UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

OUTLIER DETECTION FOR MULTICANDIDATE ELECTIONS WITH DEMOGRAPHIC GROUPS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERIO CIVIL INDUSTRIAL

PABLO UBILLA PAVEZ

PROFESOR GUÍA:
Charles Thraves Cortés-Monroy

MIEMBROS DE LA COMISIÓN:
Denis Sauré Valenzuela
Fernando Ordóñez Pizarro

Esta tesis ha sido parcialmente financiado por: FONDO PUENTE DAII - FCFM 2023
Powered@NLHPC: Esta tesis fue parcialmente apoyada por la infraestructura de supercómputo del NLHPC (ECM-02)

SANTIAGO DE CHILE
2023

DETECCIÓN DE ANOMALÍAS EN ELECCIONES CON MÚLTIPLES CANDIDATOS Y GRUPOS DEMOGRÁFICOS

La inferencia ecológica es una técnica para estimar el comportamiento individual utilizando datos agregados. Un caso particular se encuentra en las elecciones políticas, donde en cada mesa electoral conocemos los votos de los candidatos y el número de votantes de distintos grupos demográficos (como la edad, el sexo y la nacionalidad). En este trabajo, aplicamos el algoritmo EM para estimar las probabilidades de voto de los grupos demográficos para cada candidato en un distrito determinado. Desafortunadamente, el *E-step* escala exponencialmente en el número de candidatos.

En este estudio proponemos cuatro métodos polinomiales alternativos para estimar las probabilidades del paso *E-step*: (1) simulación de escenarios utilizando un método de *hit-and-run*, (2) aproximación utilizando una distribución normal multivariada con integración de Monte Carlo o (3) una distribución normal multivariada utilizando su FDP, y (4) aproximación mediante una única multinomial. Mostramos a partir de experimentos numéricos que el método de aproximación multinomial es el más rápido, ejecutándose en menos de una centésima de segundo. Además, el error absoluto promedio de la probabilidad estimada con este método es muy similar al obtenido al realizar el algoritmo EM con la probabilidad exacta.

Implementamos los métodos propuestos en la primera vuelta de la elección presidencial de Chile de 2021. Presentamos una metodología que utiliza programación entera mixta para agregar grupos y estimar intervalos de confianza en las probabilidades estimadas mediante el uso de *bootstraping*, de modo que el modelo pueda identificar correctamente las probabilidades de voto. Observamos que los distritos con más mesas electorales se benefician de conjuntos de grupos menos agregados en contraste con los distritos con menos mesas electorales. Finalmente, calculamos p-valores utilizando simulación con la aproximación multinomial, obteniendo 27 urnas electorales con un p-valor menor o igual a 10^{-8}.

OUTLIER DETECTION FOR MULTICANDIDATE ELECTIONS WITH DEMOGRAPHIC GROUPS

Ecological inference is a technique to estimate individual behaviour by using aggregated data. A particular case is found in political elections where in each ballot box we know the candidates' votes and the number of voters for different demographic group (like age, sex and nationality). In this work, we apply the EM-algorithm to estimate the voting probabilities of demographic groups for each candidate at a particular district. Unfortunately, the E-Step scales exponentially in the number of candidates.

We propose four alternative polynomial methods to estimate the E-Step probabilities: (1) sample scenarios using hit-and-run, (2) approximate using a multivariate normal with Monte Carlo integration or (3) a multivariate normal using its PDF, and (4) approximate by a single multinomial. We show from numerical computations that the multinomial approximation method is the fastest, running in less than a hundredth of a second. In addition, the mean absolute error of the estimated probability with this method is very similar to the one obtained when performing the EM-algorithm with the exact probability.

We run the proposed methods in the first round Chilean Presidential Election of 2021. We present a methodology that uses mixed integer programming to aggregate groups, and estimate confidence intervals on the estimated probabilities by using bootstrapping, so that the model can correctly identify the voting probabilities. We observe that districts with more ballot boxes benefit from less aggregated group sets in contrast to districts with less ballot boxes. Finally, we compute p-values using simulation with the multinomial approximation, obtaining 27 ballot-boxes with a p-value lower or equal than 10^{-8}.

Pa' mi weli Ruchi

Agradecimientos

Primero que todo agradecer a mi maire y mi paire, que me han apoyado siempre. También a mis hermanos Jesu y Carlos, y al resto de mi hermosa familia.

A todas las lindas amistades que he tenido a lo largo de mi vida. Mis amigos del colegio, los del Plan Común, los de Industrias, las tierras de Quilpué, los de Leiden y tantos otros que les guardo bonitos recuerdos. Al Nico que lo quiero mucho.

A la Dorkita.

Al Luca y al Pepi, que ojalá algún día se lleven bien.

A José Miguel Alvarado por motivarme a estudiar Ingeniería Industrial.

A Charles Thraves por guiarme en este proceso de tesis. A Denis Sauré y Fernando Ordóñez por participar de la comisión.

No tengo palabras para agradecerle a Linda y toda su ayuda en la gestión de titulación. A Fernanda por su buena disposición y su motivación con el magíster.

Cómo no agradecerle a *lofi hip hop radio - beats to relax/study to* que me mantuvo concentrado incontables horas.

Quedaron un poco desordenados los agradecimientos y seguro me faltó alguien, se hace lo que se puede. Saludos a quienes lean esta tesis, ojalá les guste.

TABLE OF CONTENT

Table Index

Figure Index

Chapter 1

Introduction

Election processes are a key part of the political systems in most countries. They are used to determine presidents, senators, mayors; to approve new laws or to make new constitutions. It is necessary that transparency and impartiality are upheld to ensure the seamless execution of these procedures and to ensure outcomes that most faithfully approximate the true will of the electorate.

Elections that are apparently clean may not be absent of unintended errors or fraudulent acts in low scale (Leemann & Bochsler, 2014). These cases may not be detected since they are not necessarily systematic or produced on a large scale. It is still an important task to detect such acts for various reasons: in close elections they could alter the final outcome, these actions could damage the trust in the democratic institutions, and prompt detection can improve the voting process for future elections (Fortin-Rittberger, Harfst, & Dingler, 2017).

In this study we will propose methods to determine how likely the results of an election are, thus, assessing its transparency. We will construct our models using the Chilean election system as a reference, however, these models could be applied to different scenarios with slight modifications.

In most Chilean election processes, voters are assigned to specific ballot-boxes in the area they live in. Ballot-boxes are located in what are call *mesas* (tables in Spanish), where a group of randomly selected citizens are in charge of assisting the voters, registering election attendance and counting the votes at the end of the election. These randomly selected citizens are called *vocales de mesa* (ballot-box committee), and their labour is fundamental in order

to have an honest procedure.

The demographic composition of voters between ballot-boxes may greatly differ. For example, there could be a ballot-box where mostly young people vote and another where is mostly old people. These differences between ballot-box composition could make the election results in each ballot-box completely different. If we chose to omit this difference between ballot-boxes we would make wrong estimations of how likely the results are in each. This is why our model that should take into account the possibility of having different demographic group composition for each ballot-box.

The number of candidates is also an important factor, and it may scale in some elections. In Chile (and the rest of the world) there are many election processes that have multiple-candidates, for example the election of councilors, senators and deputies. In a two-candidate election it is easier to compute probabilities associated with the voting distribution. However, when the number of candidates increases, it also increases the number of different combination outcomes. When we also consider demographic groups the number of different combination outcomes also increases.

Considering that we will have two sets of aggregate data: ballot-box voter composition and ballot-box outcome, we will be working with ecological inference, where one would like to estimate individual behavior by using aggregate data. There has been some applications of ecological inference to elections in other contexts like voting migration (Antweiler, 2007) and political studies (King, Tanner, & Rosen, 2004). Ecological inference also arises in different fields such as sociology (Duncan & Davis, 1953; Goodman, 1953, 1959; Glynn & Wakefield, 2010), geography (Cleave, Brown, & Payne, 1995; Withers, 2001; Anselin & Cho, 2002), epidemiology (Morgenstern, 1995; Jackson, Best, & Richardson, 2006; Wakefield, 2008). Most models in the literature focus on the 2×2 case, where for both classes of aggregate data there are only 2 categories. The $R \times C$ case has gained more attention due to its wide applications to different fields, being a more general case. Some of the studies have approached the $R \times C$ case with parametric approaches, where prior distributions are assumed.

The complexity of the general case is its main challenge, meaning that approximate solutions could be a promising approach to this problem. In this study we will focus on the general case with an arbitrary number of demographic groups and an arbitrary number of candidates, estimating a non-parametric distribution.

Chapter 2

Model

Consider an election process with candidates from a set $\mathcal{C}$, where $|\mathcal{C}| = C$. Voters belong to a set $\mathcal{I}$, where $|\mathcal{I}| = I$. Each voter belongs to a group from a set $\mathcal{G}$ and is assigned to a designated ballot-box from a set $\mathcal{B}$, where $|\mathcal{G}| = G$ and $|\mathcal{B}| = B$. In this context, $\mathcal{I}_g$ represents the subset of individuals belonging to group g, while $\mathcal{I}_b$ represents the subset of individuals assigned to ballot-box b, where $|\mathcal{I}_b| = I_b$ and $|\mathcal{I}_g| = I_g$.

The groups delineated by $\mathcal{G}$ could encompass demographic attributes of the voters, such as age brackets and gender. The amount of voters from group $g \in \mathcal{G}$ that vote in ballot-box $b \in \mathcal{B}$ is known and is denoted by w_{bg}. The amount of votes to candidate c in each ballot-box b is also known and is denoted by x_{bc}.

We define the probability that a voter from group $g \in \mathcal{G}$ votes for candidate $c \in \mathcal{C}$ as p_{gc}. Notably, this probability remains latent due to our access solely to the aggregate election outcome for each candidate. Our aim is to obtain an estimator $\hat{p}_{gc}$ that maximizes the likelihood of the election result given the known data.

Let us consider $\mathcal{X}$ to be the observed data: $\mathcal{X} = \{X_{bc} = x_{bc}\}_{b\in\mathcal{B},c\in\mathcal{C}}$. We may also define the unobserved data as $\mathcal{Y} = \{\mathcal{Y}_{bic}\}_{b\in\mathcal{B},i\in\mathcal{I}_b,c\in\mathcal{C}}$, where $\mathcal{Y}_{bic}$ is a binary random variable that indicates if the i-th person from ballot-box b did vote for candidate c. We will also define $Z_{bgc} = \sum_{i\in\mathcal{I}_g\cup\mathcal{I}_b} Y_{bic}$, this is the total votes that candidate c received from group g in ballot-box b. We may also use $\boldsymbol{X}_b$, $\boldsymbol{Y}_b$ and $\boldsymbol{Z}_b$ to refer to the respective multidimensional random variables of ballot-box b.

2.1. E-Step

Let $\boldsymbol{p} = \{p_{gc}\}_{g\in\mathcal{G},c\in\mathcal{C}}$. We may write the likelihood function and derive the log-likelihood function as follows:

$$\begin{aligned} l(\boldsymbol{p};\mathcal{X},\mathcal{Y}) &= \sum_{b\in\mathcal{B},i\in\mathcal{I}_b,c\in\mathcal{C}} Y_{bic}\ln(p_{g(i)c}) \\ &= \sum_{b\in\mathcal{B},g\in\mathcal{G},c\in\mathcal{C}} Z_{bgc}\ln(p_{gc}) \end{aligned}$$

We would like to know the distribution of Y_{bic} conditional to the result in ballot-box b. We will define q_{bgc} as the probability that someone from ballot-box b and group g did vote for candidate c, conditional to the result in that ballot-box. We will use use the notation $Y_{bi(g)c}$ to indicate any voter i in ballot-box b that belongs to group g. This probability may be calculated with Bayes' Theorem.

$$\begin{aligned} q_{bgc} &\coloneqq P(Y_{bi(g)c}=1|X_{b1}=x_{b1},\ldots,X_{bC}=x_{bC};\boldsymbol{p}) \\ &= \frac{P(X_{b1}=x_{b1},\ldots,X_{bC}=x_{bC}|Y_{bi(g)c}=1;\boldsymbol{p})P(Y_{bi(g)c}=1;\boldsymbol{p})}{P(X_{b1}=x_{b1},\ldots,X_{bC}=x_{bC};\boldsymbol{p})} \\ &= \frac{P(X_{b1}=x_{b1},\ldots,X_{bC}=x_{bC}|Y_{bi(g)c}=1;\boldsymbol{p})\cdot p_{gc}}{P(X_{b1}=x_{b1},\ldots,X_{bC}=x_{bC};\boldsymbol{p})} \end{aligned} \tag{2.1}$$

For some purposes in following methods we could also write the denominator using the law of total probabilities:

$$q_{bgc} = \frac{P(X_{b1}=x_{b1},\ldots,X_{bC}=x_{bC}|Y_{bi(g)c}=1;\boldsymbol{p})\cdot p_{gc}}{\sum_{c'\in\mathcal{C}} P(X_{b1}=x_{b1},\ldots,X_{bC}=x_{bC}|Y_{bi(g)c}=1;\boldsymbol{p})\cdot p_{gc'}}$$

Once we calculate q_{bgc}, we can do the E-step, considering $Q(\boldsymbol{p};\boldsymbol{p}^{(old)}) \coloneqq \mathbb{E}\left[l(\boldsymbol{p};\mathcal{X},\mathcal{Y})|\mathcal{X},\boldsymbol{p}^{(old)}\right]$.

$$\begin{aligned} Q(\boldsymbol{p};\boldsymbol{p}^{(old)}) &= \mathbb{E}\left[\sum_{b\in\mathcal{B},i\in\mathcal{I}_b,c\in\mathcal{C}} Y_{bic}\ln\left(p_{g(i)c}\right)|\mathcal{X},\boldsymbol{p}^{(old)}\right] \\ &= \sum_{b\in\mathcal{B},g\in\mathcal{G},c\in\mathcal{C}} q_{bgc}\ln\left(p_{gc}\right) \\ &= \sum_{b\in\mathcal{B},g\in\mathcal{G}} w_{bg}\sum_{c\in\mathcal{C}} q_{bgc}\ln\left(p_{gc}\right) \end{aligned}$$

In this case $\boldsymbol{p}^{(old)}$ will be a prior probability we assume for the distribution.

2.2. M-Step

In the M-Step we maximize $Q(\boldsymbol{p}; \boldsymbol{p}^{(old)})$ on $\boldsymbol{p}$. We have to add a constraint so that the probabilities add up to 1 and are non-negative. We can see that $\boldsymbol{p}^{(old)}$ (the probability we consider as a prior) is expressed using $\boldsymbol{q}$.

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{b\in\mathcal{B}, g\in\mathcal{G}} w_{bg} \sum_{c\in\mathcal{C}} q_{bgc} \ln\left(p_{gc}\right) \\ \text{subject to:} \quad & \sum_{c\in\mathcal{C}} p_{gc} = 1 \qquad \forall g \in \mathcal{G} \\ & p_{gc} \geq 0 \qquad \forall g \in \mathcal{G}, \forall c \in \mathcal{C} \end{aligned}$$

To solve this optimization we can use the Lagrangian:

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\lambda}) = \sum_{b\in\mathcal{B}, g\in\mathcal{G}} w_{bg} \sum_{c\in\mathcal{C}} q_{bgc} \ln\left(p_{gc}\right) - \sum_{g\in\mathcal{G}} \lambda_g \sum_{c\in\mathcal{C}} (p_{gc} - 1)$$

We state the first order conditions:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_{gc}} = 0 &\Leftrightarrow \frac{1}{p_{gc}} \sum_{b\in\mathcal{B}} w_{bg} q_{bgc} = \lambda_g \qquad (2.2) \\ &\Leftrightarrow p_{gc} = \frac{1}{\lambda_g} \sum_{b\in\mathcal{B}} w_{bg} q_{bgc} \\ &\Leftrightarrow \sum_{c\in\mathcal{C}} \frac{1}{\lambda_g} \sum_{b\in\mathcal{B}} w_{bg} q_{bgc} = 1 \\ &\Leftrightarrow \lambda_g = \sum_{b\in\mathcal{B}, c\in\mathcal{C}} w_{bg} q_{bgc} \\ &\Leftrightarrow \lambda_g = \sum_{b\in\mathcal{B}} w_{bg} \qquad (2.3) \end{aligned}$$

Replacing 2.3 in 2.2 we obtain:

$$p_{gc} = \frac{\sum_{b\in\mathcal{B}} w_{bg} q_{bgc}}{\sum_{b\in\mathcal{B}} w_{bg}} \qquad \forall g \in \mathcal{G}, \forall c \in \mathcal{C} \qquad (2.4)$$

2.3. EM-algorithm

The EM-algorithm consists of an iterative alternation between the E-step and the M-step. The formulation for this model goes as follows:

Algorithm 1 EM-algorithm

Input: $\mathcal{I}$, $\mathcal{G}$, $\mathcal{C}$, $\mathcal{B}$, $\boldsymbol{x}$, $\boldsymbol{w}$, ϵ
Output: $\hat{\boldsymbol{p}}$
Initialize: $\boldsymbol{p}_{gc}^{(old)} = \frac{\sum_{b\in\mathcal{B}} w_{bg}\cdot x_{bc}}{I_g}$
converge = False
while not converge **do**
  E-step: Compute q_{bgc} using $\boldsymbol{p}_{gc}^{(old)}$ as described in 2.1.
  M-step: Compute $\hat{\boldsymbol{p}}$ as described in 2.4
  if $\max_{g\in\mathcal{G},c\in\mathcal{C}}\{\hat{\boldsymbol{p}}_{gc} - \boldsymbol{p}_{gc}^{(old)}\} \leq \epsilon$ **then**
    converge = True
  $\boldsymbol{p}^{(old)} = \hat{\boldsymbol{p}}$

2.4. Log-likelihood

We can work the following expression for the log-likelihood.

$$
\begin{aligned}
l(\boldsymbol{p};\mathcal{X}) &= \sum_{b\in\mathcal{B}} \ln P(\boldsymbol{X}_b = \boldsymbol{x}_b|\boldsymbol{p}) && (2.5)\\
&= \sum_{b\in\mathcal{B}} \ln\left[\sum_{i\in\mathcal{I}_b,c\in\mathcal{C}} P(\boldsymbol{X}_b = \boldsymbol{x}_b, Y_{bg(i)c} = 1|\boldsymbol{p})\right]\\
&= \sum_{b\in\mathcal{B}} \ln\left[\sum_{i\in\mathcal{I}_b,c\in\mathcal{C}} P(Y_{bg(i)c} = 1|\boldsymbol{X}_b = \boldsymbol{x}_b;\boldsymbol{p}^{(old)})\cdot\frac{P(\boldsymbol{X}_b = \boldsymbol{x}_b, Y_{bg(i)c} = 1|\boldsymbol{p})}{P(Y_{bg(i)c} = 1|\boldsymbol{X}_b = \boldsymbol{x}_b;\boldsymbol{p}^{(old)})}\right]\\
&= \ln\mathbb{E}\left[\left(\frac{P(\mathcal{X},\mathcal{Y}|\boldsymbol{p})}{P(\mathcal{Y}|\mathcal{X},\boldsymbol{p}^{(old)})}\right)|\mathcal{X},\boldsymbol{p}^{(old)}\right]\\
&\geq \mathbb{E}\left[\ln\left(\frac{P(\mathcal{X},\mathcal{Y}|\boldsymbol{p})}{P(\mathcal{Y}|\mathcal{X},\boldsymbol{p}^{(old)})}\right)|\mathcal{X},\boldsymbol{p}^{(old)}\right] && \text{(Jansen's Inequality)}\\
&= \mathbb{E}\left[l(\boldsymbol{p};\mathcal{X},\mathcal{Y})|\mathcal{X},\boldsymbol{p}^{(old)}\right] - \mathbb{E}\left[\ln P(\mathcal{Y}|\mathcal{X};\boldsymbol{p}^{(old)})|\mathcal{X},\boldsymbol{p}^{(old)}\right]\\
&= Q(\boldsymbol{p};\boldsymbol{p}^{(old)}) - \mathbb{E}\left[\ln P(\mathcal{Y}|\mathcal{X};\boldsymbol{p}^{(old)})|\mathcal{X},\boldsymbol{p}^{(old)}\right] && (2.6)\\
&= \sum_{b\in\mathcal{B},g\in\mathcal{G}} w_{bg}\sum_{c\in\mathcal{C}} q_{bgc}\ln(p_{gc}) - \sum_{b\in\mathcal{B},g\in\mathcal{G}} w_{bg}\sum_{c\in\mathcal{C}} q_{bgc}\ln(q_{bgc})\\
&= \sum_{b\in\mathcal{B},g\in\mathcal{G}} w_{bg}\sum_{c\in\mathcal{C}} q_{bgc}\ln\left(\frac{p_{gc}}{q_{bgc}}\right)
\end{aligned}
$$

We can see that 2.6 works as a lower bound for the log-likelihood. As in each iteration we are maximizing over $\boldsymbol{p}$, we only focus in $Q(\boldsymbol{p};\boldsymbol{p}^{(old)})$, thus, increasing the lower bound for each iteration. This process is repeated until we get close enough to convergence ($\boldsymbol{p} = \boldsymbol{p}^{(old)}$), where 2.5 inequality becomes an equality.

Chapter 3

Computing the Conditional Probability

In order to perform the EM-algorithm we will need to compute the conditional probability q_{bgc} described in 2.1. In this chapter we will present an algorithm to compute this probability exactly and approximate methods that reduce its complexity.

3.1. Exact Method (EXACT)

A naive approach to calculate q_{bgc} would be to just consider each possible scenario that satisfies the observed data and compute using the law of total probabilities.

$$\begin{aligned} P(Y_{bi(g)c} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b) &= \sum_{\boldsymbol{z}_b \in \Omega_b} P(Y_{bic} | \boldsymbol{Z}_b = \boldsymbol{z}_b) P(\boldsymbol{Z}_b = \boldsymbol{z}_b) \\ &= \sum_{\boldsymbol{z}_b \in \Omega_b} \frac{z_{bgc}}{I_b} P(\boldsymbol{Z}_b = \boldsymbol{z}_b) \end{aligned}$$

Where Ω_b is the set of all possible outcomes of $\boldsymbol{Z}_b$ given the election result in that ballot-box. This can be represented by the following restrictions.

$$\begin{aligned} \sum_{g \in \mathcal{G}} z_{bgc} &= x_{bc} && \forall c \in \mathcal{C} \\ \sum_{c \in \mathcal{C}} z_{bgc} &= w_{bg} && \forall g \in \mathcal{G} \\ z_{bgc} &\geq 0 && \forall c \in \mathcal{C}\ \forall g \in \mathcal{G} \\ z_{bgc} &\in \mathbb{Z} && \forall c \in \mathcal{C}\ \forall g \in \mathcal{G} \end{aligned} \tag{3.1}$$

This method presents two complications: first, the occurrence of redundant calculations due to overlapping probabilities, and second, the lack of a straightforward means to acquire all feasible outcomes from Ω_b. To address these challenges, we introduce a recursive strategy. This methodology centers on tracking the aggregate election result up to a fixed group $f \in \mathcal{G}$.

