1}{p_1}, \ldots, \frac{1}{p_H}$, and the right-hand side is the weighted arithmetic mean of the same numbers, by inequality between these means [67] demonstrates that the claim of the theorem is true.

□

4.5.3 Comparison under misspecified response rates

In this chapter, we compare the ERRs and PS allocation methods under misspecification that is, when the true response rates differ from the ERRs used in the sample allocation ($p_h \neq r_h$). The variances were compared for all combinations of parameter values with a fixed number of strata, $H = 3$. Specifically, all possible combinations of the following parameter values were considered: all possible combinations of the values $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ for the true response rates $\{p_1, p_2, p_3\}$ and for the ERRs $\{r_1, r_2, r_3\}$. The parameter to be estimated in every stratum h ($h = 1, 2, 3$) was given values between 0 and 1, with an increment of 0.05. The size of the population $N = 10^7$,the sizes of the strata $N_1 = 2 * 10^6$, $N_2 = 3 * 10^6$, $N_3 = 5 * 10^6$, and the desired total sample size $m = 1000$ were fixed. With the different choices, a total of 15.625.000 different sets of parameters were defined. The calculations were conducted using the R statistical environment.
Figure 4.2 shows the comparison of the variances of the estimates obtained using ERRs and PS allocations. The comparison is given in terms of the total absolute misspecification of the response rates, $\sum_{h=1}^{H} |r_h - p_h|$ *(x-axis)* and of the total absolute distance of the ERRs $\{r_1, r_2, r_3\}$ from their weighted average, $\sum_{h=1}^{H} |r_h - r|$ *(y-axis)*.
The magnitude of the misspecification of the response rates appeared to have a greater impact on the relative performances of the two allocation procedures. When the total absolute misspecfication was less than 0.3, the ERR allocation almost always performed better. Meanwhile, the total absolute distance of the ERRs from their weighted average appears to have had a small and non-systematic effect.

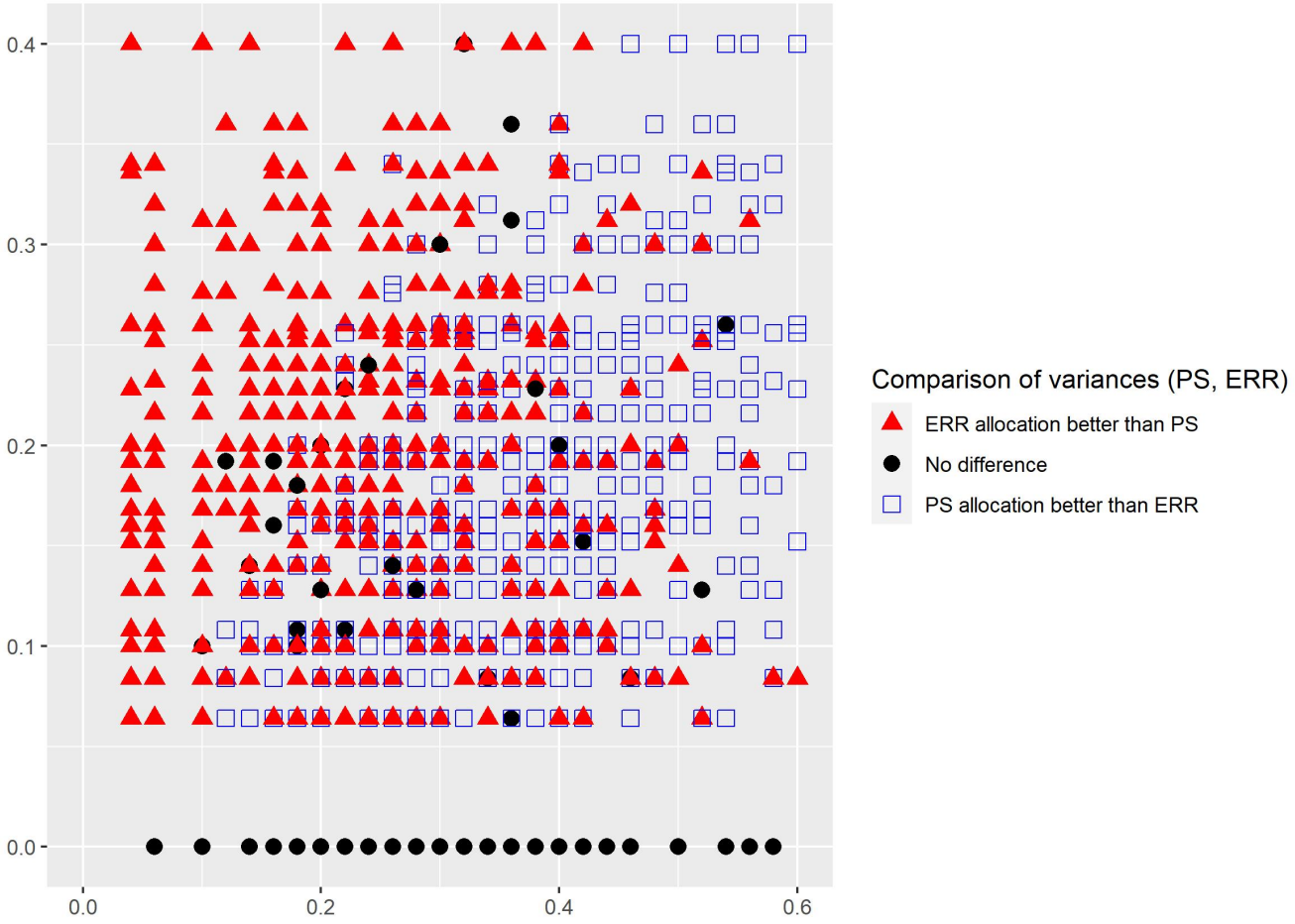


Figure 4.2: Comparison of the variances in the estimates obtained using ERR and PS allocations, in terms of the total absolute misspecification of the response rates *(x-axis:* $\sum_{h=1}^{H} |r_h - p_h|$*)* and the total absolute distance of the ERRs from one weighted average *(y-axis:* $\sum_{h=1}^{H} |r_h - r|$*).*

Source: Own figure.

Figure 4.3 shows the comparison of the variances of the estimates obtained using ERR and PS allocations in terms of the total absolute misspecification of the response rates, $\sum_{h=1}^{H} |r_h - p_h|$ *(x-axis)* and the difference in the absolute deviances of the response rates from their respective weighted averages, $\sum_{h=1}^{H}(|r_h - r| - |p_h - p|)$ *(y-axis).*

When the total absolute misspecification of the response rates was lower than 0.3, the ERR allocation yielded mostly smaller variances. Meanwhile, in the range of 0.3−0.4, the two allocations performed equally well. Most notably, an equal precision can be expected in the extreme areas of the plot.

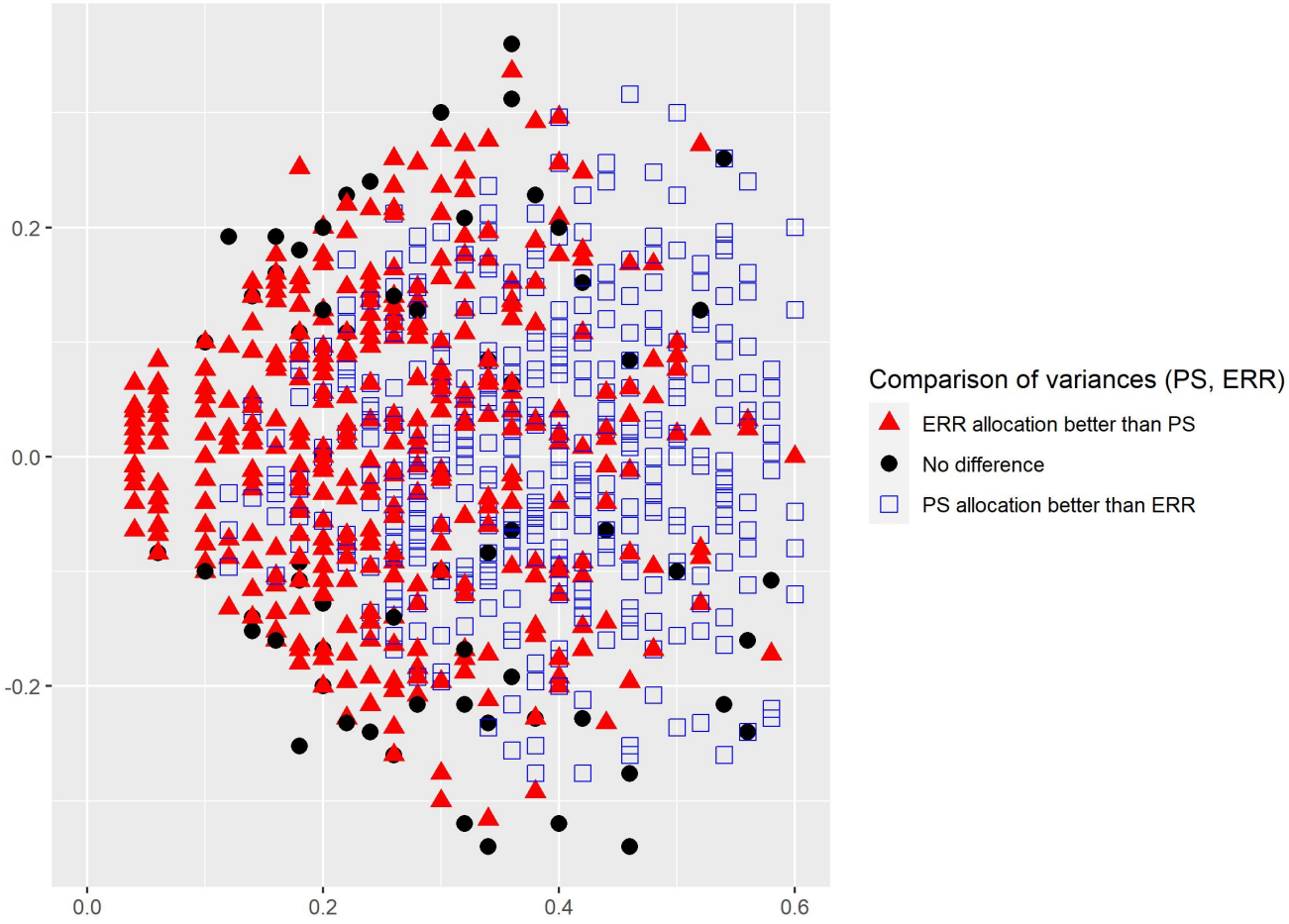


Figure 4.3: Comparison of the varianc in the estimates obtained using ERR and PS allocations, in terms of the total absolute misspecification of the response rates *(x-axis:* $\sum_{h=1}^{H} |r_h - p_h|$*)* and the difference in the absolute deviations of the response rates from their respective weighted averages *(y-axis:* $\sum_{h=1}^{H}(|r_h - r| - |p_h - p|)$*)*.

Source: Own figure.

Figure 4.4 shows the comparison of the variances of the estimates obtained using the ERR and PS allocations in terms of the total absolute distance of the response rates from their weighted average $\sum_{h=1}^{H} |r_h - r|$ *(x-axis)* and the difference in the absolute deviations of the response rates from their respective weighted averages $\sum_{h=1}^{H}(|r_h - r| - |p_h - p|)$ *(y-axis)*.
Here, the total absolute misspecification shown on the x-axes of Figures 4.2 and 4.3 was disregarded but was clearly more influential than the characteristics shown in Figure 4.4. When the difference between the total absolute deviations of the expected rates and the ERRs was less than approximately half of the latter, the ERR allocation always performed better, irrespective of whether or not the individual response rates were correctly predicted.

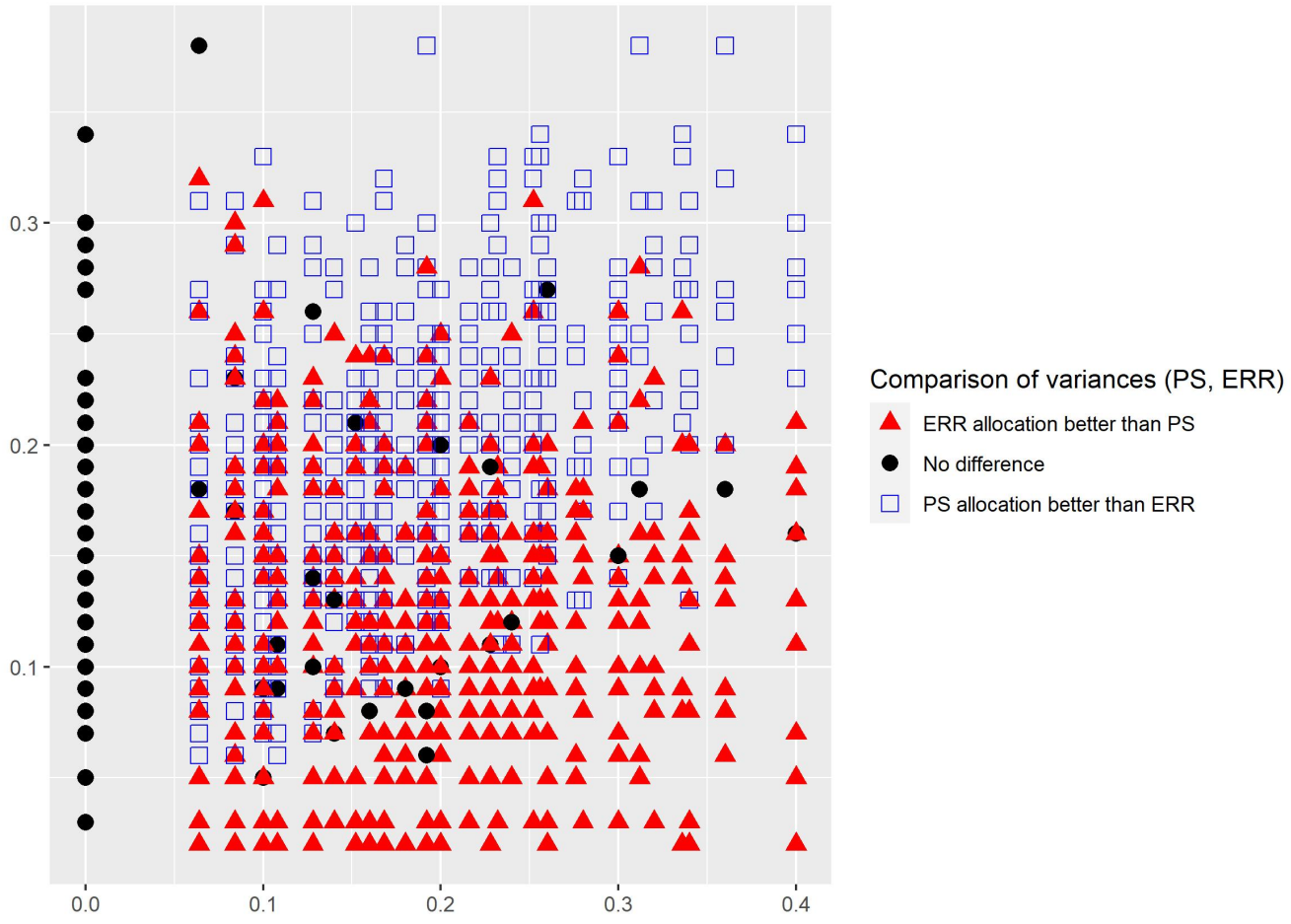


Figure 4.4: Comparison of the variances of the estimates obtained using the ERR and the PS allocations in terms of the total absolute distance of the response rates from their weighted average *(x-axis:* $\sum_{h=1}^{H} |r_h - r|$*)* and the difference in the absolute deviations of the response rates from their respective weighted averages *(y-axis:* $\sum_{h=1}^{H}(|r_h - r| - |p_h - p|)$*)*.

Source: Own figure.

4.6 Discussion

In this chapter, we demonstrated how ERRs can be utilised in the sample allocation procedure. In the process, we introduced an ERRs allocation procedure where the stratum-specific ERRs were used to determine the allocated sample sizes within each stratum. We assessed the method by comparing it with a standard proportional allocation method (PS) where stratum-specific response rates are not used. The assessment of the sample allocation procedures used a comparison of the resulting asymptotic variances based on the δ-method when assuming the expected responses were equal to the true responses. In the case of misspecified response rates, extensive enumeration was used. The first finding of the chapter is that if the stratum-specific response rates are correctly specified, ERRs allocation performs better than PS allocation in terms of the variances of the estimates. In practice, however, it may be difficult to precisely estimate the stratum-specific response rates prior to sampling. In such cases, approximate response rates based on experience need to be used. On the basis of the numerical results obtained:

(a) ERRs allocation outperforms PS allocation if the total absolute distance of the ERRs from the true response rates is moderate,

(b) The total absolute distance of the ERRs from their weighted average and the total absolute distance of the true response rates from their weighted average do not appear to affect the aforementioned finding,

(c) When the difference between the total absolute deviations of the expected and of the true response rates is less than approximately half of the latter, ERRs allocation always performs better, irrespective of whether or not the individual response rates were correctly predicted.

