Chilean Journal of Statistics
Vol. xx, No. x, Month 20xx, 1–33

Dissecting Chilean Surveys: The Missing Outcomes Case

Ernesto San Martín [1234,*], and Eduardo Alarcón-Bustamante[123]

[1]Faculty of Mathematics, Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile.
[2]Millenium Nucleus on Intergenerational Mobility: From Modelling to Policy, MOVI.
[3]Interdisciplinary Laboratory of Social Statistics, Santiago, Chile.
[4]The Economics School of Louvain, Université Catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium.

(Received: 00 Month 200x · Accepted in final form: 00 Month 200x)

Abstract

The strengths and weaknesses of two Chilean political polls and the National Socioeconomic Characterisation Survey are analyzed from a statistical modelling point of view. The rationale of the analytical strategy is based on a distinction between identified parameters and parameters of interest, which is equivalent to make a distinction between *what we can learn from the data provided by a survey* and *what we want to learn from those data*. Using partial identification techniques, each survey is analyzed at different levels according to specific subpopulations. Based on these analyses, we emphasize not only the way in which the results should be reported, but also the necessity to make explicit the uncertainty induced by the non-response rates at the survey report.

Keywords: Partial identifiability; · missing data · quantile function · ignorability condition. · Non-response.

Mathematics Subject Classification: Primary 62P25 · Secondary 62D05.

1. Introduction

Broadly speaking, public surveys are applied either to get a better gauge of citizens' political opinions (Berinsky, 2017) or to collect information that is useful for policy makers. These surveys are perceived as reliable tools as it is argued that they are applied to "representative samples". If this were the case, the analysis of the strength of a survey would be reduced to indicating how a sample design ensures access to a "representative sample". However, it is necessary to emphasize that the expression "representative sample" is *not* a statistical concept because it is logically contradictory. As a matter fact, a survey is applied to know the behavior of a population in relation to an outcome of interest. Doing so means that we have no idea about this outcome: how then can we ensure the representativeness of the survey? On the other hand, if we know this outcome at the population level, why do we need to conduct a survey?

A question then arises: how can we assess a survey? This paper intends to answer this question in a specific but quite typical case, namely when some surveyed individuals do

*Corresponding author. Email: esanmart@mat.uc.cl

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)
© Chilean Statistical Society – Sociedad Chilena de Estadística
http://www.soche.cl/chjs

not answer a specific question. Our approach is based on the tension between the following two questions: what *can* be learned from the data provided by a survey?; and, what do we *want* to learn from those data? The difference between these two question relies on the statistical concept of *identifiablity*.

As a matter of fact, a statistical model is a family of probability distributions indexed by a parameter and defined on a sample space. From a modelling point of view, a set of data is fully represented by a probability distribution that generates them. Consequently, a parameter of this distribution represents a specific characteristic of the set of data under analysis; see Fisher (1922). Technically speaking, these correspond to the *identified parameter*. However, if we attribute a characteristic to a set of data that cannot be represented by a parameter (that is, it is not a functional of the probability distribution generating the data), then we face an identification problem. Technically speaking, these correspond to a *parameter of interest*. Thus, the identified parameters summarize what can be learned from the data, whereas the parameters of interest represent what we want to learn from the data. When an injective relationship is established between them, the identification problem is solved. For details and references, see Koopmans and Reiersol (1950); San Martín (2018), San Martín et al. (2015) and San Martín and González (2022).

In this paper we will use this conceptual distinction to assess both the strengths and weaknesses of three Chilean surveys: two of political opinion (CADEM survey and the Araucanía citizen consultation), and one related to the income distribution of employees (CASEN). We will analyze the identification problem raised by missing outcome. To do that, we will use Manski's technique of partial identification, which allows us to evaluate how strong are the ignorability conditions (also known as *missing at random* condition) typically used to impute missing data. Based on this discussion, we will emphasize the way in which these survey should report their results.

Let us remark the type of conclusion that can be done from a partial identification analysis. Typically, an identification analysis allows a parameter of interest to be point identified. For instance, in a fixed effect ANOVA model, the mean of the observations nested into a same group (e.g., scores of students of a specific school) is parameterized as an addition of two parameters, namely $E(Y_{ij}) = \alpha + \theta_j$, where j labels the groups and i labels the statistical units. Let us call θ_j, *parameter of group* j), and α, *global parameter*. The group parameters are point identified if, for example, the group parameter of the first group is assumed to be equal to 0. In this case, the group parameter of a specific group is equal to the difference between the mean of that group and the mean of the first group (this explains why this identification constraint is known as *deviation from the mean*). However, a partial identification analysis provides an identification region to which the parameter of interest belongs, rather than identifying it pointwise. This is due to the fact that an identification analysis makes explicit certain assumptions (identification restrictions) under which the parameter of interest is point identified, but, in the context of application, such a restriction is incredible (Manski, 2011, 2020). Therefore, the analysis strategy consists of relaxing such assumptions in order to establish a region to which this parameter belongs. The reader may ask where is the disadvantage of accepting incredible identification constrains in order to point identify the parameters of interest. The drawback lies in the fact that scientific conclusions and/or policy recommendations depend more on such constrains than on the data and, consequently, an illusion of scientific certainty is created based only on incredible certainty.

These considerations will be illustrated through the dissection of three Chilean surveys. This paper is accordingly organized as follows. In Section 2 the political opinion survey CADEM is analyzed. Section 3 focus its attention on the National Socioeconomic Characterisation Survey CASEN. Finally, Section 4 analyze a recent citizen consultation applied in the Araucanía region in the south of Chile. In each of these sections, we provide the

corresponding methodological information of each survey and also the political and/or economical context in which the survey is used. The paper ends by a general discussion.

2. CADEM survey

We begin by dissecting the CADEM political opinion survey. After describing the purpose of the survey and summarizing the methodology used to deal with missing data, we perform a conditional identification analysis of different sub-populations of interest.

2.1 General objective and methodological information

According to the information provided on its website, the CADEM survey is one of the many services offered by the market research company *CADEM research & estrategia*. Specifically, it is related to the service called *Plaza Pública*, which describes itself as "the first and only polling platform that measures public opinion on a weekly basis to provide data and analysis on a wide range of topics of interest"[1]. This particular aspect is related to one of the general objectives of this marketing company: "We want to connect people with decision makers, through data and not from intuition, providing strategies and action plans to achieve the expected results based on a deep knowledge of the new consumer/citizen"[2].

CADEM survey delivers "reliable, timely and contingent information on the political, economic and social debate in Chile on a weekly basis". The study published by CADEM "contemplates a probabilistic survey of 700 weekly cases (with a monthly consolidation that goes from 2,800 surveys to 3,500 depending on whether the month has 4 or 5 weeks), applied 100% through cell phones, using CADEM's own database that contains more than 18 million cell phones considering both prepaid and postpaid numbers, all obtained through Random Digit Dialing and consolidated during the last four years". Its target group is, therefore, all individuals living in the national territory, Chileans and immigrants, men and women over 18 years old, inhabitants of the 15 regions of the country. This led to perform a previous stratification of the total population based on the population projections made by the National Institute of Statistics (NIS) of the Chilean Government for the year 2017 at the national level. Table 1 presents the estimated population aged 18 and over for each region of the country as of 2017 and the number of surveys proposed for each region in order to comply with the national proportionality. In addition to the distribution by region, the previous stratification considers, only as a control, the combination of sex and age variables; for more details, see CADEM (2018).

It is important to emphasize that this general information is not published week by week, except for the total number of people selected and the total number of people who agreed to answer the survey.

2.2 How are the missing responses treated?

Taking into account that the survey is conducted by telephone, the main issue is the non-response rate. CADEM is not only aware of this problem, but distinguishes three cases of non-response: cases of no contact, namely no one answers the call either because the phone is busy or out of service; cases of a non-eligible person, namely a person who answers the call, but does not satisfy the requirements of the target group; and a person who is

[1]Retrieved from https://cadem.cl/sobre-cadem/ on December 30, 2021.
[2]Retrieved from https://cadem.cl/plaza-publica/ on December 30, 2021.

Table 1. NIS population projections for 2017 and number of surveyed

Region	Population over 18 years old	Theoretical sample
XV	182,301	9
I	252,814	13
II	471,980	24
III	234,933	12
IV	595,594	30
V	1,430,182	72
VI	706,014	35
VII	804,214	40
VIII	1,634,325	82
IX	756,349	38
XIV	313,112	16
X	636,432	32
XI	80,797	4
XII	126,772	6
RM	5,713,842	287
Total	13,939,661	700

correctly selected but refuses to answer the survey. The impact of the non-response rate is assessed in the following terms:

> Estimating the magnitude of non-response is critical because of the direct relationship it may have with self-selection biases in public opinion polls. The calculation of the non-response rate is also used as a measure of validation of the results. Under the assumption that those who rejects to answer the survey are equal to those who answers it, the magnitude of the non-response rate does not offer major disadvantages, but when there is evidence that the two groups are not equivalent, the non-response can introduce serious distortions in the results (CADEM, 2018).

CADEM accordingly reports the rate of non-response. Three types of results are reported by the survey: those that make explicit the number of cases surveyed, which is approximately equal to 700; those that use a subset of these cases; and trends over time, using previous survey results. However, the impact of the non-response rate on both the results of the survey and their report are not discussed. For an example, see the survey published on the fourth week of December 2021 (CADEM, 2022).

2.3 Dissecting the CADEM survey

The objective of this section is to answer the following questions: What can be learned from the data collected by the CADEM survey? How reliable is the CADEM survey? In order to be consistent with a certain degree of reliability, how should its results be communicated?

2.3.1 Example

Let us consider the collected results during the fifth week of December 2021 (CADEM, 2022). As mentioned above, each study contains a methodological sheet, which indicates that the sampling is a probability sample with random selection of individuals and previously stratified by region; that the sample consists of 705 cases, which required making 6,401 telephone calls, so the response rate is equal to 11%. Let us focus our attention on the first question of the study:

> Do you have a very positive, positive, negative or very negative image of Gabriel Boric?

The results are the following: 63% have a very positive or positive (denoted by a) image of Gabriel Boric; 27% have a negative or very negative (denoted by b) image; and 10% do not know or non-response (denoted by c).

2.3.2 What we can learn from the data

Let M be the sample space whose components are the numbers of cellular phones. On this space we define the vector of random variables $(E, R, S, C, G) : M \longrightarrow \{0,1\}^4 \times \{1, \ldots, 15\}$, where for each $m \in M$

- $E(m) = 1$ if the person associated with cell phone m is eligible, and $E(m) = 0$ if not.
- $R(m) = 1$ if the cell phone m answers the call, and $R(m) = 0$ if not.
- $S(m) = 1$ if the person associated with cell phone m is selected, and $S(m) = 0$ if not.
- $C(m) = 1$ if the person associated with cell phone m answers the survey, and $C(m) = 0$ if not.
- $G(m) = g$ with $g \in \{1, \ldots, 15\}$ if the person associated to cell phone m belongs to region g.

From these definitions, it follows that

$$\{m \in M : S(m) = 1\} \subset \{m \in M : E(m) = 1\} \cap \{m \in M : R(m) = 1\}; \tag{2.1}$$

$$\{m \in M : S(m) = 1\} = \{m \in M : C(m) = 0\} \cup \{m \in M : C(m) = 1\}. \tag{2.2}$$

Let Y be the outcome of interest, taking values in the set $\{a, b, c\}$. The data inform about the conditional distribution of Y given $(E = 1, R = 1, S = 1, C = 1)$; that is,

$$\begin{aligned}
&P(Y = a \mid E = 1, R = 1, S = 1, C = 1) = 0.63;\\
&P(Y = b \mid E = 1, R = 1, S = 1, C = 1) = 0.27;\\
&P(Y = c \mid E = 1, R = 1, S = 1, C = 1) = 0.10;\\
&P(C = 1 \mid E = 1, R = 1, S = 1) = 0.11.
\end{aligned}$$

Both $P(Y = y \mid E = 1, R = 1, S = 1, C = 1)$ for $y \in \{a, b, c\}$, and $P(C = c \mid E = 1, R = 1, S = 1)$ for $c \in \{0, 1\}$ correspond to the identified parameter, and therefore they represent all that can be learned from the data.

2.3.3 What we want to learn from the data

The results of the CADEM survey can be interpreted conditionally to different sub-populations.