For this purpose, we introduce the notation $T_{bf}(\boldsymbol{k})$, which denotes the probability of observing the aggregate result $\boldsymbol{k} = (k_1, ..., k_C)$ within ballot-box b up to the inclusion of group f. By using this concept and applying the law of total probabilities, we arrive at the following derivation:

$$\begin{aligned} T_{bf}(\boldsymbol{k}) &= \sum_{\boldsymbol{h} \in \mathcal{H}_{bf} : \boldsymbol{h} \leq \boldsymbol{k}} T_{b,f-1}(\boldsymbol{k} - \boldsymbol{h}) \cdot \binom{w_{bf}}{h_1, ..., h_C} \prod_{c \in \mathcal{C}} p_{fc}^{h_c} \\ T_{b0}(\boldsymbol{k}) &= \mathbf{1}\{\boldsymbol{k} = 0\} \end{aligned} \tag{3.2}$$

Where $\mathcal{H}_{bf}$ is the set of all possible voting outcomes for group f. We will also define $\mathcal{K}_{bf}$ as the set of all possible voting aggregated outcomes of what has happened until group f. This last set determines all the values of $\boldsymbol{k}$ that we will need to compute with this recursive approach, starting from $f = 0$ until $f = G$. The formulation for these sets is as follows:

$$\begin{aligned} \mathcal{H}_{bf} &= \{\boldsymbol{h} \in \mathbb{Z}_+^C : \sum_{c \in \mathcal{C}} h_c = w_{bf} \ \ \forall c \in \mathcal{C}\} \\ \mathcal{K}_{bf} &= \{\boldsymbol{k} \in \mathbb{Z}_+^C : \sum_{c \in \mathcal{C}} k_c = \sum_{f' \leq f} w_{bf'}; k_c \leq x_{bc} \ \ \forall c \in \mathcal{C}\} \end{aligned} \tag{3.3}$$

We can see that calculating $P(\boldsymbol{X}_b = \boldsymbol{x}_b)$ is equivalent to calculating $T_{bG}(\boldsymbol{x}_b)$.

Furthermore, it will be necessary to compute the conditional probabilities for the case where an individual from group g has already voted for candidate c. Employing analogous

notation, we introduce $U_{bfgc}(\boldsymbol{k})$ to represent the probability of observing the partial result $\boldsymbol{k} = (k_1, ..., k_C)$ within ballot-box b up to the inclusion of group f, conditional to $Y_{bi(g)c} = 1$. This can be expressed as follows:

$$U_{bfgc}(\boldsymbol{k}) = \sum_{\boldsymbol{h} \in \mathcal{H}_{bf}: \boldsymbol{h} \leq \boldsymbol{k}} U_{b,f-1,gc}(\boldsymbol{k} - \boldsymbol{h}) \cdot \binom{w_{bf}}{h_1, ..., h_C} \cdot \left(\prod_{c' \in \mathcal{C}} p_{fc'}^{h'_c} \right) \cdot \left(\frac{h_c}{p_{fc} w_{bf}} \right)^{\mathbb{1}_{\{f=g\}}}$$

$$U_{b0gc}(\boldsymbol{k}) = \mathbf{1}\{\boldsymbol{k} = 0\}$$

To determine the probability of the complete conditional outcome, it suffices to compute $U_{bGgc}(\boldsymbol{x}_b)$ for each group and candidate. By incorporating both discussed components, we arrive at an expression for q_{bgc}.

$$q_{bgc} = \frac{U_{bGgc}(\boldsymbol{x}_b) \cdot p_{gc}}{T_{bG}(\boldsymbol{x}_b)}$$

We can compute the conditional probability for the last candidate (or an alternative candidate) using the complement probability: $U_{bGgC}(\boldsymbol{x}b) = 1 - \sum_{c<C} U_{bGgc}(\boldsymbol{x}_b)$.

To calculate $\boldsymbol{q}_b$, we propose the following algorithm employing dynamic programming.

Algorithm 2 Exact $\boldsymbol{q}_b$ computation with dynamic programming.

Initialize $\mathcal{H}_{bf}, \mathcal{K}_{bf}$ as in 3.3
$T_{bf}(\boldsymbol{k}) = 0$; $U_{bfgc}(\boldsymbol{k}) = 0$; $\forall \boldsymbol{k} \in \mathcal{K}_{bf}$
for $f = 1$ to G **do**
  for $\boldsymbol{k} \in \mathcal{K}_{bf}$ **do**
    for $\boldsymbol{h} \in \mathcal{H}_b f$ **do**
      if $h_c \leq k_c \; \forall c \in \mathcal{C}$ **then**
        $a = \binom{w_{bf}}{h_1, ..., h_C} \prod_{c \in \mathcal{C}} p_{fc}^{h_c}$
        $T_{bf}(\boldsymbol{k}) = T_{bf}(\boldsymbol{k}) + T_{bf}(\boldsymbol{k} - \boldsymbol{h}) \cdot a$
        for $c \in \mathcal{C}$ **do**
          for $g = 1$ to $f - 1$ **do**
            $U_{bfgc}(\boldsymbol{k}) = U_{bfgc}(\boldsymbol{k}) + U_{b,f-1,gc}(\boldsymbol{k} - \boldsymbol{h}) \cdot a$
          $U_{bggc}(\boldsymbol{k}) = U_{bggc}(\boldsymbol{k}) + U_{b,g-1,gc}(\boldsymbol{k} - \boldsymbol{h}) \cdot a \cdot \left(\frac{h_c}{p_{fc} w_{bf}} \right)$
Output $q_{bgc} = \frac{U_{bGgc}(\boldsymbol{x}_b) \cdot p_{gc}}{T_{bG}(\boldsymbol{x}_b)}$

We can see that for each group $f \in \mathcal{G}$ we are doing $C \cdot |\mathcal{K}_b f| \cdot |\mathcal{H}_{bf}|$ operations, thinking of a worst case scenario this could be $C \cdot (I_B)^{2C}$ operations. We conclude that the complexity of the algorithm is $O(B \cdot G \cdot C \cdot (I_B)^{2C})$.

3.2. Hit-and-Run Method (H&R)

Given the complexity of the EXACT method to compute q_{bgc} we may use an approach using simulation so that we do not need to consider each possible outcome of the election.

We shall further examine Ω_b, the set of all possible outcomes for $\boldsymbol{Z}_b = \boldsymbol{z}_b$, as previously defined in 3.1. The depiction of an outcome within ballot-box b can be illustrated using matrix representation.

		Candidate			
		1	$\cdots$	C	
Group	1	z_{b11}	$\cdots$	z_{b1C}	w_{b1}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	G	z_{bG1}	$\cdots$	z_{bGC}	w_{bG}
		x_{b1}	$\cdots$	x_{bC}	

Where the sum of row g is set to w_{bg} and the sum of column c is set to x_{bc}.

Our objective is to generate sample values for $\boldsymbol{z}_b$ while adhering to the conditions outlined in the matrix. One approach to achieve this is by iteratively updating the values within the matrix using a stochastic process. This process should ensures that we end up in a point that is independent of the initial position.

We may note that shifting two components of $\boldsymbol{z}_b$ that are neither in the same row nor column results in the alteration of two other components of $\boldsymbol{z}_b$. To illustrate, consider the scenario where we move z_{bgc} and $z_{bg'c'}$ downward by one unit; to maintain feasibility within the polytope, it necessitates moving $z_{bgc'}$ and $z_{bg'c}$ upward by one unit. Additionally, when opting to reduce a component of $\boldsymbol{z}_b$, we should verify that this modification does not infringe upon the non-negativity constraints.

We propose the following algorithm to sample values for $\boldsymbol{z}_b$.

Algorithm 3 Hit and Run

Input: $\boldsymbol{w}_b, \boldsymbol{x}_b, M, S$
Initialize $\boldsymbol{z}_b = \alpha(\boldsymbol{w}_b, \boldsymbol{x}_b)$, $\mathcal{S}_b = \phi$
for $s = 1$ to S **do**
  for $m = 1$ to M **do**
    Sample without replacement $\{g, g'\} \leftarrow \mathcal{G}$ and $\{c, c'\} \leftarrow \mathcal{C}$
    if $z_{bgc} > 0$ and $z_{bg'c'} > 0$ **then**
      $z_{bgc} \leftarrow z_{bgc} - 1$
      $z_{bg'c'} \leftarrow z_{bg'c'} - 1$
      $z_{bgc'} \leftarrow z_{bgc'} + 1$
      $z_{bg'c} \leftarrow z_{bg'c} + 1$
  $\mathcal{S}_b = \mathcal{S}_b \cup \{\boldsymbol{z}_b\}$
Output: $\mathcal{S}_b$

Where S is the number of samples, M is the step size and α is any function that gives a starting point for $\boldsymbol{z}_b$ (see Annex A.1). In Annex A.2 we discuss the election of M.

The complexity of the Hit and Run algorithm is $O(M \cdot S)$, as the size of the instance does not affect the number of operations (we always change 4 components when moving).

We could also consider methods of sampling that use polytopes defined with inequalities, where there are hit and run approaches that guarantee a uniform distribution (Mete & Zabinsky, 2012). The formulation for this approach is described in Annex A.3. It should be noted that this algorithm was not chosen as the number of iterations to get uncorrelated points was noticeably higher in this context.

As we sample we obtain a subset $\mathcal{S}_b \subseteq \Omega_b$. It is now possible to condition the event $Y_{bi(g)c} = 1$ to both $\mathcal{S}_b$ and $\boldsymbol{X}_b$. We may note that $P(\boldsymbol{X}_b = \boldsymbol{x}_b | \mathcal{S}_b) = 1$, since we built $\mathcal{S}_b$ so that all its element satisfy the given outcome of the ballot-box. Considering this method we would get the following approximation.

$$
\begin{aligned}
q_{bgc} &= P(Y_{bi(g)c} = 1 | \mathbf{X_b} = \mathbf{x_b}, \mathcal{S}_b; \boldsymbol{p}) \\
&= P(Y_{bi(g)c} = 1 | \mathcal{S}_b; \boldsymbol{p}) \\
&= \sum_{\boldsymbol{z}_b \in \mathcal{S}_b} P(Y_{bi(g)c} = 1 | \boldsymbol{Z}_b = \boldsymbol{z}_b) P(\boldsymbol{Z}_b = \boldsymbol{z}_b) \\
&= \sum_{\boldsymbol{z}_b \in \mathcal{S}_b} \frac{z_{bgc}}{w_{bg}} \cdot \prod_{g' \in \mathcal{G}} \binom{w_{bg'}}{z_{bg'1}, ..., z_{bg'C}} \prod_{c' \in \mathcal{C}} p_{g'c'}^{z_{bg'c'}}
\end{aligned}
$$

The complexity of this calculation is: $O(S \cdot B \cdot G \cdot C)$. It should be noted that the term

$P(\boldsymbol{Z}_b = \boldsymbol{z}_b)$ can be precomputed for each ballot-box and does need to be calculated in each combination.

3.3. Multivariate-Normal Approximation Method Using the CDF (MVN-CDF)

In this section we will consider a method based on the Normal-Approximation proposed by Zhengzhi Lin (Lin, Wang, & Hong, 2022). Lin's application is done for a more general case, where each component of a matrix has a distinct categorical probability, described as the Poisson Multinomial Distribution. We adapted this implementation to the case of having groups.

Since the last component of $\boldsymbol{X}_b$ is determined by the rest, we will consider $\boldsymbol{X}_b^*$ as the reduced version where we omit the last component: X_{bC}.

We will consider the following asymptotic distribution:

$$\boldsymbol{X}_b^* \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$$

Consider $\boldsymbol{p}^*$ as the reduced version of $\boldsymbol{p}$. In order to obtain the mean of the distribution we just consider that the expected voting outcome for one person of group g is the vector $\boldsymbol{p}_g^*$ (we are not considering the last candidate).

$$\begin{aligned} \boldsymbol{\mu}_b &= \boldsymbol{p}^{*T}\boldsymbol{w}_b \\ \boldsymbol{\Sigma}_b &= \operatorname{diag}(\boldsymbol{\mu}_b) - \boldsymbol{p}^{*T}\operatorname{diag}(\boldsymbol{w}_b)\boldsymbol{p}^* \end{aligned} \tag{3.4}$$

We can also get the asymptotic distribution to the conditional random variable. Using the same notation for reduced variables we get:

$$\begin{aligned} \boldsymbol{X}_b^* | Y_{bj(g)c} = 1 &\sim \mathcal{N}(\boldsymbol{\mu}_b^g, \boldsymbol{\Sigma}_b^g) + \boldsymbol{e}_c \\ \boldsymbol{\mu}_b^g &= \mu_b - \boldsymbol{p}_g^{*T} \\ \boldsymbol{\Sigma}_b^g &= \boldsymbol{\Sigma}_b - \operatorname{diag}(\boldsymbol{p}_g^{*T}) - \boldsymbol{p}_g^{*T}\boldsymbol{p}_g^* \end{aligned} \tag{3.5}$$

We can use a continuous approximation to compute the desired probability. Let us consider F_{bg} to be the cumulative density function of the multivariate normal approximation just

described (note that it does not depend on c, since it is only added as a constant). Consider the following hypercube:

$$\mathcal{A}_{bc} \coloneqq [x_{b1} - 0.5, x_{b1} + 0.5] \times ... \times [x_{bc} - 1.5, x_{bc} - 0.5] \times ... \times [x_{bC-1} - 0.5, x_{bC-1} + 0.5]$$

Then we have the following approximation.

$$\begin{aligned} P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bi(g)c} = 1) &\approx P_{NA}(\boldsymbol{x}_b - 0.5 \leq \boldsymbol{X}_b \leq \boldsymbol{x}_b + 0.5 | Y_{bi(g)c} = 1) \\ &=: F_{bg}(\mathcal{A}_{bc}) \end{aligned}$$

Using this method we will need to compute the integral $F_{bg}(\mathcal{A}_{bc})$ for all combinations of b, g, c. This integral can be computed using Monte Carlo integration methods such as the one proposed by (Genz, 1992). Using this approximation we would end up with the following result for the conditional probabilities.

$$q_{bgc} = \frac{F_{bg}(\mathcal{A}_{bc}) \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} F_{bg}(\mathcal{A}_{bc'}) \cdot p_{gc'}}$$

The complexity of this algorithm is as described in Monte Carlo integration, where we consider an error parameter ϵ, obtaining a complexity of $O(B \cdot G \cdot C^2 \cdot \epsilon^{\frac{-1}{2}})$ for running each integral. We also get a complexity of $O(B \cdot G \cdot C^3)$ for doing the Cholesky decomposition for each group. Then the complexity of the method is $O(B \cdot G \cdot C^2 \cdot [\epsilon^{\frac{-1}{2}} + C])$. It should be noted that the number of iterations needed to compute the integral may depend on the size of the instance. We are considering a scale of C^2 for the integrals as for each combination of C the integration method considers a simulation of $C - 1$ random uniform variables.

3.4. Multivariate-Normal Approximation Method Using the PDF (MVN-PDF)

We may note that in order to calculate $P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bi(g)c} = 1)$, we do not necessarily need to calculate $P(Y_{bi(g)c} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b)$, but we may calculate something proportional to it. What we are really interested in is the ratio between $P(Y_{bi(g)c} = 1 | \boldsymbol{X}_b = \boldsymbol{x}_b)$ for the different values of c for a fixed g.

Using this idea we may choose to calculate something that is easier to compute than the actual probability. Based on the approximation in 3.4, we may use the probability density function instead of the cumulative density function. This can be done as we are only interested in the ratios between values to calculate the probability, and we do not need this quantity to represent a probability on itself. We could also interpret it as taking the limit where we make the hypercube as small as possible.

Using the same notation as the previous method, let us consider f_{bg} as the probability density function for the random variable described in 3.5. Using this notion we can get this simple approximation.

$$\begin{aligned} q_{bgc} &= \frac{f_{bg}(x_{b1}, ..., x_{bc} - 1, ..., x_{bC}) \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} f_{bg}(x_{b1}, ..., x_{bc'} - 1, ..., x_{bC}) \cdot p_{gc'}} \\ &= \frac{\sqrt{(2\pi)^{C-1}|\boldsymbol{\Sigma}_b^g|} \cdot e^{-\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{e}_c - \boldsymbol{\mu}_b^g)^\top (\boldsymbol{\Sigma}_b^g)^{-1}(\boldsymbol{x}_b - \boldsymbol{e}_c - \boldsymbol{\mu}_b^g)} \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} \sqrt{(2\pi)^{C-1}|\boldsymbol{\Sigma}_b^g|} \cdot e^{-\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{e}_{c'} - \boldsymbol{\mu}_b^g)^\top (\boldsymbol{\Sigma}_b^g)^{-1}(\boldsymbol{x}_b - \boldsymbol{e}_{c'} - \boldsymbol{\mu}_b^g)} \cdot p_{gc'}} \\ &= \frac{e^{-\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{e}_c - \boldsymbol{\mu}_b^g)^\top (\boldsymbol{\Sigma}_b^g)^{-1}(\boldsymbol{x}_b - \boldsymbol{e}_c - \boldsymbol{\mu}_b^g)} \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} e^{-\frac{1}{2}(\boldsymbol{x}_b - \boldsymbol{e}_{c'} - \boldsymbol{\mu}_b^g)^\top (\boldsymbol{\Sigma}_b^g)^{-1}(\boldsymbol{x}_b - \boldsymbol{e}_{c'} - \boldsymbol{\mu}_b^g)} \cdot p_{gc'}} \end{aligned}$$

Computing the inverse covariance matrix for each combination of g and c has a complexity of $O(B \cdot G \cdot C^3)$, as inversing a matrix has a cubic order in its dimension. Then, computing q_{bgc} considering the inverse covariance matrix as a precomputed parameter has a complexity of $O(B \cdot G \cdot C)$. We see that the complexity of the entire computation is $O(B \cdot G \cdot C^3)$.

3.5. Multinomial Approximation Method (MULT)

Considering that $\boldsymbol{X}_b$ is a sum of multinomial random variables, we may consider an approximation so that $\boldsymbol{X}_b$ distributes multinomial. To get this approximation we observe that each group g has a probability of voting for candidate c, so we can weight the contribution of each group with the number of voters in that group w_{bg}. We will consider r_{bc} to be the weighted probability for candidate c in ballot-box b.

$$\begin{aligned} r_{bc} = \frac{\sum_{g \in \mathcal{G}} w_{bg} p_{bg}}{\sum_{g \in \mathcal{G}} w_{bg}} = \frac{\sum_{g \in \mathcal{G}} w_{bg} p_{bg}}{I_b} \\ \boldsymbol{X}_b \sim \text{Multinomial}(I_b, \boldsymbol{r}_b) \end{aligned} \tag{3.6}$$

We can also get approximations for the conditional probabilities. If we consider that someone from group g already voted for candidate c, then its weight in the probability is decreased by one unit. Considering r_{bgc} to be the weighted probability for candidate c in ballot-box b, when someone from group g already voted for c:

$$r_{bgc} = \frac{\sum_{g' \in \mathcal{G}} w_{bg'} p_{g'c} - p_{gc}}{\sum_{g' \in \mathcal{G}} w_{bg'} - 1} = \frac{\sum_{g' \in \mathcal{G}} w_{bg'} p_{g'c} - p_{gc}}{I_b - 1}$$

$$\boldsymbol{X}_b | Y_{bi(g)c=1} \sim \text{Multinomial}(I_b - 1, \boldsymbol{r}_b^{(g)}) + \boldsymbol{e}_c$$

Using these approximations we can work out an expression for q_{bgc}.

$$\begin{aligned}
q_{bgc} &= \frac{P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bj(g)c} = 1) \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} P(\boldsymbol{X}_b = \boldsymbol{x}_b | Y_{bj(g)c'} = 1) \cdot p_{gc'}} \\
&= \frac{\binom{I_b - 1}{x_{b1}, \ldots, x_{bc} - 1, \ldots, x_{bC}} \cdot \prod_{d \in \mathcal{C}} r_{bgd}^{x_{bd}} \cdot r_{bgc}^{-1} \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} \binom{I_b - 1}{x_{b1}, \ldots, x_{bc'} - 1, \ldots, x_{bC}} \cdot \prod_{d \in \mathcal{C}} r_{bgd}^{x_{bd}} \cdot r_{bgc}^{-1} \cdot p_{gc'}} \\
&= \frac{\frac{(I_b - 1)!}{x_{b1}! \cdot \ldots (x_{bc} - 1)! \cdot \ldots x_{bC}!} \cdot \prod_{d \in \mathcal{C}} r_{bgd}^{x_{bd}} \cdot r_{bgc}^{-1} \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} \frac{(I_b - 1)!}{x_{b1}! \cdot \ldots (x_{bc'} - 1)! \cdot \ldots x_{bC}!} \cdot \prod_{d \in \mathcal{C}} r_{bgd}^{x_{bd}} \cdot r_{bgc}^{-1} \cdot p_{gc'}} \\
&= \frac{x_{bc} \cdot \prod_{d \in \mathcal{C}} r_{bgd}^{x_{bd}} \cdot r_{bgc}^{-1} \cdot p_{gc}}{\sum_{c' \in \mathcal{C}} x_{bc'} \cdot \prod_{d \in \mathcal{C}} r_{bgd}^{x_{bd}} \cdot r_{bgc'}^{-1} \cdot p_{gc'}} \\
&= \frac{\frac{x_{bc} \cdot p_{gc}}{r_{bgc}}}{\sum_{c' \in \mathcal{C}} \frac{x_{bc'} \cdot p_{gc'}}{r_{bgc'}}} \\
&= \frac{(x_{bc} \cdot p_{gc}) / r_{bgc}}{\sum_{c' \in \mathcal{C}} (x_{bc'} \cdot p_{gc'}) / r_{bgc'}} \\
&= \frac{x_{bc} \cdot p_{gc} \cdot r_{bgc}^{-1}}{\sum_{c' \in \mathcal{C}} x_{bc'} \cdot p_{gc'} \cdot r_{bgc'}^{-1}}
\end{aligned}$$

This complexity of this algorithm is straightforward as it relies on basic operations, scaling as: $O(B \cdot G \cdot C)$.

Chapter 4

Computing the p-value

4.1. Exact Computation

Once we have our estimate for $\boldsymbol{p}$ it is in our interest to determine how likely the result of each ballot-box is. Given the result $\boldsymbol{x}_b$, we can determine $P(\boldsymbol{X}_b = x_b)$ using algorithm 2. Since we want to know how likely is this result we would also need to compute the probability for all other possible outcomes $\boldsymbol{x}_b \in R_{\boldsymbol{X}_b}$, which can be done using the same set of recursive equations.

