Digitizing Maine's Voting History with a Statistical Analysis of Error Rate

Gillian King, Class of 2022

My project concerned the digitization and analysis of error rates in Maine voting data. The work was divided into two parts: 1) digitizing voting data in Maine using an Optical Character Recognition (OCR) software and 2) performing numerous tests to assess these error rates. Part 1 used ABBYY Finereader, an OCR software, to detect "low-confidence characters" in the digitization process. These errors were then manually corrected. Since OCR software systems may not flag all incorrect entries, various statistical tests were used to correct these mistakes and analyze the error rate. In the case of this project, the error rate is defined as the ratio of incorrect entries produced by the OCR software to the total number of recognized entries.

The first test involved calculating the p-value of each row of data; a p-value "indicate[s] how incompatible the data are within a specified statistical model" (Wasserstein and Lazar, 2016, pg. 131). Each row corresponded to a town and each column represented yes/no votes for each question. Since each row of the data set follows its own statistical pattern, calculating a p-value gives an indication of how well each value fits into a given row's pattern. In Statistics, a lower p-value is typically associated with a higher likelihood that the data follows this pattern (Wasserstein and Lazar 2016).

Our team constructed a program in RStudio that plots questions against one another, and then analyzes the p-value of each row of the data set. When plotted in this way, these data form a simple linear regression. A line of best fit was added to this regression to analyze how strong the correlation was. However, calculating the p-value of each row identified individual errors in this pattern more clearly. The convention threshold p-value for accepting this data is 0.05, though there is debate in academia about the universality of this value (Wasserstein and Lazar 2016).

P-values were calculated using several built-in statistical functions in RStudio; the range of each row's p-values was typically between 0.0003-0.8. As such, these rows closely followed their predicted patterns for vote totals. Given this range, rows whose p-value was less than 0.0003 were deemed "outliers." To confirm whether a row was an outlier, the digitized row was compared to the value on the original, handwritten document.

This test was largely successful in identifying large outliers. For that reason, a follow-up statistical test was used to catch more subtle errors. The tests included the summation of each CSV column—noting whether it matched with the totals given on the original documents. If the totals did not match, the handwritten document and the converted CSV filed were reviewed alongside one another to look for individual errors.

Finally, a program was used to randomly select a row and column of the CSV file. This entry was checked against the value on the original document; this process was repeated twenty times for each data set. This second test mitigated most, if not all, of the errors for the county sheet, yet this final test would verify this fact. These three tests used in tandem mitigated enough errors to produce a high-accuracy data set to include in Professor O'Brien's digital archive of Maine voting records.

Faculty Mentor: Professor Jack O'Brien

Funded by the Kufe Family Research Fellowship

References

Wasserstein, R.L., and Lazar, N.A. (2016), "The ASA Statement on *p*-Values: Context, Process, and Purpose", *The American Statistician*, 70:2, 129-133.