First level of analysis

The first level corresponds to what we can learn from the data and it is captured by the identified parameter $P(Y = y \mid E = 1, R = 1, S = 1, C = 1)$ for $y \in \{a, b, c\}$.

Second level of analysis

A second level corresponds to focus the attention on the surveyed persons, namely $\{m \in M : S(m) = 1\}$, which by (2.1) is equivalent to $\{m \in M : E(m) = 1, S(m) = 1, R(m) = 1\}$. In this case, it is not longer possible to identified $P(Y = y \mid E = 1, S = 1, R = 1)$. As a matter of fact, by the Law of Total Probability (Kolmogorov, 1950),

$$\begin{aligned}
&P(Y = y \mid E = 1, S = 1, R = 1) = P(Y = y \mid S = 1) \quad \text{by (2.1)}\\
&= P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1) + P(Y = y \mid S = 1, C = 0)P(C = 0 \mid S = 1)
\end{aligned} \tag{2.3}$$

for each $y \in \{a, b, c\}$. In this decomposition, $P(Y = y \mid S = 1, C = 1)$ and $P(C = 1 \mid S = 1)$ are identified, whereas $P(Y = y \mid S = 1, C = 0)$ is not identified because it depends of those persons who refuse to answer the survey. Taking into account that such a probability takes values between 0 and 1, we can provide an interval of all plausible values for $P(Y = y \mid E = 1, S = 1, R = 1)$ which are compatible with the observed information: for each $y \in \{a, b, c\}$,

$$\begin{aligned} &P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1) \leq \\ &\leq P(Y = y \mid S = 1) \leq \\ &\leq P(Y = y \mid S = 1, C = 1)P(C = 1 \mid S = 1) + P(C = 0 \mid S = 1). \end{aligned} \tag{2.4}$$

Following Manski (2007), this interval corresponds to the region where $P(Y = y \mid S = 1)$ is partially identified. This interval deserves the following comments:

(1) Considering the example of Section 2.3.1, we have that $P(C = 1 \mid S = 1) = 0.11$ and $P(Y = a \mid S = 1) = 0.63$. Therefore

$$0.0693 \leq P(Y = a \mid S = 1) \leq 0.9593. \tag{2.5}$$

Thus, the survey report should be phrased in the following terms: at least 6.93% of the surveyed people have a positive or very positive image of Gabriel Boric, and at most 95.93% of the surveyed people have such positive or very positive image.

(2) This interval provides information about the uncertainty inherent to the non-response rate. In fact, the width of (2.5) is equal to $P(C = 0 \mid S = 1)$, which in this example is equal to 89%. This means that the interval is close to be uninformative.

(3) Different scenarios should be considered when reporting $P(Y = a \mid S = 1)$, $P(Y = b \mid S = 1)$ and $P(Y = c \mid S = 1)$ because these three probabilities belongs to the 2-dimensional simplex $S_3 = \{(p_1, p_2, p_3) \in [0, 1]^3 : p_1 + p_2 + p_3 = 1\}$. Thus, for instance, it can be said that 95.93% of surveyed people have a positive or very positive image of Gabriel Boric and, consequently, a 4.07% have a poor or very poor image or Gabriel Boric, or do not known or non-response, that is,

$$\begin{aligned} &1 - [P(Y = a, C = 1 \mid S = 1) + P(C = 0 \mid S = 1)] = \\ &= P(C = 1 \mid S = 1) + P(C = 0 \mid S = 1) - P(Y = a, C = 1 \mid S = 1) - \\ &P(C = 0 \mid S = 1) \\ &= P(C = 1 \mid S = 1) - P(Y = a, C = 1 \mid S = 1) \\ &= P(Y \neq a, C = 1 \mid S = 1) \\ &= P(Y \in \{b, c\}, C = 1 \mid S = 1) \\ &= P(Y = b, C = 1 \mid S = 1) + P(Y = c, C = 1 \mid S = 1), \end{aligned}$$

which is the lower bound of $P(Y \in \{b, c\} \mid S = 1)$. In the example, $P(Y = b, C = 1 \mid S = 1) = 0.0297$ and $P(Y = c, C = 1 \mid S = 1) = 0.011$.

Once the partial identification of $P(Y = y \mid S = 1)$ ($y \in \{a, b, c\}$) is established, it is possible to qualify CADEM's claims about non-responses. As it was mentioned in Section 2.2, CADEM considers that, "under the assumption that those who rejects to answer the survey are equal to those who answers it, the magnitude of the non-response rate

does not offer major disadvantages, but when there is evidence that the two groups are not equivalent, the non-response can introduce serious distortions in the results". If we consider the decomposition (2.3), the assumption advanced by CADEM corresponds to the following equality:

$$P(Y = y \mid S = 1, C = 1) = P(Y = y \mid S = 1, C = 0) \qquad \text{for all } y \in \{a, b, c\},$$

which, by definition of conditional independence, is equivalent to

$$Y \perp\!\!\!\perp C \mid \{S = 1\}; \tag{2.6}$$

here $V \perp\!\!\!\perp W \mid Z$ corresponds to the conditional independence between V and W given Z; for details and properties on conditional independence, see Florens et al. (1990, Chapter 2). This condition, typically known as *missing at random* (Rubin, 1976; Little and Rubin, 2019), is not empirically refutable because it depends on the component $P(Y = y \mid S = 1, C = 0)$ which in turn is not based on actual observations. Consequently, it is *impossible* to find out evidence establishing that "the two groups are not equivalent".

Correctly stated, condition (2.6) is an identification restriction (San Martín and González, 2022) under which $P(Y = y \mid S = 1)$ is point identified in the sense that

$$P(Y = y \mid S = 1) = P(Y = y \mid S = 1, C = 1) \qquad \text{for all } y \in \{a, b, c\}.$$

In other words, under assumption (2.6), the uncertainty induced by the non-response decreases from an interval of width $P(C = 0 \mid S = 1)$ to the singleton $\{P(Y = y \mid S = 1, C = 1)\}$. Thus, *what we want to learn from the data* coincides with *what we can learn from the data*. In passing, let us mention that condition (2.6) should be viewed as a characterization of *absence of (self-)biased* and, consequently, the identification problem induced by the non-response is exactly the same as the identification problem induced by self-selection.

Third level of analysis

A third level of analysis corresponds to focus the attention on the eligible persons, namely $\{m \in E(m) = 1\}$. In this case, the parameter of interest is given by $P(Y = y \mid E = 1)$ for $y \in \{a, b, c\}$. Let us analyze its identifiability using only the information available at the CADEM survey as published.

Using the Law of Total Probability, we have

$$\begin{aligned} P(Y = y \mid E = 1) = {} & P(Y = y \mid E = 1, R = 1)P(R = 1 \mid E = 1) + \\ & P(Y = y \mid E = 1, R = 0)P(R = 0 \mid E = 1) \end{aligned} \tag{2.7}$$

for $y \in \{a, b, c\}$. In this decomposition, $\gamma \doteq P(Y = y \mid E = 1, R = 0)$ is not identified because it is impossible to know whether a person associated with a cell phone that does not answer a call is eligible or not. On the other hand, $P(Y = y \mid E = 1, R = 1)$ can be decomposed as follows:

$$\begin{aligned} P(Y = y \mid E = 1, R = 1) = {} & P(Y = y \mid E = 1, R = 1, S = 1)P(S = 1 \mid R = 1, E = 1) + \\ & P(Y = 1 \mid E = 1, R = 1, S = 0)P(S = 0 \mid R = 1, S = 1) \\ = {} & P(Y = y \mid S = 1)P(S = 1 \mid R = 1, E = 1) + \\ & P(Y = 1 \mid E = 1, R = 1, S = 0)P(S = 0 \mid R = 1, E = 1), \end{aligned}$$

where the last equality follows from (2.1). Note that

$$\{m \in M : E(m) = 1, R(m) = 1, S(m) = 0\} = \emptyset$$

because there are no eligible persons associated with a cell phone that answered the call who are not selected. Consequently,

$$P(S = 0 \mid R = 1, E = 1) = \frac{P(S = 0, R = 1, E = 1)}{P(R = 1, E = 1)} = 0.$$

Moreover, $P(Y = 1 \mid E = 1, R = 1, S = 0)$ is a probability conditional on an event of probability 0 and, therefore, takes an arbitrary value in $[0, 1]$ (see Remark 2.1). It follows that

$$P(Y = y \mid E = 1, R = 1) = P(Y = y \mid S = 1)P(S = 1 \mid R = 1, E = 1). \tag{2.8}$$

Thus, for each $y \in \{a, b, c\}$,

$$\begin{aligned} P(Y = y \mid E = 1) &= P(Y = y \mid S = 1)P(S = 1 \mid R = 1, E = 1)P(R = 1 \mid E = 1) + \\ &\quad \gamma\, P(R = 0 \mid E = 1) \\ &= P(Y = 1 \mid S = 1)P(S = 1 \mid E = 1) + \gamma\, P(R = 0 \mid E = 1) \end{aligned}$$

for all $\gamma \in [0, 1]$. In this decomposition, $P(Y = 1 \mid S = 1)$ is partially identified by the interval (2.4); by (2.1), $P(S = 1 \mid E = 1)$ corresponds to the ratio

$$\frac{\sharp\{\text{selected persons}\}}{\sharp\{\text{eligible persons}\}},$$

which is identified; and $P(R = 0 \mid E = 1)$ corresponds to the proportion of eligible persons who did not respond to the telephone call. Taking into account that a person can be classified as eligible once he/she has answered the telephone call (see Section 2.1), then it is impossible to identify this parameter. Nevertheless, (2.1) implies that

$$\{m \in M : R(m) = 0\} \subset \{m \in M : S(m) = 0\}$$

and, therefore,

$$P(R = 0 \mid E = 1) \leq P(S = 0 \mid E = 1) = 1 - P(S = 1 \mid E = 1) = \frac{\sharp\{\text{non-selected persons}\}}{\sharp\{\text{eligible persons}\}},$$

which is identified.

Therefore, $P(Y = y \mid E = 1)$ is partially identified, where the lower bound of the identification region is given by

$$P(Y = 1 \mid S = 1, C = 1)P(C = 1 \mid S = 1)P(S = 1 \mid E = 1),$$

which by (2.1) reduces to $P(Y = y, S = 1, C = 1 \mid E = 1)$; and its upper bound is given by

$$\begin{aligned} &[\, P(Y = 1 \mid S = 1, C = 1)P(C = 1 \mid S = 1) + P(C = 0 \mid S = 1)\,] \times \\ &P(S = 1 \mid E = 1) + P(S = 0 \mid E = 1), \end{aligned}$$

which by (2.1) reduces to

$$P(Y = y, S = 1, C = 1 \mid E = 1) + P(C = 0, S = 1 \mid E = 1) + P(S = 0 \mid E = 1).$$

Using the data of the Example,

$$P(S = 1 \mid E = 1) \sim \frac{6,401}{14 \times 10^6},$$

and

$$0.0003198 \leq P(Y = a \mid E = 1) \leq 0.9999814.$$

Clearly, this interval is non-informative.

Remark 2.1 *Let $(M, \mathcal{M}, P)$ be a finite probability space. Let $\mathcal{C} = (C_1, \ldots, C_n\} \subset \mathcal{M}$ be a partition of M such that $P(C_1) = 0$ and $P(C_j) > 0$ for $j = 2, \ldots, n$. Finally, let $A \in \mathcal{M}$. In this case, the conditional probability $P(A \mid \mathcal{C})$ is a random variable defined as*

$$P(A \mid \mathcal{C}) = \sum_{j=1}^{n} P(A \mid C_j) \mathbb{1}_{C_j},$$

where $\mathbb{1}_{C_j}$ is the indicator function of the event C_j (Kolmogorov, 1950, §6); here the numbers $P(A \mid C_j)$ are computed using the following rule

$$P(A \mid C_j) = \begin{cases} \dfrac{P(A \cap C_j)}{P(C_j)} \text{ if } P(C_j) > 0; \\ \eta \in [0, 1], \text{ if } P(C_j) = 0 \end{cases} \tag{2.9}$$

with η arbitrary. This rule is a correct rule (that is, it avoids paradoxes) because it satisfies the following equality:

$$P(A) = E[P(A \mid \mathcal{C})],$$

which ensures the existence of the conditional probability. As a matter of fact, under rule (2.9), this equality reduces to the Law of Total Probability –in the general case, it corresponds to the Radon-Nikodym theorem. Moreover, the number $P(A \mid C_a)$ can be arbitrarily chosen because the random variable $P(A \mid \mathcal{C})$ does not change since $P(C_1) = 0$. For more details, see Rao (2005, Chapter 2).