We will consider the p-value of ballot-box b as the probability of obtaining something as or less probable than $\boldsymbol{x}_b$. It can be computed as follows:

$$\begin{aligned} \text{p-val}(\boldsymbol{x}_b) &= \sum_{\boldsymbol{x}'_b \in R_{\boldsymbol{X}_b} | P(\boldsymbol{x}'_b) \leq P(\boldsymbol{x}_b)} P(\boldsymbol{x}'_b) \\ &= \sum_{\boldsymbol{x}'_b \in R_{\boldsymbol{X}_b}} P(\boldsymbol{x}'_b) \cdot \mathbb{1}_{\{P(\boldsymbol{x}'_b) \leq P(\boldsymbol{x}_b)\}} \\ &= \sum_{\boldsymbol{x}'_b \in R_{\boldsymbol{X}_b}} T_{bG}(\boldsymbol{x}'_b) \cdot \mathbb{1}_{\{T_{bG}(\boldsymbol{x}'_b) \leq T_{bG}(\boldsymbol{x}_b)\}} \end{aligned} \tag{4.1}$$

Where $T_{bG}(\boldsymbol{x}'_b)$ is as described in 3.2.

4.2. Simulation Using the Multinomial Approximation

We could approximate the p-value using simulation and the approximation proposed in 3.6. We may simulate H identical, independent and identically distributed random variables following a Multinomial$(I_b, \boldsymbol{r}_b)$, obtaining: $\boldsymbol{x}_b^{(1)}, ..., \boldsymbol{x}_b^{(H)}$.Using these simulations we may calculate the p-value by counting how many of this outcomes were less probable than $\boldsymbol{x}_b$:

$$\text{p-val}(\boldsymbol{x}_b) = \frac{1}{H} \sum_{h=1}^{H} \mathbb{1}_{\{P(\boldsymbol{x}_b^{(h)}) \leq P(\boldsymbol{x}_b)\}} \tag{4.2}$$

Where we we can compute the comparison as follows.

$$P(\boldsymbol{x}_b^{(h)}) \leq P(\boldsymbol{x}_b)$$

$$\binom{I_b}{x_{b1}^{(h)}, \ldots, x_{bC}^{(h)}} \prod_{c \in \mathcal{C}} (r_{bc})^{x_{bc}^{(h)}} \leq \binom{I_b}{x_{b1}, \ldots, x_{bC}} \prod_{c \in \mathcal{C}} (r_{bc})^{x_{bc}}$$

Chapter 5

Numerical Results

In this chapter we will compare the 5 methods studied for the EM-algorithm in Chapter 3: EXACT, H&R, MVN-CDF, MVN-PDF and MULT. We will also compare both methods for computing the p-values as described in Chapter 4.

We will describe the methodology for generating instances that represent the model described in Chapter 2. We will use instances with varying value of $G \in \{2, 3, 4\}$ and $C \in \{2, 3, 5, 10\}$, keeping fixed values of $B = 50$ and $I_b = 100$. We assume that all ballot-boxes will have the same amount of voters. For each instance we will run 20 scenarios.

5.1. Numerical Instances

In Algorithm 4 we describe the methodology to generate a scenario from an instance with parameters (I, G, C, B). We also include a mixing parameter λ that will be discussed further. A scenario is completely described by the observed information $\boldsymbol{x}$ and $\boldsymbol{w}$. We also output the real probability $\boldsymbol{p}$, used only to test estimations.

Algorithm 4 Election Outcome Generation

Input: I, G, C, B, λ
Output: $\boldsymbol{w} \in \mathbb{R}^{B\times G}$, $\boldsymbol{x} \in \mathbb{R}^{B\times C}$, $\boldsymbol{p} \in \mathbb{R}^{G\times C}$
Initialize: $\omega^0 : \omega_i^0 = \lceil \frac{i\cdot G}{I} \rceil$, $\omega = \omega^0$, $N = \frac{I}{B}$, $\mathcal{I} = \{1, ..., I\}$
Sample a vector $\boldsymbol{v} \in \mathbb{N}^{\lfloor \lambda \cdot I \rfloor}$ where each component is drawn without replacement from $\mathcal{I}$.
Generate $\boldsymbol{v}' = \text{sort}(\boldsymbol{v})$
for $i = 1$ to $\lfloor \lambda \cdot I \rfloor$ **do**
    $\omega_{v_i} = \omega^0_{v'_i}$
$w_{bg} = \sum_{i=(b-1)\cdot N+1}^{b\cdot N} \mathbb{1}_{\{\omega_i = g\}}$
Simulate $\boldsymbol{p}_g$ from Dirichlet($\mathbf{1}_C$)
Simulate z_{bgc} from Multinomial($w_{bg}, \boldsymbol{p}_g$)
$x_{bc} = \sum_{g\in\mathcal{G}} z_{bgc}$

Note that we choose to simulate $\boldsymbol{p}_g$ from a Dirichlet distribution with parameter 1 on all its components. This is done so that we obtain different real probabilities for each scenario and better represent the set of different elections that may occur.

We include the λ factor as a mixing parameter that controls how groups are sorted through the ballot-boxes For the EM-algorithm to get to a good estimator of the desired probabilities we would like the groups to have a heterogeneous distribution through the ballot-boxes. The intuition behind lies in how changes in the number of voters from a group in different ballot-boxes are affecting the outcome of votes.

We will test how the distribution of groups affects the mean absolute error (MAE) of the estimated probabilities. We calculate the MAE metric considering the absolute error of all components of the probability.

$$\frac{\sum_{g\in\mathcal{G};c\in\mathcal{C}} |p_{gc} - \hat{p}_{gc}|}{G \cdot C} \tag{5.1}$$

Figure 5.1 shows how the value of λ affects the estimation of the probability $\boldsymbol{p}$ using the EM-algorithm in different instances.

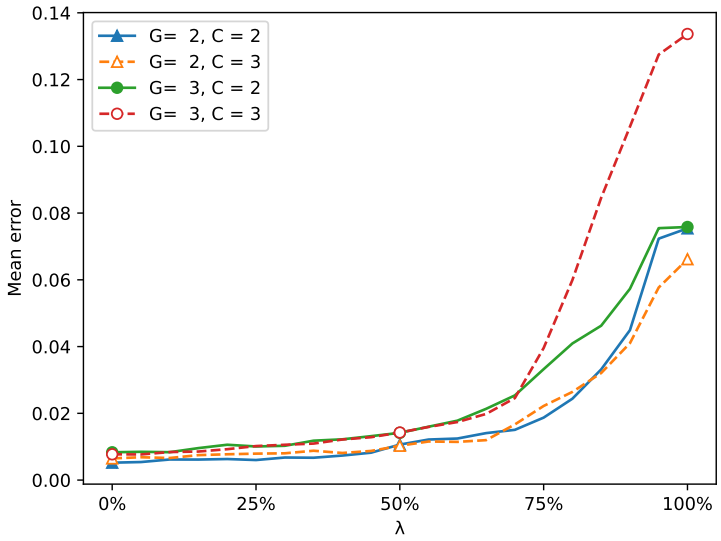


Figure 5.1: Mean absolute error over 20 scenarios for varying values of mix λ in different instances. $I_b = 100$, $B = 50$

We can observe that as the mixing parameter λ increases we get worse estimated probabilities. When we have small values of λ, ballot-boxes are more heterogeneous, thus, it is more apparent how the different composition in groups is affecting the outcome in votes.

According to these results we will consider that a mixing parameter of $\lambda = 50\%$ is sufficiently good for obtaining an estimator. We do not choose a lower value of λ since it may not be representative of what is really happening in the group assignation process of real Chilean elections.

5.2. EM-Algorithm Probability Estimation

In Annex B we present the results for the same instances considering $I_b = 200$.

5.2.1. Time Results

In Table 5.1 we present the average times that it takes the EM-algorithm to run. We set a time limit of 1 hour to each run, so we do not show results for those that did not converge. It should be noted that the number of iteration it takes the algorithm to converge varies across different methods, so the time presented takes into account how fast it converges and how fast the conditional probabilities are computed for each iteration.

Table 5.1: Mean running time over 20 scenarios in seconds for the EM-algorithm for varying instances, with fixed values of $I_b = 100$ and $B = 50$.

Instance		Method					
C	G	EXACT	H&R $S = 10^3$	H&R $S = 10^2$	MVN CDF	MVN PDF	MULT
2	2	14.510	719.939	72.335	0.851	0.140	0.003
	3	34.235	720.107	73.224	1.990	0.208	0.004
	4	51.563	723.778	72.621	3.421	0.233	0.003
3	2	549.902	709.502	71.909	3.604	0.217	0.003
	3	2,115.500	708.062	68.265	11.373	0.265	0.004
	4	2,808.304	676.996	67.471	15.927	0.320	0.005
5	2	-	683.890	67.924	12.315	0.173	0.003
	3	-	637.631	60.941	41.609	0.279	0.005
	4	-	603.475	59.644	47.176	0.281	0.004
10	2	-	594.104	60.332	61.702	0.181	0.003
	3	-	553.448	54.908	210.639	0.221	0.004
	4	-	514.156	50.178	355.949	0.308	0.004

The MULT method is the fastest and it is not highly impacted by C and G, as it is polynomial. We have a similar behavior for the MVN-PDF method. The EXACT method shows a clear exponential increase through C that makes it impossible to compute in the time limit for $C = 5$ and $C = 10$. Both H&R methods are not highly impacted by C and G, as the main time limitation is in the simulation, which will be discussed later on. The MVN-CDF method is also noticeably impacted by C but it was computable in the time limit.

We also show separate time results between simulation and EM-algorithm for the H&R methods in Table 5.2. It is clear that the bottle-neck of this algorithm comes in the simulation, as we will later show how the amount of samples affects the accuracy of the method. We can also observe that bigger instances tend to have lower simulation times due to the fact that the polytope is more restricted and it is more probable to hit the non-negativity constraints (thus, making less operations).

Table 5.2: Comparison of mean simulation time and mean EM running time in seconds over 20 scenarios for varying instances, with fixed values of $I_b = 100$ and $B = 50$.

Instance		Method			
C	G	H&R ($S = 10^2$)		H&R ($S = 10^3$)	
		Sim-time	EM-time	Sim-time	EM-time
2	2	71.869	0.466	719.469	0.471
	3	71.324	1.900	713.681	6.427
	4	70.213	2.408	702.951	20.827
3	2	70.307	1.603	704.637	4.865
	3	66.246	2.019	683.981	24.081
	4	64.956	2.515	651.587	25.409
5	2	66.042	1.882	666.138	17.752
	3	59.140	1.801	614.540	23.091
	4	57.597	2.047	581.373	22.102
10	2	58.708	1.624	578.243	15.861
	3	53.296	1.612	534.916	18.532
	4	48.476	1.701	492.142	22.014

5.2.2. Error Results

In Table 5.3 we show the MAE metric as described in 5.1. These results show that both MVN methods and the MULT method have similar MAE when compared to the EXACT method. Regarding the H&R method the magnitude of the error depends on the amount of points simulated. It is clear that bigger instances should require a bigger S, making the H&R method more demanding.

Table 5.3: Mean absolute prediction error over 20 scenarios for varying instances, with fixed values of $I_b = 100$ and $B = 50$.

Instance		Method					
C	G	EXACT	H&R $S = 10^3$	H&R $S = 10^2$	MVN CDF	MVN PDF	MULT
2	2	0.011	0.011	0.011	0.011	0.011	0.011
	3	0.014	0.014	0.016	0.014	0.014	0.015
	4	0.017	0.018	0.025	0.017	0.017	0.018
3	2	0.010	0.010	0.011	0.010	0.010	0.010
	3	0.014	0.015	0.025	0.014	0.014	0.015
	4	0.017	0.020	0.034	0.017	0.017	0.017
5	2	-	0.009	0.012	0.008	0.009	0.009
	3	-	0.016	0.025	0.013	0.013	0.013
	4	-	0.022	0.032	0.014	0.014	0.014
10	2	-	0.007	0.012	0.006	0.007	0.006
	3	-	0.015	0.020	0.009	0.009	0.009
	4	-	0.020	0.024	0.012	0.012	0.012

As we obtained similar MAE metrics for the MVN methods and the MULT method, we would recommend using the latter as it was shown to be the fastest.

5.3. p-value Estimation

In this section we will use a subset of the computed instances, considering $C = \{2, 3\}$ and $G = \{2, 3, 4\}$. We choose these ones as we need the EXACT method in order to do comparisons, where it is only reliable to run it until $C = 3$.

5.3.1. Probability Gap

It is in our interest to observe how the estimated probability $\hat{\boldsymbol{p}}$ changes the p-values obtained for each ballot-box compared to the original probability $\boldsymbol{p}$. In this case we will be comparing using the estimated $\hat{\boldsymbol{p}}$ that comes from the exact EM-method, we will call it $\hat{\boldsymbol{p}}_{(\text{EXACT})}$.

In Figure 5.2 we are showing the mean p-value absolute error and the mean absolute error between $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}_{(\text{EXACT})}$.

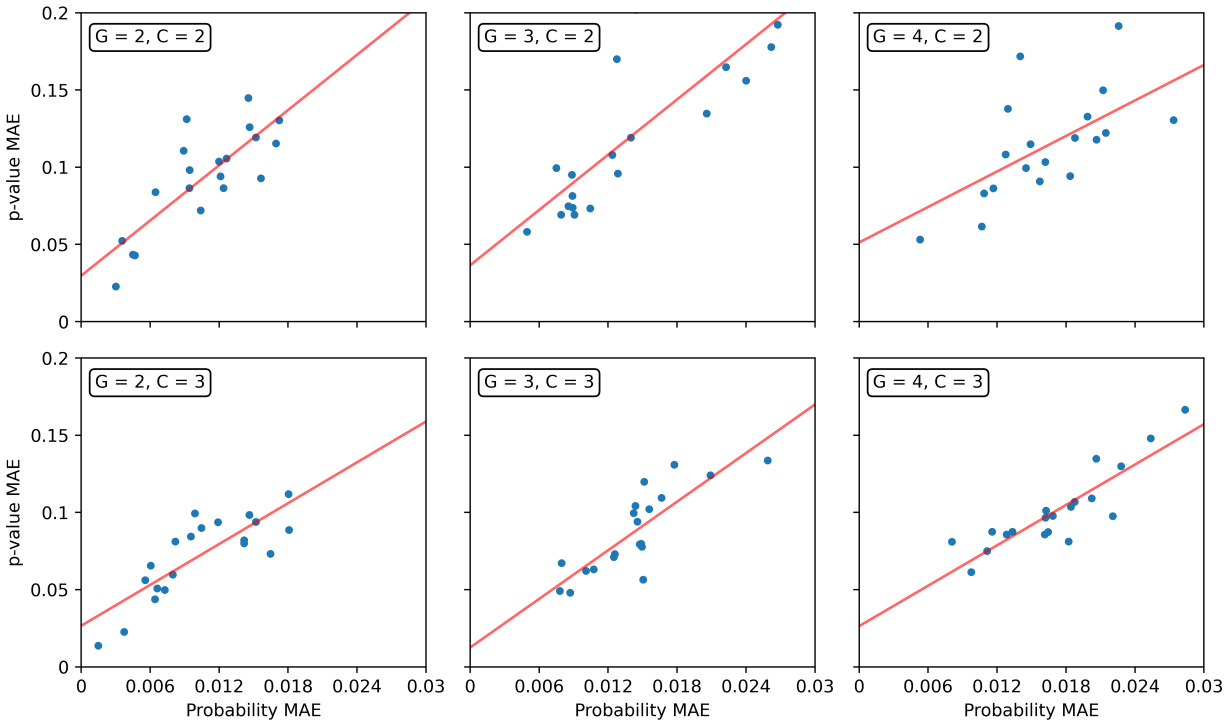


Figure 5.2: Correlation between p-values mean absolute error and estimated probability mean absolute error.

There is a clear correlation between both errors, which is the expected behavior.

We show a different approach in Figure 5.3 where we compare the p-values obtained using both probabilities $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}_{(\text{EXACT})}$. The color grading is now showing the magnitude of the MAE between both probabilities. We observe that the points with a smaller mean absolute error tend to be close to the middle, where both p-values are the same. We also observe that the p-values estimated with $\hat{\boldsymbol{p}}_{(\text{EXACT})}$ are overestimated and underestimated in similar amounts.

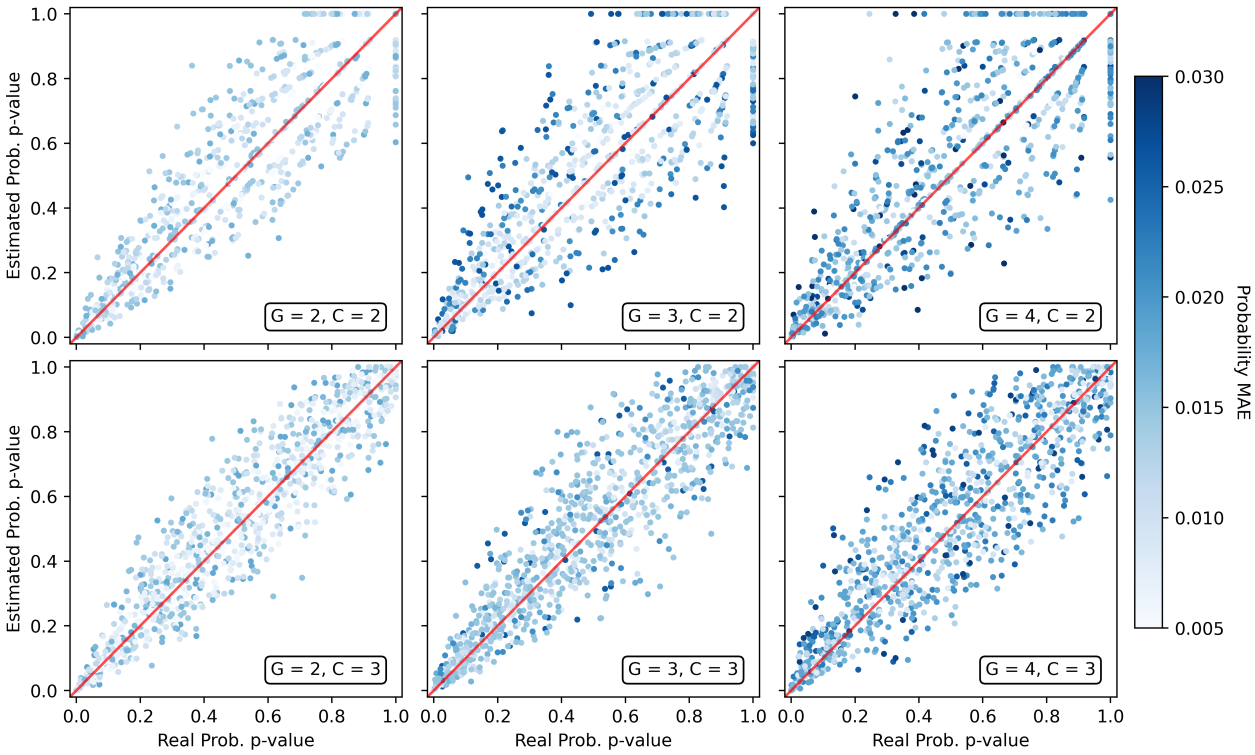


Figure 5.3: Comparison between p-values computed with real probability and p-values computed estimated probability.

5.3.2. Method Gap

We are also interested in studying how the simulation method using the multinomial approximation changes the computed p-values. For this comparison we will use the same probability $\boldsymbol{p}$ and we will only change the method used to compute the p-value in each ballot-box.

It is intuitive to think that as we do the multinomial approximation of the distribution we are altering the variance of the original distribution. To test this hypothesis we will calculate the covariance matrix for both distributions.

The covariance matrix of the original distribution was already explained in Section 3.3 and displayed in 3.4. The covariance matrix of the multinomial approximation can be easily obtained using the parameters of the approximation as follows.

$$\boldsymbol{\Sigma}_b^{(\text{Mult})} = I_b \cdot \left(\text{diag}(\boldsymbol{r}_b) - \boldsymbol{r}_b \boldsymbol{r}_b^\top\right)$$

To study how the variance changes in a ballot-box when using the multinomial approximation

we consider the ratio:

$$\frac{|\boldsymbol{\Sigma}_b^{(\text{Mult})}|}{|\boldsymbol{\Sigma}_b|}$$

In Figure 5.4 we show how this ratio affects the absolute error for different p-values. We also show in color grading the magnitude of the original p-value.

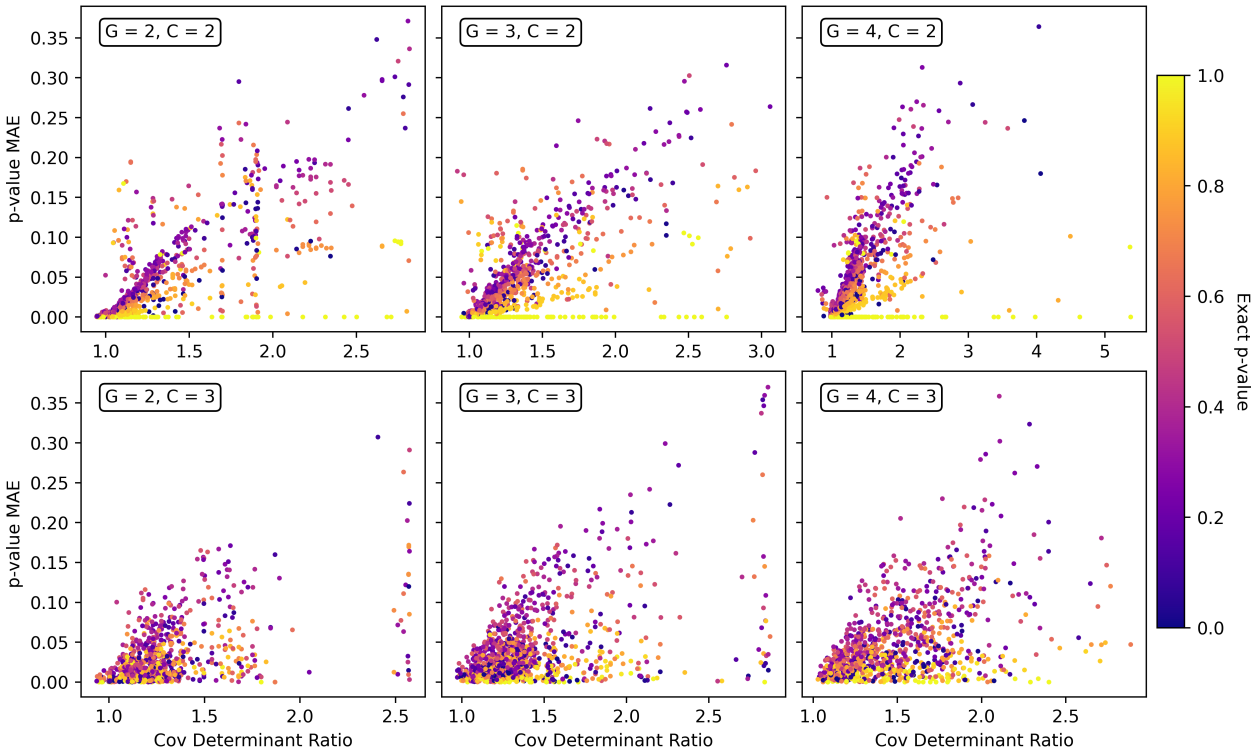


Figure 5.4: Effect of covariance determinant increase in the mean absolute error in the p-values.

