Phase 1 report

Piotr Grabowski, Jakub Kozanyi, Abel Szemler

November 2024

1 Introduction

The task of this project is to analyze whether there was election interference present, as there was a difference between how people voted online and in person. We base our findings on the comparison between election results and survey data. We then anonymize the survey data so that we can release it to the public. After that we check whether the anonymized dataset is statistically different from our original dataset.

2 Analysis methodology

For this step we used the raw survey data and the published results of the election. In the survey data we can find the names of the survey participants, their sex, date of birth, zip code, whether they voted online or on paper, which party they voted for, their marital status, education, and citizenship. The election results dataset contains how many votes each party got in each area and how many of these were cast electronically. It also includes how many votes were invalid.

To fulfill the analysis tasks we have conducted Chi square test of independence. This statistical analysis method calculates an expectation for unique combinations of two categorical variables (e.g.: number of Male with Green party preference) and then measures their statistical significance. This calculated value is what would be expected if there was no statistically significant association between the response and predictor. The expected value is calculated using the following function.

$$E_{ij} = \frac{rowtotal_j \cdot columntotal_i}{grandtotal}$$

.

Afterwards, we can calculate the chi-square statistic using the following formula.

$$x^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Once we find the x^2 statistics we can compare it against a chi-square distribution with the calculated degrees of freedom to obtain the p-value. This tells us the probability of observing a chi-square statistic as extreme as the calculated value, assuming the null hypothesis of independence is true. If the p-value is less than the significance level of 0.05, we reject the null hypothesis, suggesting an association between the variables.

2.1 Questions

(A) Is there a significant difference between the political preferences as expressed in the survey and the election results for both electronic and polling station votes?

(B) Is there a significant difference between political preferences of the voters depending on their demographic attributes recorded in the survey (that is, age, gender, education level...)?

(C) Is there a significant difference between voter's choice of the voting channel (that is, if they decide to vote either online or in person) depending on their demographic attributes recorded in the survey?

2.2 Analysis of raw data

In the following analysis we are going to use a significance level or p-value of $\alpha = 0.05$. Furthermore, the analysis is conducted on attributes; education, age, and sex. The chi-square values are referred to as x^2. Find thresholds in appendix 1.
Age column has been transformed into 5 categories ranging 0-33, 33-44, 45-55, 55-65 and 65+. Education was transformed into 4 groups; primary, vocational, bachelor, master+. These steps were done due to low amount of observations in certain groups like in education or too many different groups like in dob or age attribute. Without these transformations the statistical test would be inaccurate.

(A)

With our significance level, the critical threshold for a chi square test is 3.841, for further comparisons see appendix 1.
For the poll votes we find no significant difference between the results and the survey data with a value of 0.12.
For the e-votes we find a higher value 3.40, but still not high enough to reject the null hypothesis and state that the findings are significantly different between the survey and the results data.

(B)

There is a statistically significant association between level of education and political preference towards the Red party. $x^2_{red} = 9.86$, $x^2_{green} = 6.32$, $df = 3$
Age shows a significant association with people's party preference. It shows an especially strong association with the Red party and a slightly weaker but still significant one with the Green party. $x^2_{red} = 17.74$, $x^2_{green} = 11.37$, $df = 4$
Sex does not seem to be a significant contributor to people's choice of political party. $x^2_{red} = 3.12$, $x^2_{green} = 2.01$, $df = 1$

(C)

There is a statistically significant association between voting online and level of education of a person. $x^2_{poll} = 4.55$, $x^2_{evote} = 8.11$, $df = 3$
Age shows no association with people's choice of voting method. $x^2_{poll} = 3.57$, $x^2_{evote} = 6.37$, $df = 4$
Sex does not seem to be a significant contributor to people's choice of voting method. $x^2_{poll} = 0.01$, $x^2_{evote} = 0.02$, $df = 1$

2.3 Analysis of anonymized data

(A)

We find close to identical numbers for the test scores in the anonymized dataset as we did before $x^2_{poll} = 0.12$, $x^2_{evote} = 3.40$. There is no significant association between people's voting method and choice of party.