■

Fourth level of analysis

The non-informativity of the above identification region is primarily due to the fact that $P(S = 1 \mid E = 1)$ is extremely small, so $P(S = 0 \mid E = 1)$ is extremely large. This undesired effect could be counteracted by taking into account the information provided by the CADEM survey regarding how persons are selected: "Probabilistic sampling with random selection of individuals and previously stratified by region" (CADEM, 2022).

By the CADEM sampling design, the reasoning should be done conditionally on $\{m \in M : E(m) = 1, R(m) = 1\}$: it is impossible to know whether a person is eligible if he/she has not answered the phone call. Thus, the statement "random selection of individuals

and previously stratified by region" corresponds to the following condition:

$$P(Y = y \mid E = 1, R = 1, G, S = 1) = P(Y = y \mid E = 1, R = 1, G, S = 0),$$

which, by definition of conditional independence, is equivalent to

$$Y \perp\!\!\!\perp S \mid \{E = 1, R = 1\}, G. \tag{2.10}$$

By the Law of Total Probability, this condition implies that

$$\begin{aligned} P(Y = y \mid E = 1, R = 1, G) &= P(Y = y \mid S = 1, E = 1, R = 1, G) \\ &= P(Y = y \mid S = 1, G) \quad \text{by (2.1)}. \end{aligned} \tag{2.11}$$

Thus, in order to identify $P(Y = y \mid E = 1, R = 1)$, we marginalize with respect to G, namely

$$\begin{aligned} P(Y = y \mid E = 1, R = 1) &= \sum_{g=1}^{15} P(Y = y \mid E = 1, R = 1, G = g) P(G = g \mid E = 1, R = 1) \\ &= \sum_{g=1}^{15} P(Y = y \mid S = 1, G = g) P(G = g \mid E = 1, R = 1), \end{aligned}$$

where the last equality follows from (2.11).

In this decomposition, the conditional probability $P(G = g \mid E = 1, R = 1)$ is in principle identified, although the current information provided by CADEM does not allow to identify it. Moreover, the conditional probability $P(Y = y \mid S = 1, G = g)$ has the same identification problem that was discussed in the second level of analysis and, therefore, it is partially identified: for each $y \in \{a, b, c\}$ and $g \in \{1, \ldots, 15\}$,

$$\begin{aligned} &P(Y = y \mid S = 1, C = 1, G = g) P(C = 1 \mid S = 1, G = g) \leq \\ &\leq P(Y = y \mid S = 1, G = g) \leq \\ &\leq P(Y = y \mid S = 1, C = 1, G = g) P(C = 1 \mid S = 1, G = g) + P(C = 0 \mid S = 1, G = g). \end{aligned} \tag{2.12}$$

Therefore, the random selection of each individual in each stratum is far from helping to identify $P(Y = y \mid E = 1, R = 1)$. Furthermore, it does not help to identify $P(Y = y \mid E = 1)$ either, since $P(Y = y \mid E = 1, R = 0)$ is still unidentified.

2.4 Discussion

CADEM research & estrategia offers services that "connect people with decision makers, through data and not from intuition". Nevertheless, after dissecting the CADEM survey, we can say that this motto is far from being fulfilled. In fact, the dissection of the CADEM survey shows how weak its reliability is whatever the level of analysis.

The first level of analysis corresponds to a description of the collected data. For the sake of transparency, CADEM must not only remember for each question of the survey the total number of people who answered it, but also indicate, together with the percentages of preference for each option, the absolute frequencies. This will warn the readers and

especially the press that the results reflect the opinion of a *very small number of people*. The second level of analysis makes explicit the uncertainty induced by the non-response. CADEM should be made explicit such uncertainty by reporting both the lower and the upper bound of the identification region of $P(Y = y \mid S = 1)$. In the example, the impact of the non-response rate is dramatic, which prevents the reader from a false illusion of certainty. It should be emphasized that condition (2.6) is a plausible way to treat the non-responses. A transparent treatment of non-response should show the impact of such a condition on the conclusions of the study. As we have seen in the example, the conclusion depends much more on (2.6) than on the data itself. The third level of analysis focuses on the eligible population. Once again, for the sake of transparency, it is necessary to report both the lower and the upper bound of the identification region. The example we have used shows how uninformative the survey results are. This information is more than relevant, showing the intrinsic limits of this type of public opinion instruments.

3. CASEN survey

The National Socioeconomic Characterisation Survey (CASEN, for their initials in Spanish) is a Chilean household survey that has been applied since 1987. It is used to assess the impact of social programs on the living conditions of the population[1]. According to the Technical data sheet, the target population is *the population residing in private households throughout the national territory*. The units of analysis are families and individuals living in a household. A suitable respondent is the head of household or, alternatively, a man or woman over 18 years old.

The sampling process of the CASEN survey consists on two steps. First, blocks are chosen that correspond to sets of households; second, a household is chosen in which individuals are surveyed. Due to the pandemic by COVID19, the last version of the survey, called *2020 CASEN survey in pandemic*, was carried out in two steps: first, from the households selected in the previously mentioned sampling process, a face-to-face pre-contact was applied to obtain a contact telephone number. Second, the survey was administered by telephone.

In the 2020 CASEN in pandemic survey, 97,848 households were pre-contacted. Of these, only 86,189 households provided at least a telephone number to be contacted. Of these, 62,540 households had individuals who answered the survey, which amounted to 185,437 individuals[2]. It should be remarked that the available CASEN data set contains information of these individuals[3].

3.1 Treatment of missing outcomes in the CASEN survey

One of the objectives of the CASEN survey is to obtain an overview of the income distribution in Chile, and in particular to have an overview of poverty in the country in terms of income. However, some of the selected individuals did not answer the question on income. CASEN considers appropriate to impute these missing data, so that researchers and policy makers can use a database without missing data. The chosen imputation procedure is called *Conditional Mean Imputation*. The rationale of this technique can be summarized as follows: first, observed covariates are used to define classes. Second, individuals who did

[1]Retrieved from http://casenpandemia2020.cl/ on December 30, 2021

[2]For details, see Nota técnica N7: Desempeño del Trabajo de Campo, Casen en Pandemia en sección Notas Técnicas 2020: http://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-en-pandemia-2020

[3]The data base can be downloaded from http://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-en-pandemia-2020.

not report their income and individuals who reported it are classified in the same class if they share the characteristics of that class. For example, those people from city A, with an age range 30-35 years old who do not report the income, are classified in the same class as those people from the same city in the same age range that report the income. Third, it is computed the mean of the observed incomes conditionally on a class: the missing incomes are imputed through this mean (Little and Rubin, 2019).

More precisely, let Y be an outcome of interest, and let $\mathbf{X}$ be a set of fully observed covariates which are used to define the classes. Let Z be a binary random variable such that $Z = 1$ if the outcome is observed, and $Z = 0$ if not. The conditional mean of both respondents and non-respondents in the same class are given by $E(Y \mid \mathbf{X} = \mathbf{x}, Z = 1)$ and $E(Y \mid \mathbf{X} = \mathbf{x}, Z = 0)$, respectively. The Conditional Mean Imputation assumes that, for each $\mathbf{x}$,

$$E(Y \mid \mathbf{X} = \mathbf{x}, Z = 0) = E(Y \mid \mathbf{X} = \mathbf{x}, Z = 1). \tag{3.1}$$

This assumption is also known as *Mean Missing at Random* (Manski, 2007), *Weak Ignorability* (Imbens, 2000; Hirano and Imbens, 2004), and is equivalent to the conditional orthogonality between Y and Z given $\mathbf{X}$.

Remark 3.1 *Equation* (3.1) *is equivalent to* $E(Y \mid \mathbf{X} = \mathbf{x}, Z) = E(Y \mid \mathbf{X} = \mathbf{x})$ *for all* x*, which in turn is equivalent to the conditional orthogonality of* Y *and* Z *given* $\mathbf{X}$*. In fact, in the Hilbert space* $L^2(M, \mathcal{M}, P)$*,* Y *and* Z *are conditionally orthogonal given* $\mathbf{X}$ *if and only if*

$$Y - E(Y \mid \mathbf{X}) \perp Z - E(Z \mid \mathbf{X});$$

that is, if the correlation between both residual is equal to 0. Florens and Mouchart (1982) prove that this last condition is equivalent to $E(Y \mid \mathbf{X} = \mathbf{x}, Z) = E(Y \mid \mathbf{X} = \mathbf{x})$*. It should be remarked that this condition is implied by* $Y \perp\!\!\!\perp Z \mid \mathbf{X}$*.*

■

3.2 Dissecting the CASEN survey

3.2.1 Example

Let us focus our attention on the incomes of the salaried employees. According to the technical report *Measuring income and poverty in Chile, 2020 Casen Survey in Pandemic*[1], 45,642 individuals were considered in this category. These individuals were exposed to the following question:

> The last month, what was your net income at your main job?

The non-response rate was approximately 11.4% (40,418 valid responses); only 5,062 responses were imputed; the remaining responses (namely, 162) were kept as missing. The following covariates were used to define the classes to impute the missing incomes: X_1=Geographic location, X_2=range age, X_3=sex, X_4=educational level, X_5=category of the occupation, X_6=class of activity of the company where the individual works, and X_6=type of occupation into the company[2].

If we consider the original data (i.e., the people who reported their income), the average income is equal to 653,891.6 Chilean pesos, while the average income considering

[1]Retrieved from http://observatorio.ministeriodesarrollosocial.gob.cl on January 11, 2022

[2]For details on the imputation procedure, see the technical report: *Measuring income and poverty in Chile, Casen Survey in Pandemic 2020.*

Table 2. Quantiles of the income distribution for both original and imputed incomes

Percentage	Quantile of the original data	Quantile of the imputed data
5%	150,000	160,000
10%	230,000	242,000
25%	320,000	320,000
50%	400,000	420,000
75%	750,000	750,000
90%	1,300,000	1,300,000
95%	1,800,000	1,800,000
99%	3,500,000	3,500,000

the imputed data also was equal to 653,327 Chilean pesos. The quantiles of the income distributions for both data sets are given in Table 2. Considering the original data, it can be seen that the 5% of the surveyed individuals have an income at most equal to 150,000 Chilean pesos, while the 10% of the salaried surveyed people have an income at most equal to 230,000 Chilean pesos. When the imputed incomes are considered, these values change.

Remark 3.2 *Let Y be a real random variable. The quantile function is defined as*

$$q_X(\alpha) = inf\{t \in \mathrm{R} : P(Y \leq t) \geq \alpha\} \qquad for\ \alpha \in [0, 1].$$

This corresponds to the generalized inverse of the cumulative distribution function of Y; see Embrechts and Hofert (2013). The quantiles reported in Table 2, as in other part of this paper, were calculated using this definition (for a code, see Alarcón-Bustamante, 2022), which respects the nature of the data (the income is a discrete random variable), and not using the Hyndman and Fan (1996)'s recommendations which is used, for instance, in R Core Team (2020). ■

Table 2 shows the impact of the imputation procedure on the quantiles of the income distribution. How relevant is this impact on a global view of income distribution and poverty? Could we say that it is negligible? These questions can be answered by addressing the following one: what can we learn about the income by using the empirical evidence only? The remaining of this section is devoted to answer this question.

3.2.2 What can we learn from the data?

It was previously mentioned that the CASEN data set contains information of 185,437 individuals, that is, those individuals who answered the survey in the application step. For this reason, we will consider the sample space M as the set of these individuals. Let us define the coordinates of following random vector $(C, S, Z, Y) : M \to \{0, 1\}^3 \times R^+ \cup \{0\}$: for each $m \in M$:

- $C(m) = 1$ if the individual m answers the survey at the application step, and $C(m) = 0$ if not.
- $S(m) = 1$ if the individual is classified as a salaried employee in the application step, and $S(m) = 0$ if not.
- $Z(m) = 1$ if the individual m reports the income, and $Z(m) = 0$ if not.
- Let $Y(m)$ be the income of individual m.