We can observe that as the ratio between determinants increases we obtain higher mean absolute errors. We can also see that the error is more notable for p-values that are around the middle. This means that the main change in probability occurs to outcomes that are in the middle between the most probable and least probable outcome. The increase in variance we get through the multinomial approximation should increase the probability of this outcomes that are in the middle of the distribution.

In Figure 5.5 we show a different approach where we compare the p-value obtained with exact computation and the p-value obtained with simulation through the multinomial approximation.

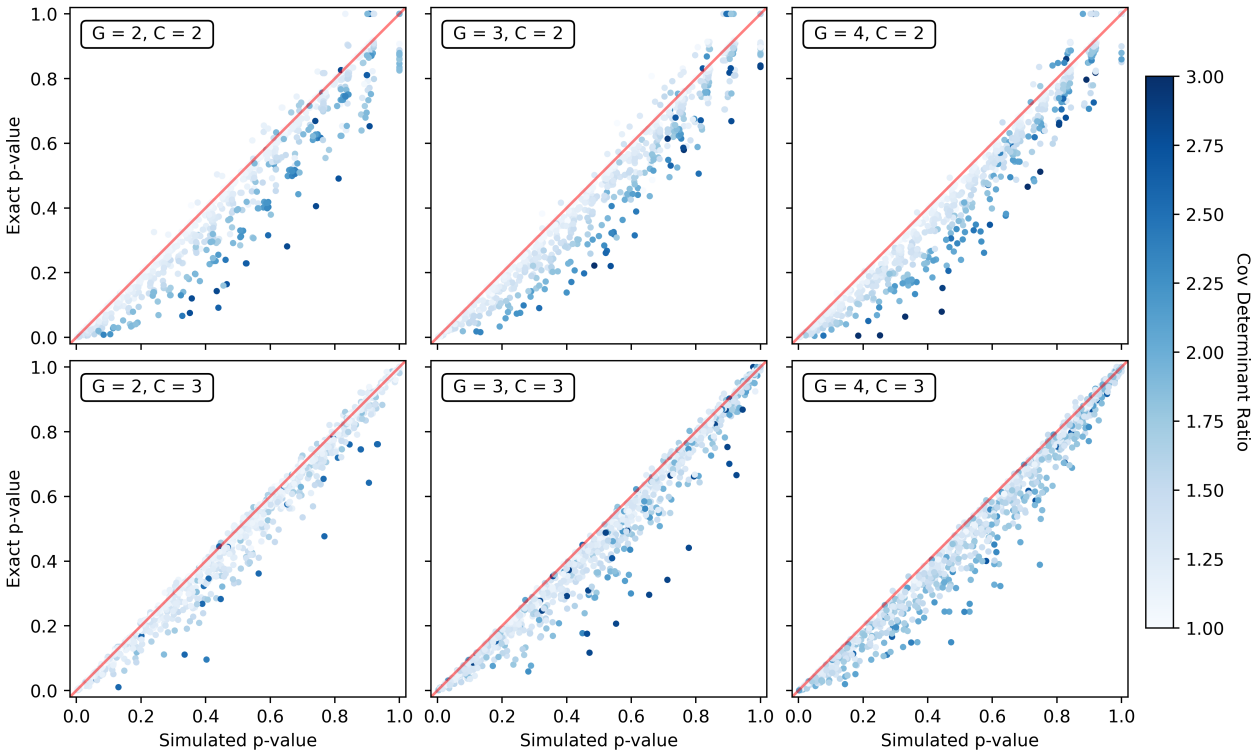


Figure 5.5: Comparison between p-values computed with the exact method and p-values simulated with the multinomial approximation.

It is clear that there is a bias that tends to increase the magnitude of the p-values computed with the approximation. We show in color grading the magnitude of the covariance determinant ratio. We can observe that those points with the biggest change in variance are the ones where the bias is more noticeable.

Chapter 6

Case Study: a Real Election

We will apply the EM-algorithm to the first round of the 2021 Chilean Presidential Election (2021-CPE) using the MULT method. This election adheres to the two-round system (TRS), which necessitates a subsequent round of voting featuring the two leading candidates if no single candidate secures an outright majority exceeding 50% during the first round.

6.1. Historical Data

The 2021-CPE election featured 7 candidates: Gabriel Boric, José Atonio Kast, Yasna Provoste, Sebastián Sichel, Eduardo Artés, Marco Enríquez-Ominami and Franco Parisi. Besides these 7 options, votes could be classified as blank votes (the ballot has no marking in it) or null votes (the ballot was filled out incorrectly). We will treat blank and null votes as a single category since different ballot-box committees may have applied varying criteria when labeling them.

Circunscripciones electorales are the territorial divisions used to assign voters to ballot-boxes. We will refer to them as districts. Across each district, there are distinct polling places, each of them with a fixed number of ballot-boxes. We will assume that two voters belonging to the same group within a district are considered comparable in terms of probability within that specific district. It is noteworthy that, in this election, the allocation of voters to ballot boxes did not consider geographical distance, a factor that was considered in subsequent elections starting from 2022.

In Figure 6.1 we show how the number of ballot-boxes varies across districts. We can observe that most districts have less than 50 ballot-boxes, which could make the estimation

of probabilities harder, as we have less information. We would expect that those districts with over a 100 ballot-boxes give us enough information for computing the estimating probabilities.

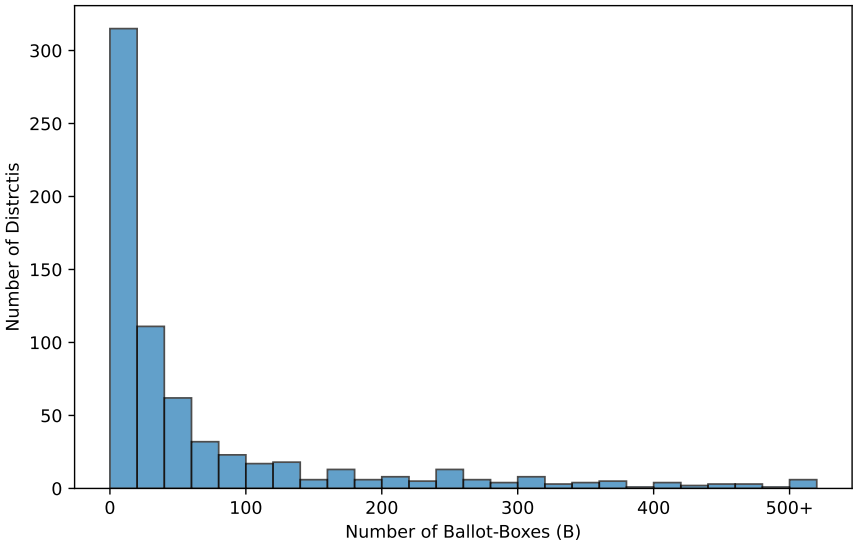


Figure 6.1: District count for ranges of length 25 for number of ballot-boxes.

The official website of the Electoral Service of Chile (SERVEL) offers detailed data of the election results (SERVEL, 2023). The detailed dataset for the votes emitted in Chile registers 7,080,276 votes emitted across 681 districts with a total of 46,639 ballot-boxes. We should note that this number may vary slightly after processing the data due to incomplete information in some ballot-boxes.

Additionally, SERVEL offers information about the aggregated demographic data of the voters in each ballot-box. There are three categorical features: sex, age range and nationality. There are 8 possible options for age range: 18-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80^{+}. For this application, we will only consider age ranges as possible groups, as there is evidence to suggest a correlation between age and political ideology (Leigh, 2005; Peterson, Smith, & Hibbing, 2020). We will not consider nationality or gender as it would further increase the number of possible groups.

In Figure 6.2 we show an example that illustrates how the group age distribution in different polling places at the district of *Puente Alto* would look. This shows at first glance that there is heterogeneity in both: districts and polling places.

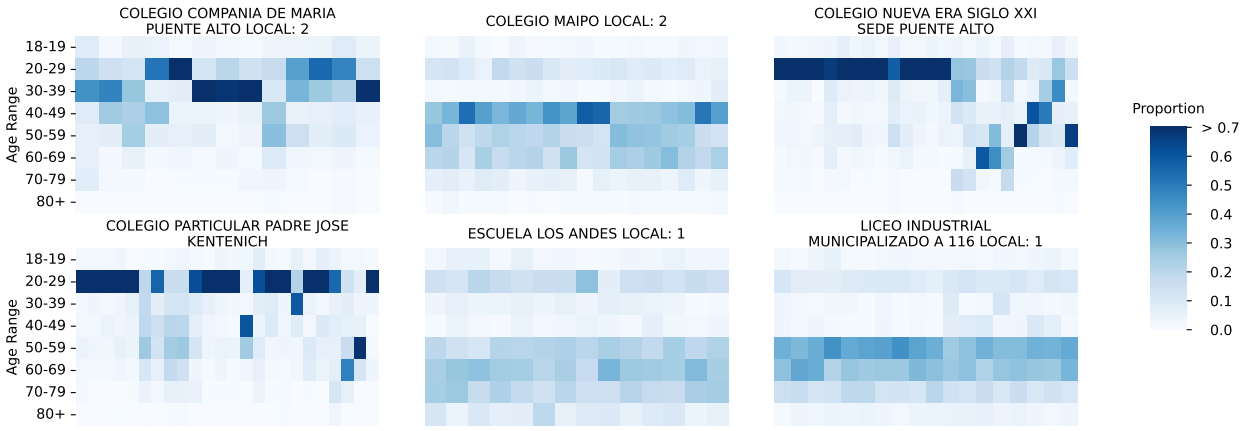


Figure 6.2: Proportions for different age ranges in polling places from the district of *Puente Alto*.

6.2. Group Aggregation

In order to run the EM-algorithm in the 2021-CPE we need to choose which groups to use. This is important because as we saw in Figure 5.1 the distribution of groups through ballot-boxes has an effect on the probability estimation.

The first group set that we could consider would be to use the full data available from age ranges, giving us the following:

$$\mathcal{G} = \{\text{18-19}, \text{20-29}, \text{30-39}, \text{40-49}, \text{50-59}, \text{60-69}, \text{70-79}, 80^{+}\}$$

However, we could aggregate some age ranges into one group aggregation a. This could be useful if the probabilities of that age range are difficult to identify given its distribution through the ballot-boxes. We may consider a group aggregation a as a set of groups form $\mathcal{G}$ where we aggregate all voters from that group. Then, a group set $\mathcal{A}$ is a set formed of group aggregations from $\mathcal{G}$.

For example, we could aggregate age ranges into 2 categories: voters up to 39 years old, and voter from 40 years old. This could be expressed as a set $\mathcal{A} = \{\text{18-39}, 40^{+}\}$. In Figure 6.3 we show how the distribution varies using this aggregation in the district of *Río Negro*.

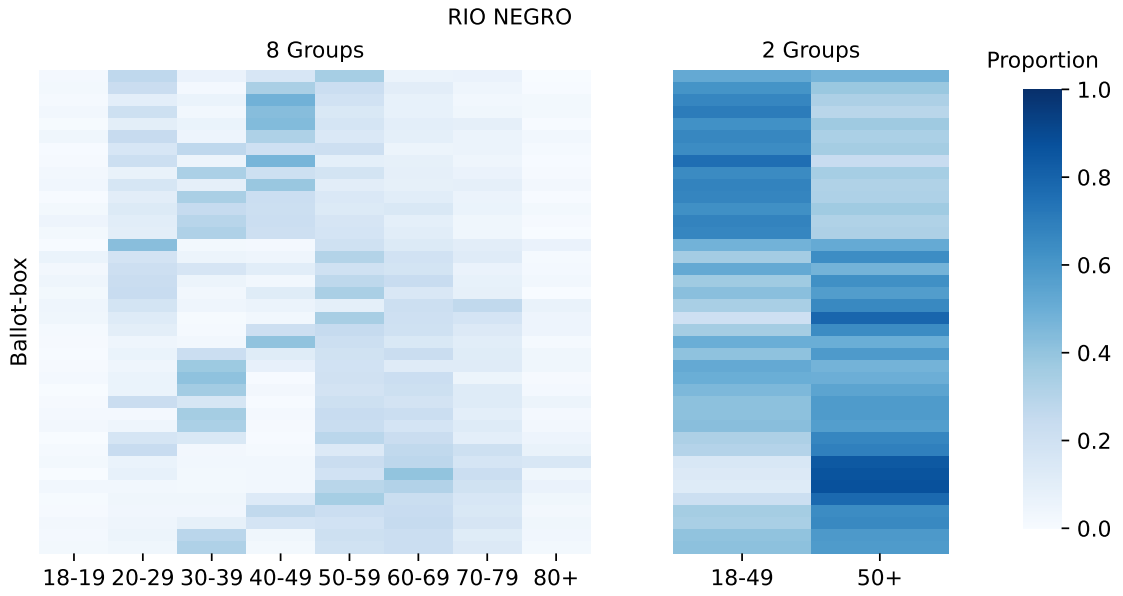


Figure 6.3: Proportions for different group aggregations in the district of *Río Negro*.

We can notice that when adding groups we may obtain distributions that are completely different. For example, we can observe that the extreme age ranges of 18-19 and 80^+ have few voters per ballot-box and show little variance accross the district. We can note that when we aggregate into 2 groups we obtain very identifiable categories, which we could expect to result in better estimations. It is not clear which aggregation is best, as we would like the estimated probabilities to better represent the reality of the district but we would not want them to be far off from the real latent probabilities.

To show an example of how the group aggregation affects the model, we will consider the original set $\mathcal{G}$ and another set $\mathcal{A} = \{\text{18-29}, \text{30-59}, 60^+\}$. We could run the EM-algorithm using both sets $\mathcal{G}$ and $\mathcal{A}$ to compare the estimated probabilities. In Figure 6.4 we show results for two different districts that noticeably vary in the number of ballot-boxes.

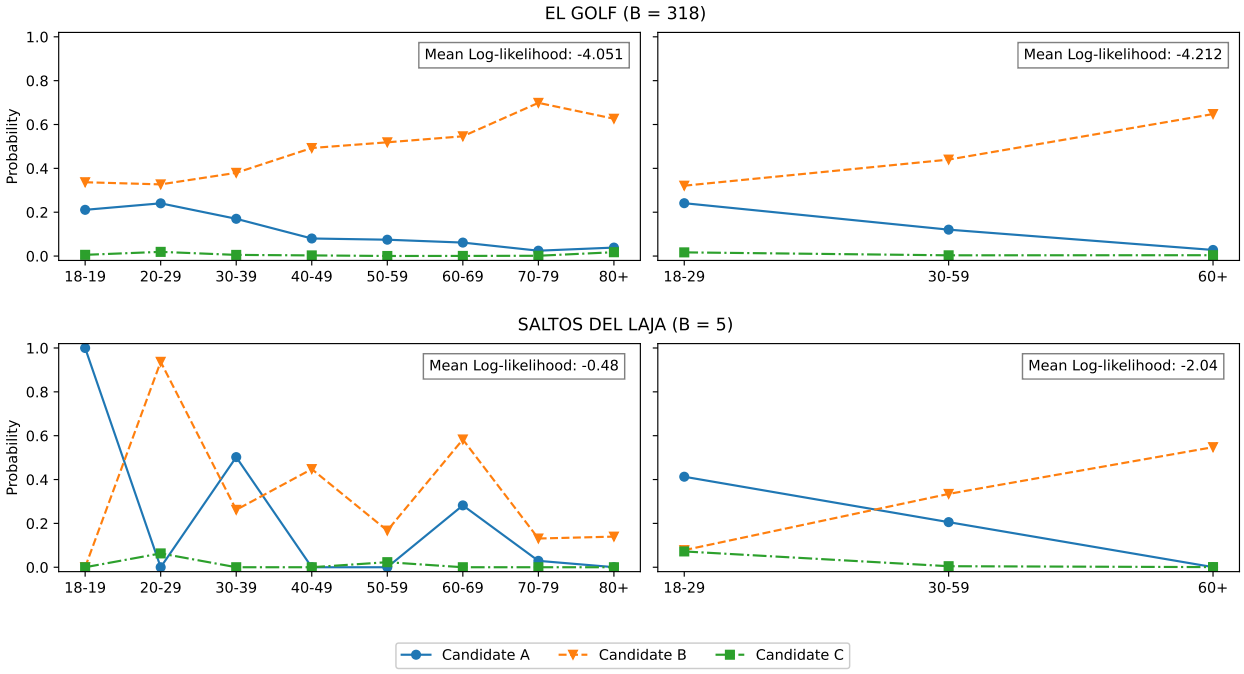


Figure 6.4: Estimated probabilities for different group aggregations.

The bottom of Figure 6.4 shows the estimated probabilities for the district *Saltos del Laja*. We observe from the bottom left panel that the estimated voting probabilities have big variations among different age ranges, specially for candidates A and B. For example, the estimated voting probability for candidate A is 1 for voters in the 18-19 age range, whereas the estimate goes to 0 for voters in the 20-29 age range. In the bottom right panel we observe the estimated voting probabilities when considering 3 aggregated age groups. In this case, the estimated probabilities among the aggregated age groups show a more regular pattern. The district of *Saltos del Laja* has only 5 ballot-boxes; as a result, it overfits the estimated probabilities when using eight demographic groups. Indeed, the log-likelihood of the bottom left case is considerably higher than the one of the bottom right panel. Note that the overfitting phenomenon, when using 8 age groups, does not occur in the district of *El Golf* (top left panel) since this district has a considerable amount of ballot boxes ($B = 318$). We observe that in the district of *El Golf* we get a similar probability in both the top left and top right panels, meaning that we should probably stay with the 8 age groups as they increase the log-likelihood.

These results show once again that is not clear which group aggregation is better, as it may depend in the group distribution and number of ballot-boxes in the district. Indeed, there

are several other possible group aggregations we could have done.. In order to approach this issue we will introduce a metric of reliability which could be obtain through bootstrapping. Besides, we could use the log-likelihood to measure how the probability is fitting to the distribution of the district.

6.2.1. Log-Likelihood

As we ran the EM-algorithm in a district, we obtain an estimated probability matrix $\hat{\boldsymbol{p}}$. We are not able to use the MAE metric anymore as we do not know the real probability. However, we are able to calculate the log-likelihood of the estimation using equation 2.6. This log-likelihood will not necessarily have an interpretation of how good the estimator is, but it could be useful to study how the estimation changes as we add more groups.

6.2.2. Bootstrapping: Estimated Probability Standard Deviation

As we ran the EM-algorithm in a district, we obtain an estimated probability matrix $\hat{\boldsymbol{p}}$. It is our objective to determine how reliable this estimation is. We could consider a bootstrapping approach to see how this estimated probability changes as we change the subset of ballot-boxes that are considered for running the EM-algorithm.

We can formulate a very simple algorithm as follows.

Algorithm 5 Bootstrapping for estimated probabilities

Input: $\mathcal{I}$, $\mathcal{G}$, $\mathcal{C}$, $\mathcal{B}$, $\boldsymbol{x}$, $\boldsymbol{w}$
Output: $\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(S)} \in [0,1]^{G \times C}$
for $s = 1$ to S **do**
  Sample: $a_1, a_2 ..., a_B$ from $\mathcal{U}\{1, B\}$ (with replacement)
  Create matrices: $\boldsymbol{x}' \in \mathbb{Z}_+^{B \times C}$, $\boldsymbol{w}' \in \mathbb{Z}_+^{B \times G}$
  Assign: $\boldsymbol{x}'_b = \boldsymbol{x}_{a_b}$, $\boldsymbol{w}'_b = \boldsymbol{w}_{a_b} \quad \forall b \in \mathcal{B}$
  Estimate $\boldsymbol{p}^{(s)}$ using $\boldsymbol{x}'$ and $\boldsymbol{w}'$ with Algorithm 1.

As we get a collection of estimated probabilities $\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(S)}$ we may use it to estimate a metric of standard deviation. This could be done for the probabilities estimated for an arbitrary group, as we are interested on observing if we should include that group or not.

The standard deviation of the estimated probability p_{gc} can be expressed as follows:

$$\text{std}(p_{gc}) \coloneqq \sqrt{\frac{1}{S-1} \cdot \sum_{s=1}^{S} (p_{gc}^{(s)} - \mu_{gc})^2}$$
$$\text{std}(\boldsymbol{p}) \coloneqq \frac{\sum_{g \in \mathcal{G}, c \in \mathcal{C}} \text{std}(\boldsymbol{p}_g)}{G \cdot C}$$

Where $\mu_{gc} \coloneqq \frac{1}{S} \sum_{s=1}^{S} p_{gc}^{(s)}$.

This metric will give us an insight on how reliable the EM-algorithm is for estimating the probabilities associated for each combination of group g and candidate c.

6.2.3. Adding More Groups: Trade-off

We will study how considering more groups affects the estimation through the log-likelihood and the standard deviation. We will consider the following sets of aggregated groups.

- 1 Group: $\mathcal{A} = \{18^+\}$
- 2 Groups: $\mathcal{A} = \{\text{18-49}, 50^+\}$
- 4 Groups: $\mathcal{A} = \{\text{18-29}, \text{30-49}, \text{50-69}, 70^+\}$
- 8 Groups: $\mathcal{A} = \{\text{18-19}, \text{20-29}, \text{30-39}, \text{40-49}, \text{50-59}, \text{60-69}, \text{70-79}, 80^+\}$

We will select districts with varying number of ballot-boxes through Chile. For each selected district we are going to run the EM-algorithm with the 4 different group aggregations. For each group aggregation we will compute both the log-likelihood and the standard deviation, obtaining the result displayed in Figure 6.5.

6.2.4. Best Group Aggregation

Given the results in Figure 6.5, we would like to choose the best group combination for each district. We know that we would like to maximize the log-likelihood but without sacrificing so much standard deviation in the estimator. In this particular application we will consider a threshold (γ) for the standard deviation each group voting probability must satisfy given a group set. We will choose this threshold to be $\gamma = 0.05$.

