**Capstone Project: Apply Inferential Statistics**

Jake Schroeder

**Introduction**

The data being analyzed in this project is Google Trends data, aggregated annually by mean, and annual population forecast data from the Federal government of the United States of America. With the goal of understanding the relationship between populations of certain age groups and their potential impacts on popularity of keywords in the banking sector, this section of the project analyzes the data using a two-sample permutation t-test.

**Permutation Test Explained**

This experiment confirms data integrity by leveraging random sampling. We utilize a function that draws pairs of random samples using Numpy. We provide the size of the Numpy arrays (which must match) and are returned random indices for selection. It is from these random indices that random sampling occurs.

We repeat the random sampling process 10,000 times for each age group, for each keyword. Each random sample is fed into a Pearson Correlation Coefficient function, which returns a Numpy array of the resulting coefficients for all 10,000 trials. We record the p-value of each age cohort as a single entry in a Numpy array containing entries for all ages compared to a same keyword.

**T-Test Explained**

*Our t-test is under the null hypothesis that the strong correlations examined were a fluke*; that is, the permutation test will show weak reproducibility of the initial strong correlations. Reproducibility in this case is measured by p-value at the 95 percent confidence interval.

*The p-value in this experiment is calculated by finding the percentage of observations that exceeded the initial observation made by the actual data.*

For example, if the original data suggests “bankruptcy” is positively correlated at strength of .5 to the 47-year-old age cohort, we only count correlations for this cohort that are stronger when compared in absolute terms.

Indeed, correlations are measured by their absolute value - this enables us to measure total strength of correlations, independent of whether those correlations are positive or negative.

**Analysis Process**

Each keyword is analyzed against the entire group of age cohorts. This analysis is then stored in a distinct pandas DataFrame with the mean correlation coefficient as the top value, and the p-value as the bottom value. When observed at a glance, we see the data as columns of age cohorts, with two rows of mean correlation and p-value, accordingly.

**Analysis Results Summarized**

Though there is much data to parse through, a simple analysis of each resulting DataFrame proves the null hypothesis to be incorrect. To reiterate, the null hypothesis states that the strong correlations suggested by the initial data were not normal given the weak relationship that we assume these two sets of data to share.

With high p-values across the board (very few p-values remotely near 0.05), the reproducibility of strong correlations is evident. Therefore, we can say with high confidence that the Google Trends and population demographic data share a strong relationship worth further investigation.

**Other Notes:**

See iPython Notebook for Details