From these definitions it follows that

$$\begin{aligned} &\text{(i) } \{m \in M : S(m) = 1\} \subset \{m \in M : C(m) = 1\}; \\ &\text{(ii) } \{m \in M : Z(m) = 1\} \subset \{m \in M : S(m) = 1\} \cap \{m \in M : C(m) = 1\}. \end{aligned} \tag{3.2}$$

From the CASEN survey, the information summarized in Table 3 is available. This shows that the following conditional probabilities are identified:

$$P(S = 1 \mid C = 1) = 0.246; \quad P(Z = 1 \mid S = 1, C = 1) = 0.885544.$$

Furthermore, the conditional distribution of the income $P(Y \leq y \mid Z = 1, C = 1, S = 1)$ is identified, which is depicted in Figure 1. In particular, the average income $E(Y \mid Z = 1, C = 1, S = 1)$ is identified, and it is equal to 653,891.6 Chilean pesos.

Table 3. Total of individuals by random variable – 2020 CASEN survey

Event	Cardinality
$\{m \in M : C(m) = 1\}$	185,437
$\{m \in M : S(m) = 1\}$	45,642
$\{m \in M : Z(m) = 1\}$	40,418

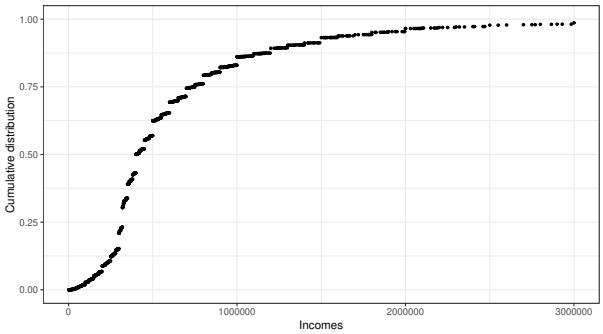

Figure 1. The observed income distribution $P(Y \leq y \mid Z = 1, C = 1, S = 1)$

3.2.3 What we want to learn from the data

Analogous to the analysis of the CADEM survey, the results of the CASEN survey can be interpreted conditionally to different sub-populations. This is the content of this section.

First level of analysis

The first level corresponds to what we can learn from the data. This level is accordingly captured by the identified parameters above described. Regarding the distribution of the reported incomes, Figure 1 shows that the slope of the curve rapidly increases for lower incomes. As a matter of fact, until 75% of the salaried employees, there are non-dramatic changes in the income, so there is a low variability. In contrast, in the 25% of employees with highest incomes this slope increase slowly, which means that there is a great variability among the incomes.

Second level of analysis: Surveyed salaried employees

The second level of analysis is focused on the parameter of interest $P(Y \leq y \mid C = 1, S = 1)$, that is, the income distribution of the salaried employees who answered the survey. The objective of this section is to make explicit the impact of the non-response rate on the income distribution, the average income and the corresponding quantiles. By doing so, it will be appreciated how strong is the Conditional Mean Imputation implemented by the CASEN survey.

Income distribution:

Let us start by the income distribution. Using the Law of Total Probability, we have that

$$P(Y \leq C = 1, S = 1) = P(Y \leq y \mid C = 1, S = 1, Z = 1)P(Z = 1 \mid C = 1, S = 1) + P(Y \leq y \mid C = 1, S = 1, Z = 0)P(Z = 0 \mid C = 1, S = 1). \tag{3.3}$$

In this decomposition, both $P(Y \leq y \mid C = 1, S = 1, Z = 1)$ and $P(Z = z \mid C = 1, S = 1)$, $z \in \{0, 1\}$, are identified, whereas $P(Y \leq y \mid C = 1, S = 1, Z = 0)$ is not identified because it depends on the employees who did not report their income. Instead of using an ignorability condition (as the Conditional Mean Imputation), the relevant question is what can be learned about $P(Y \leq| C = 1, S = 1)$ without introducing additional assumptions. Taking into account that $P(Y \leq y \mid C = 1, S = 1, Z = 0) \in [0, 1]$, it is possible to bound $P(Y \leq y \mid C = 1, S = 1)$ as follows:

$$\begin{aligned} &P(Y \leq y \mid C = 1, S = 1, Z = 1)P(Z = 1 \mid C = 1, S = 1) \leq \\ &\leq P(Y \leq| C = 1, S = 1) \leq \\ &\leq P(Y \leq y \mid C = 1, S = 1, Z = 1)P(Z = 1 \mid C = 1, S = 1) + P(Z = 0 \mid C = 1, S = 1), \end{aligned} \tag{3.4}$$

where $P(Z = 1 \mid C = 1, S = 1) = 0.866$. This identification region, depicted in Figure 2, includes an infinite number of income distributions that are compatible with the observations. Moreover, it reflects the uncertainty induced by the non-response rate: in fact, the width of this interval is equal to the non-response rate, namely $P(Z = 0 \mid C = 1, S = 1)$.

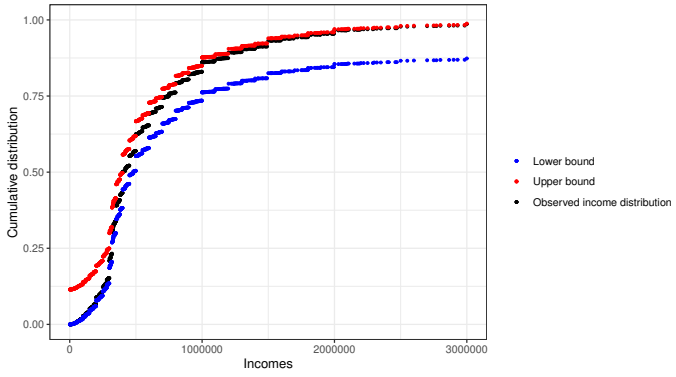

Figure 2. Identification region for $P(Y \leq y \mid C = 1, S = 1)$

Average income

At the second level, the average income corresponds to the conditional expectation $E(Y \mid C = 1, S = 1)$, which is decomposed as

$$E(Y \mid C = 1, S = 1) = E(Y \mid C = 1, S = 1, Z = 1)P(Z = 1 \mid C = 1, S = 1) + E(Y \mid C = 1, S = 1, Z = 0)P(Z = 0 \mid C = 1, S = 1).$$

In this decomposition, $E(Y \mid C = 1, S = 1, Z = 1)$ and $P(Z = z \mid C = 1, S = 1)$, for $z \in \{0, 1\}$, are identified, whereas $E(Y \mid C = 1, S = 1, Z = 0)$ is not identified because it depends on the employees who did not report their income. However, this last conditional expectation could be partially identified provided the support of Y is bounded. Although theoretically the support of Y is bounded, in practice the lower bound is known, whereas the upper bound is finite but *unknown*: how large is it? 5,000,000 Chilean pesos? 25,000,000 Chilean pesos? There is no way to answer this question and, therefore, there is no way to provide a partial identification region for $E(Y \mid C = 1, S = 1)$. For additional discussion on partial identifiability of a conditional expectation, see Alarcón-Bustamante et al. (2020).

Quantiles of $P(Y \leq y \mid C = 1, S = 1)$:

Although the first moment of the income distribution $P(Y \leq y \mid C = 1, S = 1)$ is not even partially identified, it is possible to learn from the respective quantiles, and to appreciate the impact of the non-response rate on them. The quantiles of the income distribution $P(Y \leq y \mid C = 1, S = 1)$ are given by

$$q_{Y|C=1,S=1}(\alpha) = \inf\{t \in \mathrm{R} : P(Y \leq t \mid S = 1, C = 1) \geq \alpha\} \quad \text{for } \alpha \in [0, 1].$$

This quantile function is non identified because it is defined in terms of a non identified probability distribution, namely $P(Y \leq t \mid S = 1, C = 1)$. However, using the bounds in (3.4), it is possible to partially identified the quantile function $q_{Y|C=1,S=1}$ by using the quantiles of the income distribution $P(Y \leq y \mid S = 1, C = 1, Z = 1)$: for $\alpha \in (0, 1)$,

$$\begin{aligned} & q_{Y|C=1,S=1,Z=1}\left(\frac{\alpha - P(Z = 0 \mid C = 1, S = 1)}{P(Z = 1 \mid C = 1, S = 1)}\right) \leq \\ & \leq q_{Y|C=1,S=1}(\alpha) \leq \\ & \leq q_{Y|C=1,S=1,Z=1}\left(\frac{\alpha}{P(Z = 1 \mid C = 1, S = 1)}\right). \end{aligned} \tag{3.5}$$

For a proof, details and reference, see San Martín and González (2022, Section 4).

The identification region (3.5) shows the impact of the non-response rate on the quantile function of $P(Y \leq y \mid C = 1, S = 1)$ in the sense that one of the bounds of the quantile function is *non-informative* for some values of α. As a matter of fact,

- If $\alpha \leq P(Z = 0 \mid C = 1, S = 1)$, then the lower bound in (3.5) is equal to the minimum of the support of the conditional distribution $P(Y \leq y \mid C = 1, S = 1, Z = 1)$ and, therefore, it is non-informative.
- If $\alpha \geq P(Z = 1 \mid C = 1, S = 1)$, then the upper bound in (3.5) is equal to the maximum of the support of the conditional distribution $P(Y \leq y \mid C = 1, S = 1, Z = 1)$ and, therefore, it is non-informative.

Therefore, the quantile function of $P(Y \leq y \mid C = 1, S = 1)$ is informative (that is, provides values in the interior of the support of $P(Y \leq y \mid S = 1, C = 1, Z = 1)$) in the following two cases:

(1) If $P(Z = 0 \mid S = 1, C = 1) < P(Z = 1 \mid S = 1, C = 1)$ or, equivalently, the non-response rate among the employees individuals is smaller than 50%, then the quantile function $q_{Y|C=1,S=1}$ is informative for all

$$\alpha \in [\, P(Z = 0 \mid S = 1, C = 1),\, P(Z = 1 \mid S = 1, C = 1)\,].$$

(2) If $P(Z = 0 \mid S = 1, C = 1) > P(Z = 1 \mid S = 1, C = 1)$ or, equivalently, the non-response rate among the employees individuals is greater than 50%, then the quantile function $q_{Y|C=1,S=1}$ is informative for all

$$\alpha \in [0, P(Z = 1 \mid C = 1, S = 1)\,] \cup [\, P(Z = 0 \mid C = 1, S = 1),\, 1].$$

Let us illustrate this result with the data of the Example. In this case, $P(Z = 0 \mid C = 1, S = 1) = 0.114456$; the corresponding identification regions of the quantile $q_{Y|C=1,S=1}(\alpha)$ for some values of α are summarized in Table 4. We also summarize the quantiles of the income distribution with imputations, thereafter called *CASEN income distribution* and denoted as $\widetilde{q}_{Y|C=1,S=1}(\alpha)$. It should be noted that the CASEN income distribution almost overlapped with the distribution of observed incomes. Furthermore, the CASEN income distribution is in the interior of the identification region (3.4), as theoretically expected; see Figure 3.

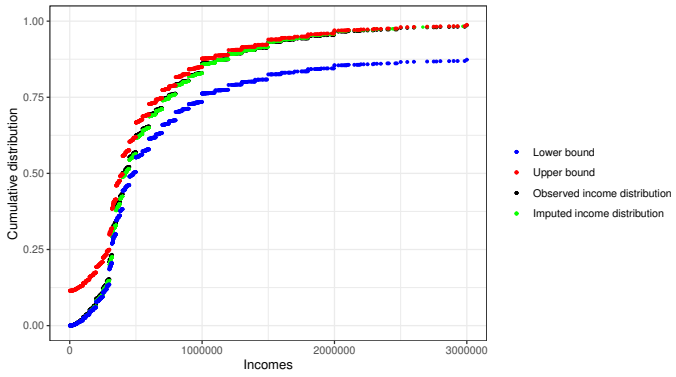

Figure 3. Identification region for $P(Y \leq y \mid C = 1, S = 1)$ and CASEN income distribution

Table 4 deserves the following comments:

(1) For α smaller than the non-response rate, the income of employees can be much lower than the income that can be deduced from the CASEN income distribution. In other words, the non-response rate has such an impact that it is not possible to know how poor the "poorest of the income of employees" are.
(2) For α greater that the response rate, the income of employees can be much higher than the income that can be deduced from the CASEN income distribution. In other words, the response rate has such an impact that it is not possible to know how rich the "richer of the income of employees" are.
(3) It can be remarked that for (some) $\alpha_1 \leq \alpha_2$, the identification region of $q_{Y|C=1,S=1}(\alpha_1)$ at least intersects the identification region of $q_{Y|C=1,S=1}(\alpha_2)$. This clearly increases the uncertainty of the conclusions that can be drawn using the partially identified income distribution and which, on the other hand, is rendered invisible when using the CASEN income distribution.

