

Capstone 1 Statistical Data Analysis

James Olmstead

Context and Data:

My question for my Capstone project is to see if we can use survey data to create categorize the respondents based on how they answered questions on political preferences in the United States. The dataset in question can be found at the following website, which also contains a link to download the documentation as a pdf: <https://www.voterstudygroup.org/publication/2018-voter-survey-1>

A Question of Affiliation

After cleaning the data, I decided to perform some exploratory analysis. Because of the enormous number of survey questions, there were many different directions for me to go. Since I am looking to categorize respondents by political labels, I decided to take a closer look at how the respondents identified themselves. This portion of the survey asked them to give a yes/no answer on if they identified with a certain political affiliation. Those affiliations were Libertarian, Socialist, Green, Environmentalist, Liberal, Moderate, Conservative, Radical, Progressive, Traditional, Christian, Feminist, Fundamentalist, and None.

Analysis

In order to examine these affiliations in more detail, I decided to see what sort of overlaps there were between the affiliations. After some research, it was clear the inferential statistic of choice was to be the [chi-squared statistic](#), which is best used for categorical variables. This test allows me to check how statistically significant the relationship is between 2 categories. After getting the chi-squared statistic and associated p-value, I am able to get the Cramer's V, which is the actual measure of the strength of the relationship.

Results

A great way to visualize the results is a coefficient matrix. This makes it easy to see the pair-wise test results. These matrixes are on the second page of this document as Figures 1 and 2.

Observations and Conclusion:

Most nearly every pair-wise test was statistically significant, with a few exceptions mostly being tests involving Libertarian, Moderate, and Fundamentalist. In addition, most of the results for the associations themselves are to be expected. The strongest associations are between Green/Environmentalist and Liberal/Progressive both having Cramer's V values above .5. On the other hand, associations between Traditional/Christian wasn't as high as I would have guessed (only .29) while Environmentalist/Feminist is higher than I thought (.33) since there is no inherent similarity besides being associated with left-wing politics. It is also important to note the relatively strong association for Conservative/Liberal (.36) is likely a negative one i.e. those who identify as Conservative don't identify as Liberal and vice versa.

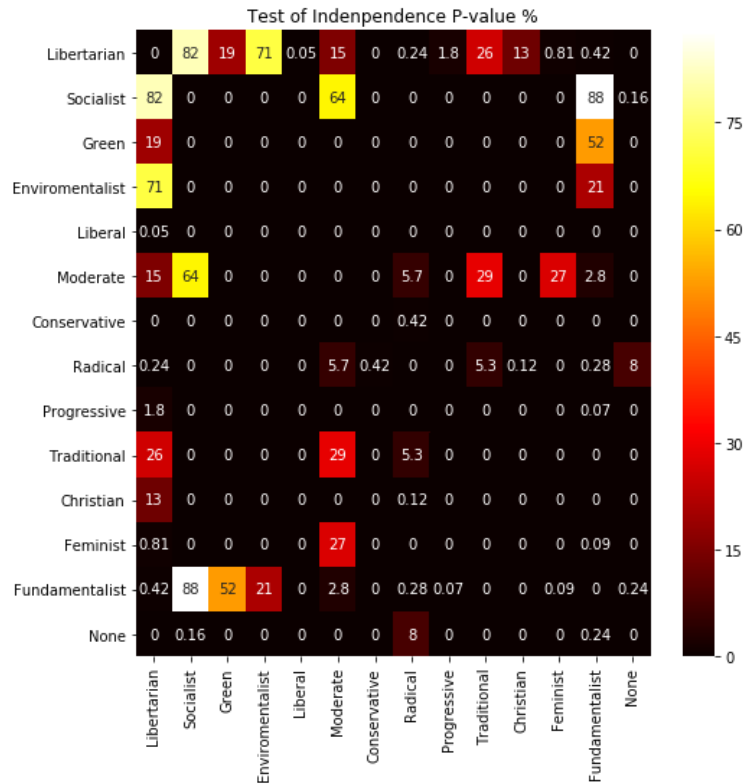


Fig. 1

Pairwise p -values for the chi-squared statistic. Any value below 5% is considered statistically significant.

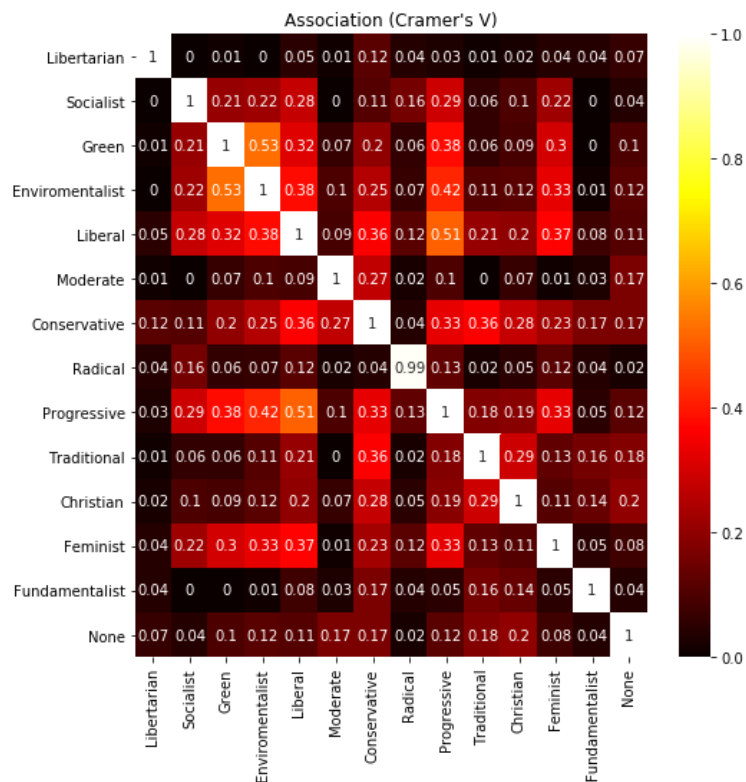


Fig. 2

Cramer's V association. Values closer to 1 signify a closer association. Keep in mind association can be positive or negative in direction. (E.g. answering 'No' to conservative may be strongly associated with 'Yes' for liberal.)