In this chapter, statistics other than proportions were not investigated, because of the problematic nature of the distributional assumptions including the homogeneity of variances which would have to be made. However, given that the proportions were obtained as averages of specific indicators, we expect similar results to hold in more general cases.

When examining the variances in situations where response rates are inaccurately specified, considering factors such as the number of strata and their variability can substantially increase the complexity of the analysis. We aim to explore this further in a more general context for future research.

The allocation method described in this chapter may be applied to other sampling designs which include separate allocation steps for subsamples. These designs include multistage sampling where sample allocations taking into consideration the different ERRs in the different primary sampling units may be applied.

Chapter 5

A new scheme for assessing survey quality

The ERRs allocation provided an improvement to the sample component in the survey estimates. In addition to the sample component, there is another source of error, which is related to the quality of the measurement. In this chapter, we provide a new aspect for evaluating the quality of surveys. The framework presented here assesses the uncertainty with which a survey yields the result. This uncertainty is defined as the extent to which the results would be different if the original survey were repeated. The aspect presented here can be added to the TSE framework (introduced in Chapter 3.2.1). Within the basic concept of TSE, errors are defined as the deviation of a survey response from its underlying true value [60]. However, the concept of an underlying true value is problematic due to quality errors in complete enumerations, which motivates the analysis of replication surveys.

Our approach models the precision of the values found in a survey compared to a potential replication of the survey. We define nonresponse uncertainty (NU) and measurement uncertainty (MU), which refer to the sources of difference between two replications of surveys and can be linked to nonresponse and measurement error in the TSE. Unlike general methods that assess the reliability and validity of a given question, this new scheme assesses the uncertainties of the survey as a whole.

After a short introduction to measurement error in Chapter 5.1, the quality issues of the theoretical population values are outlined in Chapter 5.2. Then we present the replication survey framework and show how the total difference between two replication surveys can be decomposed in theory into NU and MU (Chapter 5.3). The new approach is also illustrated with a case study: two replications of the ESS are compared for some selected variables both at the sample and individual levels (Chapter 5.4).

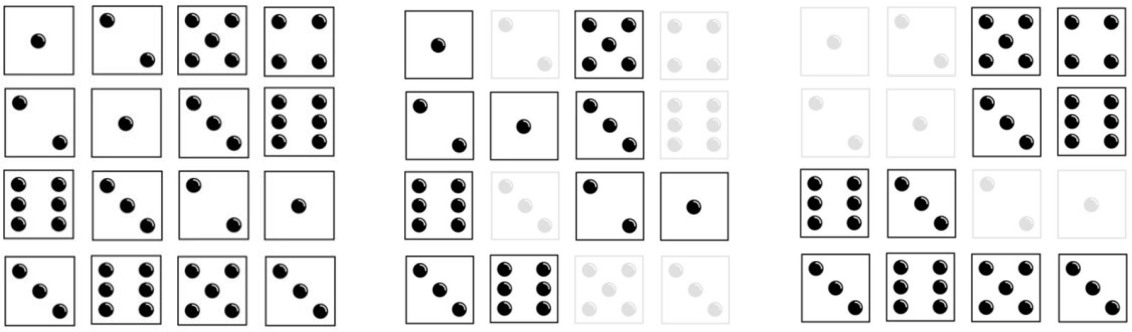
This chapter is accepted as:
B. Szeitl & T. Rudas (2024) "Assessing survey quality with a replication survey: nonresponse uncertainty and measurement uncertainty in the ESS" *Methods, data, analysis (MDA)*

The case study presented in this chapter is based on the OTKA research (nr. K 125162) "How to Approach the Unreachable? Pilot analysis of unit non-response and ineligible populations in empirical surveys by re-contacting". The research was led by Tóth István György and Blanka Szeitl at Tárki Social Research Institute.

5.1 Introduction

A major part of the inaccuracy of survey estimates is due to different types of measurement error [18]. The measurement error is between the construct (μ_i) and measurement (Y_i) sections in the TSE framework. It is one of the biggest challenges of human population surveys because it includes several human factors in the answering process (cognitive functions, social desirability, etc.) whose pattern is difficult to detect. In this chapter, we outline the most relevant methods for analyzing measurement error.

Drawing on the comparison to a traditional roll of dice, the concept of measurement error can be visualized in Figure 5.1. In the case of measurement error, we differentiate between random error and systematic bias. Figure 5.1a represents the total sample, i.e. the answers of 16 individuals to a selected question. Figure 5.1b represents the random error: There are six cases in which we cannot observe the true answers, but the measurement mechanism is independent of the values. Systematic bias in measurement procedures can be found in Figure 5.1c. The inaccurately measured values are influenced by the actual values of the experiments: this mechanism tends to affect the measurement of lower values, thus introducing bias to the measurement. Unlike nonresponse error, in the case of measurement error, we are unaware that there is a misrepresentation and there is a possibility that we consider the measured value as the true value for the specific respondent.



(a) n (b) random error (c) systematic bias

Figure 5.1: Measurement error using the analogy of the dice roll

Source: Own figure.

In the analysis of measurement error, we investigate how the responses of observed individuals vary and how this affects the accuracy of estimates [45]. The existing literature proposes various solutions to mitigate these effects during the data collection stage (random response, item count techniques, high level of anonymity [18, 45]). During data processing and analysis, measurement error is considered part of the overall quality of the survey questions [45]. There are techniques to estimate or predict the quality of the question, such as the Multitrait-Multimethod (MTMM) approach and the Survey Quality Program (SQP). Additionally, structural equation models can be used to model measurement error [43]. However, it should be noted that the exploration of measurement error within survey quality aspects is relatively limited, constraining the available solutions [41].

Before correcting for measurement error, we need to establish a model for the reliability and validity of the measurements [45]. In the measurement model, we investigate the relationship between two latent variables (f_1 and f_2). Their relationship is represented by the correlation coefficient $\rho(f_1, f_2)$, which can be estimated by the observed correlation coefficent of the measured variables, $\rho(y_1, y_2)$ [18]. The relationship between f_1 and y_1 and between f_2 and y_2 will not be perfect due to measurement errors (e_1 and e_2). The standardized effect of the variable of interest f_i on y_i is called the quality coefficient (q_i). If the latent variables and the errors are uncorrelated and the variables are standardized, the variance of the observed variables is 1, and it follows that:

$$var(y_i) = q_i^2 + var(e_i). \tag{5.1}$$

Because of this result, the quality of the ith question is $q_i^2 = 1-$var(errors) and the coefficient q_i is called the quality coefficient [45]. It can also be shown that:

$$\rho(y_1, y_2) = \rho(f_1, f_2)q_1q_2 \tag{5.2}$$

When the quality of the two variables decreases, the correlation coefficient $\rho(y_1, y_2)$ also decreases, but at a much faster rate. If the quality of the measurements is equal to 0.5 (the average quality in survey research, Alwin, 2007), then the quality coefficients q_i are 0.7 and the expected correlation between the observed variables will be only half the size of the correlation between the variables of interest. If the quality coefficients decrease to 0.6, then this correlation will be as small as one-third of the true value [45].
It is evident that measurement models are built on the assumption of an inherent true value, known as the latent variable. Following the quality indicators theory, these latent variables can be quantified, and the theory of variables is evaluated in relation to this latent variable. Chapter 5.2 outlines the potential issues associated with this concept.

5.2 The illusion of true population values

In the context of sample surveys, it is assumed that the population value being estimated is known and can be measured for all members of the population. However, the true value of population parameters can only be obtained through complete enumerations, i.e. censuses. Census data are crucial in sampling design and during the assessment and correction of the sample in certain aspects. The TSE approach to survey quality assumes that there exists a true population value that is not known and the goal is to estimate it [18]. In the TSE framework, each component of errors and biases is defined relative to the assumed true value, which in practice also represents the census value.

Censuses, on the other hand, face issues related to their true completeness. Some of these problems are theoretical, while others are relevant to the practice of survey sampling. The core of the theoretical problems is related to the superpopulation concept, which was previously discussed in Chapter 3.1.3.1. Based on this concept, a truly complete census cannot be achieved, as the population is not fixed and is constantly changing. According to the superpopulation concept, the census is also a survey, only with a very large sample size [54]. Practical issues with census values

are related to the accuracy of census data, particularly concerning the inclusion of all individuals in the enumeration and the accuracy of their responses to census questionnaires [68]. Census coverage differs from that of sample surveys because censuses tend to cover the entire population without sampling. Consequently, they are prone to nonresponse errors typical of sample surveys. Inaccurate responses in censuses lead to the same measurement errors seen in sample surveys, making it unnecessary to examine them separately in the context of censuses [68].

On the contrary, census coverage is a significant topic in the current statistical literature. This chapter presents some recent findings to establish a foundation for the concept of replication surveys.

The precision of census data is thoroughly examined by different institutions, including the United States Census Bureau (US CB), the Office for National Statistics in the United Kingdom (ONS), and the United Nations Economic Commission for Europe (UNECE). These entities offer comprehensive recommendations to guarantee the accuracy of censuses [69]. The techniques proposed by these bodies for evaluating quality utilize sample surveys to calculate potential under and overrepresentation in census data.

For the 2021 Census, the ONS has estimated coverage using a Census Coverage Survey (CCS), which was a post-census sample survey. The data was used in a capture-recapture structure linked to the census data, on which logistic regression models were used to estimate the effect of different demographic characteristics on the probability of response/coverage [68]. Their results suggest that the probability of response could not only be subject to non-response error, but also to systematic bias similar to non-response bias, as it varied somewhat for the main demographic variables such as gender and age (Figure 5.2 and 5.3).

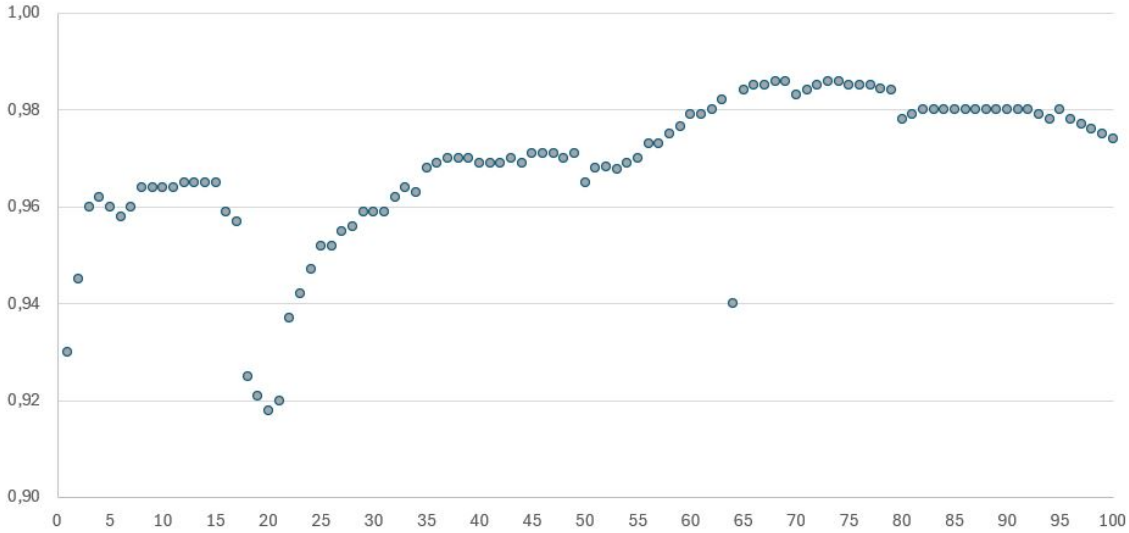


Figure 5.2: Age-gender undercoverage probabilities (female) in England and Wales

Source: Office for National Statistics, Census 2021, Coverage estimation for Census 2021 in England and Wales Methodology for coverage estimation of Census 2021 in England and Wales. pp. 9.

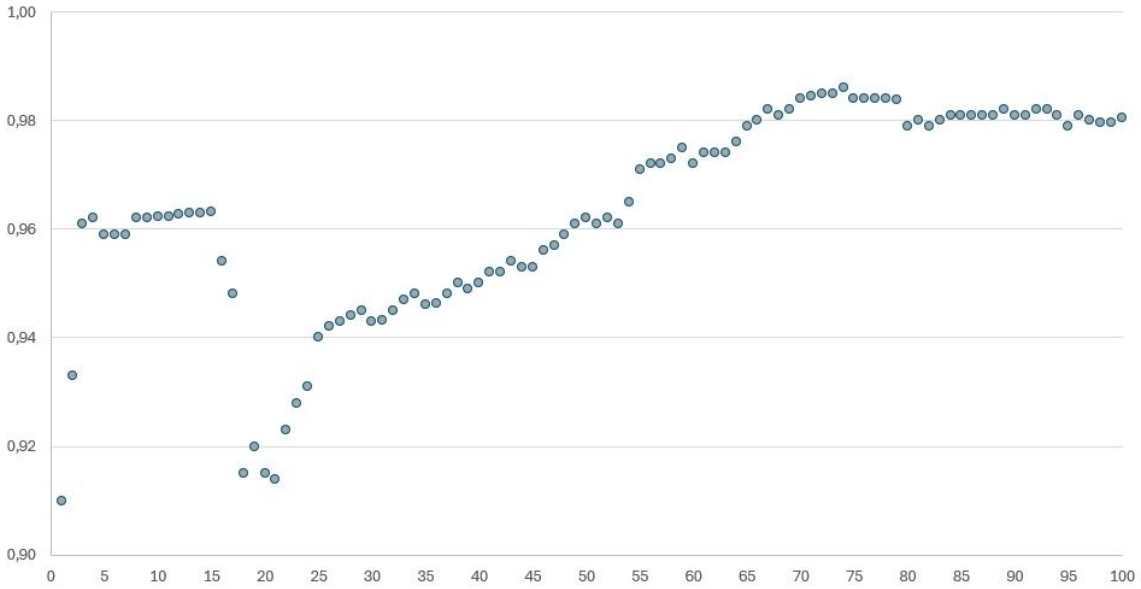


Figure 5.3: Age-gender undercoverage probabilities (male) in England and Wales

Source: Office for National Statistics, Census 2021, Coverage estimation for Census 2021 in England and Wales Methodology for coverage estimation of Census 2021 in England and Wales. pp. 10.

The CB uses an additional survey to improve the accuracy of their censuses. In contrast to the ONS, the CB's supplementary survey is conducted before the census data collection process, serving as a preparation tool. This survey also assesses public attitudes towards the census, which is a valuable indicator of the likelihood of encountering incorrect, false, or incomplete data [70]. The American Statistical Association (ASA) guidelines for handling incomplete responses highlight that this issue exhibits a demographic pattern, with notable differences between Black and Non-Black populations. To address this, the ASA references two pertinent benchmark data sources, namely the Demographic Analysis (DA) and the Population Estimates Program, both of which involve the collection of sample survey data [71].

A complete enumeration of the population eliminates any sampling errors; however, it does not guarantee error-free estimates. Various factors such as measurement errors, nonresponse errors, coverage errors, and processing errors can still be significant, and in some cases even more severe compared to sample surveys [72]. Even with advances in technology, these issues cannot be completely eliminated and may even introduce additional problems in data collection. Therefore, it is important to question the concept of true values in survey research. Especially since the census data, which is a benchmark for sample surveys, are in most cases also validated by additional (or even previous) sample surveys.

In the following chapters, we propose an alternative approach to assess survey quality: instead of an illusion of a true population parameter, our method addresses the issue of survey quality through replication surveys.

5.3 Theory

5.3.1 Replication surveys

A replication survey attempts to revisit the original survey, including not only the previously successful respondents, but also the part of the sample that could not be reached in the original survey [73]. According to the definition of replication surveys, it has two main objectives: (1) it acts as a reliability check of the original study or (2) it detects the validity and inconsistency of the measurements [73]. In the following, replication surveys are considered to be used for (1) purpose. It is also common to assess the reliability of the measurements in test-retest and longitudinal studies [74, 75, 43]. In these cases, the entire initial sample is not revisited, but only the successful sample from the original survey. The primary aim of these researches is to reduce measurement error in longitudinal studies. A great illustration is the dependent interview technique (DI) and the proactive dependent interview technique (PDI). DI and PDI use information from responses provided in previous interview rounds to modify the phrasing and routing of the questions in subsequent survey waves, as well as to facilitate edit checks within the interview [76]. Both approaches are widely used in panel surveys to achieve greater longitudinal consistency, lower levels of random measurement error, and to reduce the response burden [77].