The previous conclusions allow us to understand the meaning of ignorability conditions, such as the Conditional Mean Imputation technique or, more generally, Missing at Random

Table 4. Quantiles of both the partial identified income distribution and the CASEN income distribution

	$q_{Y\|C=1,S=1}(\alpha)$		
α	LB	UB	$\widetilde{q}_{Y\|C=1,S=1}(\alpha)$
0.05	1,200	170,000	160,000
0.10	1,200	250,000	242,000
$P(Z=0 \mid C=1,S=1)$	1,200	265,000	250,000
0.25	300,000	320,000	320,000
0.50	400,000	480,000	420.000
0.75	700,000	1,000,000	750,000
0.80	800,000	1,300,000	865,172
$P(Z=1 \mid C=1,S=1)$	1,100,000	25,000,000	1,200,000
0.90	1,200,000	25,000,000	1,300,000
0.95	1,800,000	25,000,000	1,800,000
0.99	3,500,000	25,000,000	3,500,000
1.00	25,000,000	25,000,000	25,000,000

conditions. These conditions come from the identification restriction

$$Y \perp\!\!\!\perp Z \mid C=1, S=1, \mathbf{X}$$

which, by definition of conditional independence, is equivalent to

$$\begin{aligned} P(Y \leq y \mid C=1,S=1,\mathbf{X}) &= P(Y \leq y \mid C=1,S=1,Z=1,\mathbf{X}) \\ &= P(Y \leq y \mid C=1,S=1,Z=0,\mathbf{X}). \end{aligned}$$

These equalities means that the missing observations do not provide more relevant information about the output Y, being the only "statistical job" to carefully estimate $P(Y \leq y \mid C=1,S=1,Z=1,\mathbf{X})$ –this is the standard procedure.

3.3 Discussion

One of the objectives of the CASEN survey is to obtain an overview of the income distribution of employees and, in particular, to have a look at the incomes of the lowest paid employees as well as those of the highest paid. For this purpose, the self-reported income of survey respondents who fall into the category of salaried employees is used. However, individuals who are exposed to the survey are not required to report their income. As a consequence, the survey includes a non-response rate which, for the 2020 CASEN survey in pandemic, is equal to 11.4456%. Before providing an overview of the distribution of incomes, CASEN applies statistical techniques designed to impute missing income, specifically the Conditional Mean Imputation technique.

Our dissection of the CASEN survey aims to make explicit the policy meaning of this imputation technique. To this end, a partial identification analysis was developed to show the impact of the non-response rate on both the mean of the distribution of the income distribution of employees and on the respective quantiles. One of the main conclusions is that "the poor may be poorer" than what can be asserted from the CASEN income distribution, and that "the rich may be richer" than what can be stated from it.

With this conclusion in mind, it is possible to assess the sense of the imputation technique used by CASEN: the Conditional Mean Imputation technique corresponds to an assumption of *income homogeneity*. As a matter of fact, it is assumed that, among individuals with characteristics $\mathbf{X}=\mathbf{x}$ who did not report their income, there is no relevant

income information that was not accessed: all the effectively relevant information has already been observed in those who did report their income. Consequently, the income of an employee who did not report it should be related to the average income of all employees sharing the same characteristics $\mathbf{X} = \mathbf{x}$. The partial identification analysis show how heterogenous could be the income distributions of employees. Policy decisions should be aware on this uncertainties.

4. The Araucanía citizen consultation

The Araucanía citizen consultation is of special political interest given the ongoing violent conflicts in the region. This is the main motivation for having chosen to analyze it. But there is also a relevant methodological aspect: the information provided by the consultation can be related to the national referendum held in 2020. We will study how plausible this relationship is, and how it affects the conclusions that can be drawn.

4.1 Historical and economical context

The capital of the Araucanía region, Temuco, is located 620 kilometers to south of Santiago, the capital of Chile. The Araucanía Region is known for being the original area of the Mapuche People (in the 16th century called "Araucanos"), possibly the only indigenous people with whom the Crown of Spain made a Capitulation of Peace, known as the *Paces de Quilín*, made on January 5 and 6, 1614. This treaty established the Biobío River as the border, south of which "the Mapuches lived independently for two hundred and forty years, until 1881" (Bengoa, 2007). In 1881, "Manuel Recabarren, Minister of the Interior [at the time], led Chilean troops to the south and, together with General Gregorio Urrutia, advanced hundreds of kilometers along the border and militarily occupied the area" (Bengoa, 2016). This completed the occupation of Araucanía by the Chilean government.

The Araucanía Region, in addition to the Biobío, Los Ríos and Maule regions, develop the country's forestry industry: "the forestry sector represents 1.9% of the domestic GDP, reaching in 2017 USD 5,196 million (3,373 billion of Chilean pesos). Biobío region represents 60.0% of the forestry GDP, followed by La Araucanía region with 10.5%, and Los Ríos, and Maule regions with 10.1% each. Regarding the participation of the three forestry subsectors included in the sectorial GDP, the paper, and pulp industry, as well as products derived from paper represents 44.3%, forestry participates with 29.4%, and the wood products, and wood industry represent 26.3%" (Instituto Forestal, 2021).

Many of the conflicts in the area are due to the presence of forestry companies, whose worldview on nature and its resources is not entirely shared by the Mapuche people's worldview. In addition, part of the forestry exploitation takes place on what were once Mapuche lands, which has triggered a series of territorial claims (Andrade, 2019).

4.2 Recent political context

On October 12, 2021, the President of the Chilean Republic declared a *state of emergency* for the provinces of Biobío and Arauco, in the Biobío region, and in the provinces of Cautín and Malleco, in the Araucanía Region, for a 15 days period (Diario Oficial de la República de Chile, 2021). According to the Chilean Constitution, this is one of its prerogatives, and it may declare such state of emergency for no more than 15 days. Once a state of emergency is declared, the respective zones will be under the immediate dependence of the Chief of National Defense appointed by the President of the Republic, who will assume the direction and supervision of his jurisdiction with the powers and obligations

established by law (Constitución de la República de Chile, 2005, At.42). By declaring a state of emergency, the President of the Republic may restrict the freedom of locomotion and assembly (Constitución de la República de Chile, 2005, At.43).

Among the reasons that led to this decision, the Diario Oficial de la República de Chile (2021) mentions the following ones:

(1) An increase of violence acts linked to drug trafficking, terrorism and organized crime, committed by armed groups that have not only made attempts on the lives of members of the Law Enforcement and Security Forces, but have also attacked people and destroyed facilities and machinery used in industrial, agricultural and commercial activities.
(2) Since 2018, there has been an increase in crimes and offenses against persons and against property; against public order, including attacks against authority, attacks and threats against prosecutors of the Public Prosecutor's Office and the Judiciary.
(3) There has been a 116% increase in reported incidents related to crimes contemplated in Law No. 17,798 on Arms Control, including the seizure of weapons and ammunition.
(4) The number, magnitude and seriousness of the crimes and facts indicated, committed in the provinces of the regions of Biobío and Araucanía, imply a serious alteration of public order –understood as the "situation that allows the peaceful exercise of rights and the fulfillment of obligations, ensuring peaceful coexistence"– in the terms established in Article 42 of the Constitution of the Chilean Republic, which allows the enactment of the state of emergency constitutional exception with respect to such areas of the national territory, provided for in said article.

As it was mentioned above, the state of emergency may not be extended for more than fifteen days, notwithstanding that the President of the Republic may extend it for the same period. However, for successive extensions, the President shall always require the consent of the National Congress, specifically the Senate (Constitución de la República de Chile, 2005, Art.42). Until January 2022, the National Congress has approved the extension of the state of emergency for 6 consecutive times[1]. It should be mentioned that the official account of the recent conflicts in La Araucanía does not relate these conflicts to the territorial claims of the Mapuche people.

4.3 Organization of the Araucanía consultation and results

In order to know the opinion of the citizens of the 32 communes of La Araucanía regarding the renewal of the state of emergency in the region, the Regional Intendancy and the Association of Municipalities of La Araucanía organized a citizen consultation, which took place on November 5, 6 and 7, 2021. The consultation was carried out electronically, and all persons over 18 years old registered in the electoral registry in any of the 32 municipalities may participate from a computer, cell phone or another device connected to the internet[2].

The citizen consultation was limited to the following question:

Do you agree with Congress extending the state of emergency in the Araucanía Region?

The results of the consutlation are summarized in Table 5.

[1]For details, see https://www.senado.cl/senado/site/cache/search/pags/search164185188127928.html. Retrieved on January 10, 2021.
[2]Retrieved from https://www.consultaaraucania.cl/ on January 10, 2022.

Table 5. Results of the Araucanía consultation

Option	Votes	% wrt the consultation	% wrt electoral roll
Yes	118,258	81.56	13.34
No	26,655	18.38	3.01
Blank votes	54	0.04	0.01
Null votes	27	0.02	0.00
Total	144,994	100	16.36

4.4 How these results were used?

Sections 4.1 and 4.2 attempt to illustrate the complexity of the political situation in the Araucanía region. This complex context may explain why successive extensions of the state of emergency have been subject to lively debate. In fact, those extensions did not achieve unanimity in the Senate: they were approved not more than 2 or 3 votes in favor. Let us mention the Senate session of November 24, 2021, where the extension of the state of emergency was approved by 16 votes in favor, 14 against, and one abstention. Among the reasons that were mentioned for approving the extension, the Araucanía consultation was explicitly mentioned as an important factor. This was stated by Senator Francisco Chahuán, from the right coalition Chile Vamos, who affirmed that "the state of exception has generated greater tranquility. Attacks against property and arson crimes have decreased. We must listen actively and in La Araucanía there was a citizen consultation that supported this measure"[3]. These expressions are in line with the assessment made by the Governor of La Araucanía, Luciano Rivas, independent, near to the Chile Vamos coalition, at the end of the consultation: "With great respect, but also with great strength, we ask politicians, especially all the deputies and senators of Chile, that our voice be heard, do not turn a deaf ear"[4].

As mentioned by Governor Rivas[1], the Araucanía citizen consultation was one of the first, if not the first, non-binding consultations to be held in Chile. This, added to the complex political situation in the Araucanía region, could explain the interest that this consultation aroused, especially in the relationship that its results have with recent elections, namely the 2020 national referendum on the possibility of a new constitution and the 2021 governor elections. One of these studies is the one conducted by Cayul et al. (2021), which was initially published in the electronic journal CIPER[2]. This study analyzes the representativeness at the municipality level of the Araucanía Consultation on three axes: Mapuche population, rurality, and population that voted for the *non-approval* option in the 2020 national referendum. According to the authors, "these axes are fundamental to establish whether or not there is a bias in the results, since it analyzes the cultural, socioeconomic and political dimension". To achieve this objective, the authors analyze, on the one hand, the participation in the second round of the election of Regional Governors in Araucanía with the percentage of Mapuche population, the percentage of rural population and the percentage of non-approval in the 2020 referendum; and, on the other hand, the participation in the citizen consultation in Araucanía with the same percentages already mentioned. The choice of the regional governors is due to the fact that in that election "a similar universe of approximately 125,000 people participated".

