Final project assignment (Phase 1)

Analysis

We decided to use a chi-squared test for all three questions. The chi-squared test is a statistical method used to assess whether there is a significant association between two categorical variables. It is suitable for our analysis as it helps determine whether the observed frequencies in each category differ from expected frequencies under the null hypothesis. This test is particularly effective for analyzing the relationship between demographic factors, voting preferences, and voting methods, which are categorical in nature.

The scipy.stats chi-squared test computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table, providing a method to assess the association between different variables.

Question A
In our analysis, we began by creating a contingency table for polling votes and one for e-votes. Each table compared the number of votes for each party (Green and Red) between the survey and the actual election results.

The hypothesis for Question A is, "There is no significant difference in political preferences in polling votes and e-votes between the survey and the actual results."

Upon conducting the chi-squared test, we found that the p-value for polling votes was 0.477 and for e-voting was 0.268. Since both p-values are greater than the significance level of 0.05, we conclude that there is no statistically significant difference between the political preferences expressed in the survey and the actual election results for electronic and polling station votes. This indicates that the variations observed in the survey data compared to the actual results are likely due to random chance rather than indicating a true difference in voter preferences between the survey and the actual election.

Question B
In question A we tested for differences between the surveyed results and the actual results. Here we will test for association between demographic value and political preference.

We constructed a contingency table for sex and political preference, education_level and political preference, and lastly age and political preference.
The p-value for sex is 0.0155 which indicates that there is statistically significant association between sex and political preference.

The p-value for education is 0.0019 which also indicates that there is statistically significant association between education level and political preference.

The p-value for age is 0.0000 (rounded) which again suggests that there is statistically significant association between the age of people and their political preference.

Question C

To test if there is significant difference between voting channel and demographic attributes we used the chi-squared test again. We are testing the hypothesis of association of demographic attribute and choice of voting channel, that is if the p-value is above 0.05 we can reject association of the attribute and voting channel preference. However, if the p-value is below 0.05 we can accept statistical significance.

We created contingency tables between sex and voting channel, education and voting channel, and lastly age and voting channel.

The p-value for sex vs voting channel is 0.3860, which means that there is no statistical significance in the association of sex and choice of voting channel.

The p-value for education vs voting channel is 0.5470, which also indicates that there is no statistical significance between a person's education level and their preferred voting channel.

The p-value for age vs voting channel is 0.0000 (rounded), which does indicate that there is a statistical significance between the age of people and their preferred voting channel, because it is commonly known that older people tend to be less technically savvy.

Anonymization method

Removal of (direct) identifiers
In our endeavor to achieve k=2 anonymity within our small dataset, the first critical step was to remove direct identifiers. Names are a direct form of identification and their removal significantly reduces the risk of individual identification. Citizenship information can lead to the identification of individuals, especially in our dataset where only 9 different nationalities were present resulting in limited diversity in citizenship. By stripping these elements from the dataset, we anonymized individuals represented, laying a foundational layer of privacy protection.
We decided to exclude the zip codes from our dataset. Given our small dataset where all zip codes correspond to the municipality of Copenhagen, they don't significantly contribute to the analysis and insights we aim to derive.

Age Group Categorization:
To further anonymize the dataset, we categorized ages into broader groups: 18-40, 40-70, and 70+. The resulting age groups were rather broad to ensure that there were no uniquely identifiable records in the dataset.

Marital Status Simplification & Education Level Generalization:
The dataset initially contained detailed marital statuses, including divorced, widowed, married/separated, and never married. To achieve k=2 anonymity, we simplified these categories into a binary classification: married or not married.” Married/separated”became “Married” and the rest are classified as “Not married”. This generalization helps prevent the identification of individuals based on their marital status. We also generalized the education levels of individuals into three broad categories: low, middle, higher education and not stated. The mappings looks as follows:

```
education_bins = {
    'Primary education': 'Lower Education',
    'Upper secondary education': 'Lower Education',
    'Vocational Education and Training (VET)': 'Middle Education',
    'Short cycle higher education': 'Middle Education',
    'Vocational bachelors educations': 'Middle Education',
    'Bachelors programmes': 'Higher Education',
    'Masters programmes': 'Higher Education',
    'PhD programmes': 'Higher Education',
    'Not stated': 'Not Stated'
}
```

Suppression
After all these steps, we encountered a single row that prevented us from achieving k=2 anonymity. To handle this, we found similar rows and suppressed their marital_status, so three males above 70 years of age have a “*” as their marital_status.