Our method is not concerned with how to reduce errors but about presenting a new concept of measuring uncertainties. In our approach, uncertainties are defined relative to the potential repetition of an original study without any relevant changes in the research design. Replication surveys serve as a substitute for the notion of inherent true value, which is a fundamental aspect of the TSE framework.

5.3.2 Decomposition of the total difference of answers

In this chapter, we formally show the decomposition of the total difference of the answers from two replications of a survey. We consider the cases of continuous variables by decomposing the mean, and correlation coefficient, and discrete variables by decomposing the relative frequency of the i^{th} category, and χ^2-test statistics for independence. In the following, we present NU and MU in the joint attempted sample of potential first and second replications of a survey. The following notations are used for different groups of answers (Figure 5.4): set A denotes the answers of the total completed sample of the first replication of a survey, set B concerns the answers of the total completed sample of the second replication of a survey, set C denotes the answers of the group of those who responded only to the first replication of a survey, and set D denotes the answers of those who responded only to the second replication of a survey. Sets E and F refer to answers from those, who responded to both replications. Set E concerns the answers of the first survey, and set F concerns the answers of the second survey.

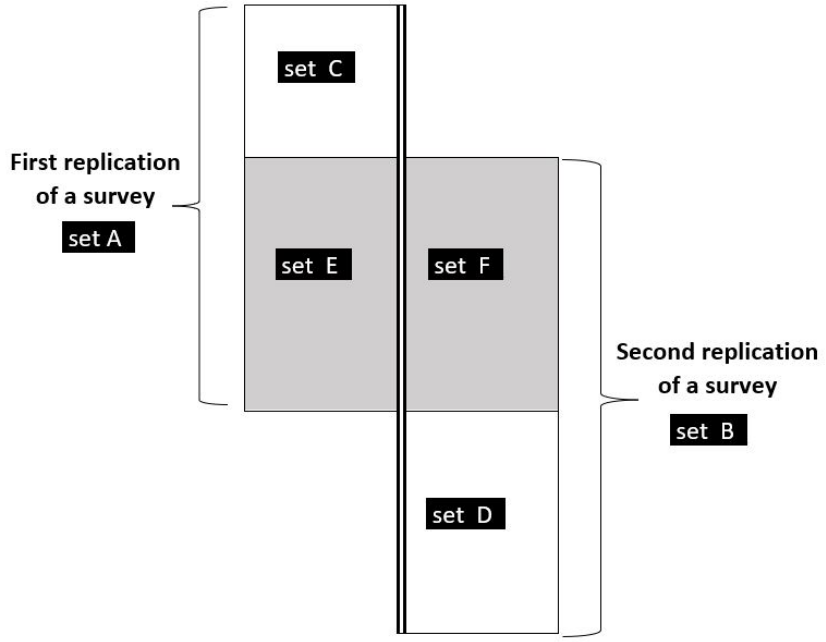


Figure 5.4: Different groups of answers in replication surveys

Source: Own figure.

NU depends on whether individuals who respond to the first replication of a survey give different answers from those who respond to the second replication of a survey, and thus NU captures the uncertainty of the base of responders/non-responders. Since there are respondents who did not participate in either replication of a survey and are therefore not included in this analysis, a particular kind of NU is studied: only in relation to the two replications of the survey. MU is the difference between the answers of the two replications of a survey from the same respondents. MU is defined both at the individual/respondent level and at the sample level. If MU occurs, there is an observation gap between the answers obtained in the first replication of a survey and the answers obtained in the second replication of a survey. Theorems 5.3.2.1 - 5.3.2.4 includes the decompositions of the mean, correlation coefficient, relative frequency of the i^{th} category of a discrete variable, and χ^2-test statistics for independence between two discrete variables.

Theorem 5.3.2.1 (Decomposition of the total difference of the mean)**.** *Let X denote a variable measured in both replications and $\bar{X}_A, \bar{X}_B, \bar{X}_C, \bar{X}_D, \bar{X}_E, \bar{X}_F$ be the mean of X in sets $A - F$ respectively. Let m_A, m_B, m_C, m_D, m_E, and m_F denote the sample sizes for each set, respectively. The decomposition of the difference of $\bar{X}$ between the first and second replications of a survey can be written as the weighted average of the differences between sets C and D and the weighted average of the differences between sets E and F. The decomposition of the difference is:*

$$\bar{X}_A - \bar{X}_B = \frac{m_C + m_D}{m_A + m_B}\left(\bar{X}_C - \bar{X}_D\right) + \frac{m_E + m_F}{m_A + m_B}\left(\bar{X}_E - \bar{X}_F\right), \tag{5.3}$$

where $\bar{X}_C - \bar{X}_D$ is the NU and $\bar{X}_E - \bar{X}_F$ is the MU.

Theorem 5.3.2.2 (Decomposition of the total difference of the correlation coefficient)**.** *Let X and Y denote two variables measured in both replications and let $r(X,Y)_A$, $r(X,Y)_B$, $r(X,Y)_C$, $r(X,Y)_D$, $r(X,Y)_E$, $r(X,Y)_F$ be the correlation coefficients of X and Y in sets A–F, respectively. The decomposition of the difference of $r(X,Y)$ between the first and the second replications is obtained as the weighted average of the Fisher's z-scores [78] of the differences between set C and set D and the difference between sets E and F. The Fisher's z-transformed correlation coefficients are denoted by $r'(X,Y)_A$, $r'(X,Y)_B$, $r'(X,Y)_C$, $r'(X,Y)_D$, $r'(X,Y)_E$, $r'(X,Y)_F$ in sets A–F, respectively. Following the standard Fisher's z-score method in Alexander [78], the decomposition of the difference is:*

$$\begin{aligned} r(X,Y)_A - r(X,Y)_B =& \frac{m_C + m_D}{m_A + m_B} \left(r'(X,Y)_C - r'(X,Y)_D\right) \\ &+ \frac{m_E + m_F}{m_A + m_B} \left(r'(X,Y)_E - r'(X,Y)_F\right), \end{aligned} \tag{5.4}$$

where $r'(X,Y)_C - r'(X,Y)_D$ is the NU and $r'(X,Y)_E - r'(X,Y)_F$ is the MU.

Theorem 5.3.2.3 (Decomposition of the total difference in the relative frequency of the i^{th} category)**.** *Let g_{iA}, g_{iB} denote the relative frequencies in the two replications and let ν_{iA}, ν_{iB}, ν_{iC}, ν_{iD}, ν_{iE}, ν_{iF}, the number of cases of category i. The decomposition of the difference in the relative frequency of a given category between the first and second replications is obtained as the weighted average of the differences between set C and set D and the difference between sets E and F. The decomposition of the difference is:*

$$g_{iA} - g_{iB} = \frac{m_C + m_D}{m_A + m_B} \left(\frac{\nu_{iC}}{m_C} - \frac{\nu_{iD}}{m_D}\right) + \frac{m_E + m_F}{m_A + m_B} \left(\frac{\nu_{iE}}{m_E} - \frac{\nu_{iF}}{m_F}\right), \tag{5.5}$$

where $\left(\frac{\nu_{iC}}{m_C} - \frac{\nu_{iD}}{m_D}\right)$ is the NU and $\left(\frac{\nu_{iE}}{m_E} - \frac{\nu_{iF}}{m_F}\right)$ is the MU.

Theorem 5.3.2.4 (Decomposition of the total difference in the χ^2-test statistics for the independence)**.** *Let X and Y denote two variables measured in both replications, let r denote the number of response categories of variable X and let ν denote the number of response categories of variable Y. The observed frequencies of each cell (ij) in sets $A-F$ are denoted with O_{ij_A}, O_{ij_B}, O_{ij_C}, O_{ij_D}, O_{ij_E}, O_{ij_F}, respectively and the expected frequencies of each ij cell are denoted with E_{ij_A}, E_{ij_B}, E_{ij_C}, E_{ij_D}, E_{ij_E}, E_{ij_F} for all sets respectively. If identical marginal distributions are assumed for X and Y, between sets C and E and sets D and F, the decomposition of the*

difference is:

$$\begin{aligned}\chi^2{}_A - \chi^2{}_B = \sum_{i=1}^{r}\sum_{j=1}^{\nu}\Bigg(&\left[\frac{1}{E_{ij_C}}\left(O_{ij_C}^2 - 2O_{ij_C}\right) - \frac{1}{E_{ij_D}}\left(O_{ij_D}^2 - 2O_{ij_D}\right)\right] + \\ &\left[\frac{1}{E_{ij_C}}\left(O_{ij_E}^2 - 2O_{ij_E}\right) - \frac{1}{E_{ij_D}}\left(O_{ij_F}^2 - 2O_{ij_F}\right)\right] + \\ &\left[\frac{1}{E_{ij_C}}\left(2O_{ij_C}O_{ij_E}\right) - \frac{1}{E_{ij_D}}\left(2O_{ij_D}O_{ij_F}\right)\right] + \\ &\left[\frac{1}{E_{ij_E}}\left(O_{ij_C}^2 - 2O_{ij_C}\right) - \frac{1}{E_{ij_F}}\left(O_{ij_D}^2 - 2O_{ij_D}\right)\right] + \\ &\left[\frac{1}{E_{ij_E}}\left(O_{ij_E}^2 - 2O_{ij_E}\right) - \frac{1}{E_{ij_F}}\left(O_{ij_F}^2 - 2O_{ij_F}\right)\right] + \\ &\left[\frac{1}{E_{ij_E}}\left(2O_{ij_C}O_{ij_E}\right) - \frac{1}{E_{ij_F}}\left(2O_{ij_D}O_{ij_F}\right)\right]\Bigg)\end{aligned} \tag{5.6}$$

If identical marginal distributions are not assumed, the expected frequencies of set $C + E$ and set $D + F$ cannot be given as a sum of the expected frequencies of the separate sets, but can be written as $E_{ij_C} + E_{ij_E} - \left(\frac{O_{i\cdot C}}{m_C} - \frac{O_{i\cdot C}+O_{i\cdot E}}{m_C+m_E}\right)$ and $E_{ij_D} + E_{ij_F} - \left(\frac{O_{i\cdot D}}{m_D} - \frac{O_{i\cdot D}+O_{i\cdot F}}{m_D+m_F}\right)$. In this case, the decomposition of the difference of the χ^2 test statistics becomes more complex, which will not be discussed further in this chapter.

It can be seen, that the total difference between the responses obtained in two replications of a survey can be decomposed exclusively into NU and MU in the case of continuous variables regarding the mean, and correlation coefficients, in the case of discrete variables regarding relative frequency and the χ^2-test statistics for independence. This means that if a survey is repeated, the total difference is due to a change in the respondent base and to the different answers of those who respond to both surveys.

5.4 Case study

5.4.1 Data and methods

The effect of NU and MU is illustrated with a case study using the 8^{th} wave of the ESS, which was performed twice in Hungary within the framework of a methodological research. The attempted sample size in the original ESS survey was $4,000$ addresses, of which, for budgetary reasons, $3,000$ addresses were randomly selected as the attempted sample in the replication survey of the ESS. Hereafter, the original ESS survey will be referred to as the first replication of the survey and the replication survey will be referred to as the second replication of the survey. The word "original" is not used in the new terminology to express that none of the surveys can measure the original response, each one measures a response that has some deviation from the true value. For the second replication of the ESS, all the first fieldwork conditions and standards regarding data collection were applied and the allocations of the addresses to interviewers were independent in the two replications. A shortened version of the questionnaire was used in the second replication of the survey by skipping some thematic block of questions. The wording and the order of the questions remained the same to minimize the contextual effect. The list of the repeated questions can be found in the Appendix Table 8.1.

The uncertainties are defined by comparing the answers of the first replication of the ESS and the second replication of the ESS. In the following, we present NU and MU in the common attempted sample of the first and second replications of the ESS ($n = 3,000$). Figure 5.5 presents the different sub-sets and their sample sizes in the replication of the ESS. Following the notations presented in Section 5.3.2 the sets are as follows: set A denotes the answers of the total completed sample of the first replication of the ESS, set B concerns the answers of the total completed sample of the second replication of the ESS, set C denotes the answers of the group of those who responded only to the first replication of the ESS, and set D denotes the answers of those who responded only to the second replication of the ESS. Sets E and F refer to answers from those, who responded to both replications. Set E concerns the answers of the first replication of the ESS, and set F concerns the answers of the second replication of the ESS. Compared to Figure 5.4, Figure 5.5 is supplemented with the unsuccessful addresses in both replications of the ESS.

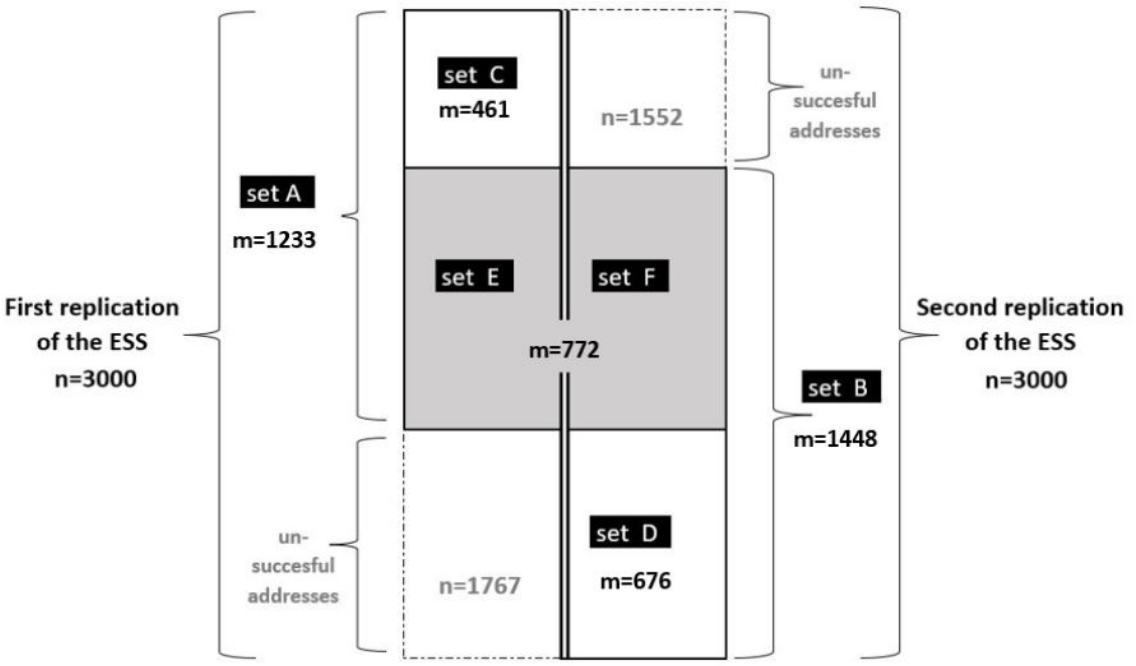


Figure 5.5: Different groups of answers and sample sizes in the replications of the ESS

Note: The columns on either side of the figure represent the two replications of the ESS in the same structure in which the different groups of answers of replication surveys in general are presented in Figure 5.4

The first ESS replication was fielded between May and September 2017, while the second replication was conducted between November 2018 and January 2019. The magnitudes of the uncertainties are only meaningful for variables that are stable over time. Therefore, variables were selected for the analysis, for which it can be assumed that their real values are stable in 20 months of time between the two replications. This ensures that the differences observed between the two replications are not due to a real change in opinions/attitudes, but are the results of measurement uncertainty. The selection procedure for the stable variables was based on the following two criteria:

(i) variables for which the real change in answers is not possible or extremely unlikely (i.e. logically assumed to be stable); or

(ii) variables for which ESS time series data (available for Hungary between 2016-2022, which can be found in the Appendix Figure 8.1) do not show statistically significant nonstationarity (i.e. may be considered stable in time). Stationarity is tested with the Dickey-Fuller test[1].

In Table 5.1 the variables selected for the analysis based on these criteria can be found. The distribution of all the variables included in the analysis is shown in the Appendix Table 8.2 for the total samples and for different subsets respectively.