We are able to reproduce the third analysis by considering the data summarized at

[3]Retrieved from https://www.senado.cl/estado-de-excepcion-constitucional on January 10, 2021.
[4]Retrieved from https://assets.eldesconcierto.cl/2021/11/Copia-de-Copia-de-Discurso-Consulta-Araucani%CC%81a.pdf on January 11, 2022.
[1]See his speech of November 7, 2021 in https://assets.eldesconcierto.cl/2021/11/Copia-de-Copia-de-Discurso-Consulta-Araucani%CC%81a.pdf.
[2]At https://www.ciperchile.cl/2021/11/10/consulta-ciudadana-en-la-araucania/

Table 6. Figures 4 and 5 reproduce their analysis. Cayul et al. (2021) conclude that "those municipalities with a higher percentage of votes for the non-approval to a new Constitution also had a higher participation in both the citizen consultation and in the second round of governors' elections, but the effect is significantly lower in the latter. That is, there would be a political bias of those who participate in the consultation". The final conclusion of the study is the following:

> We observe then that, when comparing two elections with a similar participation rate, the people who participate in them are very different. While participation in the consultation was higher in urban, non-Mapuche municipalities that voted for non-approval, these same biases are not observed in the second round of governors election.
>
> Electors, then, are not representative at the municipal level, and this suggests that the consultation is not necessarily representative of the population of Araucanía. This implies that the interpretation of the results should be done with caution, and without extrapolating conclusions for the entire region, especially given the relevance that has been sought to give to the consultation.

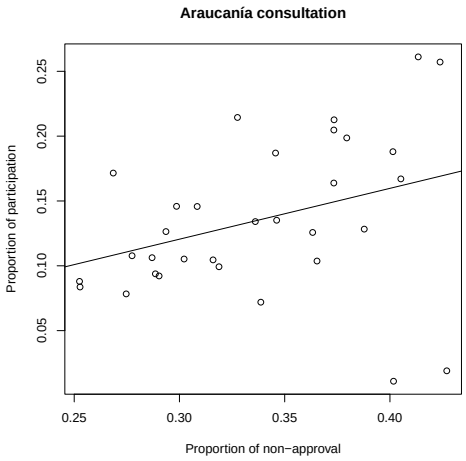

Figure 4. Relationship between 2020 referendum and 2021 Araucanía consultation participation

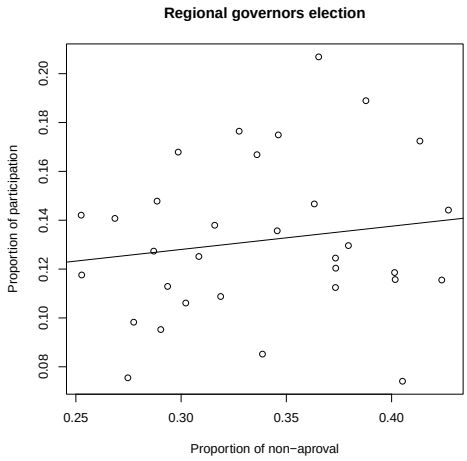

Figure 5. Relationship between 2021 governor election and 2021 Araucanía consultation participation

Table 6. 2020-2022 elections in the Araucanía Region by municipality

		2021 electoral roll	**2020 referendum for a new constitution**				**2021 regional governors election**				**2022 Araucanía consultation**
c	Municipality	Total	Approve	Non-approve	Null votes	Blank votes	Tuma	Rivas	Null Votes	Blank votes	Total of votes
1	Angol	46,976	12,017	7,349	97	57	2,050	3,924	85	30	9,331
2	Carahue	24,663	4,978	2,220	39	25	1,552	1,481	45	9	3,595
3	Cholchol	11,149	2,951	1,197	51	19	815	804	24	5	1,046
4	Collipulli	22,172	5,159	3,075	65	41	1,120	1,500	32	17	4,714
5	Cunco	20,105	4,090	1,914	47	31	1,041	1,108	24	14	1,996
6	Curacautín	21,097	3,491	2,378	33	20	599	911	39	14	3,524
7	Curarrehue	8,521	1,740	891	12	12	337	371	11	7	613
8	Ercilla	8,036	1,521	1,020	26	26	433	508	8	4	1,511
9	Freire	22,892	5,198	2,401	51	20	1,578	1,535	31	14	2,394
10	Galvaino	12,092	2,472	995	46	24	877	629	24	10	1,285
11	Gorbea	15,651	3,290	2,084	25	20	1,170	1,743	34	10	2,008
12	Lautaro	34,520	8,037	4,788	68	22	1,615	2,186	66	15	5,656
13	Loncoche	23,234	6,125	2,072	27	13	1,411	1,280	32	9	1,944
14	Lonquimay	12,115	1,963	1,318	42	26	852	523	11	16	133
15	Los Sauces	7,464	1,409	1,050	24	12	346	712	12	6	142
16	Lumaco	9,187	1,711	1,206	31	31	478	1,080	18	8	2,399
17	Melipeuco	7,723	1,542	584	15	18	335	236	10	2	605
18	Nueva Imperia	30,850	8,765	3,218	83	34	2,407	1,849	64	22	5,292
19	Padres las Casas	59,516	16,892	7,020	142	67	3,063	3,522	107	28	7,523
20	Perquenco	6,814	1,778	757	11	7	721	400	14	9	994
21	Pitrufquén	23,226	5,566	2,947	44	10	1,529	2,494	28	12	3,139
22	Pucón	29,594	8,176	3,140	36	23	1,054	1,776	59	19	3,189
23	Purén	12,240	2,657	1,583	30	19	648	824	38	14	2,506
24	Renaico	9,806	2,640	1,080	17	8	430	477	20	7	904
25	Saavedra	13,239	2,960	1,000	34	27	831	983	42	25	1,165
26	Temuco	238,028	78,332	38,159	360	169	15,375	25,684	740	200	51,039
27	Teodoro Scmidt	14,052	3,026	1,742	28	23	1,304	1,559	32	12	1,457
28	Toltén	10,554	2,126	1,213	30	29	526	1,001	15	6	1,327
29	Traiguén	18,595	4,144	2,189	34	35	1,012	1,445	48	18	3,477
30	Victoria	32,607	7,060	5,193	76	44	1,418	2,264	60	25	8,385
31	Vilcún	24,718	5,971	3,022	70	56	1,614	2,455	41	14	3,315
32	Villarrica	56,170	15,087	6,534	89	40	2,301	3,537	95	27	5,911

4.5 Dissecting the use of Araucanía citizen consultation

4.5.1 Statement of the problem

The previous analysis consists in comparing two or more elections that share a common electoral roll. Now, for each election, there are specific probability distributions that are identified, namely (i) the distributions of participation and non-participation, and (ii) the distribution of preferences conditionally on the electors participating in the election. More precisely, let M be the sample space composed of the electors, and define the following random variables on M:

- $V_1(m) = 1$ if the elector m participated at the 2020 referendum, and $V_1(m) =$ if not.
- $V_2(m) = 1$ if the elector m participated at the 2021 governor election, and $V_2(m) =$ if not.
- $V_3(m) = 1$ if the elector m participated at the 2021 citizen consultation, and $V_3(m) =$ if not.
- Let Y_1 be the preference at the 2020 referendum, namely $\mathcal{Y}_1 = \{\text{approve}, \text{non-approve}, \text{blank vote}, \text{null vote}\}$.
- Let Y_2 be the preference at the 2021 governor election, namely $\mathcal{Y}_2 = \{\text{Tuma}, \text{Rivas}, \text{blank vote}, \text{null vote}\}$.
- Let Y_3 be the preference at the 2021 citizen consultation, namely $\mathcal{Y}_3 = \{\text{yes}, \text{no}, \text{blank vote}, \text{null vote}\}$.
- Let C be the municipality in which each elector is registered. C takes 32 different values because there are 32 municipalities; see Table 6.

If we consider each election separately, then the identified parameters are the following:

$$P(Y_i = y_i \mid V_i = 1, C = c) \quad (y_i, c) \in \mathcal{Y}_i \times \{1, \ldots, 32\}; \qquad P(V_i = 1 \mid C = c) \quad c \in \{1, \ldots, 32\}$$

for $i = 1, 2, 3$.

If these elections are jointly used, it should be verified if the set of electors is the same, that is, if the following equality holds:

$$\begin{aligned} &\{m \in M : V_1(m) = 1\} \cup \{m \in M : V_1(m) = 0\} = \\ &= \{m \in M : V_2(m) = 1\} \cup \{m \in M : V_2(m) = 0\} \\ &= \{m \in M : V_3(m) = 1\} \cup \{m \in M : V_3(m) = 0\}. \end{aligned}$$

Certainly, the Chilean Electoral Service (SERVEL for the initials in Spanish) has access to this information and it can verify such equality. In what follows, we will assume that this equality is fulfilled.

The analysis described in Section 4.4 consists in comparing

$$\{P(Y_1 = \text{non-aproval} \mid V_1 = 1, C = c) \, : \, c \in \{1, \ldots, 32\}\}$$

with

$$\{P(V_3 = 1 \mid C = c) \, : \, c \in \{1, \ldots, 32\}\}.$$

However, for each $c \in \{1, \ldots, 32\}$, $\{m \in M : V_1(m) = 1, C(m) = v\}$ is not necessarily equal to $\{m \in M : V_3(m) = 1, C(m) = c\}$.

A fair comparison needs to use the same electors, which in turn lead to consider

$$\{P(Y_1 = \text{non-aproval} \mid V_1 = 1, V_3 = 1, C = c) \,:\, c \in \{1, \ldots, 32\}\}$$

and

$$\{P(V_3 = 1 \mid V_1 = 1, C = c) \,:\, c \in \{1, \ldots, 32\}\}.$$

This is due to the fact that the political behavior of those who participate in both elections is not necessarily the same as the political behavior of those who participate in one, or the other, or both. It is, therefore, necessary to identify $P(Y_1 = \text{non-aproval} \mid V_1 = 1, V_3 = 1, C = c)$ and $P(V_3 = 1 \mid V_1 = 1, C = c)$ for each c.

4.5.2 Partial identification of $P(V_1 = v_1, V_3 = v_3 \mid C = c)$

Both $P(Y_1 = \text{non-aproval} \mid V_1 = 1, V_3 = 1, C = c)$ and $P(V_3 = 1 \mid V_1 = 1, C = c)$ require the identifiability of $P(V_1 = v_1, V_3 = v_3 \mid C = c)$ for $(v_1, v_3) \in \{0, 1\}^2$. Taking into account that $P(V_1 = v_1 \mid C = c)$ and $P(V_3 = v_3 \mid C = c)$ are identified, the way to relate them to the joint distribution $P(V_1 = v_1, V_3 = v_3 \mid C = c)$ is through the Fréchet inequality (Fréchet, 1960a,b), namely for each $c \in \{1, \ldots, 32\}$

$$\begin{aligned} &\max\{1, P(V_1 = v_1 \mid C = c) + P(V_3 = v_3 \mid C = c) - 1\} \leq \\ &\leq P(V_1 = v_1, V_3 = v_3 \mid C = c) \leq \\ &\leq \min\{P(V_1 = v_1 \mid C = c), P(V_3 = v_3 \mid C = c)\} \quad \forall\, (v_1, v_3) \in \{0, 1\}^2 \end{aligned} \tag{4.1}$$

It should be emphasized that these bounds are the best ones; see the constructive proof in Fréchet (1960a). Thus, for $(v_1, v_3) = (1, 1)$, it follows that

$$\begin{aligned} &\max\{0, P(V_1 = 1 \mid C = c) - P(V_3 = 0 \mid C = c)\} \leq \\ &\leq P(V_1 = 1, V_3 = 1 \mid C = c) \leq \min\{P(V_1 = 1 \mid C = c), P(V_3 = 1 \mid C = c)\}. \end{aligned} \tag{4.2}$$

For municipality c the lower bound is informative (that is, greater than 0) if $P(V_1 = 1 \mid C = c) > P(V_3 = 0 \mid C = c)$, that is, if the rate of participation at the 2020 referendum is greater than the rate of non-participation at the 2021 citizen consultation; or, equivalently, if the rate of non-participation at the 2020 referendum is smaller than the rate of participation at the 2021 citizen participation. Table 7 summarizes the results, where LB_{13} is the lower bound of (4.2) and UB_{13} is the corresponding upper bound. It can be seen that, for each municipality, the lower bound is always 0, which means that a plausible assumption is that none of those who participated in one election participated in the other. Another plausible assumption is that

$$P(V_1 = 1, V_3 = 1 \mid C = c) = P(V_3 = 1 \mid C = c), \tag{4.3}$$

that is, the rate of joint participation is equal to the rate of participation at the 2021 citizen consultation. In this case, $P(V_1 = 0, V_3 = 1 \mid C = c) = 0$, that is, no elector did not participate at the 2020 referendum and participated at the 2021 citizen consultation. Certainly this conclusion may seem implausible, which in turn would imply that (4.3) is implausible as an assumption.