We would like to try each possible group set for each district, however, this is not possibly as the number of different group combinations scale exponentially. Consider we have G

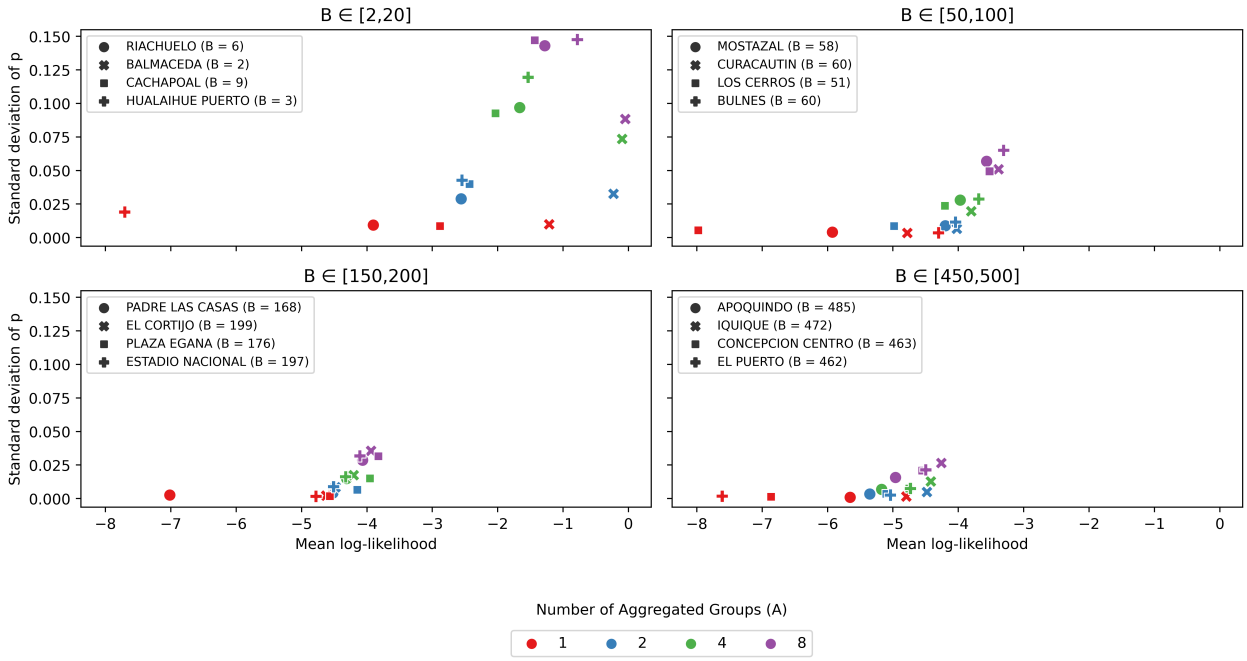


Figure 6.5: Standard deviation of estimated probability vs log-likelihood over districts of different ballot-boxes number.

possible age ranges and we can only merge neighbor age ranges, the total number of group sets of size A we can make are:

$$\binom{G-1}{A-1}$$

As trying each feasible group set is not a reliable option. We will propose an iterative approach to obtain a good group set for a given district. Firstly, we will determine a way to choose a group set given its size: A.

We know that we would like groups to be heterogeneous accross ballot-boxes in order to make good estimators, so we will choose the group set that maximizes the sum of the standard deviation of groups distribution in that set. This would be done through mixed-integer programming, where we set the number of groups as a constraint.

$$\begin{aligned} &\max_{\boldsymbol{a}} \sum_{m\in\mathcal{M}} a_m \cdot \sigma_m \\ &\text{subject to}: \sum_{m\in\mathcal{M}|g\in m} a_m = 1 \qquad \forall g \in \mathcal{G} \\ &\sum_{m\in\mathcal{M}} a_m = A \\ &a_m \in \{0,1\} \qquad \forall m \in \mathcal{M} \end{aligned} \tag{6.1}$$

The decision variable $\boldsymbol{a}$ represents the following:

$$a_m = \begin{cases} 1 & \text{Group combination } m \text{ is included in the set} \\ 0 & \text{Otherwise} \end{cases}$$

We are considering σ_m to be the standard deviation of group combination m. Then, $\mathcal{M}$ is the set of all possible group combinations given the full set $\mathcal{G}$. For the particular structure of this application we can define it as follows:

$$\mathcal{M} = \{m \in \mathcal{P}(\mathcal{G}) \mid \forall g_1, g_2 \in m, \{g \in \mathcal{G} \mid g_1 \leq g \leq g_2\} \subseteq m\} \tag{6.2}$$

We are considering the power set of $\mathcal{G}$ but only taking those subsets that satisfy that each element has at least one neighbor in the set. Let us note that we are considering $\mathcal{G} = \{1, 2, ..., G\}$ so that each index corresponds to an age range in ascending order.

Let us take into account that the cardinality: $|\mathcal{M}| = \sum_{k=1}^{G} = \frac{G \cdot (G+1)}{2}$, scales as $O(G^2)$. For the application in the 2021-CPE we have $|\mathcal{M}| = 36$ which is a reasonable instance of the problem.

Now that we have a way of choosing a group set given a size A, we will propose an iterative approach to get the final set for each district. The idea is that we would like to have as much groups as possible, as it is shown to improve the log-likelihood. This way we will start with the biggest set possible $\mathcal{G}$ (as it is the original set) and observe how the standard deviation of the estimation is using Algorithm 5. If the standard deviation is below the threshold then we choose that set, else we try with a smaller set of size $G - 1$. As we have different choices for a set of this size we solve the optimization problem described in 6.1. If this new solution does not satisfy the threshold we keep decreasing the set size. If no group set satisfies the threshold we just use the group set of size 1 where we merge all age ranges into one.

Algorithm 6 Choosing a group set

Input: $\mathcal{I}$, $\mathcal{G}$, $\mathcal{C}$, $\mathcal{B}$, $\boldsymbol{x}$, $\boldsymbol{w}$, γ
Output: $\mathcal{A}$
for $A = G$ to 1 **do**
  Solve the optimization problem described in 6.1, input $\mathcal{G}$, A
  $\mathcal{A} = \{m \in \mathcal{M} | a_m = 1\}$
  $w'_{ba} = \sum_{g \in a} w_{bg}, \forall a \in \mathcal{A}$
  Apply bootstrapping using Algorithm 5, input $\mathcal{I}$, $\mathcal{A}$, $\mathcal{C}$, $\mathcal{B}$, $\boldsymbol{x}$, $\boldsymbol{w}'$
  if $\text{std}(p_{ac}) \leq \gamma, \forall a \in \mathcal{A}, \forall c \in \mathcal{C}$ **then**
    Stop

6.3. Results

We will evaluate the discussed models and methodology for each district of the 2021-PGE. We will first choose aggregated group sets according to algorithm 6, then we will perform the EM-algorithm according to 1 and finally we will compute the p-values for each ballot-box using the multinomial approximation as described in 4.2 using a precision up to 10^9 samples. We should take into account that these p-values may have biases according to the discussion in Section 5.3.

6.3.1. Group Aggregation Sets

As we run the methodology established in 6 we obtained a group set $\mathcal{A}$ for each district in the election. We can study some statistics related to the chosen sets.

In Figure 6.6 we observe the fraction of districts where the original age range is selected as a single group. We can notice that the age ranges of 70-79 and 80^+ are almost never selected as a single group, as their probabilities must be hard to identify. A similar phenomenon occurs with the younger age ranges of 18-19 and 20-29, but they do get selected as a single groups in some more instances. We observe that those age ranges in the middle: 30-39, 40-49 and 50-59; are the age ranges that tend to be selected as a single group, as they probably are well represented through the district and the estimated probability does not radically change when doing bootstrapping.

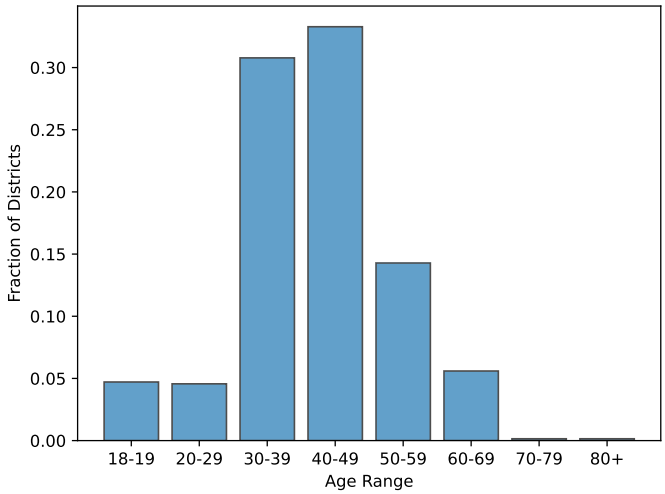
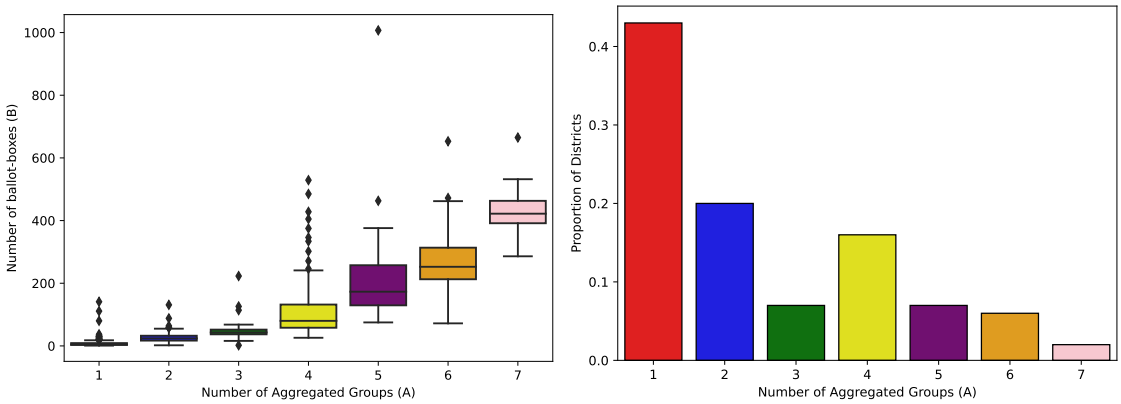


Figure 6.6: Fraction of districts where the age-range is included as a single group.

We can also study the size of the chosen group sets. In Figure 6.7 we display some statistics of the sets over the districts. In Figure 6.7.a we can observe how the number of ballot-boxes in a district affects the number of aggregated groups selected. We can observe that when there are more ballot-boxes the methodology tends to result in bigger sets, meaning that it is feasible to identify the probability of more age-ranges. In Figure 6.7.b we observe the proportion of districts that had each aggregated group size. We can observe that most districts get a set with a single group, this is most likely due to many districts containing few ballot-boxes, meaning that a lot of groups would be hard to identify. We can see that in general there are less districts as we increase the aggregated group set size. This indicates that there are few districts where the distribution of groups among the ballot-box is reliable enough to identify multiple age-range probabilities. We should also note that no district satisfied the threshold when trying the original group set where every age range is treated individually.



(a) Number of ballot-boxes for number of aggregated groups quartiles.

(b) Porportion of districts for number of aggregated groups.

Figure 6.7: Statistics for aggregated group sets

6.3.2. Probabilities

As we run the EM-algorithm in each district we will obtain a probability matrix estimation $\hat{\boldsymbol{p}}$. We would like this estimator to be a good representation of how voters actually vote in the given district.

In Figure 6.8 we observe a graded representation of how the estimated probabilities would look for the districts of *Calama*, *Maipú* and *Villarrica*. We can first notice that this districts vary in the number of selected age ranges. We can note that *Maipú* is the district with the higher number of ballot-boxes in 6.8, however we should note it was in the higher end of the distribution for districts with 6 age ranges according to Figure 6.7.a. Regarding the probabilities, we can observe decently regular patterns accross different candidates. We can see that *Candidate 1* and *Candidate 8* have low probabilities for all age groups accross all age ranges. Regarding *Candidate 2* we can observe higher voting probabilities in the district of *Calama* where there is quadratic correlation with age, in constrast with the other districts of *Maipu* and *Villarrica* where *Candidate 2* is less voted and it shows a decreasing correlation with age. In regards to *Candidate 3* we see a common decreasing correlation between probabilities and age in all districts, where the voting probabilities are higher in the district of *Maipú*. An opposite pattern is find in *Candidate 4* where there is an increase in the probabilities with age, with the higher probabilities found on the district of *Villarrica*.

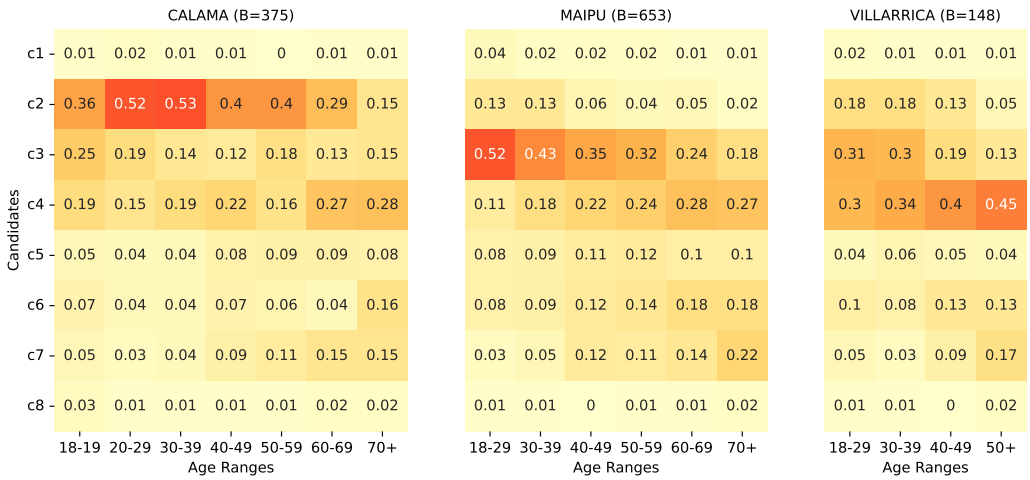


Figure 6.8: Probability estimation for the districts of *Calama*, *Maipú* and *Villarrica*

6.3.3. p-values and Outliers

We present some general results for the p-values obtained. We should note that obtaining a low p-value is an indicator that a ballot-box is an outlier but it does not states what happened in that ballot-box. Some reasons a ballot-box could show a low p-value are the following:

1. There were errors in the data, this may include: vote count, group count or ballot-box number mislabeling.

2. There was relevant information about the voter assignation of that ballot-box that was not considered in the model. This could happen if a particular set of voters (particularly different from the rest of the district) where assigned to a specific ballot-box or to a specific polling place due to some arbitrary criteria.

3. The vote count was altered and does not reflect the actual results of that ballot-box.

To visualize how a ballot-box with a low p-value looks, we present the scheme displayed in Figure 6.9. We show statistics for all ballot-boxes in the polling place in order to identify patterns related to the low p-values. The matrices displayed in Figure 6.9 represent the following from left to right: the vote count for each candidate, the vote difference to the expected value (rounded for visualization), the number of voters for each selected age range and the p-values. The difference to the expected value for a candidate c in ballot-box b, considering a estimated matrix probability $\hat{\boldsymbol{p}}$ and an aggregated group set $\mathcal{A}$ can be calculated

as follows:

$$x_{bc} - \mathbb{E}(X_{bc}) = x_{bc} - \sum_{a \in \mathcal{A}} w_{ba} \cdot \hat{p}_{ac}$$

In Figure 6.9 we can observe that all ballot-boxes, besides those labeled 142 and 144, show normal p-values of at least 1 in a hundred. The ballot-box labeled 144 shows a p-value of less than 10^{-9} (less than 1 in a billion), meaning that the computed value with simulation was 0. In this ballot-box we can see an important deviation from the expected value in the votes for candidates 3, 6 and 8. We can also observe that this ballot-box was primarily composed by voters in the 18-29 range.

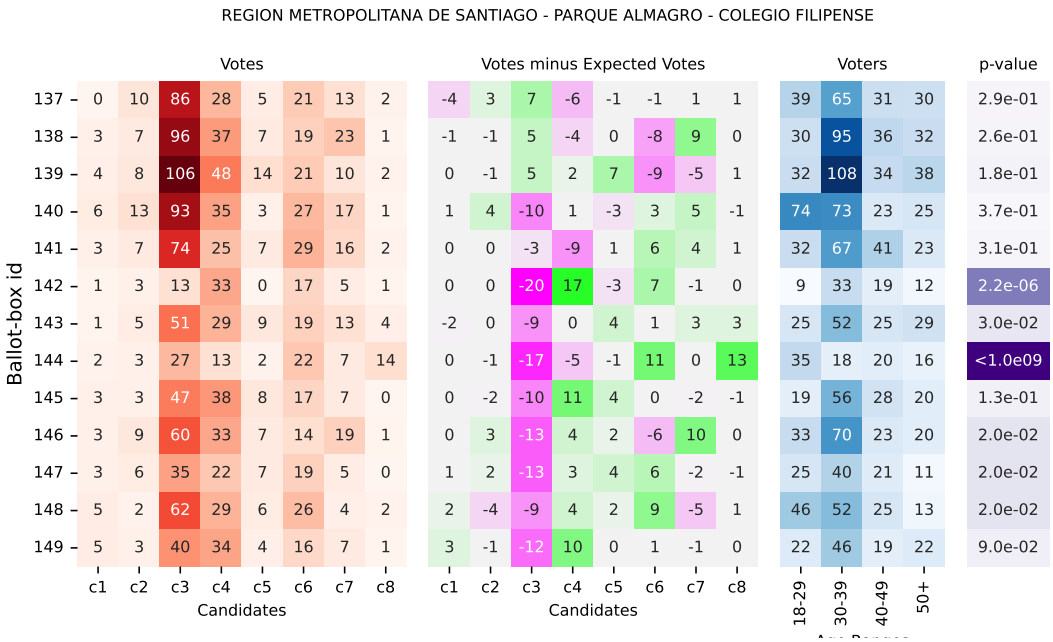


Figure 6.9: Results in the polling place of: *Colegio Filipense*

In Annex C we show a detailed result for the polling places containing the lowest encountered p-values in the range $\leq 10^{-8}$.

In Table 6.1 we present the p-value count for different ranges and in Figure 6.10 we present a scatter plot for the p-value and number of votes of each ballot-box. .

We should note that by construction, p-values should approximate to a Uniform$(0, 1)$ distribution. We observe a big concentration of p-values in the interval $[0.1, 1]$, at around 80%. We can note that the percentages do not match the expected quantity of each range. This could be attributed to different reasons: there is not enough information to get to good

estimators in some polling places, and the present outliers could alter both the probability distribution (specially in small polling places) and the interval count. We can observe that the count until 10^{-2} matches relatively well the expected results, however, from 10^{-3} we observe that the count does not decrease as it should.

We also observe that there are 15 ballot-boxes that have a p-value of less or equal than 10^{-9}. For those values that are strictly less than 10^{-9} it could be possible to further simulate samples in order to get an estimation to the order of magnitude. To further estimate, we would need to increase the sample size by 10 every time, making it computationally expensive. Even though we are not able to exactly compute these p-values, the chosen sample size is enough to determine those ballot-boxes as outliers.

Table 6.1: p-value count in the 2021-GCE.

By Intervals			Cumulative		
Range	Total	%	Range	Total	%
$(10^{-1}, 10^{0}]$	36584	78.50	$\leq 10^{0}$	46605	100.00
$(10^{-2}, 10^{-1}]$	7471	16.03	$\leq 10^{-1}$	10021	21.50
$(10^{-3}, 10^{-2}]$	2055	4.41	$\leq 10^{-2}$	2550	5.47
$(10^{-4}, 10^{-3}]$	259	0.56	$\leq 10^{-3}$	495	1.06
$(10^{-5}, 10^{-4}]$	111	0.24	$\leq 10^{-4}$	236	0.51
$(10^{-6}, 10^{-5}]$	54	0.12	$\leq 10^{-5}$	125	0.27
$(10^{-7}, 10^{-6}]$	28	0.06	$\leq 10^{-6}$	71	0.15
$(10^{-8}, 10^{-7}]$	16	0.03	$\leq 10^{-7}$	43	0.09
$(10^{-9}, 10^{-8}]$	12	0.03	$\leq 10^{-8}$	27	0.06
$(10^{-\infty}, 10^{-9}]$	15	0.03	$\leq 10^{-9}$	15	0.03

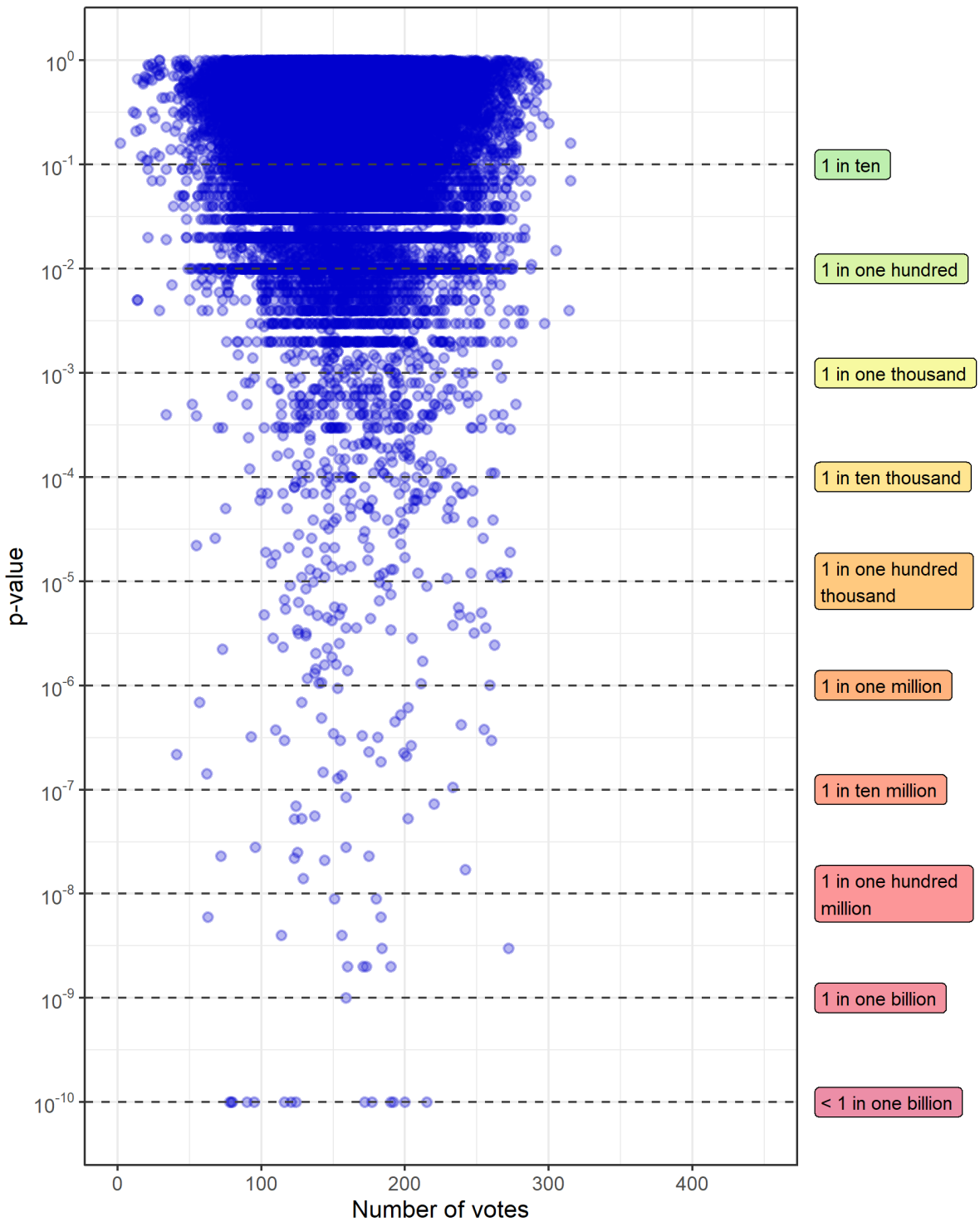


Figure 6.10: Estimated p-value accross all ballot-boxes.