[1]The Dickey-Fuller test is used to examine if a unit root is present in an autoregressive model. The alternative hypothesis of the test is to determine whether the model is stationary [79].

variable	type	levels of measurement	criteria	Augmented Dickey-Fuller test
Level of respondent's education	factual	ordinal scale (1-14)	i	-
Level of mother's education	factual	ordinal scale (1-14)	i	-
General trust	attitude	ordinal scale (1-10)	ii	stable
Religiosity	attitude	ordinal scale (1-10)	ii	stable

Table 5.1: Selected stable variables

Figure 5.6 shows the distribution of selected variables in the first (set A) and second (set B) replications of the ESS. In the case of all variables, significant differences can be found when comparing the results of the two replications of the survey. The χ^2-test is significant for the level of education ($p = 0,000$), mother's level of education ($p = 0,000$), general trust ($p = 0,000$), and religiousness ($p = 0,005$). It can be seen that the variables selected as stable on the basis of the criteria summarized in Table 5.1 still showed a difference in the replications.

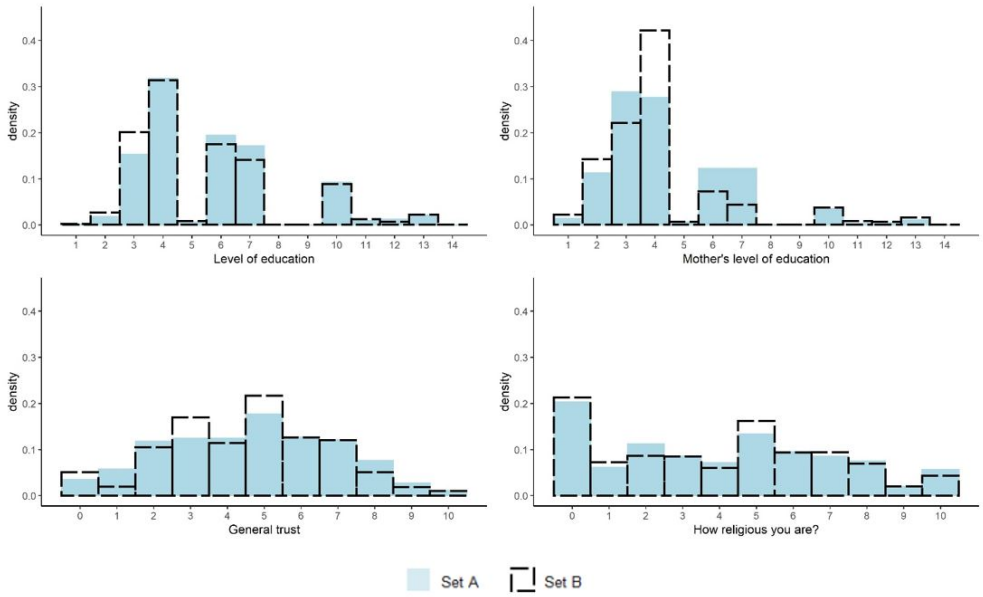


Figure 5.6: The total differences between the two replications of the surveys (set A – first replication; set B – second replication)

Source: Own figure.

In Chapter 5.3.2, the theory of the decomposition of the total difference between replication surveys was presented. Based on this theory, the total difference between the two replications of the ESS in Figure 5.6 is the sum of NU and MU. In the following, NU and MU are analyzed separately.

5.4.2 Nonresponse uncertainty

NU is related to the fact that the completed samples of the surveys are usually different. In Figure 5.7, the distributions of the variables in set C and set D can be found. Among the univariate distributions, there is no significant difference; the overall answers among sets C–D are similar. χ^2-tests were used to compare the

distributions. Based on the test results no significant difference is detected in the case of the respondent's level of education ($p = 0,337$), general trust ($p = 0,218$), and religiousness ($p = 0,128$). The distributions of the mother's level of education differed significantly ($p = 0,000$) between set C and set D. This means that regarding the distribution of the variables, in the case of 3 out of the 4 variables, there is no significant difference between the individuals who responded only to the first replication of the survey and those who responded only to the second replication of the survey.

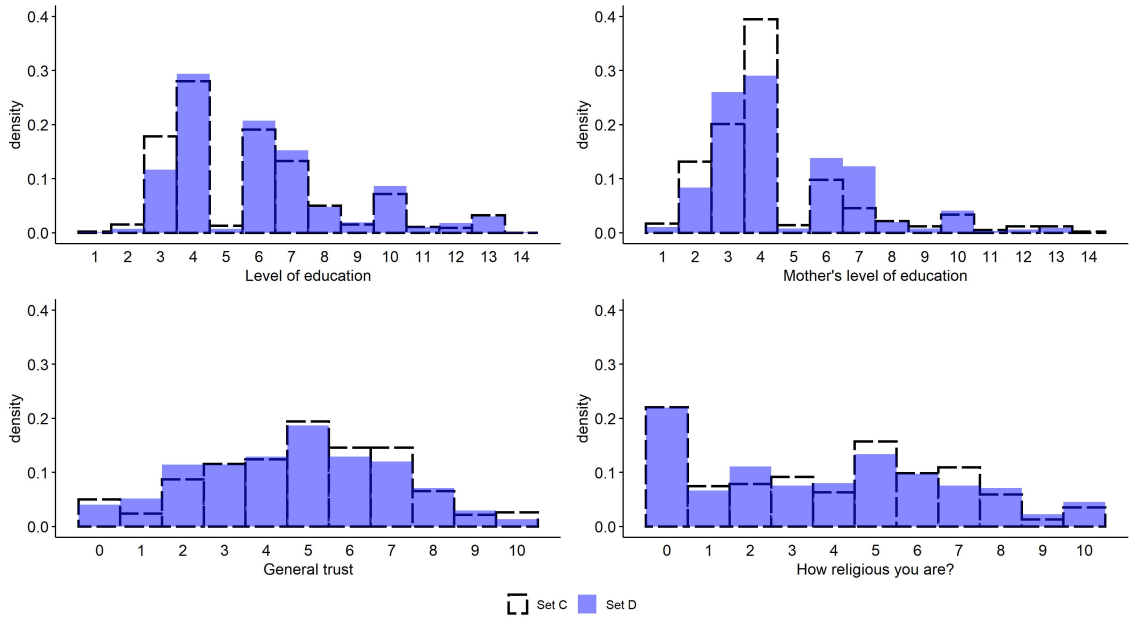


Figure 5.7: Distributions of answers of those who responded only to one survey

Source: Own figure.

5.4.3 Measurement uncertainty

MU is the uncertainty between the answers of those, responding to both replications of the ESS. Figure 5.8 shows the distributions of the variables obtained from the first replication (set E) and from the second replication (set F). Set E and Set F are composed of the same individuals: set E represents their first answers and set F represents their second answers. Overall, similar results are found when examining the distributions from the two surveys. χ^2-tests were used to compare the distributions, based on which no significant differences are detected in the case of the respondent's level of education ($p = 0,186$), but significant differences are detected in the case of the mother's level of education ($p = 0,000$), general trust ($p = 0,000$), and religiousness ($p = 0,006$). This means that at the sample level, MU is not relevant in the case of the respondent's level of education but relevant in the case of the other variables, thus, asking the same individuals twice results in a significantly different distribution.

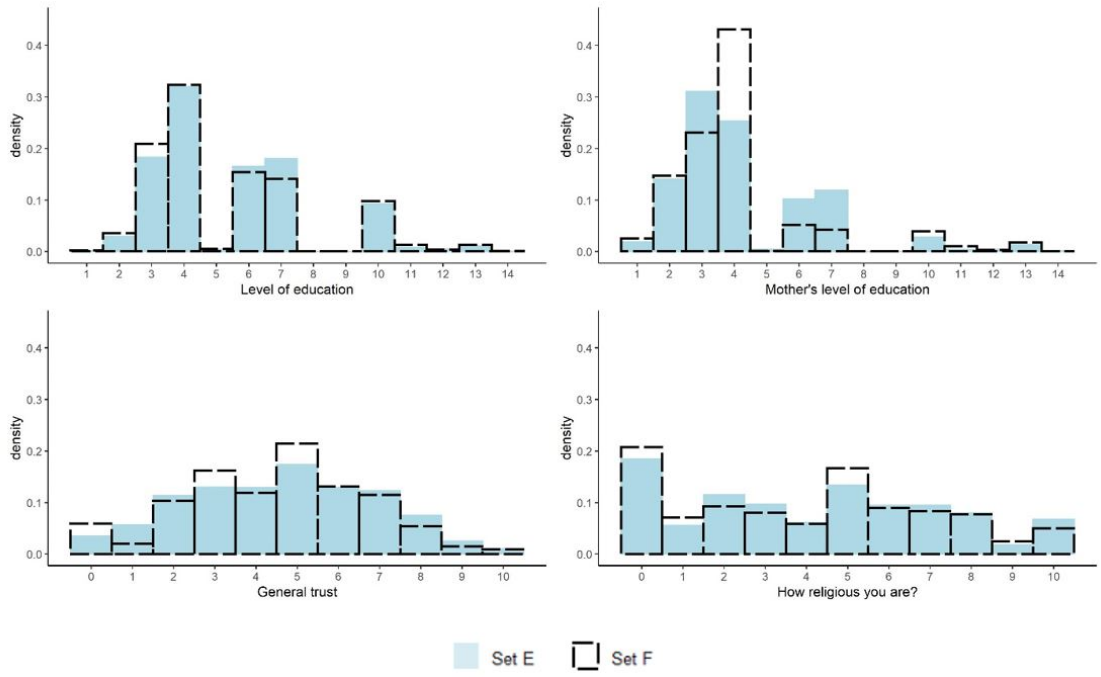


Figure 5.8: Distributions of answers of those who responded to both surveys

Source: Own figure.

In the following, this difference is presented at the individual level. Even if there was a short time between the two replications of the ESS, it is hard to exclude the real changes in the answers. Out of the four variables under consideration, there are two factual variables for which the probability of a real change is very low: the level of education and the mother's level of education. Since 20 months of time passed between the two replications of the survey, it is unlikely that a respondent's or the mother's highest level of education would increase by one category (a category covers an average of 4 years). Moreover, it is logically impossible for respondents or their mothers to have a lower level of education a year and a half later.

Figure 5.9 presents the differences between the first and second answers relative to the answers from the first replication of the ESS. If there were a real change in answers, the distributions of the difference would be skewed towards the high values (positive changes in the answers). It can be seen that for each variable the distributions of the difference are symmetric. The figure also shows that reporting of the unlikely or logically impossible changes in answers to education questions is common not only for the full sample but also for the sub-sample of respondents over 45 years of age (yellow charts), for whom a change in the highest level of education, and mother's highest level of education is even less likely due to advancing age. This underlines the fact that, although a real change cannot be completely excluded, it can rather be said that the difference is mainly due to MU. This uncertainty measured for factual variables is assumed to be present as large or even larger volumes in the case of the attitude variables. The figure also shows that for each variable, the mean of the difference (μ) is around 0.

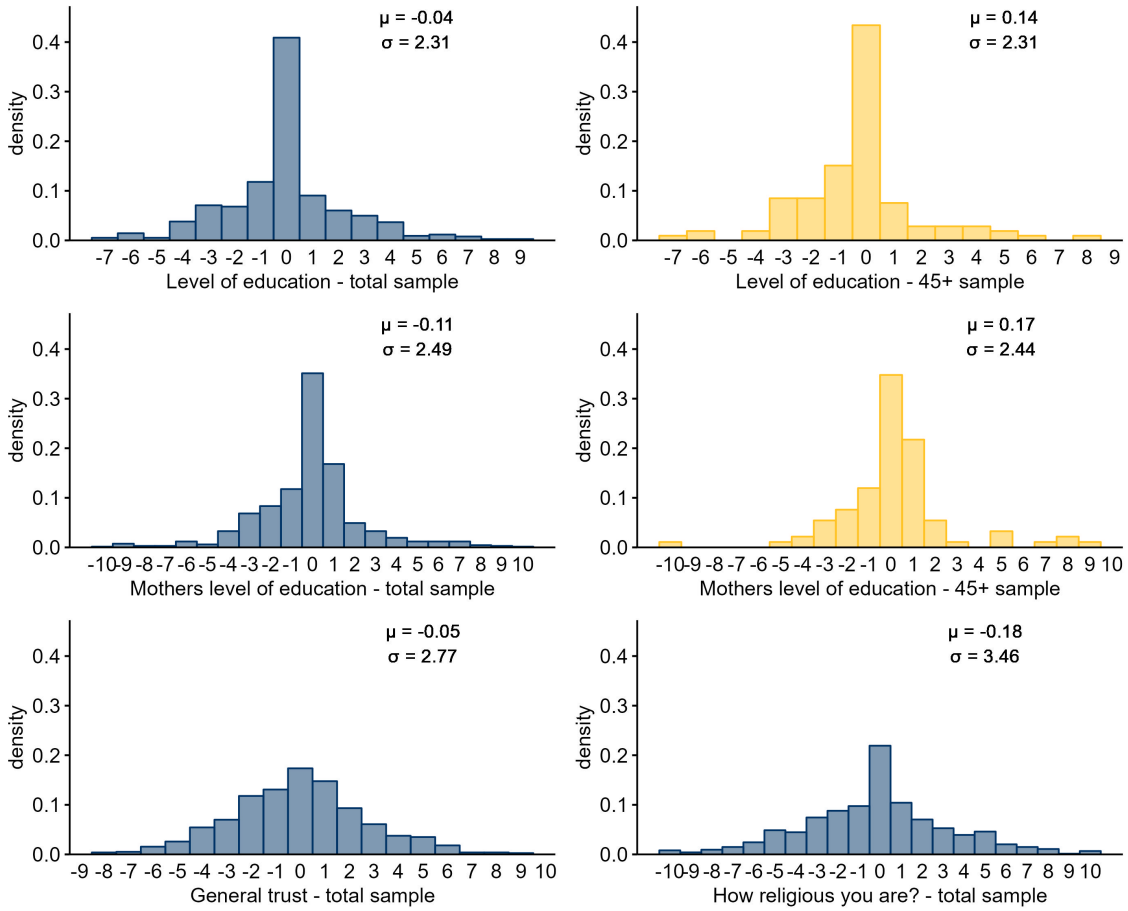


Figure 5.9: Difference between first and second answers (set E − set F)

Source: Own figure.

5.4.3.1 Regression to the mean

When repeated measurements are made on the same subject, regression to the mean (RTM) always occurs [80]. In the following, we show to what extent it can be detected in the case of replication surveys. The RTM was first discussed by Francis Galton in the late 1800s. Galton's famous example was the average height of fathers and their sons. He found that tall fathers had (on average) sons who were shorter than them, whereas short fathers had (on average) sons who were taller than them. That is, sons with fathers at the extreme tails of the distribution had heights closer to the population mean height. Galton assumed that the measurements are related to common genetic material and are observed with random errors. If values are observed with random errors, the observed values deviate from the true values in a positive or negative direction. The observed values in the extreme positive tail of the normal distribution during the first measurement are more likely to deviate from the true value by a positive error, whereas the observed values in the extreme negative tail of the normal distribution during the first measurement are more likely to deviate from the true value by a negative error. Due to the properties of the normal density, in the second observation, the former values will tend to be smaller, whereas observed values in the extreme negative tail of the distribution tend to be

greater during the second measurement.

Therefore, the natural change in repeated measurement is due to the characteristics of the normal density function, and it can be observed between any two variables when observations are measured with random errors, the variables are associated, and the joint distribution is normal.

In most analyzes, RTM is presented between continuous variables [81, 82, 83]; and we have not found treatments for the case of ordinal-scale variables. In this chapter, we show that RTM can also occur between ordinal-scale variables, and we illustrate it with the repeated answers of the two surveys.

Theorem 5.4.3.1 (Regression to the mean [84])**.** *Let X_1 and X_2 be random variables with joint distribution function F. Assume that X_1 and X_2 have the same marginal distribution and let μ denote their common mean. The distribution F exhibits regression to the mean if, for all $c > \mu$,*

$$\mu \leq \mathbf{E}[X_2|X_1 = c] < c,$$

and for all $c < \mu$,

$$\mu \geq \mathbf{E}[X_2|X_1 = c] > c.$$

To determine the extent of the deviation caused by the RTM the correlation coefficient (ρ) is involved [85]. The RTM effect is

$$\mathbf{E}[X_2|X_1 = c] = \mu + \rho(c - \mu).$$

If ρ is 1, then RTM is not present at all. The smaller the correlation between X_1 and X_2, the stronger the RTM effect, since X_2 is expected to be even closer to the mean. If ρ is 0, the difference between the two measurements is entirely due to the RTM phenomenon. If X_1 is large ($c > \mu$) then X_2 is expected to be smaller (if only $0 \leq \rho < 1$), and if X_1 is smaller than the mean, then X_2 is expected to be larger. In both cases, X_2 is expected to be closer to the mean than X_1. For ordinal-scale variables, the number of response categories is finite, thus in the negative/positive extreme ends of the scale, tending toward the mean definitely occurs because there is no room to take smaller/greater values. In the case of ordinal-scale variables, the strength of the association between the two measurements is also a key aspect: if there is no association between the first and the second measurement, the difference is entirely due to the RTM, while if the strength of the association is the strongest possible, then RTM is not present. The change in RTM effect between the two endpoints depends on the properties of the chosen association measure.