Table 7. Bounds of joint participation ratios

c	**Municipality**	$P(V_1 = 1 \mid C = c)$	$P(V_2 = 1 \mid C = c)$	$P(V_3 = 1 \mid C = c)$	LB_{13}	UB_{13}	LB_{23}	UB_{23}
1	Angol	0.42	0.13	0.20	0.00	0.20	0.00	0.13
2	Carahue	0.29	0.13	0.15	0.00	0.15	0.00	0.13
3	Cholchol	0.38	0.15	0.09	0.00	0.09	0.00	0.09
4	Collipulli	0.38	0.12	0.21	0.00	0.21	0.00	0.12
5	Cunco	0.30	0.11	0.10	0.00	0.10	0.00	0.10
6	Curacautín	0.28	0.07	0.17	0.00	0.17	0.00	0.07
7	Curarrehue	0.31	0.09	0.07	0.00	0.07	0.00	0.07
8	Ercilla	0.32	0.12	0.19	0.00	0.19	0.00	0.12
9	Freire	0.34	0.14	0.10	0.00	0.10	0.00	0.10
10	Galvaino	0.29	0.13	0.11	0.00	0.11	0.00	0.11
11	Gorbea	0.35	0.19	0.13	0.00	0.13	0.00	0.13
12	Lautaro	0.37	0.11	0.16	0.00	0.16	0.00	0.11
13	Loncoche	0.35	0.12	0.08	0.00	0.08	0.00	0.08
14	Lonquimay	0.28	0.12	0.01	0.00	0.01	0.00	0.01
15	Los Sauces	0.33	0.14	0.02	0.00	0.02	0.00	0.02
16	Lumaco	0.32	0.17	0.26	0.00	0.26	0.00	0.17
17	Melipeuco	0.28	0.08	0.08	0.00	0.08	0.00	0.08
18	Nueva Imperia	0.39	0.14	0.17	0.00	0.17	0.00	0.14
19	Padres las Casas	0.41	0.11	0.13	0.00	0.13	0.00	0.11
20	Perquenco	0.37	0.17	0.15	0.00	0.15	0.00	0.15
21	Pitrufquén	0.37	0.17	0.14	0.00	0.14	0.00	0.14
22	Pucón	0.38	0.10	0.11	0.00	0.11	0.00	0.10
23	Purén	0.35	0.12	0.20	0.00	0.20	0.00	0.12
24	Renaico	0.38	0.10	0.09	0.00	0.09	0.00	0.09
25	Saavedra	0.30	0.14	0.09	0.00	0.09	0.00	0.09
26	Temuco	0.49	0.18	0.21	0.00	0.21	0.00	0.18
27	Teodoro Scmidt	0.34	0.21	0.10	0.00	0.10	0.00	0.10
28	Toltén	0.32	0.15	0.13	0.00	0.13	0.00	0.13
29	Traiguén	0.34	0.14	0.19	0.00	0.19	0.00	0.14
30	Victoria	0.38	0.12	0.26	0.00	0.26	0.00	0.12
31	Vilcún	0.37	0.17	0.13	0.00	0.13	0.00	0.13
32	Villarrica	0.39	0.11	0.11	0.00	0.11	0.00	0.11

4.5.3 Partial identification of $P(V_3 = 1 \mid V_1 = 1, C = c)$

From (4.2) it can be deduced the identification region for $P(V_3 = 1 \mid V_1 = 1, C = c)$, namely

$$\max\left\{0, \frac{P(V_1 = 1 \mid C = c) - P(V_3 = 0 \mid C = c)}{P(V_1 = 1 \mid C = c)}\right\} \leq \tag{4.4}$$
$$\leq P(V_3 = 1 \mid V_1 = 1, C = c) \leq \min\left\{1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)}\right\}.$$

For each municipality c, the lower bound is informative if $P(V_1 = 1 \mid C = c) > P(V_3 = 0 \mid C = c)$, whereas the upper bound is informative (that is, smaller than 1) if $P(V_3 = 1 \mid C = c) < P(V_1 = 1 \mid C = c)$, that is, if the rate of participation at the citizen consultation is smaller than the rate of participation at the 2020 referendum. Table 8 shows the corresponding lower and upper bound. It can be seen that the lower bound is uninformative, whereas the upper bound is informative: it corresponds to the ratio of participation at the citizen consultation given that electors participated at the 2020 referendum.

4.5.4 Partial identification of $P(V_2 = 1, V_3 = 1 \mid C = c)$

Following the arguments developed in Section 4.5.2, it follows that

$$\max\{0, P(V_2 = 1 \mid C = c) - P(V_3 = 0 \mid C = c)\} \leq \tag{4.5}$$
$$\leq P(V_2 = 1, V_3 = 1 \mid C = c) \leq \min\{P(V_2 = 1 \mid C = c), P(V_3 = v_3 \mid C = c)\}.$$

Table 7 shows the corresponding lower and upper bounds. Lower bounds are always uninformative because, for each municipality, the rate of participation at the 2021 gov-

Table 8. Partial identification of $P(V_3 = 1 \mid V_1 = 1, C = c)$ and $P(Y_1 = \text{non-approve} \mid V_1 = 1, V_3 = 1, C = c)$

		$P(V_3 = 1 \mid V_1 = 1, C = c)$		$P(Y_1 = \text{non-approve} \mid V_1 = 1, V_3 = 1, C = c)$	
c	**Municipality**	LB	UB	LB	UB
1	Angol	0.00	0.48	0.00	0.72
2	Carahue	0.00	0.50	0.00	0.61
3	Cholchol	0.00	0.25	0.05	0.38
4	Collipulli	0.00	0.57	0.00	0.85
5	Cunco	0.00	0.33	0.00	0.47
6	Curacautín	0.00	0.60	0.00	0.99
7	Curarrehue	0.00	0.23	0.14	0.44
8	Ercilla	0.00	0.58	0.00	0.94
9	Freire	0.00	0.31	0.00	0.46
10	Galvaino	0.00	0.36	0.00	0.44
11	Gorbea	0.00	0.37	0.02	0.61
12	Lautaro	0.00	0.44	0.00	0.66
13	Loncoche	0.00	0.24	0.02	0.33
14	Lonquimay	0.00	0.04	0.37	0.41
15	Los Sauces	0.00	0.06	0.39	0.45
16	Lumaco	0.00	0.81	0.00	1.00
17	Melipeuco	0.00	0.28	0.00	0.38
18	Nueva Imperia	0.00	0.44	0.00	0.47
19	Padres las Casas	0.00	0.31	0.00	0.42
20	Perquenco	0.00	0.39	0.00	0.49
21	Pitrufquén	0.00	0.37	0.00	0.54
22	Pucón	0.00	0.28	0.00	0.38
23	Purén	0.00	0.58	0.00	0.89
24	Renaico	0.00	0.24	0.06	0.38
25	Saavedra	0.00	0.29	0.00	0.35
26	Temuco	0.00	0.44	0.00	0.58
27	Teodoro Scmidt	0.00	0.30	0.08	0.52
28	Toltén	0.00	0.39	0.00	0.59
29	Traiguén	0.00	0.54	0.00	0.75
30	Victoria	0.00	0.68	0.00	1.00
31	Vilcún	0.00	0.36	0.00	0.52
32	Villarrica	0.00	0.27	0.04	0.41

ernor election is smaller than the rate of non-participation at the 2021 citizen consultation. This means that, although the overall participation rates in both elections are very similar (16% for the citizen consultation, 14% for the governor election), a plausible assumption is that there are no electors who participated in both elections. On the other hand, sometimes the upper bound is equal to $P(V_2 = 1 \mid C = c)$, sometimes to $P(V_3 = 1 \mid C = c)$. In the first case, namely when it is assumed that $P(V_2 = 1, V_3 = 1 \mid C = c) = P(V_2 = 1 \mid C = c)$, then there are no electors who participated in the 2021 governors election and did not participate in the 2021 citizen consultation. In the second case, namely $P(V_2 = 1, V_3 = 1 \mid C = c) = P(V_3 = 1 \mid C = c)$, then there are no electors who did not participate in the 2021 governors election and who participated in the 2021 citizen consultation. Again, it can be stated that these assumptions may not seem entirely plausible, which in turn shows that it is possible to have turnout rates in both elections lower than the upper bound. By passing, this jeopardizes the argument according to which the governor election and the citizen consultation can be compared because their rate of participation are similar.

4.5.5 Partial identification of $P(Y_1 = y \mid V_1 = 1, V_2 = 1, C = c)$

For each municipality c, the conditional probability $P(Y_1 = y \mid V_1 = 1, V_2 = 1, C = c)$ can not vary arbitrarily because it is related to the identified conditional probability $P(Y_1 = y \mid V_1 = 1, C = c)$ through the following decomposition:

$$\begin{aligned} P(Y_1 = y \mid V_1 = 1, C = c) = {} & P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c) P(V_3 = 1 \mid V_1 = 1, C = c) + \\ & P(Y_1 = y \mid V_1 = 1, V_3 = 0, C = c) P(V_3 = 0 \mid V_1 = 1, C = c). \end{aligned}$$

In this decomposition, $\gamma_c \doteq P(Y_1 = y \mid V_1 = 1, V_3 = 0, C = c)$ is non identified, whereas $P(V_3 = 1 \mid V_1 = 1, C = c)$ and $P(V_3 = 0 \mid V_1 = 1, C = c)$ are partially identified by (4.5).

Let $C \in \{1, \ldots, 32\}$ and $p_c \doteq P(V_3 = 0 \mid V_1 = 1, C = c)$ be fixed. It follows that $P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c)$ belongs to the set

$$\left\{ \frac{P(Y_1 = y \mid V_1 = 1, C = c) - \gamma_c p_c}{1 - p_c} \ : \ \gamma_c \in [0, 1] \right\}$$

which reduces to the interval

$$A_{p_c} \doteq \left[\frac{P(Y_1 = y \mid V_1 = 1, C = c) - p_c}{1 - p_c}, \ \frac{P(Y_1 = y \mid V_1 = 1, C = c)}{1 - p_c} \right].$$

Now, if $p_{1,c} < p_{2,c}$, then

$$A_{p_{1,c}} \subset A_{p_{2,c}}.$$

Therefore, for each $c \in \{1, \ldots, 32\}$

$$\begin{aligned} P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c) &\in \bigcup_{p_c \in [l_c, u_c]} A_{p_c} \\ &= A_{u_c}, \end{aligned}$$

where $[l_c, u_c]$ is given by (4.4). It follows that, for each $c \in \{1, \ldots, 32\}$, $P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c) \in$ belongs to an identification region where the lower bound is given by

$$\max \left\{ 0, \frac{P(Y_1 = y \mid V_1 = 1, C = c) - \min \left\{ 1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)} \right\}}{1 - \min \left\{ 1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)} \right\}} \right\},$$

and the upper bound is given by

$$\min \left\{ 1, \frac{P(Y_1 = y \mid V_1 = 1, C = c)}{1 - \min \left\{ 1, \frac{P(V_3 = 1 \mid C = c)}{P(V_1 = 1 \mid C = c)} \right\}} \right\}.$$

Table 8 shows the corresponding lower and upper bound of $P(V_3 = 1 \mid V_1 = 1, C = c)$ and $P(Y_1 =$ non-approve $\mid V_1 = 1, V_3 = 1, C = c)$. It can be observed the uncertainty induced by the joint participation in both elections. In particular, four municipalities have an extreme uncertainty because the width of their identification regions is at least equal to 0.9: Curacautín, Ercilla, Lumaco and Victoria. In these municipalities, the proportion of non-approval conditionally on the participation at both the 2020 referendum and the 2021 consultation is any value. Moreover, the conditional probability to participate at the citizen referendum given that electors participated at the 2020 referendum is, respectively, 0.6 0.58, 0.81 and 0.68. On the other hand, two municipalities, Lonquimay and Los Sauces, have the smaller uncertainty and, consequently, the rate of approval conditionally on the joint participation is less uncertainty: between 0.37 and 0.41 for Lonquimay; and between 0.39 and 0.45 for Los Sauces. Nevertheless, the rate of participation at the 2021 consultation given participation at the 2020 referendum are quite small: 0.04 and 0.06, respectively.