6.4. General Recommendations

Identifying and addressing outlier ballot-boxes in future elections in Chile could help improve the overall integrity and transparency of the electoral process. Here are some actions that SERVEL could take in the short and medium term with the information the methods provide:

1. Hold the publication of official results for that ballot-box .
2. Verify that protocols where followed in that ballot-box and in the polling place containing that ballot-box.
3. Verify that the results in that ballot-box matches the official record.
4. Change the ballot-box committee for future elections.
5. Assign watchmen from the main parties to guard the counting of that ballot-box for future elections.

If these methods were to be put into practice, it would be beneficial to incorporate the voter assignation method directly into the model. This would assist in improving the identification of the territorial units for conducting the probability estimation and would also consider any arbitrary decisions that could have done in the allocation of voters.

Chapter 7

Conclusions

Through this study we analyzed different methods for computing the EM-algorithm in order to do ecological inference in the context of Chilean elections. It was shown that the EXACT method fails to compute medium and large instances, meaning it would be advisible to utilize approximated methods from $C = 3$ onwards. The amount of groups G does not affect the EXACT method exponentially, as the number of different combinations does not increase monotonically with G, so it could be possible to compute instances with a lot of demographic groups if the number of candidates is small. The MULT method was shown to be the fastest so it would be recommended to use it in most instances, specially if C is high. The MVN-PDF method is not as fast, but it is still polynomial and not highly impacted by the size of the instance, so it would also be a advisible to use it. The MVN-CDF gets similar estimations compared to the MULT and MVN-PDF methods, but is also highly impacted by the size of the instance so it would not be a good choice. The H&R method needs a lot of samples in order to have a good estimation in big instances meaning it is not reliable. It should be noted that these methods could be used for other applications where the distribution corresponds to a sum of multinomial random variable.

We observed that the methodology proposed for the EM-algorithm works best in heterogeneous ballot-boxes, so a previous assessment of the group distribution should be considered. This can be analyzed through the standard deviation and correlation of different groups in a ballot-box. It should be noted that in the case of having different amounts of voters per ballot-box, this analysis should be done through the proportion of voters from the group instead of the quantity, so that ballot-boxes are comparable.

In regards to the p-values, we showed that there is an important gap in the calculation when using the estimated probabilities from the EM-algorithm. This means that even though the MAE of the estimated probability obtained could be small, the error when computing the p-value could be amplified, as we are considering a sum of probabilities from a slightly different distribution. We also showed that there is a gap when using the multinomial approximation, specially when the variance of the original distribution is amplified. This shows that p-values should be carefully computed as there are errors attributed to both the estimated probability and the method used to compute the p-values.

We finally showed a methodology for aggregating groups in order to get reliable estimations. We used the 2021-PGE as a case study to test the proposed methods in this work. We showed that as there were more ballot-boxes in a district it was convenient to consider bigger sets (less aggregated). Using these group sets we computed estimated probabilities for each district and p-values for each ballot-box. We obtain 27 ballot-boxes with a p-value lower or equal than 10^{-8}. We discussed possible explanations to such low p-values: mislabeling, unmatched data, manipulation, among others. We stressed the fact that these models do not give an explanation to the low p-values, but they can work as an indicator to better improve the election process.

7.1. Future Work

This study leaves some open questions that would be interesting to address in the future.

- Consider different simulation approaches for the H&R method. As we have shown, there is an increase in the MAE of the estimation as we consider less samples, meaning it would be necessary to find a faster sampling method. It would be interesting to study how important the correlation between obtained points affects the estimation, so that we could allow a smaller step-size in some situations. It may even be possible to study a deterministic approach to come up with a subset of Ω_b that is representative of the aggregate outcome. It could also be interesting to study sampling approaches that take into account the original probability distribution and how it affects the final estimation.

- Consider different methods to approximate the p-values. As we have shown, the multinomial approximation has a bias estimating the p-values when there are big changes in

the original variance. We could study how simulations coming from the multivariate normal approximation compute the p-value, as it maintains the original covariance matrix. However, it should be noted that this method is considerably slower, and as we are interested in studying the presence of outliers, we do need a big sample size.

- Address special cases for elections in the Chilean context. The methods studied assume that voters from the same group are comparable within a district. This assumption may not be true for recent elections where the ballot-box assignation is based on distance. This factor could be taken into account in the probability distribution. It could also be interesting to study a model that uses different districts for the same ecological inference process. This would be particularly useful for small districts with a small number of ballot-boxes, where the estimation is less reliable. A general method could take into account both: aggregate demographic information from voters within a district (as shown in the application of this study) and aggregate demographic information from districts (such as average salary, unemployment rates and geographic location).

Bibliography

Anselin, L., & Cho, W. K. T. (2002). Spatial effects and ecological inference. *Political analysis*, *10*(3), 276–297.

Antweiler, W. (2007). Estimating voter migration in canada using generalized maximum entropy. *Electoral Studies*, *26*(4), 756–771.

Cleave, N., Brown, P., & Payne, C. (1995). Evaluation of methods for ecological inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(1), 55–72.

Duncan, O., & Davis, B. (1953). An alternative to ecological inference. *American Social Review*, *18*, 665–666.

Fortin-Rittberger, J., Harfst, P., & Dingler, S. C. (2017). The costs of electoral fraud: establishing the link between electoral integrity, winning an election, and satisfaction with democracy. *Journal of Elections, Public Opinion and Parties*, *27*(3), 350–368.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, *1*(2), 141–149.

Glynn, A. N., & Wakefield, J. (2010). Ecological inference in the social sciences. *Statistical methodology*, *7*(3), 307–322.

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American sociological review*, *18*(6), 663.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, *64*(6), 610–625.

Jackson, C., Best, N., & Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in medicine*, *25*(12), 2136–2159.

King, G., Tanner, M. A., & Rosen, O. (2004). *Ecological inference: New methodological*

strategies. Cambridge University Press.

Leemann, L., & Bochsler, D. (2014). A systematic approach to study electoral fraud. *Electoral studies*, *35*, 33–47.

Leigh, A. (2005). Economic voting and electoral behavior: How do individual, local, and national factors affect the partisan choice? *Economics & Politics*, *17*(2), 265–296.

Lin, Z., Wang, Y., & Hong, Y. (2022). The poisson multinomial distribution and its applications in voting theory, ecological inference, and machine learning. *arXiv preprint arXiv:2201.04237*.

Mete, H. O., & Zabinsky, Z. B. (2012). Pattern hit-and-run for sampling efficiently on polytopes. *Operations Research Letters*, *40*(1), 6–11.

Morgenstern, H. (1995). Ecologic studies in epidemiology: concepts, principles, and methods. *Annual review of public health*, *16*(1), 61–81.

Peterson, J. C., Smith, K. B., & Hibbing, J. R. (2020). Do people really become more conservative as they age? *The Journal of Politics*, *82*(2), 600–611.

SERVEL. (2023). *Resultados electorales históricos.* https://www.servel.cl/centro-de-datos/resultados-electorales-historicos. (Accesed: 2023-08-15)

Wakefield, J. (2008). Ecologic studies revisited. *Annu. Rev. Public Health*, *29*, 75–90.

Withers, S. D. (2001). Quantitative methods: advancement in ecological inference. *Progress in Human Geography*, *25*(1), 87–96.

Annex A

Hit and Run Method

A.1. Starting Point for the Hit-and-Run Algorithm

Algorithm 7 Starting Point for Hit and Run

Input: $\boldsymbol{w}_b \in \mathbb{Z}_+^G, \boldsymbol{x}_b \in \mathbb{Z}_+^C$
Initialize $\boldsymbol{z}_b = \mathbf{0} \in \mathbb{Z}_+^{G \times C}$
for $g = 1$ to G **do**
  for $c = 1$ to C **do**
    $z_{bgc} = \min\{w_{bg}, x_{bc}\}$
    $w_{bg} = w_{bg} - z_{bgc}$
    $x_{bc} = x_{bc} - z_{bgc}$

A.2. Choosing the Step-Size

For algorithm 3 to work properly we would like to choose a step-size M so that the sampled points are independent from each other (in other words, we would like the starting point of each iteration not to determine the outcome).

For this study we will do an empiric evaluation on the effect of the step-size. We will evaluate 4 instances with varying values of G and C. For each instance we will simulate 20 scenarios containing 50 ballot-boxes with 100 voters each.

We will apply the algorithm using a step-size of $M = 100$ and $S = 10000$ for each ballot-box. Let us consider $\vec{\mathcal{S}}(b)$ as the ordered array containing all the points sampled in the algorithm (even if repeated). We will evaluate how correlated sampled points are using the following.

$$\rho_{m \cdot M}(b) = \text{corr}(\vec{\mathcal{S}}_{m \to S}(b), \vec{\mathcal{S}}_{1 \to S-m}(b))$$

What we are doing is looking how one point correlates to the point obtained m iterations after, this allows us to evaluation the correlation for step-sizes in multiples of M.

Using this procedure we obtain the following average result for varying step-sizes over the different instances.

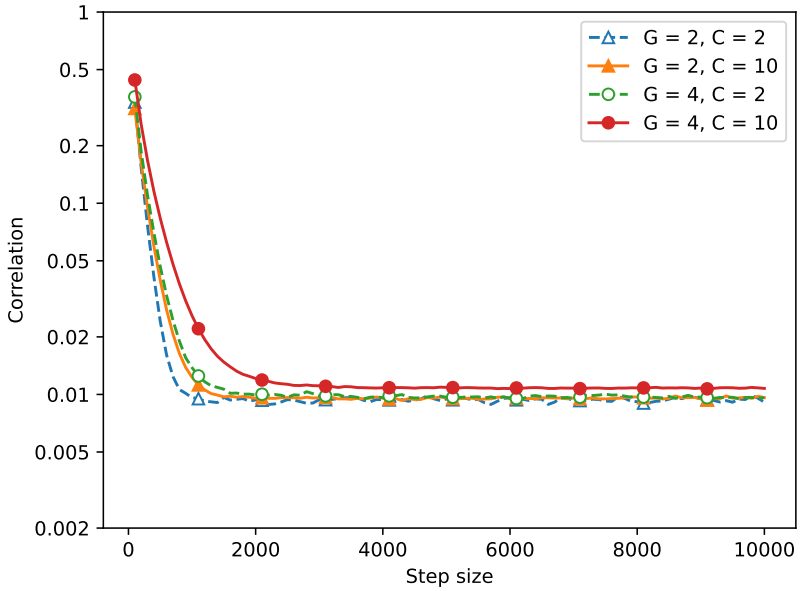


Figure A.1: Mean correlation for various step-sizes over different instances. The graph is log-scaled.

We observe a notable decrease for all instances when augmenting the step-size from 100 to 1000. Then we observe small decreases for some instances until they all hit a plateau around a step-size of 3000. This last value will be used as the step-size when running this method through this study

A.3. Another Formulation for the Discrete Polytope

$$\sum_{g=1}^{G-1} z_{bgc} \leq x_{bc} \qquad \forall c \in \{1, ..., C-1\}$$

$$\sum_{c=1}^{C-1} z_{bgc} \leq w_{bg} \qquad \forall g \in \{1, ..., G-1\}$$

$$\sum_{g=1}^{G-1} \sum_{c=1}^{C-1} z_{bgc} \geq I_b - x_{bC} - w_{bG}$$

$$z_{bgc} \geq 0 \qquad \forall c \in \{1, ..., C-1\} \; \forall g \in \{1, ..., G-1\}$$

$$z_{bgc} \in \mathbb{Z} \qquad \forall c \in \{1, ..., C-1\} \; \forall g \in \{1, ..., G-1\}$$

Annex B

EM-algorithm Probability Estimation with 200 Voters per Ballot-Box

B.1. Time results

Table B.1: Mean running time over 20 scenarios in seconds for the EM-algorithm for varying instances, with fixed values of $I_b = 200$ and $B = 50$.

Instance		Method					
C	G	EXACT	H&R $S = 10^3$	H&R $S = 10^2$	MVN CDF	MVN PDF	MULT
2	2	59.634	740.965	74.230	1.385	0.176	0.003
	3	100.076	736.357	75.210	2.469	0.242	0.004
	4	188.472	698.575	75.254	3.620	0.250	0.005
3	2	-	745.531	72.961	6.076	0.220	0.003
	3	-	718.089	70.551	12.326	0.249	0.004
	4	-	727.985	72.246	17.310	0.318	0.005
5	2	-	711.375	71.098	19.955	0.208	0.003
	3	-	576.660	67.952	42.570	0.265	0.003
	4	-	676.125	65.974	58.629	0.303	0.005
10	2	-	670.993	66.676	111.144	0.233	0.002
	3	-	534.038	61.340	204.632	0.266	0.003
	4	-	596.844	58.734	327.127	0.334	0.004

Table B.2: Comparison of mean simulation time and mean EM running time in seconds over 20 scenarios for varying instances, with fixed values of $I_b = 200$ and $B = 50$.

Instance		Method			
		H&R ($S = 10^2$)		H&R ($S = 10^3$)	
C	G	Sim-time	EM-time	Sim-time	EM-time
2	2	73.409	0.820	740.046	0.919
	3	72.915	2.295	722.202	14.155
	4	72.732	2.522	686.560	12.015
3	2	71.105	1.855	733.173	12.359
	3	68.670	1.881	699.799	18.290
	4	69.916	2.331	703.638	24.347
5	2	69.471	1.627	693.218	18.157
	3	65.756	2.196	567.549	9.111
	4	64.044	1.929	652.590	23.535
10	2	65.161	1.515	655.398	15.595
	3	59.866	1.474	525.634	8.404
	4	57.256	1.479	578.101	18.743

B.2. Error results

Table B.3: Mean absolute prediction error over 20 scenarios for varying instances, with fixed values of $I_b = 200$ and $B = 50$.

Instance		Method					
C	G	EXACT	H&R $S = 10^3$	H&R $S = 10^2$	MVN CDF	MVN PDF	MULT
2	2	0.009	0.009	0.009	0.009	0.009	0.009
	3	0.011	0.011	0.013	0.011	0.011	0.011
	4	0.012	0.013	0.024	0.012	0.012	0.013
3	2	-	0.009	0.010	0.009	0.009	0.009
	3	-	0.013	0.027	0.011	0.011	0.012
	4	-	0.022	0.040	0.013	0.013	0.013
5	2	-	0.008	0.013	0.007	0.007	0.007
	3	-	0.016	0.028	0.009	0.009	0.009
	4	-	0.025	0.036	0.011	0.011	0.011
10	2	-	0.008	0.013	0.005	0.005	0.005
	3	-	0.017	0.021	0.007	0.007	0.007
	4	-	0.022	0.026	0.008	0.008	0.008

Annex C

Detailed Results for Polling Places with Low p-value Ballot-Boxes

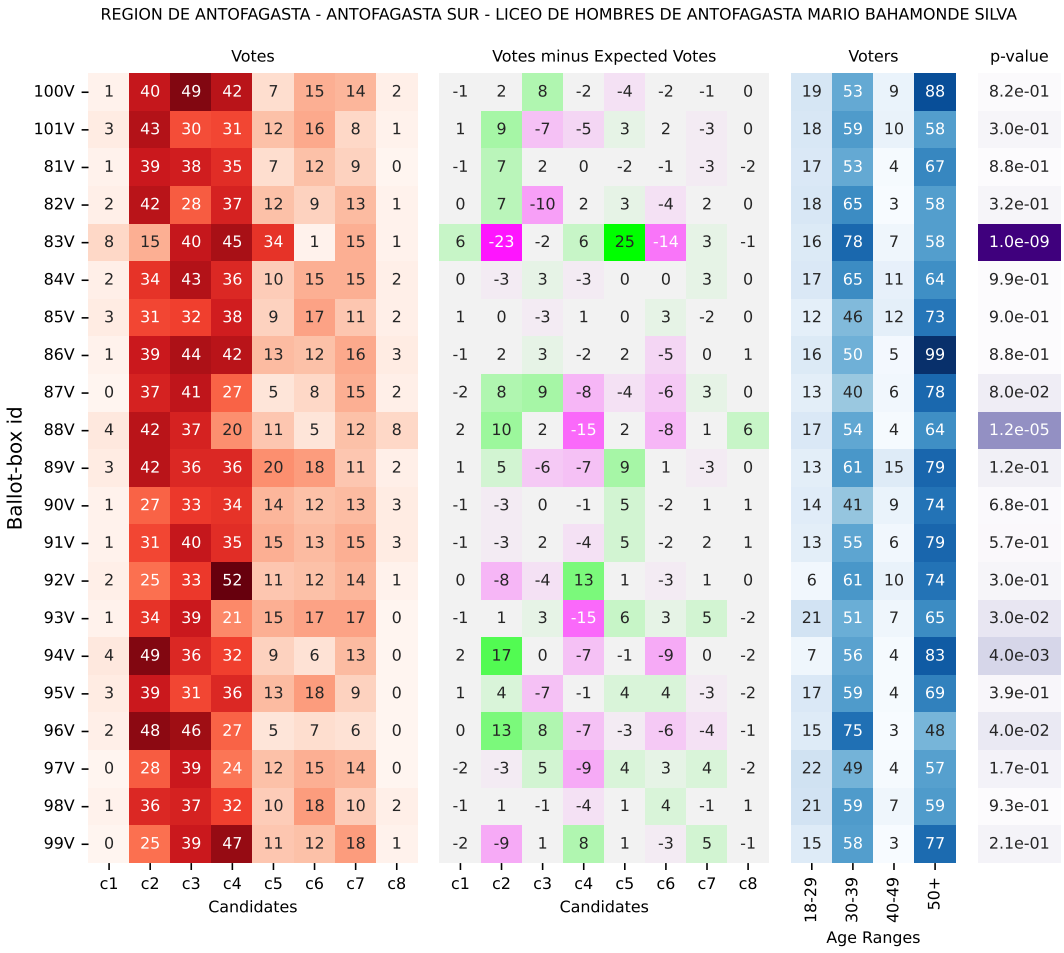


Figure C.1: Results in the polling place of: *Liceo de Hombres de Antofagasta Mario Bahamonde Silva*

