1 Israel Election Dataset - Analysis and Methods

1.1 Introduction

We describe here the data for elections for the 22nd Knesset in Israel, with votes per city. (**Remark:** by city here we mean any 'yeshuv' which can be a city, village, kibutz etc.).

We also describe the data analysis and tools used to answer different questions about the data.

1.1.1 Notations:

- Our dataset is a matrix $\boldsymbol{N} \in \mathbb{R}_{\boldsymbol{C} \times \boldsymbol{K}}$ where K is the number of parties and C is the number of cities. n_{ij} is the number of voters for party j in city i. In addition, we have the following:

- Let $n_{i\bullet} = \sum_{j=1}^{K} n_{ij}$ be the total number of legal votes ('kolot ksherim') in city i. $\tilde{n}_{i\bullet}$ is the total number of eligible voters in city i ('baalei zhut bhira'). From these, we can calculate the voting turnout at city i: $v_i = \frac{n_{i\bullet}}{\tilde{n}_{i\bullet}}$. (In the data file you are given both $\tilde{n}_{i\bullet}$ and the voting turnout, but you need to re-calculate v_i without 'kolot psulim').

- Similarly the number of total votes for party j is $n_{\bullet j} = \sum_{i=1}^{C} n_{ij}$.

- Let $n = \sum_{i=1}^{C} n_{i\bullet} = \sum_{i=1}^{C} \sum_{j=1}^{K} n_{ij}$ be the total votes across all cities. Similarly, let $\tilde{n} = \sum_{i=1}^{C} \tilde{n}_{i\bullet}$ be the total number of eligible votes in Israel (having 'zhut bhira')

- Let also z_i is the number of bad votes 'psulim' in city i.

1.2 Computing Parties Vote Share

1. The fraction of votes for party j in the elections is $p_j \equiv \frac{n_{\bullet j}}{n}$. The vector $p = (p_1, ..., p_K)$ represents the share of votes for each party, such that s_i, the number of seats in the parlament for each party, is approximately $s_i \approx 120 p_i$ (the exact relationship is much more complicated, and includes rounding, thresholding small parties due to 'ahuz hasima', the Badder-Offer law etc.).

2. Since the elections are meant to represent the opinions of all citizens in the country, voting turnout may be an issue as it can distort the actual preferences of the citizens - that is, if the turnout for the potential voters of party i is much larger than the turnout of teh potential voters of party v_j, then the share of votes p_i may be much higher for the first party compared to p_j for the second, even if in the general population the situation is reversed.

3. A natural question which we would like to answer is: can we infer from the elections results the actual preferences in the population? A followup question is: if every citizen in the coutry would have voted, would we see a significantly different result in the elections?

4. Denote by $\tilde{n}_{\bullet j}$ the (unknown) total number of votes for party j if every citizen actually voted. Similarly, denote by q_j the (unknown) share of votes for the party in this situation. Our goal will be to esitmate the q_j values from the election results.

1.3 A Statistical Model for Voting

We assume that each person in Israel decides in advance which party he/she prefers. Then, on election day, people from city i who prefer party j vote

with probability p_{ij}. Therefore, the number of actual voters for party j in village i is $n_{ij} \sim Binom(\tilde{n}_{ij}, p_{ij})$. Both $\tilde{n}_{ij}$ and p_{ij} are unknown parameters, and we will try to estimate them from the data in order to make a correction for the total number of votes. The problem is that the number of unknown parameters is $K \times C$which is on the order of the data size, and we have no hope of estimating the parameters reliably. Therefore, we need to make additional assumptions in order to estimate parameters.

1.4 Simulation Study

Our goal is to estimate the unknown q_j values (partie's proportion in the popoulation) from the observed p_j values (parties proportion in the election). For the real data, we don't know how good will our estimates be.

However, we can make different assumptions on the voting probabilities of individuals, and evaluate the performance of different corrections under these assymptions in a simulation study. The high-level description for a simulation study is as follows:

1. Choose values for the real numbers of voters $\tilde{n}_{ij}$ and voting probabilities p_{ij} , and compute the parties proportions q_j from the $\tilde{n}_{ij}$ values.

2. Simulate (many times) the observed number of voters in the election n_{ij} using $n_{ij} \sim Binom(\tilde{n}_{ij}, p_{ij})$

3. Apply a correction (see next section) to get estimators $\hat{\tilde{n}}_{ij}$ and subsequentially estimators $\hat{q}_j$ for the population proportions

4. Compare the true values q_j to the estimated values $\hat{q}_j$: Compute the empirical bias, variance and mean-suared error of the estimators $\hat{q}_j$.

1.5 Estimating total votes

We propose here different estimators for the votes distribution if everybody voted. The estimators differ in their assumptions, computation and statistical properties.