(B)

Education is not significant, but close, for both parties. $x^2_{red} = 3.05$, $x^2_{green} = 4.76$, $df = 2$
Age is statistically significant for the Green party and close to being significant for the Red party. $x^2_{red} = 6.62$, $x^2_{green} = 10.33$, $df = 3$
There is no significant association between sex and people's choice of party. $x^2_{red} = 2.01$, $x^2_{green} = 3.12$, $df = 1$

(C)

Education shows no statistically significant connection to voting method. It is another case of sharp decrease in chi-statistic. $x^2_{poll} = 0.24$, $x^2_{evote} = 0.42$, $df = 2$
Age has no statistically significant implication for people's choice of voting method. There is a notable sharp decrease in the statistical significance of this attribute. $x^2_{poll} = 0.24$, $x^2_{evote} = 0.43$, $df = 3$
There is no significant association between sex and people's choice of voting method. $x^2_{poll} = 0.01$, $x^2_{evote} = 0.02$, $df = 1$

2.4 Statistical difference between datasets

We have observed a sharp decreases in the statistical significance of the education (pre-anonymization: $x^2_{poll} = 4.55$, $x^2_{evote} = 8.11$, post-anonymization: $x^2_{poll} = 0.24$, $x^2_{evote} = 0.42$) and age (pre-anonymization: $x^2_{poll} = 3.57$, $x^2_{evote} = 6.37$, post-anonymization: $x^2_{poll} = 0.24$, $x^2_{evote} = 0.43$) group in connection with people's choice of voting method. This decrease is less observable for the party preference, where the chi test scores follow a similar pattern in the anonymized as the not-anonymized datasets. We have decided that these values are acceptable as this level of anonymization was necessary to achieve appropriate levels of privacy metrics.

3 Anonymization Methods

To ensure privacy and reduce re-identification risks, we applied several anonymization techniques to the raw survey dataset. These techniques included the removal of direct identifiers, generalization of quasi-identifiers, and suppression of sensitive information where needed, as outlined in the following sections.

3.1 Removal of Direct Identifiers

The dataset initially contained a direct identifier, *name*, which we removed in the initial anonymization step. This eliminated explicit identification of individuals.

3.2 Generalization of Quasi-Identifiers

We generalized quasi-identifiers to reduce their granularity, minimizing the risk of re-identification by increasing the size of attribute groups. The transformations included:

- **Age Group**: Dates of birth were converted to age categories: *18-35*, *36-65*, and *65+*. These categories retained age-related patterns while preventing exact age-based re-identification.
- **Region**: ZIP codes were grouped into broader regional categories, either *Region A* or *Region B*. This approach enhanced privacy by reducing specificity.
- **Marital Status**: Marital status categories were consolidated. *Never married*, *Divorced*, and *Widowed* were grouped into *Single*, while *Married/separated* was grouped into *Married*. This consolidation masked distinctions within marital status.
- **Education**: Education levels were generalized into broader categories. *Primary education*, *Upper secondary education*, and *Not stated* were grouped

as *Basic Education*, while all higher education categories—including *Vocational Education and Training (VET)*, *Short cycle higher education*, *Vocational bachelor's education*, *Bachelor's programmes*, *Master's programmes*, and *PhD programmes*—were grouped under *Higher Education*. This generalization preserves the distinction between basic and higher education levels while enhancing privacy.

- **Citizenship**: Citizenship was generalized to either *Domestic* (for Denmark) or *Foreign*, removing specific national information and reducing sensitivity.

3.3 Local Suppression for High-Risk Groups

We applied local suppression to records in high-risk groups that retained uniqueness following generalization. In such cases, we set quasi-identifiers, such as *age group*, *region*, *education*, and *citizenship*, to "Unknown" for these records. This targeted suppression minimized overall information loss while effectively reducing re-identification risk within these groups.

3.4 Information Loss and Consistency Analysis

To assess the effectiveness of anonymization, we calculated several metrics:

- **Suppression Rate**: The suppression rate, representing the percentage of cells set to "Unknown," was calculated at 10.5%, indicating moderate data alteration.
- **Categorical Consistency**: Categorical consistency between the original and anonymized data was calculated to be 89.5%, indicating that the majority of values aligned closely with the original dataset, preserving analytical validity.

These anonymization techniques were designed to achieve a balance between privacy and data utility, allowing for meaningful analysis while protecting individual privacy.

4 Summary

We did not find a significant difference between votes cast electronically and on paper. We used multiple anonymization methods to ensure that a potential adversary cannot identify survey participants. Statistical analysis of our anonymized dataset does provide a bit different results to the original dataset, but we decided that the changes are small enough for the anonymization to still be considered valid.

5 Appendix

5.1 appendix 1

Degrees of freedom (df)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Figure 1: chi-square distribution table