Figure 5.10 illustrates the appearance of RTM for the selected ordinal-scale variables of the ESS with a scatter plot of the change in answers (y-axis = answer from the first survey − answer from the second survey and x-axis = answer from the first survey). Respondents whose responses in the first survey were unusually low tended to give a higher response in the second survey (the difference between the second and first answers is negative), while respondents whose first responses were unusually high in the first survey tended to give a lower answer in the second survey (the difference between the first and second answers is positive), which means that the inconsistency of the responses that seemed like a random fluctuation is, in fact, an obligate change due to the RTM phenomenon.

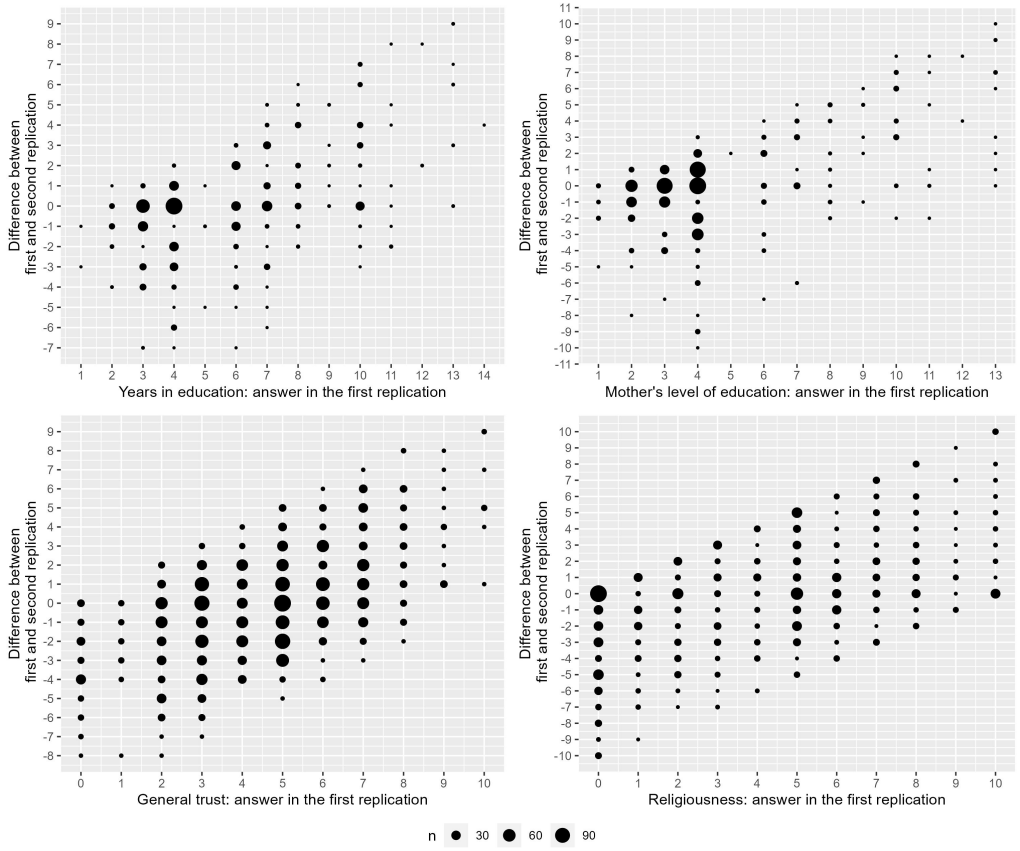


Figure 5.10: Difference between the first and second answers (first answers on the x-axis, and the difference between the first and second answers is on the y-axis)

Source: Own figure.

A figure showing the differences observed by the response categories is provided in the Appendix Figure 8.2. Regarding the potential demographic characteristics of the inconsistency in the answers, in the Appendix Figure 8.3 the observed differences are presented based on the gender of the respondents. It can be seen that no patterns appear to be found along certain values of the questions or along the gender variable.

In the case of general trust and religiousness, the RTM phenomena is presented from a different perspective (Figure 5.11 and Figure 5.12). In the case of general trust, the mean is 4.54 and the mode is value 5. In observing the RTM phenomenon the mean as the reference point to which the observations may regress toward is considered to be 5. In Figure 5.11, the values on the x-axis represents the first answers' absolute difference from the mean value (0 represents value 5 as answer) and the values on the y-axis represents the second answers' absolute difference from the mean value (0 represents value 5 as answer). The size of the bubbles represent the proportion of the given pattern. It can be seen, that in the case of answers with a greater difference relative to the mean value (values 3, 4, 5 on the x-axis) there is a higher share of those regression back towrd the mean in the case of their second answers. In the case of religiousness, the mean is 4.02 and the mode is value 0. This distribution is skewed to the right, which make the picture concerning RTM more complex. However, in observing the RTM phenomenon the mean as the reference point to which the observations may regress toward is considered to be 4. In Figure 5.12, the values on the x-axis represents the first answers' absolute

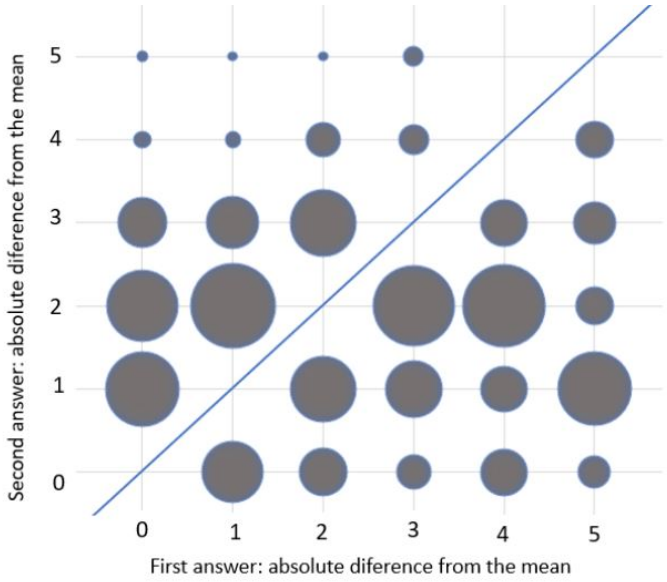


Figure 5.11: General trust: Responses relative to the mean values

Source: Own figure.

difference from the mean value (0 represents value 4 as answer) and the values on the y-axis represents the second answers' absolute difference from the mean value (0 represents value 4 as answer). The size of the bubbles represent the proportion of the given pattern. It can be seen, that in the case of answers with a greater difference relative to the mean value (values 4, 5, 6 on the x-axis) there is a higher share of those regression back towrd the mean in the case of their second answers.

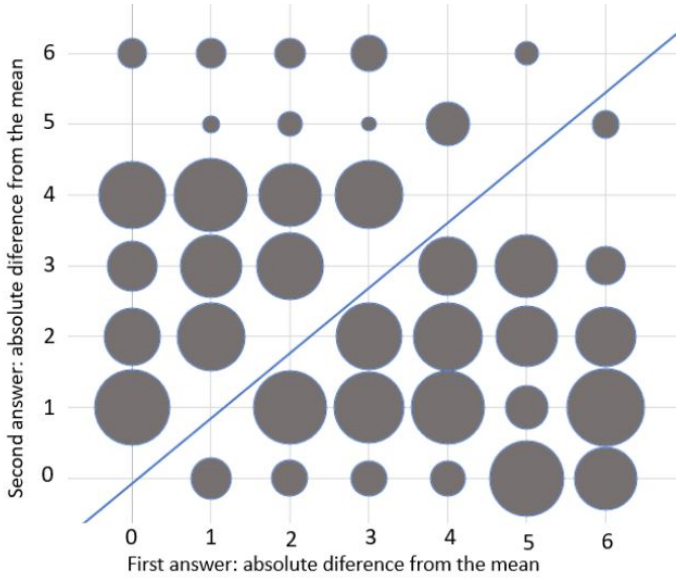
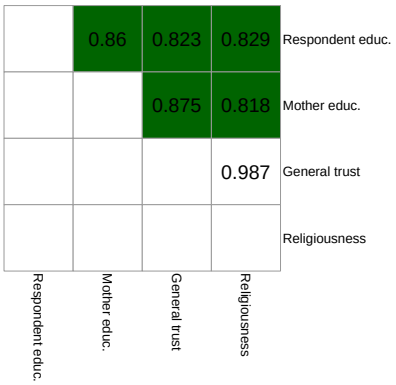


Figure 5.12: How religiousness you are? - Responses relative to the mean values

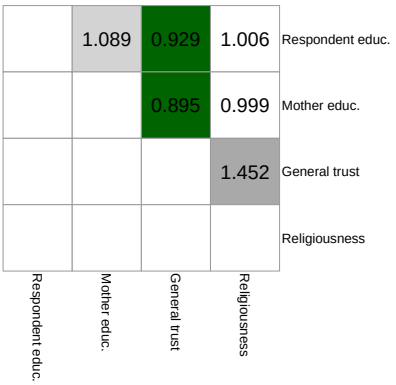
Source: Own figure.

5.4.3.2 Joint effects of the uncertainties

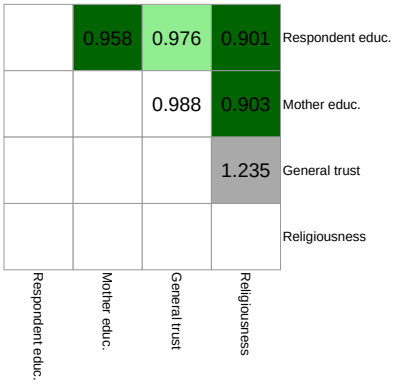
In this chapter, we present the extent of NU and MU when measuring the association between two ordinal-scale variables. The chosen association measure is Cramér's V, which also takes into account the different sample sizes of the different sets and the number of categories of different variables. It is based on Pearson's chi-squared statistic, which is decomposed into NU and MU following Chapter 5.3.2. The extent of NU is measured by the relative difference in the association measurement between the group of those who responded only to the first replication and those who responded only to the second replication of the ESS. The extent of MU is measured by the relative difference in the association measurement between the first and second answers of those, who responded to both replications of the ESS. The joint effect of NU and MU is measured by the extent to which the association measure differs between the total respondent base of the first and second replications of the ESS. In Figure 5.13 the proportional change in the Cramér's V values relative to the Cramer's V value in the first replication of the survey can be found.



(a) set C − set D



(b) set E − set F



(c) set A − set B

Figure 5.13: Differences in Cramér's V values

Source: Own figure.

It can be seen that in the case of NU (Figure 5.13a), most pairs of variables the change in the Cramer's V value is around $0,82$ and $0,99$ (i.e. the associations are weaker for the second replication than for the first replication of the ESS). In the case of MU (Figure 5.13b), the associations were both weaker and stronger for the

second replication than for the first replication of the ESS (the changes are between $0,89$ and $1,45$). As a result, the joint effect of NU and MU (Figure 5.13c) is minor: between the first and the second replications of the ESS, the total change in the associations is between 0.90 and $1,23$.

5.5 Discussion

When considering bias in survey data, there is usually a strong emphasis on comparing the respondent base with the population and identifying problems arising from "missing" individuals. Our results draw attention to another dimension: the instability of the respondent base and the instability in their answers are also relevant. In this chapter, we discussed a new aspect of assessing the quality of survey values involving instability with uncertainties. The new aspect can be added to the TSE framework by considering survey uncertainties compared to a potential replication of the survey. The chapter discussed the new approach in theory and with a case study of the 8^{th} wave of the ESS, which was performed twice in Hungary. We do not suggest that surveys should be replicated as a way of collecting data, but rather point out a possible new outlook on evaluation that could be the subject of further methodological research. The first finding of the chapter was that the total difference between two survey replications is the sum of NU and MU; therefore, the total difference was the combination of uncertainty about the respondent bases and uncertainty about the answers obtained. We found that for the univariate analysis, NU was negligible but relevant for the multivariate analysis. For MU, we compared the answers of those who responded to both surveys. The second finding of the study was that although the first and second answers generally resulted in the same distribution, on an individual basis, respondents appeared to be inconsistent with their answers. This phenomenon was explained with RTM, which occurs because values are observed with random errors. The third finding of the study was that, in multivariate analysis, both NU and MU are relevant, but their joint impact cause minor differences at the total sample level. The limitation of the method is related to the characteristics of the variables considered. Excluding real change is not always possible, but, in general, our results are applicable when data collections are relatively close in time, the information that the variable measures can consider stable, and when there are no other external effects that affect the population opinion. Another limitation of the study arises from the questionnaire: as a shortened version of the original questionnaire was used in the replication, the effect of the changed context may have some influence on the response process even if the order and wording of the questionnaire remained exactly the same. The analysis used the standard, validated questions of the ESS, however, their inherent measurement errors may be a factor if the respondents provide inconsistent answers. This effect would be difficult to distinguish and is beyond the scope of the chapter.

5.6 Further research: Models for measurement uncertainty

In this chapter, we present configurations that may cause the observed changes between the first and second replications of the survey. The content of this chapter

serves as a foundation for future research and will help shape the inquiries we aim to explore in the future.

We consider two models for the observational process and we examine the models for the correlation coefficients of a given variable which is observed in the surveys. The first model assumes that the results from the two surveys are observations with errors relative to the supposed true values of the variables, as in the theory of total survey error. The second model does not rely on the true values of the variables, rather it assumes that the second replication of the survey yields result with errors relative to the first replication of the survey.

Let X and Y be the supposed true values. Given that the same sample is observed twice, only MU is present. X_i and Y_i measure X and Y in the i-th surveys. We assume that the expected values exist and the standard differences are finite. According to the first model, denote by ϵ_i and δ_i independent normally distributed errors representing the difference from the true values in the i^{th} surveys. Then $X_1 = X + \epsilon_1$, $Y_1 = Y + \delta_1$, $X_2 = X + \epsilon_2$, $Y_2 = Y + \delta_2$. Since MU is a nonsystematic error, ϵ_i and δ_i are independent of X and Y, respectively. In Chapte 5.4.3.1, we found that the two surveys resulted in the same distributions at the sample level; thus, we assume that $\mathbf{E}(\epsilon_i) = \mathbf{E}(\delta_i) = 0$. In this model, the difference between the correlation coefficients can be written as follows:

$$\big|r(X + \epsilon_1, Y + \delta_1)\big| - \big|r(X + \epsilon_2, Y + \delta_2)\big|. \tag{5.7}$$

We expand the expression as follows:

$$\begin{aligned} &\left|\frac{\mathbf{E}((X+\epsilon_1) - \mathbf{E}(X+\epsilon_1))((Y+\delta_1) - \mathbf{E}((Y+\delta_1))}{(\mathbf{D}(X) + \mathbf{D}(\epsilon_1) + 2cov(X,\epsilon_1))(\mathbf{D}(Y) + \mathbf{D}(\delta_1) + 2cov(Y,\delta_1))}\right| \\ &- \left|\frac{\mathbf{E}((X+\epsilon_2) - \mathbf{E}(X+\epsilon_2))(Y - \mathbf{E}(Y+\delta_2))}{(\mathbf{D}(X) + \mathbf{D}(\epsilon_2) + 2cov(X,\epsilon_2))(\mathbf{D}(Y) + \mathbf{D}(\delta_2) + 2cov(Y,\delta_2))}\right| \end{aligned} \tag{5.8}$$

Then, we can simplify (5.8) as follows:

$$\begin{aligned} &\left|\frac{\mathbf{E}((X - \mathbf{E}(X))((Y) - \mathbf{E}(Y))}{(\mathbf{D}(X) + \mathbf{D}(\epsilon_1))(\mathbf{D}(Y) + \mathbf{D}(\delta_1))}\right| \\ &- \left|\frac{\mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))}{(\mathbf{D}(X) + \mathbf{D}(\epsilon_2))(\mathbf{D}(Y) + \mathbf{D}(\delta_2))}\right| \end{aligned} \tag{5.9}$$

The numerators in (5.9) are the same and positive, so to identify under what conditions the correlation coefficients increase, decrease or remain the same, it suffices to observe the following quotient:

$$A = \frac{(\mathbf{D}(X) + \mathbf{D}(\epsilon_1))(\mathbf{D}(Y) + \mathbf{D}(\delta_1))}{(\mathbf{D}(X) + \mathbf{D}(\epsilon_2))(\mathbf{D}(Y) + \mathbf{D}(\delta_2))} \tag{5.10}$$

If the value of $A = 1$, there is no difference in the correlation coefficients (for example, if we further assume that $\mathbf{D}(\epsilon_1) = \mathbf{D}(\epsilon_2)$ and that $\mathbf{D}(\delta_1) = \mathbf{D}(\delta_2)$). If the value of $A > 1$, the correlation coefficients in the second replication of the survey will be smaller than in the first replication of the survey (for example if $\mathbf{D}(\epsilon_1) < \mathbf{D}(\epsilon_2)$

and $\mathbf{D}(\delta_1) < \mathbf{D}(\delta_2)$). If the value of $A < 1$ the correlation coefficients in the second replication of the survey will be greater than in the first replication of the survey (for example if $\mathbf{D}(\epsilon_1) > \mathbf{D}(\epsilon_2)$ and $\mathbf{D}(\delta_1) > \mathbf{D}(\delta_2)$). Although this model with assuming observations with errors relative to the supposed true values may seem reasonable, it did not provide a natural explanation for the phenomenon of decreasing correlation coefficients.