1. We can first do the following simple correction: if in city i the voting turnout was v_i, this means that every vote actually counted in this city reprsents not one but v_i^{-1} votes from the populatio of the city. We can thus give weights to the votes in each city. We get the following esitmator for the votes in a city:

$$\hat{\tilde{n}}_{ij} = \frac{n_{ij}}{v_i} \tag{1}$$

$$\hat{\tilde{n}}_{\bullet j} = \sum_{i=1}^{C} \hat{\tilde{n}}_{ij} = \sum_{i=1}^{C} n_{ij} v_i^{-1} \tag{2}$$

$$\hat{q}_j = \frac{\hat{\tilde{n}}_{\bullet j}}{\sum_{k=1}^{K} \tilde{n}_{\bullet k}} = \frac{\hat{\tilde{n}}_{\bullet j}}{\tilde{n}} = \frac{\sum_{i=1}^{C} n_{ij} v_i^{-1}}{\sum_{i=1}^{C} \sum_{j=1}^{C} n_{ij} v_i^{-1}} \tag{3}$$

   This estimator adjusts the voting in each city according to the voting turnout. We used this estimator in class, and the results were shown. A main problem with this adjustment is that it assumes that all voters in a city are equally likely to vote. But what if the voters of a certain party are more/less likely to vote?

2. To develop the next estimator, we will assume that the voter turnout for each **party** is a **constant** (rather than for each **city**). Let α_j be the voter turnout for party j, i.e. we assume $p_{ij} = \alpha_j$ Then we assume $n_{ij} \sim Binom(\tilde{n}_{ij}, \alpha_j)$ and have:

$$\tilde{n}_{\bullet j} \approx n_{\bullet j} \alpha_j^{-1} \tag{4}$$

   and therefore, if we knew the α_j values, we could use the estimator:

$$\hat{q}_j = \frac{n_{\bullet j} \alpha_j^{-1}}{\sum_{k=1}^{K} n_{\bullet k} \alpha_k^{-1}} \tag{5}$$

   We next need to estimate the α_js from the data. After we do so, we can just plug in the estimators $\hat{\alpha}_j^{-1}$ into the above equation to get:

$$\hat{q}_j = \frac{n_{\bullet j} \hat{\alpha}_j^{-1}}{\sum_{k=1}^{K} n_{\bullet k} \hat{\alpha}_k^{-1}}. \tag{6}$$

   How would we estimate the parties voting turnout? The idea is simple: if in cities where a party is strong we see higher voting turnouts, then the voting turnout for the voters of this parties is high (and the same for lower turnouts indicating a lower turnout for the party). To translate this idea into mathematical formulation, we would like the cities voting turnout v_i to be explained by the parties voting turnouts α_j. That is:

$$v_i \approx \frac{n_{i\bullet}}{\sum_{j=1}^{K} n_{ij}\alpha_j^{-1}} \tag{7}$$

since $v_i = \frac{n_{i\bullet}}{\tilde{n}_{i\bullet}}$, we can formulate a least-squares problem:

$$(\hat{\alpha}_1^{-1}, ..., \hat{\alpha}_K^{-1}) = argmin_{\alpha_1^{-1},..,\alpha_K^{-1}} \sum_{i=1}^{C} (\sum_{j=1}^{K} n_{ij}\alpha_j^{-1} - \tilde{n}_{i\bullet})^2. \tag{8}$$

That is, the inverse turnout parameters α_j^{-1} can be obtained as the least squares solution of a linear regression problem with desing matrix N and outcome vector $y = \tilde{n}$. The least-squares solution is therefore:

$$\hat{\alpha}^{-1} = [N^T N]^{-1} N^T \tilde{n} \tag{9}$$

and our estimator for q is

$$\hat{q} = \frac{n \odot [N^T N]^{-1} N^T \tilde{n}}{\|n \odot [N^T N]^{-1} N^T \tilde{n}\|_1}. \tag{10}$$

where $\odot$ is the Hadamard entrywise product, and $|| \cdot ||_1$ is the L_1-norm: $||x||_1 = \sum_i |x_i|$ The estimator we get here is different from the one in Eq. 3 above. How does it perform with respect to the actual election results compared to the previous estimator?

Remark: The least-squares criteria is not the only one which we can use for estimation, even if our assumption that $p_{ij} = \alpha_j$ is correct. For example, we

can alternatively want to fit the actual number of voters in each city, but rather the proportion how voted in each city

$$(\hat{\alpha}_1^{-1}, ..., \hat{\alpha}_K^{-1}) = argmin_{\alpha_1^{-1},..,\alpha_K^{-1}} \sum_{i=1}^{C} (\sum_{j=1}^{K} \frac{n_{ij}}{n_{i\bullet}} \alpha_j^{-1} - v_i^{-1})^2 \tag{11}$$