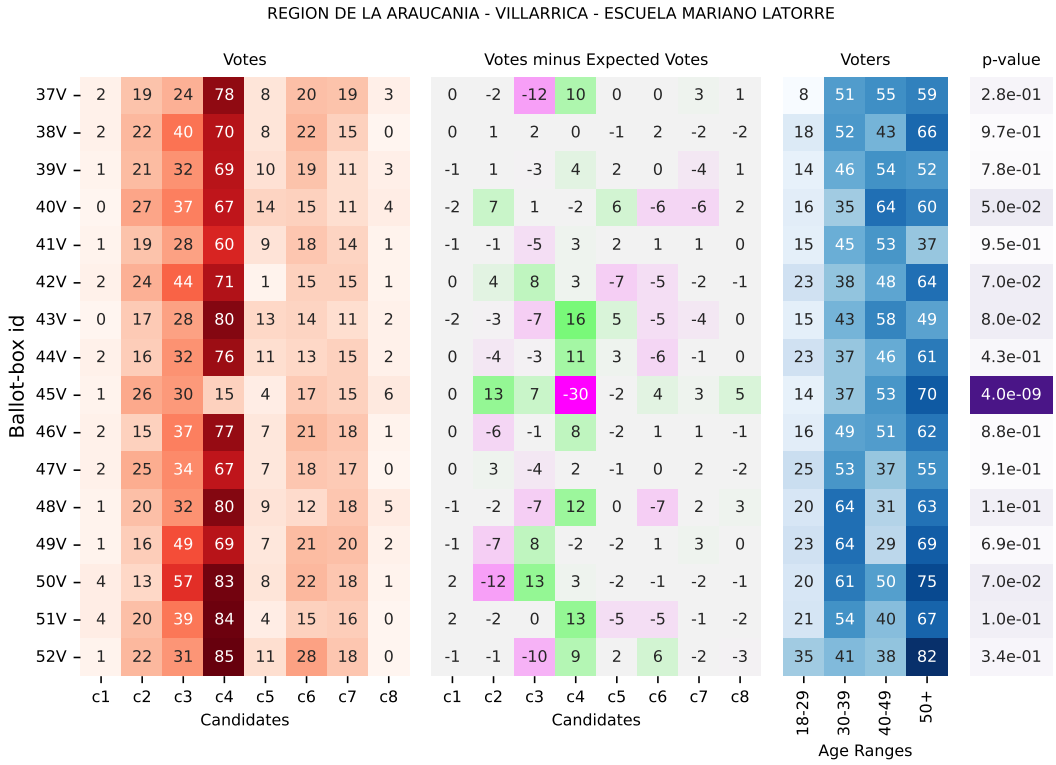


Figure C.2: Results in the polling place of: *Escuela Mariano Latorre*

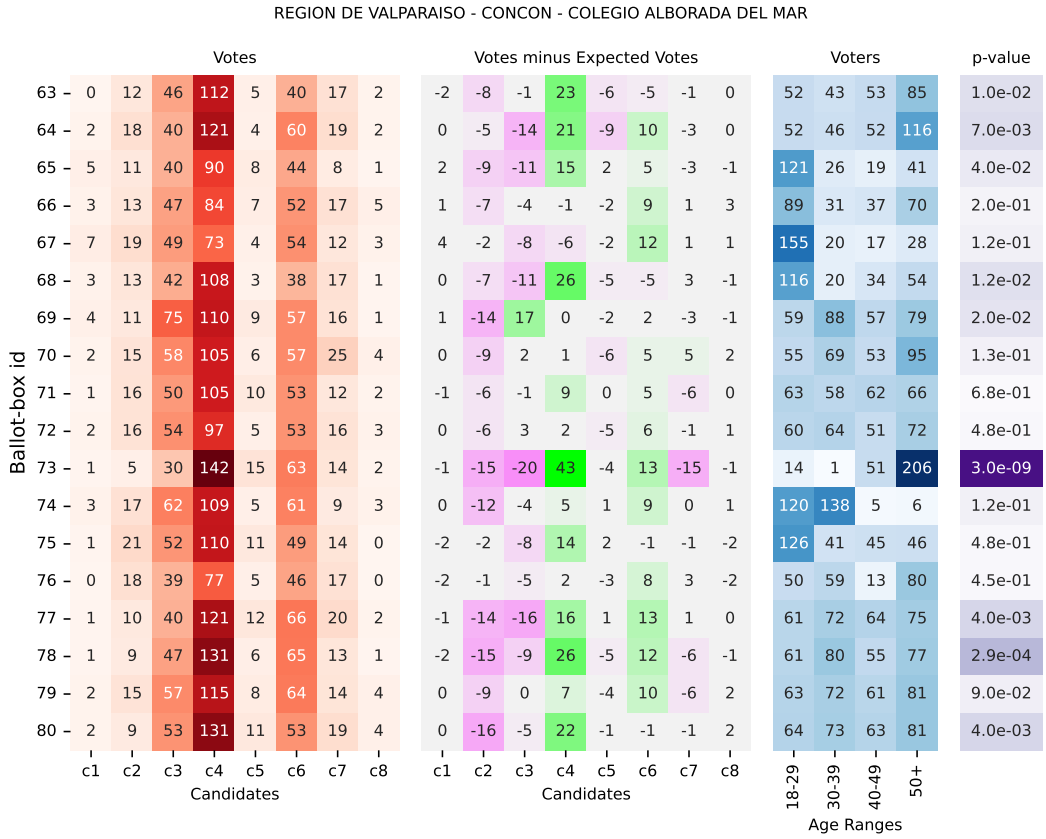


Figure C.3: Results in the polling place of: *Colegio Alborada del Mar*

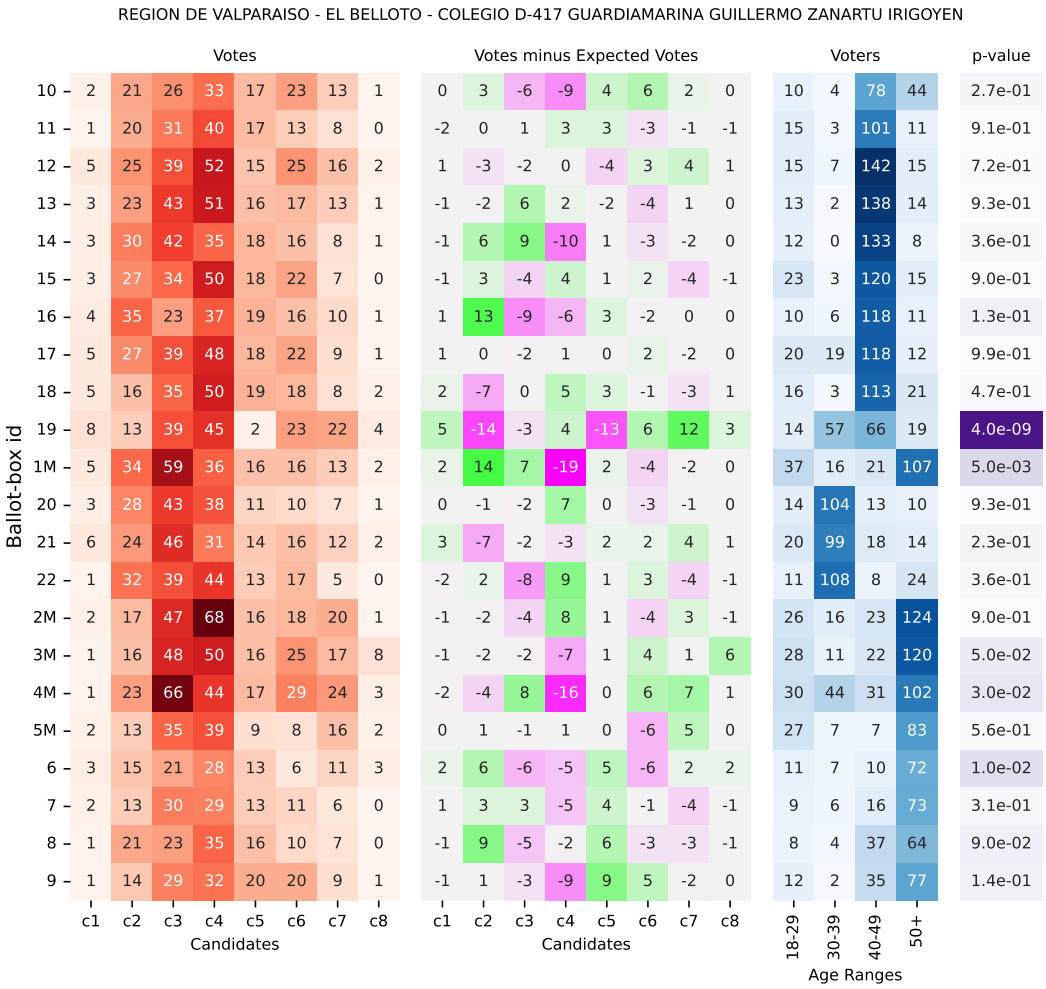


Figure C.4: Results in the polling place of: *Colegio D-417 Guardiamarina Guillermo Zañartu Irigoyen*

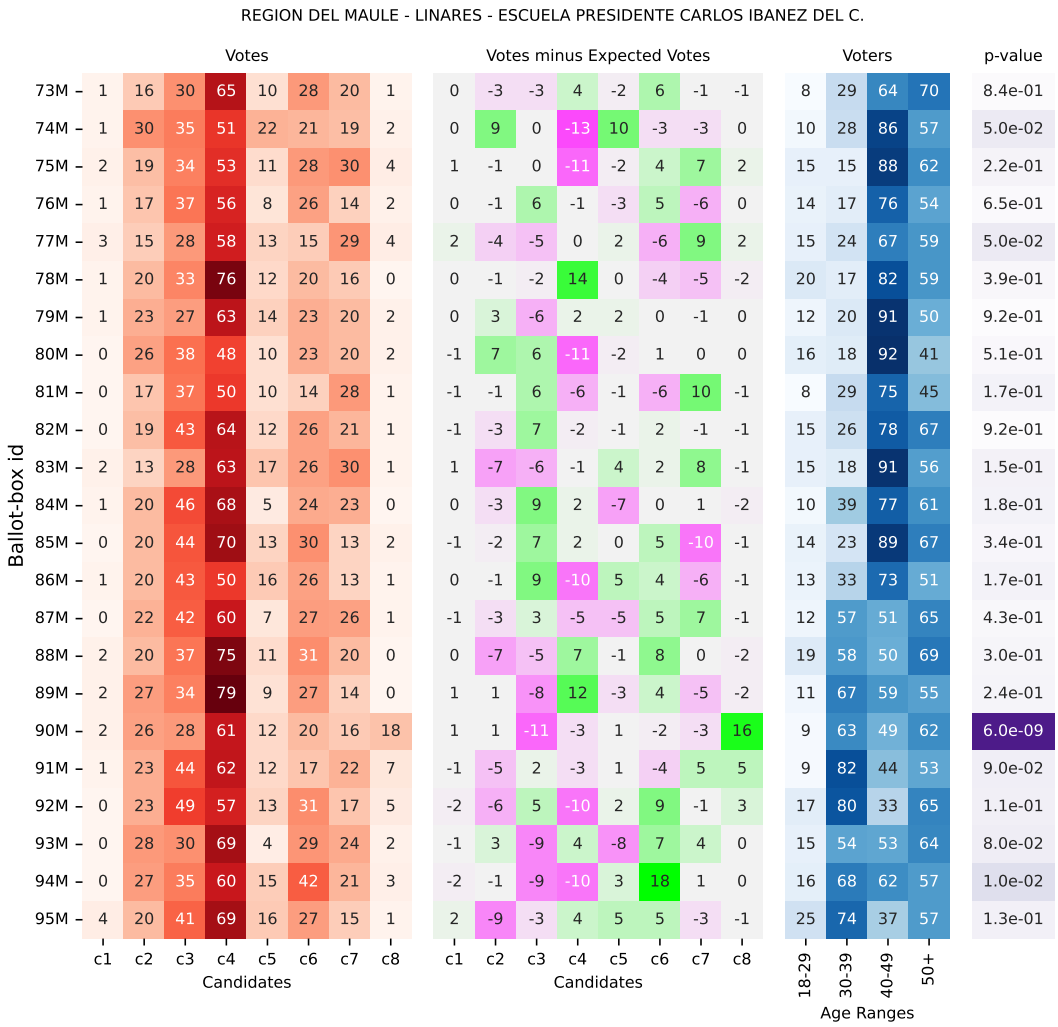


Figure C.5: Results in the polling place of: *Escuela Presidente Carlos Ibáñez del C.*

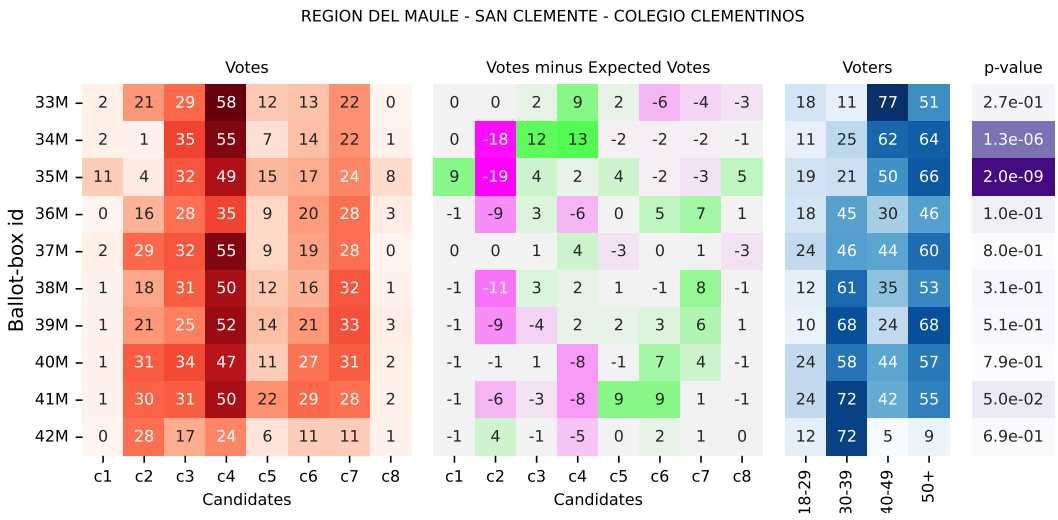


Figure C.6: Results in the polling place of: *Colegio Clementinos*

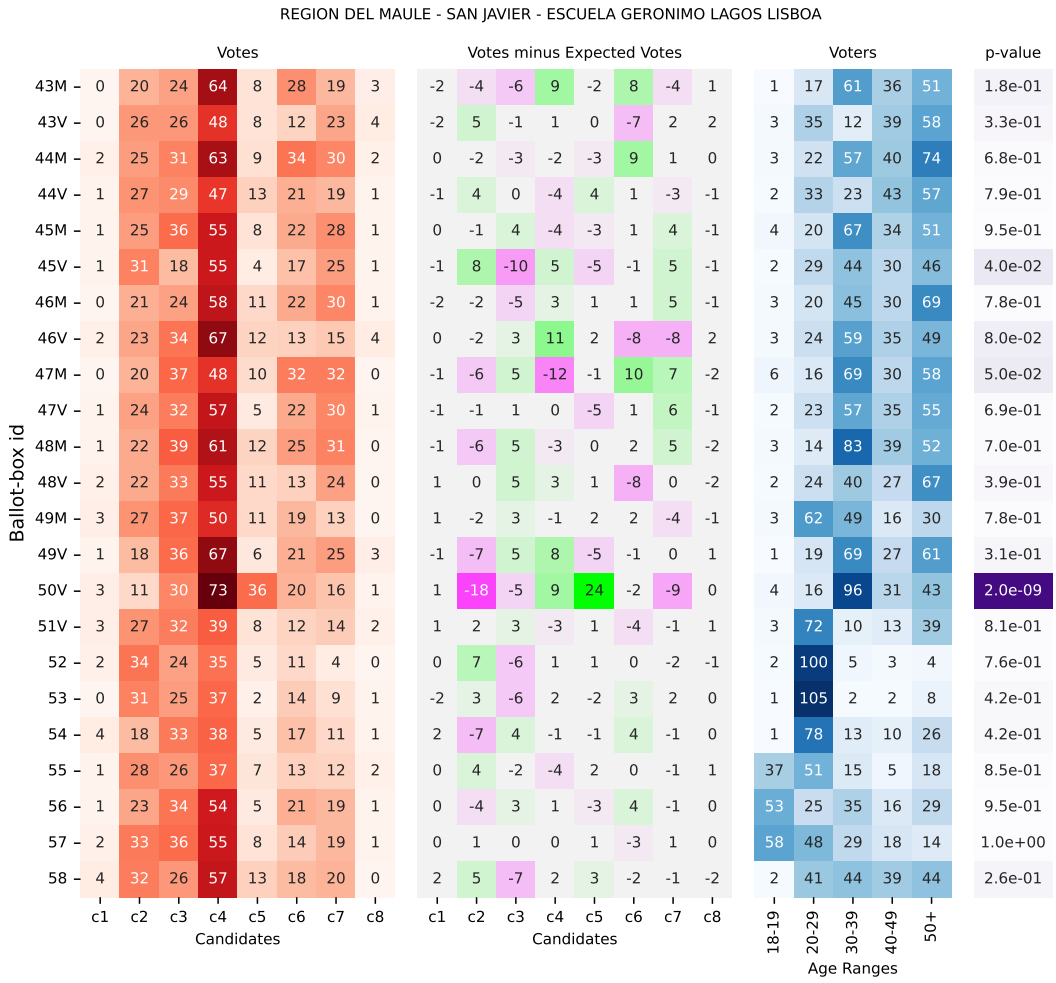


Figure C.7: Results in the polling place of: *Escuela Gerónimo Lagos Lisboa*

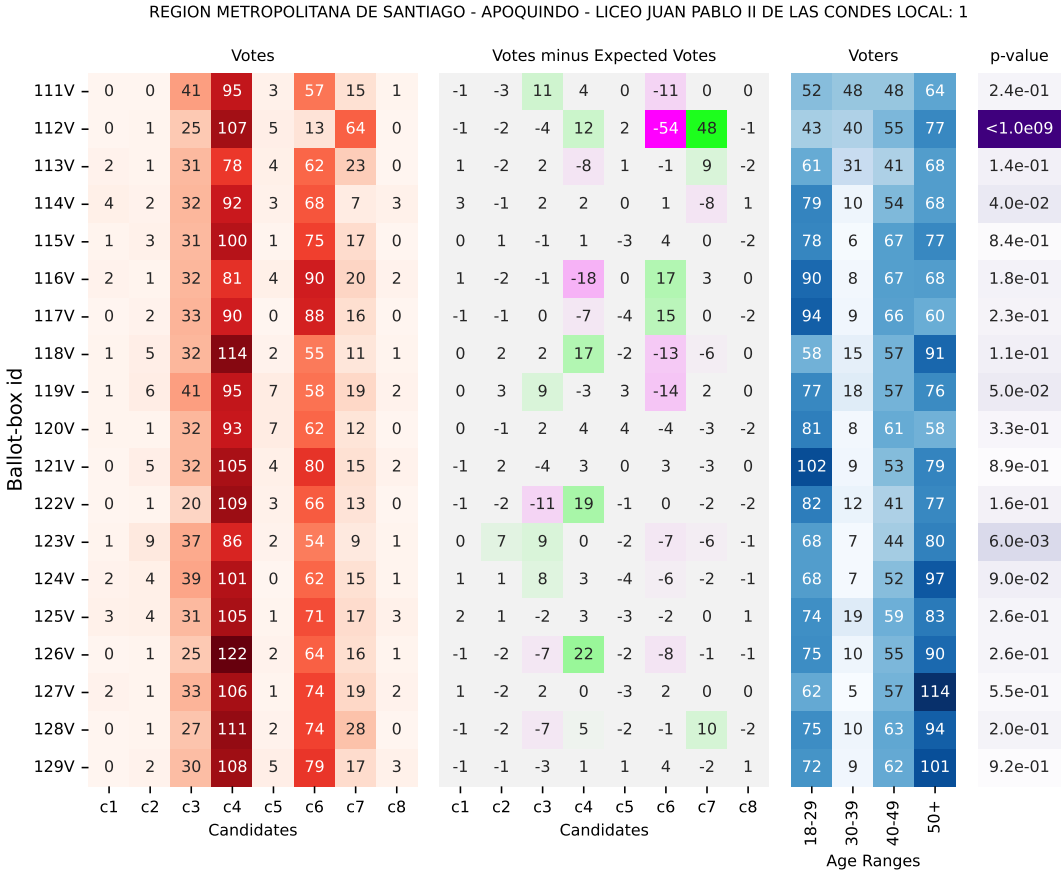


Figure C.8: Results in the polling place of: *Liceo Juan Pablo II de Las Condes Local: 1*

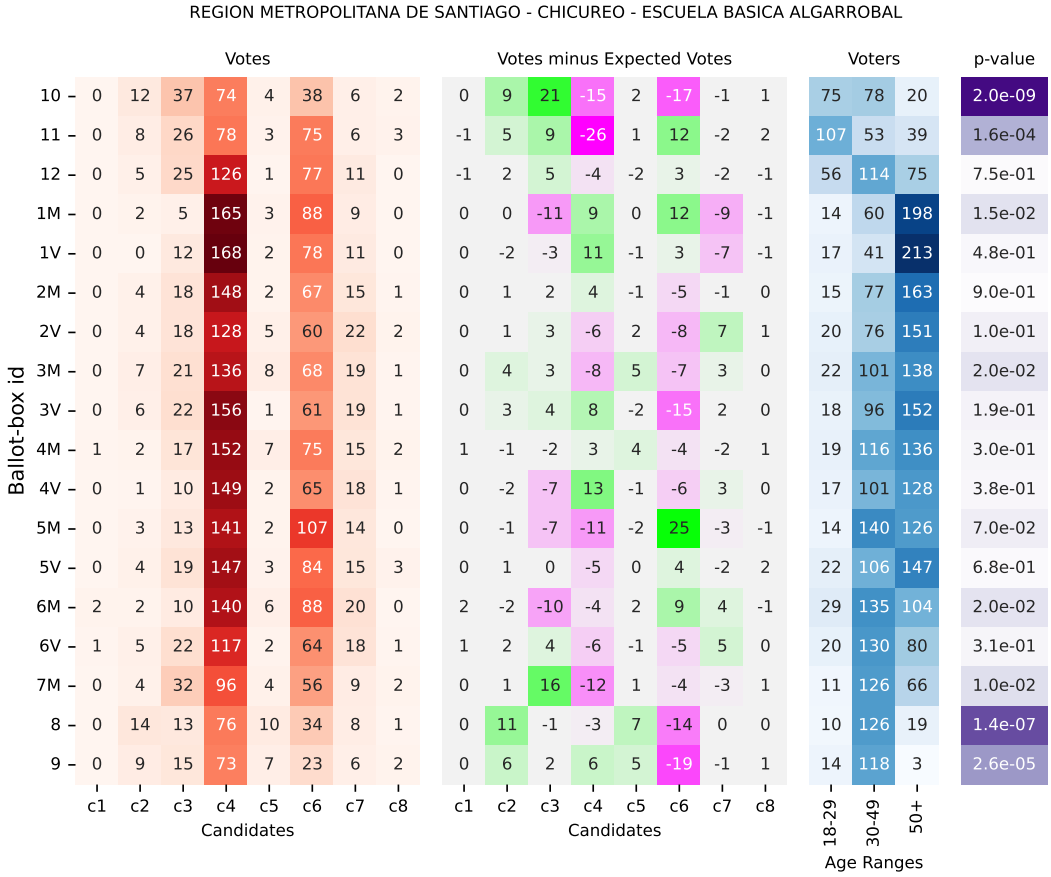


Figure C.9: Results in the polling place of: *Escuela Básica Algarrobal*

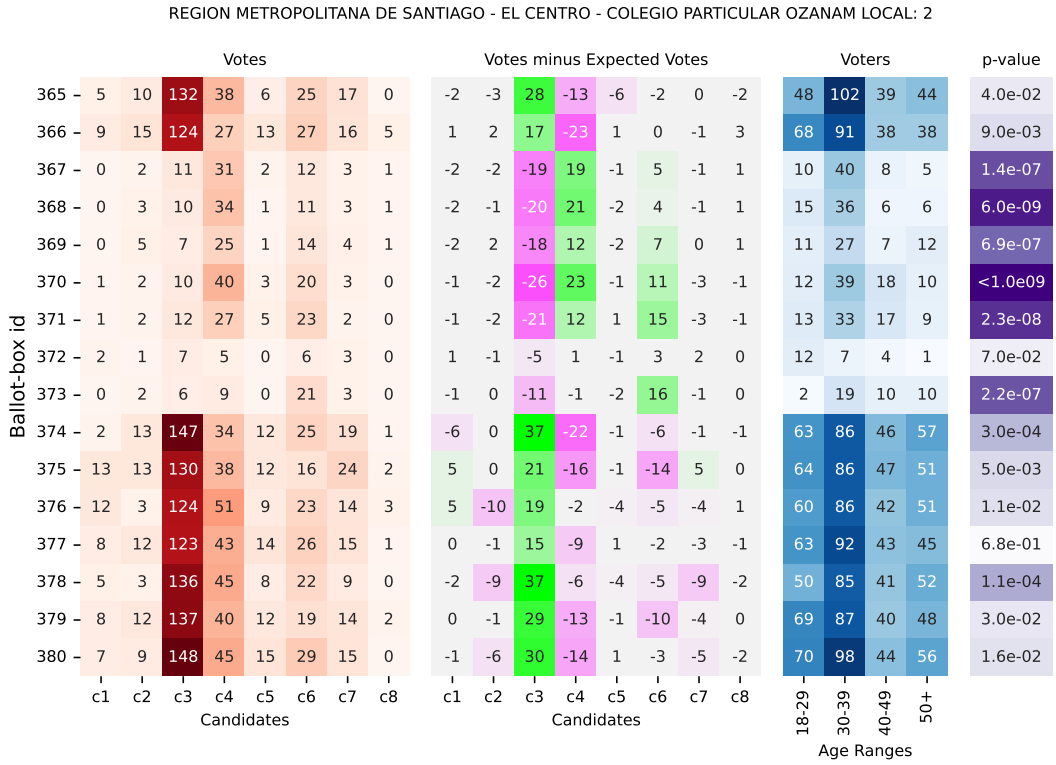


Figure C.10: Results in the polling place of: *Colegio Particular Ozanam Local: 2*