In the second model, MUs are observed in the second replication of the survey relative to the first one, which are denoted by ϵ and δ. Then, $X_2 = X_1 + \epsilon$, and $Y_2 = Y_1 + \delta$. As in the case of the first model, here, ϵ and δ are independent of X_1 and Y_1, respectively, since these uncertainties are nonsystematic. From our previous finding in Chapter 5.4.3, $\mathbf{E}(\epsilon) = \mathbf{E}(\delta) = 0$. In this model, the difference between the correlation coefficients can be written as follows:

$$\big|r(X_1, Y_1)\big| - \big|r(X_1 + \epsilon, Y_1 + \delta)\big|. \tag{5.11}$$

We expand the expression as follows:

$$\begin{aligned} &\left|\frac{\mathbf{E}((X_1 - \mathbf{E}(X_1))(Y_1 - \mathbf{E}(Y_1))}{\mathbf{D}(X_1)\mathbf{D}(Y_1)}\right| \\ &- \left|\frac{\mathbf{E}((X_1 + \epsilon) - \mathbf{E}(X_1 + \epsilon))((Y_1 + \delta) - \mathbf{E}(Y_1 + \delta))}{(\mathbf{D}(X_1) + \mathbf{D}(\epsilon) + 2cov(X_1, \epsilon))(\mathbf{D}(Y_1) + \mathbf{D}(\delta) + 2cov(Y_1, \delta))}\right| \end{aligned} \tag{5.12}$$

Then, we can simplify (5.12) as follows:

$$\begin{aligned} &\left|\frac{\mathbf{E}((X_1 - \mathbf{E}(X_1))(Y_1 - \mathbf{E}(Y_1))}{\mathbf{D}(X_1)\mathbf{D}(Y_1)}\right| \\ &- \left|\frac{\mathbf{E}((X_1) - \mathbf{E}(X_1))((Y_1) - \mathbf{E}(Y_1))}{(\mathbf{D}(X_1) + \mathbf{D}(\epsilon))(\mathbf{D}(Y_1) + \mathbf{D}(\delta))}\right| \end{aligned} \tag{5.13}$$

The numerators in (5.13) are the same and positive, so to identify under what conditions the correlation coefficients change, is suffices to observe the following quotient:

$$B = \frac{\mathbf{D}(X_1)\mathbf{D}(Y_1)}{(\mathbf{D}(X_1) + \mathbf{D}(\epsilon))(\mathbf{D}(Y_1) + \mathbf{D}(\delta))} \tag{5.14}$$

If the value of $B = 1$ there is no difference in the correlation coefficients (for example, if we further assume that ϵ and δ are constants thus $\mathbf{D}(X_1) = \mathbf{D}(X_1) + \mathbf{D}(\epsilon)$ and $\mathbf{D}(Y_1) = \mathbf{D}(Y_1) + \mathbf{D}(\delta)$). However, if this is the case, ϵ and δ are constant 0, when the second replication of the survey results in observations without any MUs. If the value of $B > 1$, the correlation coefficients in the second replication of the survey will be smaller than in the first replication of the survey; however, since standard differences cannot be negative, based on this model, if any MU exists, the correlation coefficients will decrease.

With the second model of the observational process, we can describe decrease in the absolute value of the correlation coefficients. This phenomenon can be linked to Shannon's theory of entropy [86], which proposes a measure for the amount of uncertainty or entropy encoded in a random variable. This theory is employed to model communication systems in which a message is transmitted through a noisy channel from a source to a destination. In the second model, the source is the result of the first survey, the destination is the result of the second survey, and the noisy channel is the MU between the two data collections. The definition of Shannon's entropy can be extended to two random variables as joint entropy. The joint entropy is the sum of the univariate entropies only if the variables are independent; in all other cases, it depends on several other parameters that are not closely related to the topic of this thesis but highlights the complexity of the issue.

Chapter 6

Summary

Survey statistics deal with data from questionnaire surveys. The thesis focuses on a special field of this, sample-based human population surveys. In human population surveys, the measurement is that the sampled individuals answer a question and these answers are used to estimate a population parameter.

Human population surveys are the main basis for public, economic, commercial, and political decisions, thus the reliability of the estimates is crucial. Election forecasts based on survey data have been predicting election outcomes with increasing error over the last 20 years, suggesting methodological reassessment and improvement of data collection and estimation procedures (Chapter 2). This increasing inaccuracy is the main motivation for the thesis. The reliability of human population surveys depends on two factors: (1) how well the sample represents the population under study, and (2) how well the responses reflect the true population values.

The thesis presented the mathematical foundations of human population surveys (Chapter 3) and outlined the main sampling procedures. From a mathematical point of view, one of the most important properties of survey research is that the sample is drawn from finite populations and, therefore, the classical limit theorems of mathematical statistics do not apply. The thesis describes the finite population application of the central limit theorem and the underlying super-population concept (Chapter 3.1).

In collecting data from people, in addition to the mathematical aspects, there are human factors to consider, as individuals may refuse to be included in the sample or may deliberately or accidentally give false answers to questions. These phenomena hinder data collection from being carried out in precise rules of mathematical statistics and lead to biases in the estimates.

The thesis presented the total survey error framework and a structure of how the total potential error can be divided into a set of mathematical and non-mathematical factors (Chapter 3.2). Mathematical factors can be well controlled by precise sample selection and appropriate post-correction procedures. Therefore, it can be assumed that errors related to human factors are more responsible for the current inaccuracies.

Two sources of error have been addressed in detail in the thesis: (1) the nonresponse error associated with the composition of the sample and (2) the measurement error associated with the accuracy of the answers. In this thesis, we provide a new correction method and a new assessment scheme to deal with these errors.

Chapter 4 presented a new sample allocation method to efficiently correct for bias due to sample composition. Our method allocates sub-sample sizes from population strata inversely proportional to expected response rates (ERRs). Based on the results of the evaluation of the new method, we can say that if response rates are correctly predicted, the allocation of ERR always results in a lower variance in the estimates than the allocation method currently in practice. This is proven by theorems on mean inequality.

Simulations have also been used to illustrate that even for miscalculated ERRs, the new allocation method often performs better. Important considerations here are, for example, the difference between expected and actual response rates and the dispersion of the actual response rates of the strata.

Our new allocation method addresses the problem that a decreasing response rate in survey research results in a biased sample composition, and therefore increases the inaccuracy of the estimates. This effect can be controlled by our method.

In the thesis, a new scheme for survey assessments is introduced: the scheme of replication surveys (Chapter 5). We show that the replication survey framework provides a unique opportunity to simultaneously investigate the factors most responsible for the error of the estimates, nonresponse and measurement error. This would not be possible in any other way. The main difference between the replication survey framework and the traditional approach (e.g. the total survey error framework) is that the quality of a survey is assessed by the extent to which the same result would be obtained if the data collection were repeated. The total survey error framework measures the total error of a survey estimate relative to the true population value of the parameter under investigation. The new approach presented in this chapter assumes that, although the true population value exists, it cannot be measured by surveys. Every survey measures the true population value with error; therefore, treating a census value as a true parameter is illusory and results in a false benchmark. The structure of repeated surveys is first theoretically described, and the theorems were introduced to show that the variance of a repeated survey can be decomposed into a part due to sample composition (nonresponse uncertainty) and a part due to response uncertainty (measurement uncertainty). Theorems are presented for both discrete and continuous variables: for continuous variables, theorems are presented for the decomposition of the mean and the correlation coefficient, while for discrete variables, theorems are presented for the decomposition of the relative frequency of the ith category of a variable and the decomposition of the χ^2-test statistics.

We then present a case study along which the framework can be demonstrated. This case study summarized the results of a 2019 OTKA research in survey methodology with the participation of the author. The research repeated the 8^{th} wave of the ESS in Hungary and compared the results found along the sample composition and measurement dimensions. We found that the measurement error is more significant than the sample composition error component: respondents gave different answers even on objective parameters such as the respondent's or mother's education level for the first and second data collection. We also found that the regression to the mean (RTM) phenomenon is present for the measurement error. The results on RTM have so far only been related to continuous variables, but the results of the thesis show that it is also relevant for categorical variables.

The thesis also reported results that motivate further research on the topic. The author's further plans include: testing the repeated survey framework for more than two replicates and examining measurement uncertainty on a set of responses; building multivariate models to examine measurement error and, based on this, developing new data correction (post-stratification) procedures; including the mode of data collection in the studies (most notably, taking into account face-to-face, telephone and online data collections); and modeling uncertainty using the concept of entropy.

The thesis is based on the following three published papers

1. B. Szeitl and T. Rudas (2022): Reducing Variance with Sample Allocation Based on Expected Response Rates in Stratified Sample Designs, Journal of Survey Statistics and Methodology, Volume 10, Issue 4, September 2022, Pages 1107–1120,
https://doi.org/10.1093/jssam/smab021

2. B. Szeitl and T. Rudas (2024): Assessing survey quality with a replication survey: nonresponse uncertainty and measurement uncertainty in the ESS, Methods, data, analyses (MDA)

3. Messing, V., Ságvári, B., Szeitl, B. (2022): Is "push-to-web" an alternative to face-to-face survey?: Experiences from a "push-to-web" hybrid survey in Hungary. (In Hungarian) STATISZTIKAI SZEMLE (0039-0690): 100/3 pp 213-233 (2022)

Further publications of the author are

- Buda, J., Hajdu, G., Szeitl, B., Janky, B. (2023): A new method for the imputation of key indicators based on separate high-quality survey data. INTERNATIONAL JOURNAL OF PUBLIC OPINION RESEARCH (accepted)

- Messing, V., Ságvári, B., Szeitl, B. (2023): Respondings as expected? The effects of survey mode on estimates of sensitive attitudes in self-completion and face-to-face interviews of the European Social Survey. SURVEY RESEARCH METHODS (under review)

- Szeitl, B., Tóth, I. (2021): Revisiting the ESS R8 sample a year after – Lessons from a re-contact survey to test patterns of unit non-response in Hungary. Survey Methods: Insights from the Field. https://surveyinsights.org/?p=14864

- Simonovits, G., Kates, S., Szeitl, B. (2019): Local Economic Shocks and National Election Outcomes: Evidence from Hungarian Administrative Data. POLITICAL BEHAVIOR 41(2), 337–348.

Conference presentations by the author:

- European Social Survey Conference (*accepted*): *Responding as expected? The effects of survey mode on estimates of sensitive attitudes in self-completion and face-to-face interviews of the European Social Survey.* 8-10 July 2024

- Comparative Survey Design and Implementation Workshop: *Sampling innovations and data collection challenges in probability panels.* 18-20 March 2024

- General Online Research Conference (GOR): *Probability theory in survey methods (workshop).* 20-23 February 2024
- European Survey Research Association (ESRA): *The risk of nonresponse bias in online and mixed-method surveys.* 17-21 July 2023
- Workshop on data collection in survey methodology, Central Statistical Office of Hungary: *Biasing estimates from online survey data collections based on the characteristics of the Hungarian population. (In Hungarian)* 7 November 2023
- Workshop on Survey Climate and Trust in Scientific Surveys, University of Kassel: *The prevalence and potential bias of online surveys.* 4-5 October 2022
- General Online Research Conference (GOR): *Quantifying nonresponse and measurement uncertainty in surveys based on a replication of the European Social Survey.* 7-9 September 2022
- European Young Statistician Meeting (EYSM): *Controlling Unit-Nonresponse Bias During Within-Household Selection With Optimal Allocation and New Specification of Kish Grid.* 29 July - 2 August 2019
- European Survey Research Association (ESRA): *Controlling Unit-Nonresponse Bias During Within-Household Selection with New Allocation Involving Response Rates.* 15-19 July 2019

Chapter 7

Összefoglalás

A survey statisztika kérdőíves kutatásokból származó adatokkal foglalkozik. A dolgozat ennek egy speciális ágára a mintavételen alapuló lakossági survey adatfelvételekre fókuszált. A lakossági survey adatgyűjtések esetében a mérés az, hogy a mintába került személyek megválaszolnak egy kérdést és ezek a válaszok alapján becsüljük meg a populációs paramétert.

A lakossági survey kutatások legfontosabb alapjául szolgálnak a közéleti, gazdasági, kereskedelmi és politikai döntéseknek, így a becslések megbízhatósága kulcsfontosságú. A survey adatokon alapuló választási előrejelzések az elmúlt 20 évben egyre nagyobb hibával jelzik elő a választások kimenetelét, amiből arra következtethetünk, hogy indokolt az adatgyűjtési és becslési eljárások módszertani újraértékelése és fejlesztése (2. fejezet). Ez a növekvő pontatlanság tekinthető a dolgozat alapjául szolgáló kutatások legfőbb motivációjának. A lakossági survey-ek megbízhatósága leginkább két tényezőn múlik: (1) mennyire jól reprezentálja a minta a vizsgált populációt, illetve, hogy (2) mennyire jól tükrözik a válaszok a valódi populációs paramétereket.

A dolgozat bemutatta a lakossági survey-ek matematikai alapjait (3. fejezet) és vázolta a legfontosabb mintavételi eljárásokat. Matematikai szempontból a survey kutatások egyik legfontosabb tulajdonsága az, hogy a mintát véges populációkból vesszük, ezért a matematikai-statisztika klasszikus határérték tételei nem érvényesek. A dolgozat ismertette a centrális határeloszlás tétel véges populációs alkalmazását és az ennek alapját képező szuper-populációs koncepciót (3.1. fejezet).

Azáltal, hogy emberektől gyűjtünk adatokat, a matematikai szempontokon túl, számos emberi tényezővel kell számolnunk, hiszen a potenciális válaszadók visszautasíthatják a mintába kerülést, illetve megtehetik azt is, hogy egy adott kérdésre szándékosan vagy véletlenül nem a valódi választ adják. Ezek a jelenségek gátolják azt, hogy az adatgyűjtést a matematikai statisztika precíz szabályait betartva vigyük véghez és torzításokat eredményeznek a becslések esetében.

A dolgozat bemutatta a total survey error keretrendszerét és azt, hogy az összes potenciális hiba hogyan osztható fel a matematikai és az emberi tényezők csoportjára (3.2. fejezet). Látható, hogy a matematikai tényezők precíz mintaválasztással és megfelelő utólagos korrekciós eljárásokkal jól kontrollálhatóak. Emiatt feltehető, hogy a jelenleg tapasztalható pontatlanságokért inkább az emberi tényezők csoportjába tartozó hibák a felelősek.

A dolgozat két hibaforrással foglalkozott részletesen: (1) a mintakompozíciós tényezőkhöz kapcsolható nonresponse error-ral, és (2) a válaszok pontosságához

kapcsolható measurement error-ral. A dolgozatban ezen hibák kezelésére nyújtottunk korrekciós módszert, illetve alternatív szemléletet.

A 4. fejezet egy új minta allokációs módszert mutatott be arra vonatkozóan, hogy a minta-összetételből fakadó torzítást hatékonyan korrigáljuk. Módszerünk a a populációs rétegekből a várt válaszadási arányokkal (expected response rates, ERRs) fordítottan arányosan jelöli ki az alminták méretét. Az új módszer kiértékelésének eredményei alapján azt mondhatjuk, hogy amennyiben a válaszadási arányokat sikerül helyesen előre jelezni az ERR allokáció minden esetben alacsonyabb varianciát eredményez a becslésekben mint a jelenleg gyakorlatban lévő allolációs megoldás. Ezt a középérték egyenlőtlenégre vonatkozó tételek bizonyítják.

Szimulációk segítségével azt is illusztráltuk, hogy még hibás ERR-ek esetében is sokszor jobban teljesít az új allokációs módszer. Ekkor fontos szempont például az, hogy mekkora a különbség a várt és a valódi válaszadások között (azaz, hogy mekkor a hiba mértéke) és az, hogy mekkora a rétegek valódi válaszadási arányainak szórása.