4.6 Discussion

The partial identification analysis shows the impact of the uncertainty due to the joint participation in both elections, namely 2020 referendum and 2021 consultation, on the proportion of electors who chose the *non-approve* option at the 2020 referendum. This impact can be diminished if, for each municipality, the joint distribution $P(V_1 = v_1, V_3 = v_3 \mid C = c)$ with $(v_1, v_3) \in \{0,1\}^2$ were known. This seems to be feasible for the Chilean Electoral Service, without having to transgress elector identity protection. If this were the case, then $P(V_3 = 1 \mid V_1 = 1, C = c)$ would be identified. However, this fact does not ensure that $P(Y_1 = y \mid V_1, = 1, V_3 = 1, C = c)$ is point identified because $P(Y_1 = y \mid V_1 = 1, V_3 = 0, C = c)$ is not identified given the secrecy of the vote. Consequently, following the arguments developed in Section 4.5.5, $P(Y_1 = y \mid V_1, = 1, V_3 = 1, C = c)$ belongs to an identification interval with a lower bound given by

$$\max\left\{0, \frac{P(Y_1 = y \mid V_1 = 1, C = c) - P(V_3 = 1 \mid V_1 = 1, C = c)}{P(V_3 = 0 \mid V_1 = 1, C = c)}\right\}$$

and an upper one given by

$$\min\left\{1, \frac{P(Y_1 = y \mid V_1 = 1, C = c)}{P(V_3 = 0 \mid V_1 = 1, C = c)}\right\}.$$

It can be deduced that this interval is informative (that is, strictly included in $[0,1]$) if

$$P(V_3 = 1 \mid V_1 = 1, C = c) < P(V_3 = 0 \mid V_1 = 1, C = c),$$

which is a surprising result.

It could be argued that, under "mild conditions", it is possible to ignore joint participation, and thus argue for the reliability of studies such as the one reported in Section 4.4. The partial identification analysis developed in Section 4.5.5 shows that there are two possible assumptions that could be made: the first one would be to assume that $P(Y_1 = \text{non-approve} \mid V_1 = 1, V_3 = 0) = 0$, that is, that no elector who participated in the 2020 referendum and did not participate in the 2021 citizen consultation chose the option *non-approve*. It should be mentioned that this assumption is quite strong and hard to believe. A second assumption would be $Y_1 \perp\!\!\!\perp V_3 \mid \{V_1 = 1\}, C$, which is equivalent to the following two equivalent conditions:

$$\begin{aligned} &P(Y_1 = y \mid V_1 = 1, V_3 = 1, C = c) = P(Y_1 = y \mid V_1 = 1, C = c); \\ &P(V_3 = v_3 \mid Y = y, V_1 = 1, C = c) = P(V_3 = v_3 \mid V_1 = 1, C = c) \quad v_3 \in \{0,1\}. \end{aligned}$$

The last condition means that, once an elector of a specific municipality participated at the 2020 referendum, the participation at the 2021 consultation does not depend on the preference at the 2020 referendum: again hard to believe.

Finally, it should be emphasized that the previous analysis of partial identification is applicable to critically assess the comparisons over time of political surveys as they would be correctly done if made conditionally on joint participation.

5. Concluding remarks

This paper illustrates a traditional service that Applied Statistics can render to society. In fact, during the XIX century, statistics was considered as "statistics is the science of social

facts, expressed in numerical terms", as expressed by Moreau de Jones (1847), or as the prospectus of the Statistical Society of London stated, "Statistics [...] may be sad [...] to be ascertaining and bringing together of those «facts which are calculated to illustrate the condition and prospects of society;»and the objective of Statistical Science is to consider the results which they produce, with the view to determine those principles upon which the well-being of society depends" (Journal of the Statistical Society of London, 1838). As it is well known, these considerations go back to Süßmilch (1998) and his idea of seeking order in the figures that summarize the profile of a *state* – hence the term *Statistics*.

These original ideas show clearly the need of every statesman for statistics in order to "illustrate, with new or more accurate data, a multitude of issues that arise every day, stimulating public opinion, being the subject of parliamentary discussions, and forming problems whose solution can only be offered by Statistics" Moreau de Jones (1847). The two surveys analyzed in this paper, as well as the citizen consultation, are examples of the scenario described by Moreau de Jones. As a matter of fact, the socioeconomic survey CASEN is used by policy makers either in order to assess social policies or to have a global view of poverty or income distribution. Stake-holders, as the press or politicians, use the two political opinion polls (CADEM and Araucanía citizen consultation) either to influence citizens' political opinion or to justify political arguments at the parliament.

We complement Moreau de Jones' scenario by making explicit new frontiers of what statisticians and social scientists call *data of good quality*. From a statistical point of view, we focus our assessment of the surveys on the correct way of communicating their results, so that the uncertainty induced by non-responses is made explicit. The results can be reported at different levels depending on the population of interest to which the results are to be generalized. The advantage of this strategy is that it makes explicit how this uncertainty could be reduced, which part of it can not be reduced unless very strong assumptions are introduced. The price to be paid in the face of these strong assumptions is the drawing of non-credible conclusions –that is, Law of Decreasing Credibility (Manski, 2013). For instance, in the Araucanía citizen consultation, the uncertainty of the option at the 2020 referendum conditionally on the joint participation at both the 2020 referendum and 2021 consultation can decreased whether the Chilean Electoral Service provides information on such joint participation. However, the uncertainty can not decrease to a point value because of the secrecy of the vote.

At the methodological level, the assessment or dissection of the Chilean surveys was performed making a distinction between identified parameters and parameters of interest: what we can learn from the data is represented by the identified parameters, while what we want to learn from the data is represented by the parameters of interest. In (almost) all empirical research there is a gap between both types of parameters; it is quite relevant to highlight the difference and to study their possible relationships –which is equivalent to solving an identification problem. The way Clifford (1982) expresses himself is illuminating and perhaps summarizes the perspective developed in this paper:

> Anyone who has tried to make sense to real data will, sooner or later, have come across the problem of nonidentifiability. Broadly speaking this means that their first explanation of the data is not the only one. The existence of alternative explanations becomes important when decisions have to be made and particularly so when different explanations suggest completely different courses of action.

The identification regions we established for each of the Chilean survey contain such different substantive explanations.

Supplementary material

The computational routine implemented in R is available online at https://github.com/edalarconb?tab=repositories

Author contributions Conceptualization, E.S., E.A.; methodology, E.S., E.A.; software, E.S., E.A.; validation, E.S., E.A.; formal analysis, E.S., E.A.; investigation, E.S., E.A.; data curation, E.S., E.A.; writing—original draft preparation, E.S., E.A.; writing—review and editing, E.S., E.A.; visualization, E.S., E.A. supervision, E.S., E.A. All authors have read and agreed to the published version of the manuscript.

Acknowledgements The authors would like to thank the Editors-in-Chief and the anonymous reviewers for their valuable comments and suggestions which improved substantially the quality of this paper.

Funding This research was supported by the Agencia Nacional de Investigación y Desarrollo (ANID) Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI) NCS2021072. The first author was also partially funded by the FONDECYT Project No. 1181261, whereas the second author was partially funded by the Postdoctoral FONDECYT grant No. 3220422.

Conflicts of interest The authors declare no conflict of interest.

References

Alarcón-Bustamante, E., 2022. EmpiricalEvidence: an R package for empirical research. Available from https://github.com/edalarconb.

Alarcón-Bustamante, E., San Martín, E., and González, J., 2020. Predictive validity under partial observability. In Wiberg, M., Molenaar, D., González, J., Böckenholt, U., and Kim, J.-S., editors, *Quantitative Psychology*, pages 135–145, Cham. Springer International Publishing.

Andrade, M. J., 2019. La lucha por el territorio mapuche en Chile: una cuestión de pobreza y medio ambiente. *L'Ordinaire des Amériques*.

Bengoa, J., 2007. *El Tratado de Quilín*. Catalonia, Santiago, Chile.

Bengoa, J., 2016. Sarmiento y sarmientadas. In Domingo Faustino Sarmiento, editor, *Conflicto y Armonía de las Razas en América*, pages 5–14. Akal-Inter Pares, Santiago, Chile.

Berinsky, A. J., 2017. Measuring public opinion with surveys. *Annual Review of Political Science*, 20, 309–329.

CADEM, 2018. Diseño metodológico de Plaza Pública Cadem 2018.

CADEM, 2022. Encuesta Plaza Pública. Quinta Semana de Diciembre.

Cayul, P., Durán, E., and Jaimovich, D., 2021. ¿Es representativa la consulta ciudadana en la Araucanía? (Is the citizen consultation in Araucanía representative?).

Clifford, P., 1982. Some General Comments on Nonidentifiability. In LeCam, L. and Neyman, J., editors, *Probability Models and Cancer*, pages 81–83. North-Holland Publishing Company, Amsterdam, NL.

Constitución de la República de Chile, 2005. Decreto 100. fija el texto refundido, coordinado y sistematizado de la Constitución Política de la República de Chile.

Diario Oficial de la República de Chile, 2021. Declara estado de excepción constitucional de emergenia en las zonas del territorio nacional que indica.

Embrechts, P. and Hofert, M., 2013. A Note on Generalized Inverses. *Mathematical Methods of Operations Research*, 77, 423–432.

Fisher, R. A., 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222, 309–368.

Florens, J. P. and Mouchart, M., 1982. A Note on Noncausality. *Econometrica*, 50, 583–592.

Florens, J. P., Mouchart, M., and Rolin, J.-M., 1990. *Elements of Bayesian Statistics*. Marcel Dekker, New York.

Fréchet, M., 1960a. Les tableaux dont les marges sont données. *Trabajos de Estadística*, 11, 1–18.

Fréchet, M., 1960b. Sur les tableaux dont les marges et des bornes sont données. *Revue de l'Institut International de Statistique*, 28, 10–32.

Hirano, K. and Imbens, G. W., 2004. The propensity score with continuous treatments. In Shewhart, W. A. and Wilks, S. S., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, chapter 7, pages 73–84. Wiley Series in Probability and Statistics.

Hyndman, R. J. and Fan, Y., 1996. Sample Quantiles in Statistical Packages. *The American Statistician*, 50, 361–365.

Imbens, G. W., 2000. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706–710.

Instituto Forestal, 2021. *Anuario Forestal. Boletín Estadístico Número 174*. Instituto Forestal, Santiago, Chile.

Journal of the Statistical Society of London, 1838. Introduction. *Journal of the Statistical Society of London*, 1, 1–5.

Kolmogorov, A. N., 1950. *Foundations of the Theory of Probability*. Chelsea Publishing Company, New Yok.

Koopmans, T. and Reiersol, O., 1950. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21, 165–181.

Little, R. J. and Rubin, D. B., 2019. *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons.

Manski, C. F., 2007. *Identification for Prediction and Decision*. Harvard University Press.

Manski, C. F., 2011. Policy Analysis with Incredible Certitude. *The Economic Journal*, 121:F261–F289.

Manski, C. F., 2013. *Public Policy in an Uncertain World*. Harvard University Press.

Manski, C. F., 2020. The lure of incredible certitude. *Economics & Philosophy*, 36, 216–245.

Moreau de Jones, A., 1847. *Éléments de Statistique Comprenant les Principes Généraux de cette Science, et un Apery Historiques de ses Progrès*. Guillaumn et Cie., Libraires, Paris.

R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, M., 2005. *Conditional Measures and Applications. Second Edition*. Chapman & Hall/CRC, New Yok.

Rubin, D. B., 1976. Inference and missing data. *Biometrika*, 63, 581–592.

San Martín, E., 2018. Identifiability of structural characteristics: How relevant is it for the bayesian approach? *Brazilian Journal of Probability and Statistics*, 32, 346–373.

San Martín, E. and González, J, 2022. A Critical View on the NEAT Equating Design: Statistical Modelling and Identifiability Problems. *Journal of Educational and Behavioral Statistics*, Accepted for publication.

San Martín, E., González, J., and Tuerlinckx, F., 2015. On the unidentifiability of the fixed-effects 3pl model. *Psychometrika*, 80, 450–467.

Süßmilch, J. P., 1998. *L'Ordre Divin dans les changements de l'espèce humaine, démontré par la naissance, la mort et la propagation de celle-ci. Texte intégral de lédition de*

1741 traduit et annoté par Jean-Marc Rohrbasser. À l'Institut National D'Études Démographiques, Paris.