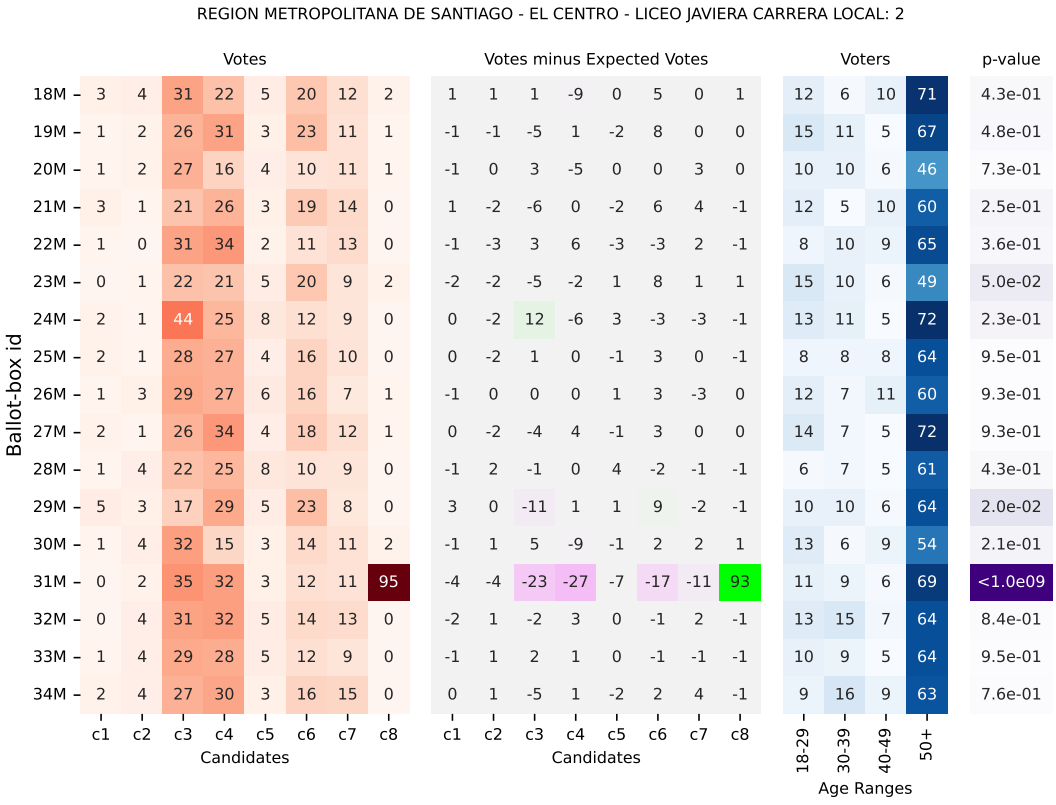


Figure C.11: Results in the polling place of: *Liceo Javiera Carrera Local: 2*

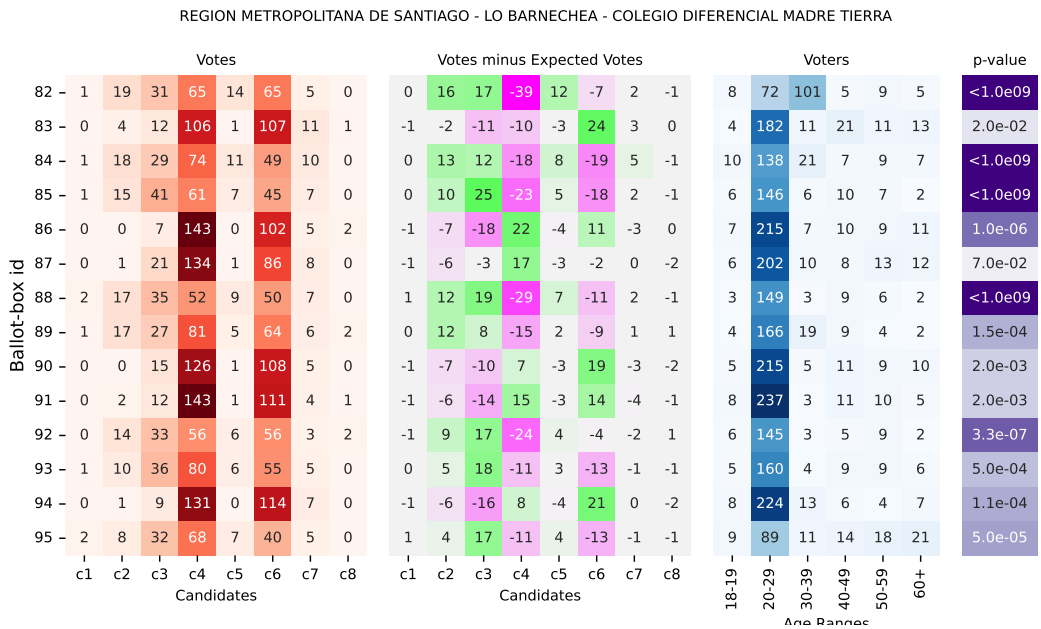


Figure C.12: Results in the polling place of: *Colegio Diferencial Madre Tierra*

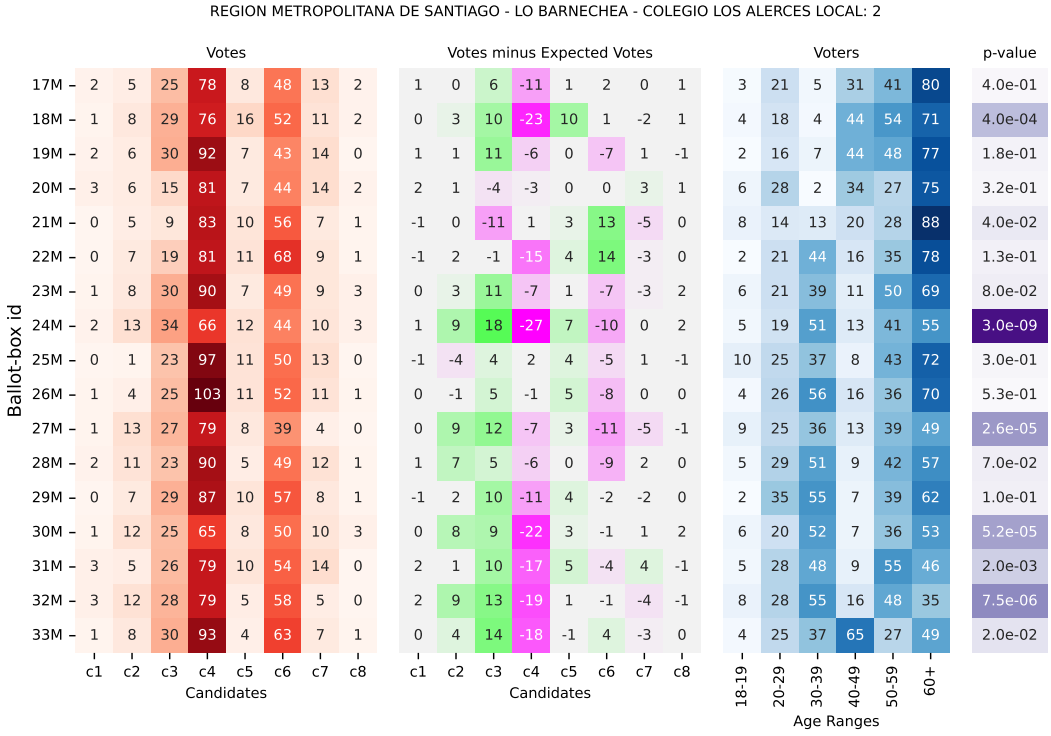


Figure C.13: Results in the polling place of: *Colegio Los Alerces Local: 2*

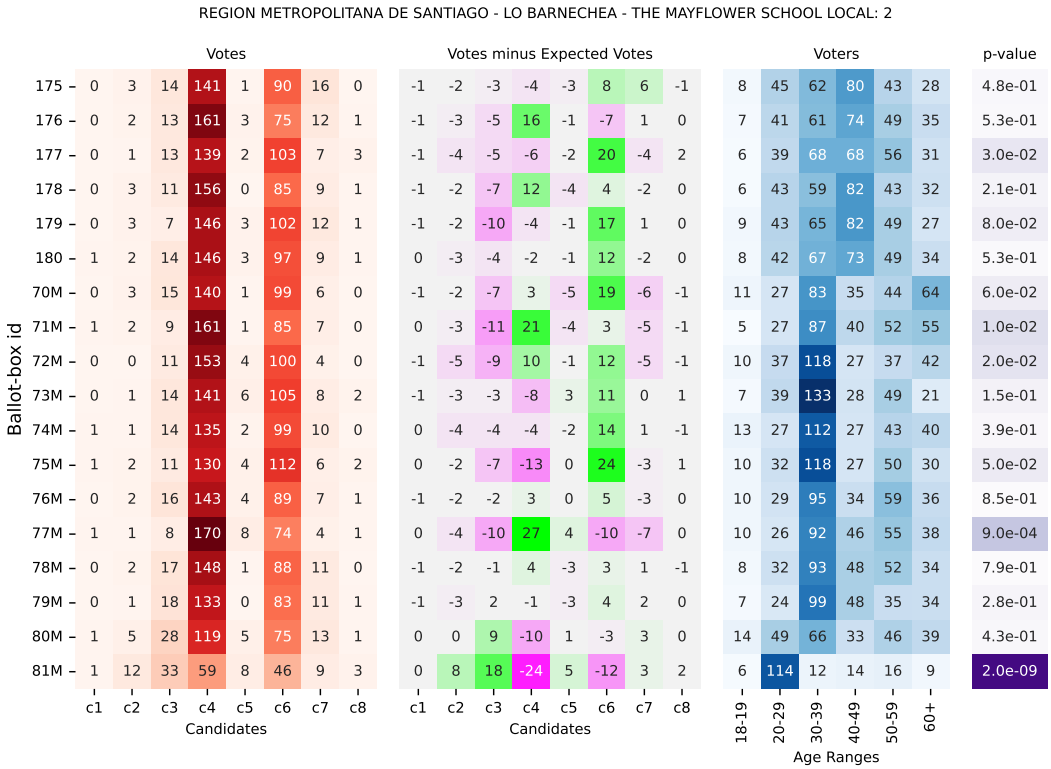


Figure C.14: Results in the polling place of: *The Mayflower School Local: 2*

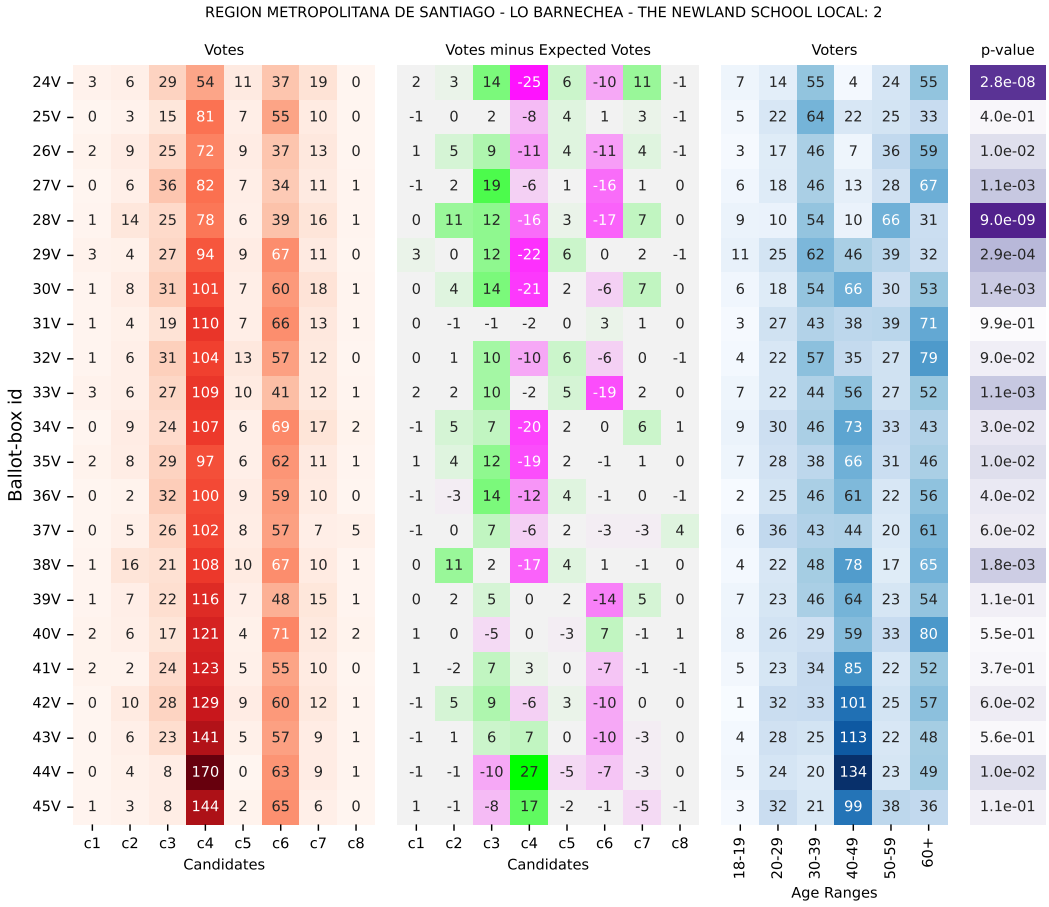


Figure C.15: Results in the polling place of: *The Newland School Local: 2*

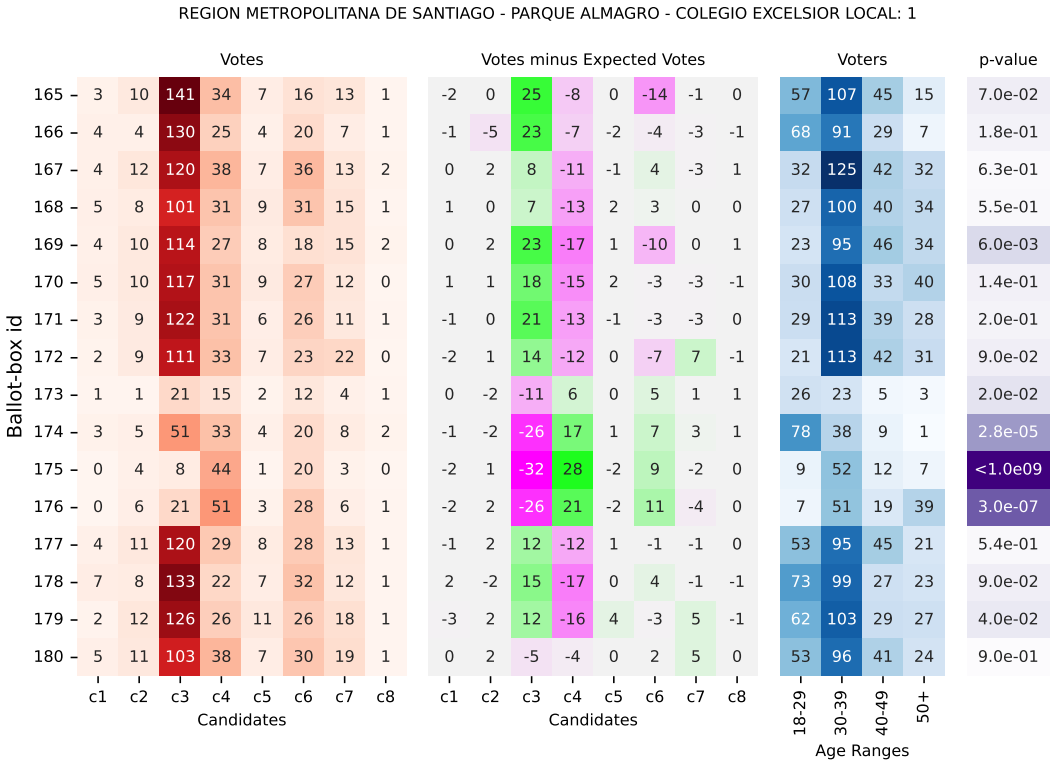


Figure C.16: Results in the polling place of: *Colegio Excelsior Local: 1*

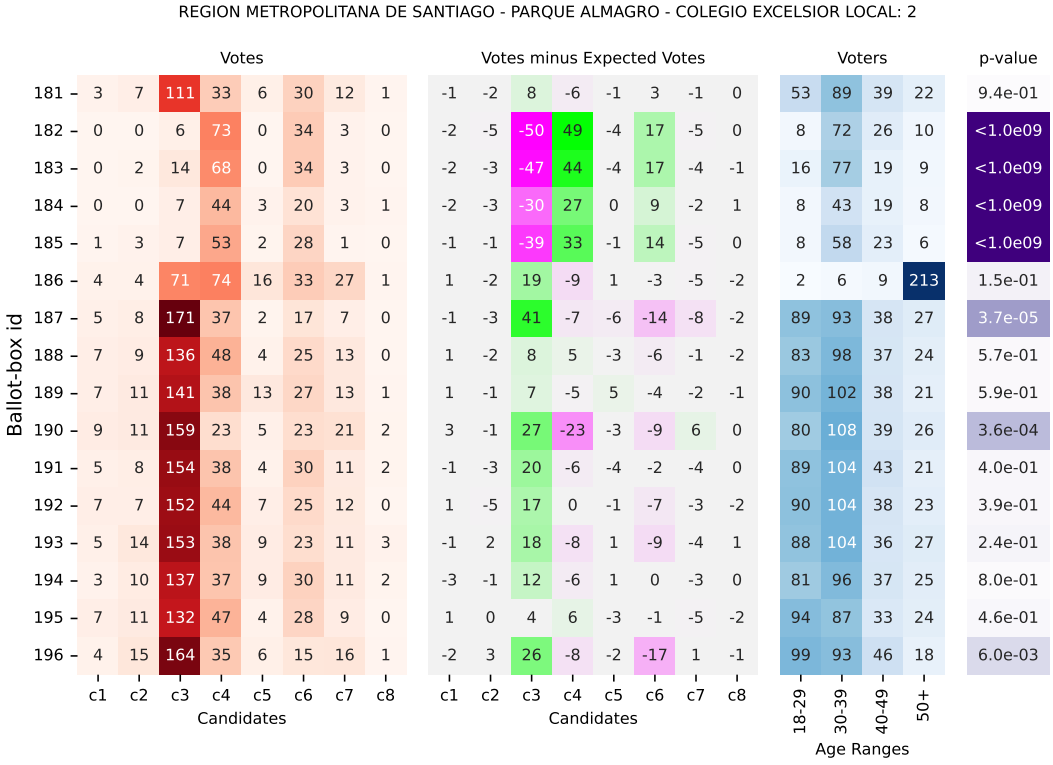


Figure C.17: Results in the polling place of: *Colegio Excelsior Local: 2*

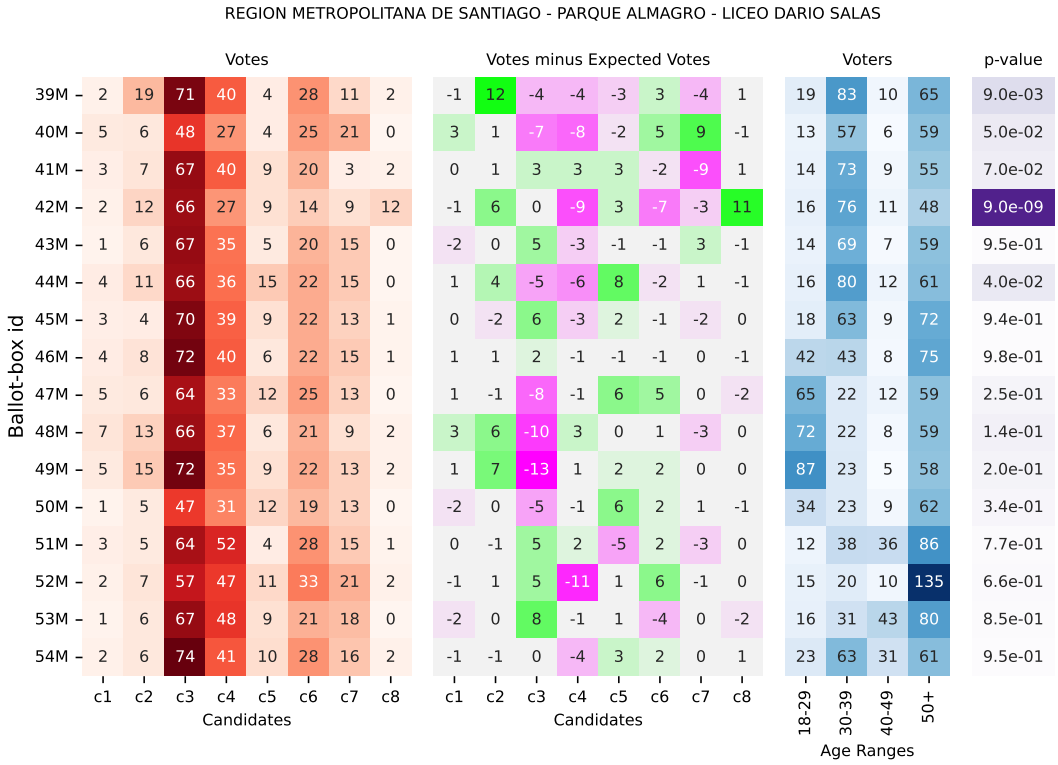


Figure C.18: Results in the polling place of: *Liceo Darío Salas*

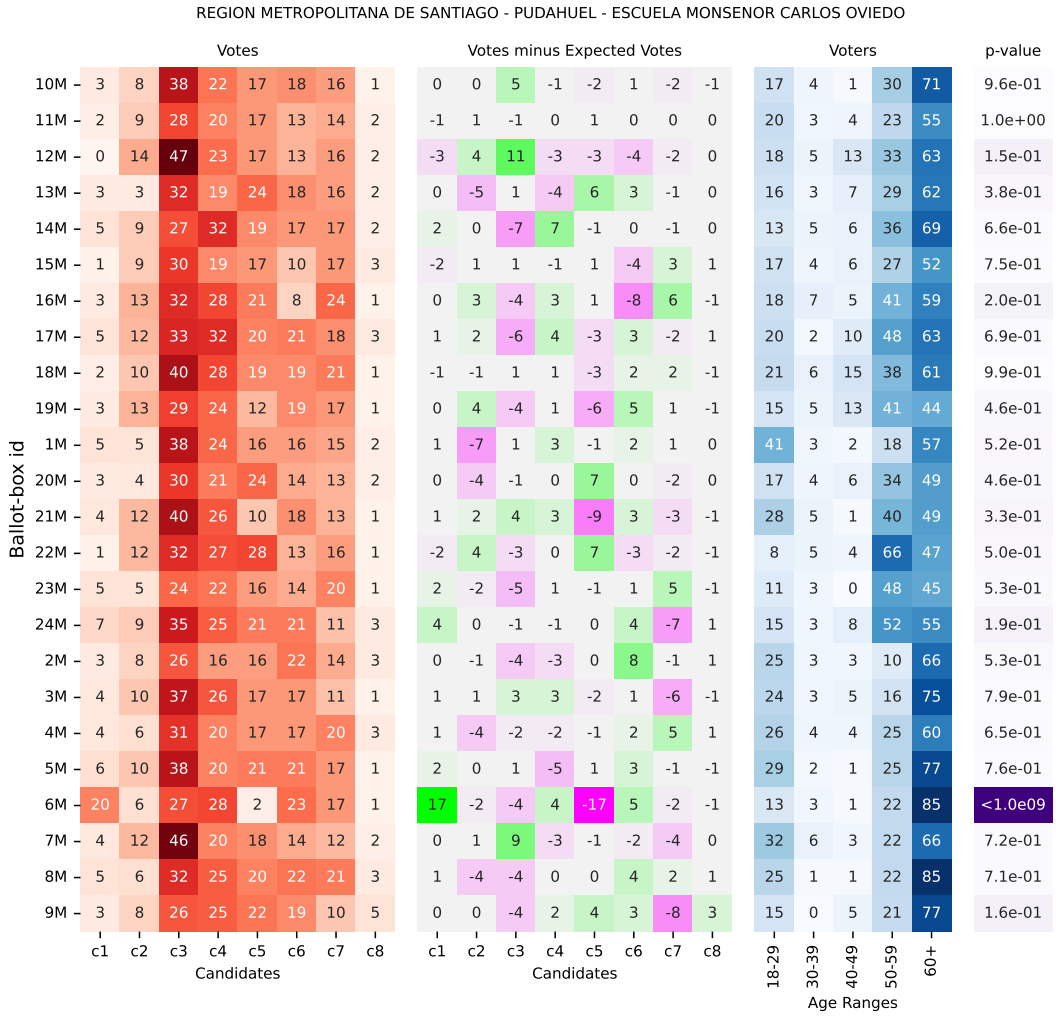


Figure C.19: Results in the polling place of: *Escuela Monseñor Carlos Oviedo*