Az új allokációs módszerünk arra a problémára nyújt megoldást, hogy az adatfelvételek esetében tapasztalható csökkenő válaszadási arány torz minta-kompozíciót eredményez és emiatt növeli a becslések pontatlanságát. Módszerünkkel ez a hatás kontrollálható.

A mérési hibával kapcsolatban a dolgozat egy új szemléletet vezetett be (5. fejezet). Azt mutattuk be, hogy a megismételt adatgyűjtések (replication survey) keretrendszere egyedülálló lehetőséget nyújt arra, hogy a becslések hibájáért leginkább felelős tényezőket, a nonrepsonse és a measurement error-t egyidőleg vizsgáljuk. Erre semmilyen más formában nem lenne lehetőség. A megismételt survey-ek keretrendszere leginkább abban különbözik a tradicionális szemlélettől (például a total survey error keretrendszerétől), hogy egy survey minőségét úgy ítéli meg, hogy mennyiben kapnánk azonos eredményt akkor, ha az adatgyűjtést megismételnénk. A total survey error framework egy survey becslés teljes hibáját a vizsgált paraméter valódi populációs értékéhez képest határozza meg. A dolgozatban bemutatott új szemlélet abból indul ki, hogy habár a valódi populációs érték létezik, az survey módszerekkel nem mérhető. Minden mérés hibával méri a valódi populációs értéket, emiatt egy cenzus során mért érték valódi paraméterként kezelése csak illúzió és hamis viszonyítási pontot eredményez. A megismételt survey-ek struktúráját elsőként elméletben ismertettük és bevezettük az arra vonatkozó tételeket, hogy egy survey ismétlése esetén az eltérés felbontható egy mintakompozícióból (nonresponse uncertainty) és egy válaszadási bizonytalanságból (measurement uncertainty) eredő részre. A tételek diszkrét és folytonos változók esetére is mutatnak be releváns eredményeket: folytonos változók esetére az átlag és a korrelációs együttható dekompozíciójára vonatkozó tételek, míg diszkrét változók esetében egy változó i-edik kategóriájának relatív gyakoriságára és a χ^2 próbastatisztika dekompozíciójára vonatkozó tételek szerepelnek.

Ezután egy esettanulmányt mutattunk be, mely mentén a keretrendszer demontrálható. Ez az esettanulmány a szerző részvételével zajló 2019-es OTKA kutatás eredményit összegezte. A kutatás során az ESS 8. magyarországi hullámát megismételtük és a talált eredményeket összevetettük a minta-kompozíciós és mérési dimenziók mentén. Azt találtuk, hogy a mérési hiba sokkal jelentékenyebb, mint a mintakompozícióból álló hibarész: a válaszadók még az olyan objektív paraméterekkel

kapcsolatban is különböző válaszokat adtak az első, illetve a második adatfelvétel esetében, mint például a válaszadó, vagy az anyja iskolai végzettsége. Azt találtuk továbbá, hogy a mérési hiba esetében jelen van az átlaghoz való visszatérés (regression to the mean, RTM) jelensége. Az RTM-mel kapcsolatos eredmények eddig kizárólag folytonos változókra vonatkoztak, a dolgozat eredményei alapján viszont látszik, hogy ez kategórikus változók esetében is releváns.

A dolgozat számos olyan eredményt is közölt, amely indokolttá teszi a téma további kutatását. A szerző további tervei között szerepel: a megismételt survey keretrendszer vizsgálata kettőnél több ismétlés esetére, illetve a mérési bizonytalanság vizsgálata válaszok sorozatán; többváltozós modellek építése a mérési hibák elemzésére és ennek alapján új adatkorrekciós (utólagos rétegzési) eljárások fejlesztése; az adatgyűjtési mód bevonása a vizsgálatokba (leginkább a személyes, a telefonos és az online adatgyűjtések figyelembe vétele); valamint a bizonytalanságok modellezése az entrópia fogalmával.

A disszertáció a szerző három publikációján alapul:

1. B. Szeitl and T. Rudas (2022). Reducing Variance with Sample Allocation Based on Expected Response Rates in Stratified Sample Designs, Journal of Survey Statistics and Methodology, Volume 10, Issue 4, September 2022, Pages 1107–1120,
https://doi.org/10.1093/jssam/smab021

2. B. Szeitl and T. Rudas (2024). Assessing survey quality with a replication survey: nonresponse uncertainty and measurement uncertainty in the ESS, Methods, data, analyses (MDA), *elfogadott*

3. Messing, V., Ságvári, B., Szeitl, B. (2022): Is "push-to-web" an alternative to face-to-face survey?: Experiences from a "push-to-web" hybrid survey in Hungary. (In Hungarian) STATISZTIKAI SZEMLE (0039-0690): 100/3 pp 213-233

A szerző további publikációi:

- Buda, J., Hajdu, G., Szeitl, B., Janky, B. (2023). A new method for the imputation of key indicators based on separate high-quality survey data. INTERNATIONAL JOURNAL OF PUBLIC OPINION RESEARCH (elfogadott)

- Messing, V., Ságvári, B., Szeitl, B. (2023): Respondings as expected? The effects of survey mode on estimates of sensitive attitudes in self-completion and face-to-face interviews of the European Social Survey. SURVEY RESEARCH METHODS (elfogadott)

- Szeitl, B., Tóth, I. (2021): Revisiting the ESS R8 sample a year after – Lessons from a re-contact survey to test patterns of unit non-response in Hungary. Survey Methods: Insights from the Field. https://surveyinsights.org/?p=14864

- Simonovits, G., Kates, S., Szeitl, B. (2019). Local Economic Shocks and National Election Outcomes: Evidence from Hungarian Administrative Data. POLITICAL BEHAVIOR 41(2), 337–348.

A szerző konferencia-megjelenései:

- European Social Survey Conference (*elfogadott*): *Responding as expected? The effects of survey mode on estimates of sensitive attitudes in self-completion and face-to-face interviews of the European Social Survey.* 2024. július 8-10
- Comparative Survey Design and Implementation Workshop: *Sampling innovations and data collection challenges in probability panels.* 2024. március 18-20
- General Online Research Conference (GOR): *Probability theory in survey methods (workshop).* 2024. február 20-23
- European Survey Research Association (ESRA): *The risk of nonresponse bias in online and mixed-method surveys.* 2023. július 17-21
- Adatgyűjtés-módszertani workshop, Központi Statisztikai Hivatal (KSH): *Az online survey adatgyűjtésekből származó becslések torzítása a magyar lakosság jellemzői alapján.* 2023. november 7.
- Workshop on Survey Climate and Trust in Scientific Surveys, University of Kassel: *The prevalence and potential bias of online surveys.* 2022. október 4-5
- General Online Research Conference (GOR): *Quantifying nonresponse and measurement uncertainty in surveys based on a replication of the European Social Survey.* 2022. szeptember 7-9
- European Young Statistician Meeting (EYSM): *Controlling Unit-Nonresponse Bias During Within-Household Selection With Optimal Allocation and New Specification of Kish Grid.* 2019. július 29 - augusztus 2
- European Survey Research Association (ESRA): *Controlling Unit-Nonresponse Bias During Within-Household Selection with New Allocation Involving Response Rates.* 2019. július 15-19

Bibliography

[1] Wayne A. Fuller. *Sampling statistics*. John Wiley Sons, Inc., Hoboken, New Jersey, 2009.

[2] *Statistics Canada. Quality Guidelines. Catalogue No. 12-539-XIE. Third Edition, October 1998.*

[3] Yadolah Dodge. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 2003.

[4] H. O. Hartley. Statistics as a science and as a profession. *Journal of the American Statistical Association*, 75(369):1–7, 1980.

[5] Yves Tillé. *Sampling and estimation from finite populations*. John Wiley Sons Ltd, 2020.

[6] W. G. Hansen, M. H. Madow. *On the History of Statistics and Probability*, chapter Some important events in the historical development of sample survey. New York: Marcel Dekker, 1974.

[7] Hald Anders. *A History of Mathematical Statistics from 1750 to 1930*. John Wiley Sons Ltd, 1998.

[8] Hald Anders. *A History of Probability and Statistics and Their Applications before 1750*. John Wiley Sons Ltd, 2003.

[9] Tabak John. *Probability and Statistics: The Science of Uncertainty*. Facts On File, Inc, 1998.

[10] Carl-Erik Särndal. Sample survey theory vs. general statistical theory: Estimation of the population mean. *International Statistical Review / Revue Internationale de Statistique*, 40(1):1–12, 1972.

[11] V. P. Godambe. Foundations of survey-sampling. *The American Statistician*, 24(1):663–685, 1970.

[12] Eurostat. Quality report on european statistics on international trade in goods – 2018-2021 data – 2022 edition. Technical report, Eurostat, 2021.

[13] Központi Statisztikai Hivatal. A ksh 2020–2021. Évi tevÉkenysÉge. Technical report, Központi Statisztikai Hivatal, 2021.

[14] ESOMAR. Global market research 2021. Technical report, ESOMAR, 2021.

[15] Pew Research Center. How public polling has changed in the 21st century. Technical report, Pew Research Center, Pew Research Center.

[16] Mellon J. Prosser C. he twilight of the polls? a review of trends in polling accuracy and the causes of polling misses. *Government and Opposition*, 53(4):757–790, 2018.

[17] AAPOR Executive Council. 2020 pre-election polling: An evaluation of the 2020 general election polls. Technical report, American Association for Public Opinion Research, 2020.

[18] Robert Groves, Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey Methodology*, volume 2. John Wiley Sons, 01 2009.

[19] Loosveldt G. Vandenplas C. Stoop I. Beullens, K. Response rates in the european social survey: Increasing, decreasing, or a matter of fieldwork efforts? *Survey Methods: Insights from the Field.*, 2018.

[20] Bates N. Burt G. Silberstein A. Atrostic, B. K. Nonresponse in u.s. government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, 17(2):209–226, 1970.

[21] De Heer W. De Leeuw, E. D. *R. M. Groves, D. A. Dillman, E. Eltinge, R. J. A. Little (Eds.), Survey Nonresponse*, chapter Trends in Households Survey Nonresponse: A Longitudinal and International Comparison. New York: Wiley, 2002.

[22] Murtaugh M. A. Edwards S. Slattery M. L. Rogers, A. Contacting controls: Are we working harder for similar response rates, and does it make a difference? *American Journal of Epidemiology*, 160(1):85–90, 2004.

[23] Presser S. Singer E. Curtin, R. Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, 69(1):87–98, 2005.

[24] E. Singer. Nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5):637–645, 2006.

[25] Handbook of Survey Research. *Survey Nonresponse. In P. Marsden J. D. Wright (Eds.).* Bingley: Emerald Group Publishing Limited, 2010.

[26] Cobben F. Schouten B. Bethlehem, J. G. *Handbook of Nonresponse in Household Surveys.* John Wiley, 2011.

[27] Williams D. Brick, J. M. Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1):36–59, 2013.

[28] F. Kreuter. Facing the nonresponse challenge. *The Annals of the American Academy of Political and Social Science*, 645(1):23–35, 2013.

[29] Bruce Meyer, Wallace Mok, and James Sullivan. Household surveys in crisis. *Journal of Economic Perspectives*, 29:199–226, 11 2015.

[30] Brick J. M. Williams, D. Trends in u.s. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, 9(1):1–26, 2017.

[31] É. Havasi. Válaszmegtagadó háztartások. *Statisztikai Szemle*, 75(10):831–843, 1997.

[32] European Central Bank. The household finance and consumption survey: Methodological report for the second wave, ecb statistics paper series no. 17. Technical report, European Central Bank, 2016.

[33] I. Gy. Szeitl, B. Tóth. Revisiting the ess r8 sample a year after – lessons from a re-contact survey to test patterns of unit non-response in hungary. *Survey Methods: Insights from the Field.*, 2021.

[34] T. W. Smith. The hidden 25 percent: An analysis of nonresponse on the 1980 general social survey. *Public Opinion Quarterly*, 47(3):386–404, 1983.

[35] I. Stoop. *G. Loosveldt, M. Swyngedouw and B. Cambré (Eds.) Measuring Meaningful Data in Social Research*, chapter No time, too busy. Time strain and survey cooperation., pages 301–314. Acco, 2007.

[36] J. Goyder. *The silent minority: Nonrespondents on sample surveys.* Boulder: Westview Press, 1987.

[37] Sturgis P. Purdon S. Campanelli, P. *Can you hear me knocking? An investigation into the impact of interviewers on survey response rates.* London: GB National Centre for Social Research, 1997.

[38] I. A. L. Stoop. *The hunt for the last respondent: Nonresponse in sample surveys.* The Hague: Sociaal en Cultureel Planbureau, 2005.

[39] Lepkowski J. M. Tucker, C. Telephone survey methods: Adapting to change. *Advances in Telephone Survey Methodology*, page 1–26, 2008.

[40] Presser S. Singer, E. *In J. M. Lepkowski,, N. C. Tucker, J .M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, R. L. Sangster. Advances in telephone survey methodology*, chapter Privacy, confidentiality, and respondent burden as factors in telephone survey nonresponse., page 447–470. New Jersey: John Wiley Sons., 2008.

[41] D.T. Oberski. *Measurement errors in comparative surveys.* PhD thesis, University of Tilburg, 2011.

[42] Duane F. Alwin. *Margins of Error: A Study of Reliability in Survey Measurement.* New York: Wiley, 2007.

[43] W. Saris and I. Gallhofer. Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, 1:29–43, 2007.

[44] L. J. Rips Tourangeau, R. and K. Rasinski. *The Psychology of Survey Responses.* New York: Cambridge University Press, 2000.

[45] Saris W. Loewe G. Ochoa C. Revilla, M. Can a non-probabilistic online panel achieve question quality similar to that of the european social survey? *International Journal of Market Research*, 57(3):395–412, 2015.

[46] Gy. Ziermann M. Éltető, Ö. Meszéna. *Sztochasztikus módszerek és modellek*. Közgazdasági és Jogi Könyvkiadó, Budapest, 1982.

[47] Thulin G. Backlund S. Atmer, J.G. Coordination of samples with the jales technique. *Statistik Tidskrift*, (13):443–450, 1975.

[48] A. B. Sunter. List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, (26):261–268, 1977.

[49] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer-Verlag New York, 2005.

[50] The ESS Sampling and Weighting Expert Panel. European social survey round 11 sampling guidelines: Principles and implementation. Technical report, European Social Survey, 2022.

[51] Ding P. Li, X. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112:1759–1769, 2017.

[52] Ed Stanek. Ideas on superpopulation models and inference. c00ed62.

[53] W. G. Cochran. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17:164–177, 1946).

[54] W. E. Deming and F. Stephan. On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36:45–49, 1941.

[55] W. G. Madow and L. H. Madow. On the theory of systematic sampling. *The Annals of Mathematical Statistics*, 15:1–24, 1944.

[56] W. A. Ericson. Subjective bayesian models in sampling finite population i. *Journal of the Royal Statistical Society B*, 31:195–234, 1969.

[57] P. Erdös and A. Rényi. On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 4:49–6, 1959.

[58] J. Hájek. Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5:361–374, 1960.

[59] W.G. Cochran. *Sampling Techniques. 3rd Edition*. John Wiley Sons, New York, 1977.

[60] P. Paul Biemer. Total survey error - design, implementation, and evaluation. *Public Opinion Quarterly*, 70:817–848, 5 2010.

[61] Hajdu G. Szeitl B. Janky B. Buda, J. A new method for the imputation of key indicators based on separate high-quality survey data. international journal of public opinion research. *under review*.

[62] D.E. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[63] Guillaume Osier. Unit non-response in household wealth surveys. Statistics Paper Series 15, European Central Bank, 2016.

[64] Katharine Abraham, Aaron Maitland, and Suzanne Bianchi. Non-response in the american time use survey: Who is missing from the data and how much does it matter? *Public Opinion Quarterly*, 70:676–703, 10 2006.

[65] Michael D. Larsen. Proportional allocation to strata. In P. J. Lavrakas, editor, *Encyclopedia of Survey Research Methods*, pages 630–630. SAGE Publications, Inc., 2008.

[66] Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis*. Springer-Verlag New York, 2007.

[67] Peter Southcott Bullen. *Handbook of Means and Their Inequalities*. Dordercht London: Kluwer Academic Publishers, 2003.

[68] Office for National Statistics (ONS). Coverage estimation for census 2021 in england and wales. Technical report, Office for National Statistics, 2022.

[69] United Nations. New frontiers for censuses beyond 2020. Technical report, United Nations, 2020.

[70] Deborah M. Stempowski and Maryann M. Chapin. A multidimensional quality assessment of the united states 2020 census. Technical report, United States Census Bureau, 2022.

[71] American Statistical Association 2020 Census Quality Indicators Task Force. 2020 census quality indicators: A report from the american statistical association. Technical report, American Statistical Association, 2020.

[72] Swensson B. Wretman J. Särndal, C.-E. *Model assisted survey sampling*. Springer-Verlag Publishing, 1992.

[73] Michael A. La Sorte. Replication as a verification technique in survey research: A paradigm. *The Sociological Quarterly*, 13(2):218–227, 1972.

[74] M. Hout and O. P. Hastings. Reliability of the core items in the general social survey: Estimates from the three-wave panels, 2006–2014. *Sociological Science*, 3:971–1002, 2016.

[75] R. Tourangeau. Survey reliability: Models, methods, and findings. *Journal of Survey Statistics and Methodology*, 9:961–991, 10 2021.

[76] B. Oberski D. Pavlopoulos D. Pankowska, P. Bakker. Dependent interviewing: A remedy or a curse for measurement error in surveys? *Survey Research Methods*, 15(2):135–146, 2021.

[77] Eckman S. Jäckle, A. Is that still the same? has that changed? on the accuracy of measuring change with dependent interviewing. *Journal of Survey Statistics and Methodology*, 8(4):706–725, 2019.

[78] Alexander E. A. A note on averaging correlations. *Bulletin of the Psychonomic Society*, 28(4):335–336, 1990.

[79] D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 366(74):427–431, 1979.

[80] Adrian Barnett, Jolieke van der Pols, and Annette Dobson. Regression to the mean: What it is and how to deal with it. *International journal of epidemiology*, 34:215–20, 03 2005.

[81] Samuel Schwab Simon Held, Leonhard Pawel. Replication power and regression to the mean. *Significance*, 17:10–11, 2020.

[82] A. Hrobjartsson and P.C. Gotzsche. Is the placebo powerless? an analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine*, 344:1594–1602, 2001.

[83] G.S. Kienle and H. Kiene. The powerful placebo effect: fact or fiction? *Journal of Clinical Epidemiology*, 50:1311–1318, 1997.

[84] Myra L. Samuels. Statistical reversion toward the mean: More universal than regression toward the mean. *The American Statistician*, 45:344–346, 11 1991.

[85] RT Mee and TC Chua. Regression toward the mean and the paired sample t test. *Am Statistician*, 45(1):39–42, 1991.

[86] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

List of Figures

Chapter 8

Appendix

Var	Label
ppltrst	Most people can be trusted, or you can't be too careful
pplfair	Most people try to take advantage of you, or try to be fair
pplhlp	Most of the time people helpful or mostly looking out for themselves
trstlgl	Trust in the legal system
trstplc	Trust in the police
trstplt	Trust in politicians
stflife	How satisfied with life as a whole
stfeco	How satisfied with present state of economy in Hungary
stfgov	How satisfied with the Hungarian government
edlvbhu	Highest level of education
eduyrs	Years of full-time education completed
edulvlfa	Father's highest level of education
edulvlma	Mother's highest level of education
hhmmb	Number of people living regularly as member of household
marital	Legal marital status
mainact	Main activity, last 7 days
wkhct	Total contracted hours per week in main job overtime excluded
uempnyr	Become unemployed in the next 12 months, how likely
hincsrc	Main source of household income
hinctnt	Household's total net income, all sources
hincfel	Feeling about household's income nowadays
rlgdgr	How religious are you
rlgatnd	How often attend religious services apart from special occasions

Table 8.1: List of common variables in the two replications of the ESS

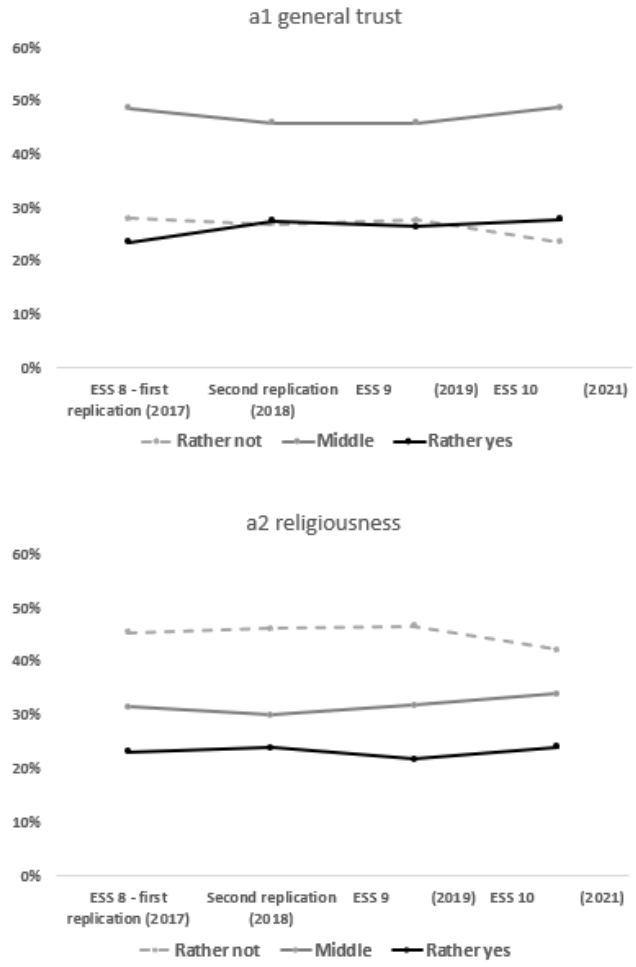


Figure 8.1: Attitude variables selected based on the ESS time series

Var	Type	Question	Category	n=1233	n=1448	n=461	n=676
a1	attitude	Most people try to take advantage of you, or try to be fair?	Most people try to take advantage of me	5	3.5	5	4
			1	2.1	5.8	2.4	5.2
			2	9.8	11.8	8.7	11.4
			3	14.9	12.6	11.5	11.4
			4	11.8	12.4	12.4	12.9
			5	20.8	17.6	19.3	18.6
			6	13.3	12.6	14.5	12.9
			7	13	12.1	14.5	12
			8	5.6	7.7	6.5	7.1
			9	1.9	2.7	2.2	3
			Most people try to take advantage of me	1.6	1.1	2.6	1.3
			missing	0.2	0.1	0.4	0.3
a2	attitude	Regardless of whether you belong to a particular religion, how religious would you say you are?	Not at all religious	20.9	19.5	21.9	21.4
			1	7.1	5.9	7.4	6.5
			2	8.6	11	7.8	10.8
			3	8.4	8.4	9.1	7.4
			4	5.9	6.8	6.3	7.8
			5	16.1	13	15.6	13
			6	9.2	9.3	9.8	9.5
			7	9.2	8.4	10.8	7.4
			8	6.9	7.4	5.9	7
			9	2	1.9	1.3	2.2
			Very religious	4.4	5.6	3.5	4.4
			missing	1.4	2.7	0.7	2.5
f1	factual	What is the highest level of education you have successfully completed?	Did not attend any school at all	0.2	0.1	0.2	0.1
			Primary school (1-4 classes) or equivalent	2.6	1.8	1.5	0.7
			Primary school (5-7 classes) or equivalent	18.7	14.5	17.8	11.5
			Completed Primary School or equivalent	29	29.6	28	29.1
			Certificate of Trade school	0.8	0.6	1.3	0.7
			Incompleted Secondary School	16	17.8	19.1	20.6
			Completed secondary school or equivalent	13.1	16	13.2	15.1
			With certificate of Intermediate Technological Educational Institute or equivalent, no university degree	6	4.5	5	4.7
			Higher form of vocational education	1.2	1.7	1.5	1.9
			Attended some years of Higher Education (at least 1 year) but not holding a Diploma	8.3	8.6	7.2	8.6
			Diploma in College	1.1	0.9	1.1	1
			Diploma in University	0.6	1.1	0.9	1.8
			Post-Graduate Diploma holder	1.9	1.9	3.3	3
			PhD holder	0.1	0.1	0	0.1
			missing	0.4	0.8	0	0.9
f2	factual	What is the highest level of education your mother successfully completed?	Did not attend any school at all	1.9	1.5	1.5	1
			Primary school (1-4 classes) or equivalent	12.4	10.9	11.9	8.1
			Primary school (5-7 classes) or equivalent	19.3	27.6	18.2	25.3
			Completed Primary School or equivalent	36.7	26	35.8	28.3
			Certificate of Trade school	0.6	0.6	1.3	0.7
			Incompleted Secondary School	6.1	11.5	8.9	13.5
			Completed secondary school or equivalent	3.8	11.6	4.1	12
			With certificate of Intermediate Technological Educational Institute or equivalent, no university degree	2.2	1.4	2	1.9
			Higher form of vocational education	0.9	0.9	1.1	0.7
			Attended some years of Higher Education (at least 1 year) but not holding a Diploma	3.2	3.3	3	4
			Diploma in College	0.7	0.3	0.4	0.3
			Diploma in University	0.6	0.5	1.1	0.6
			Post-Graduate Diploma holder	1.4	1.1	1.1	0.9
			PhD holder	0.1	0.1	0.2	0
			missing	10.1	2.8	9.3	2.7

Table 8.2: Summary of the variables involved

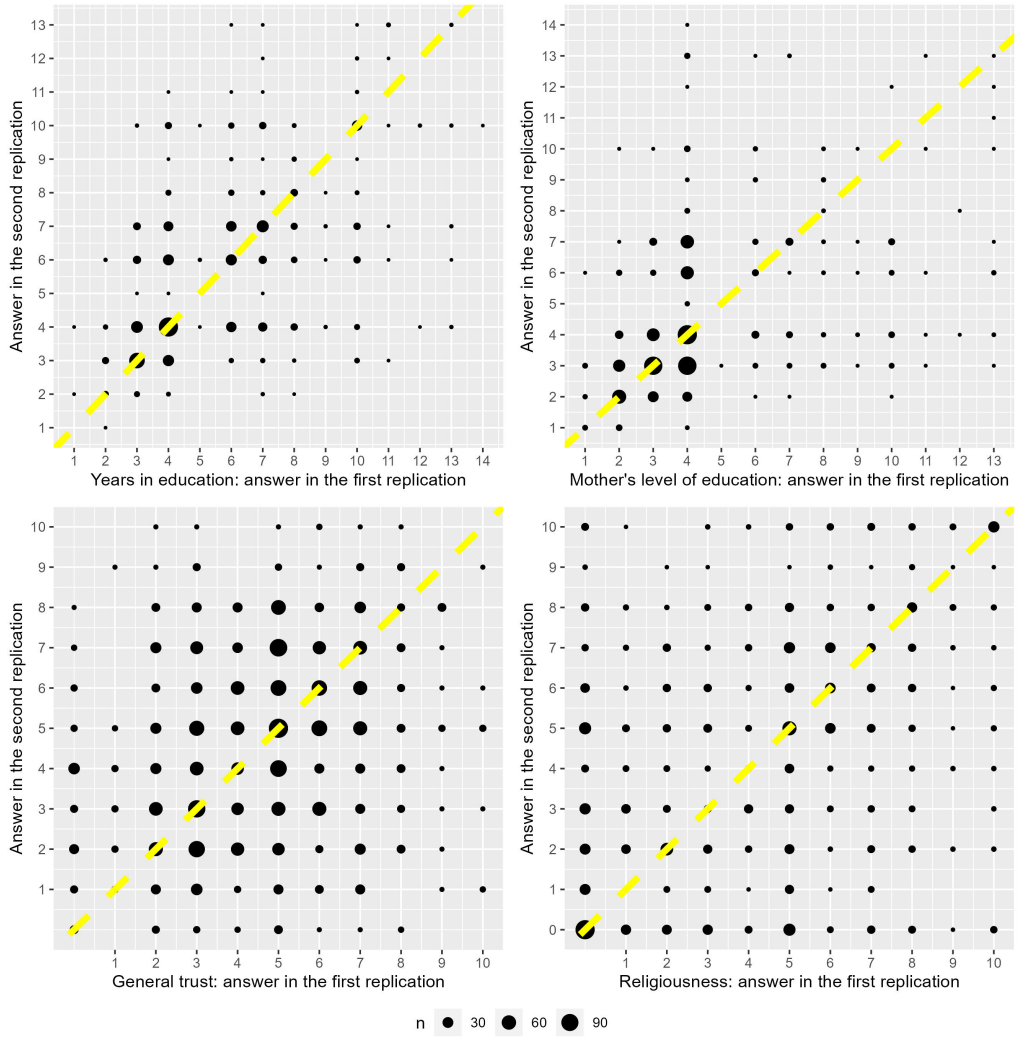


Figure 8.2: The differences in the answers between the first answers (x-axis) and the second answers (y-axis) of those responded to both surveys

Source: Own figure.

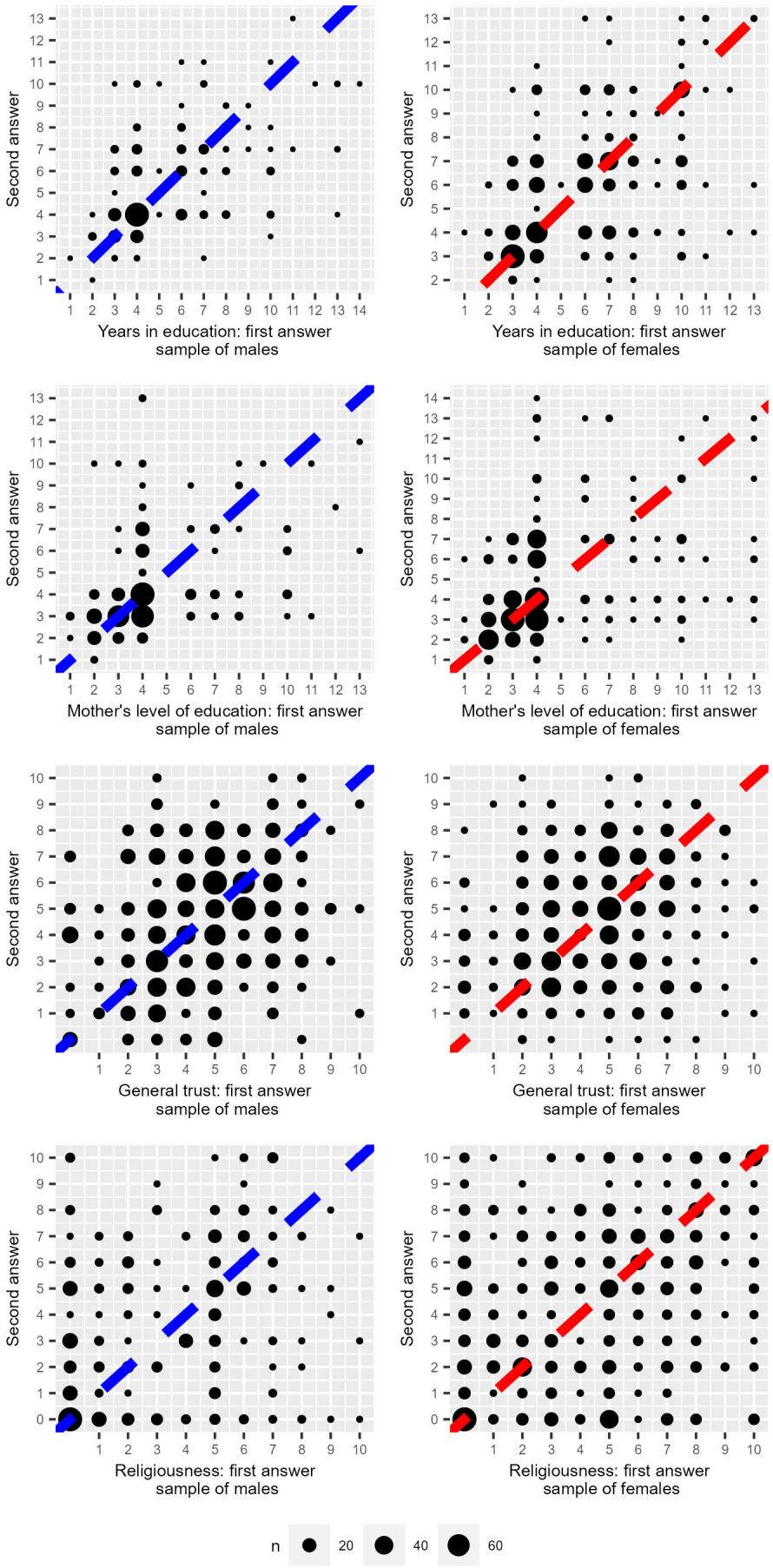


Figure 8.3: The differences in the answers between the first answers (x-axis) and the second answers (y-axis) of those responded to both surveys based on the gender of the respondent

Source: Own figure.