Statistical Data Analysis

Capstone 1: NYC Crime Data 2019
Jupyter notebook: https://github.com/elizabeamedalla/Capstone-1

The data we have in this capstone consists of categorical value. The most efficient way to apply inferential statistics for our data is through the Pearson Chi-squared Test. Chi-squared test measures how expectations compare to actual observed data (or model results). The data used in calculating a **chi-square statistic** must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample.

The first test I did was the Chi-square Goodness-of-fit Test. This tests whether the population mean of two variables is the same. For this analysis, the two variables are NYC census data in 2019 by race and crime victim data by race in NYC 2019. My **null hypothesis** states that the population means between the population of different races in NYC in 2019 **is the same as the sample mean** of different races of crime victims in NYC 2019. My **alternative hypothesis** states that **t**he population mean between the population of different races in NYC 2019 **is different from the sample mean** of different races of crime victims in NYC 2019.

Using the chi-square goodness-of-fit test, I tested the p-value and critical values using my sample data and population data. First, I got the population of races in NYC in 2019 using various census data and used this as my population. Second I used the victim race data in my dataset as my sample. Third, I created tables for my population (NYC 2019 census) and sample data (crime victims) by race. Fourth, I applied the chi-squared formula to calculate my chi-square statistic. I assigned my observed data as my sample table and I calculated my expected data as the product of the population ratios and the sample size. Fifth, I calculated the p-value and critical value using my chi-squared statistic. Based on the p-value, I concluded that **I reject the null hypothesis** and determine that the sample mean of the population mean is different from the population mean of crime victims by race in NYC 2019.

The second test I did was the Chi-square Independence Test. This tests whether two variables are independent of each other. For this analysis, I used the victim race sample mean and crime type sample mean. My **null hypothesis** is that the victim race sample mean is related to the offense type sample mean. My **alternative hypothesis** is that the victim race sample mean is not related to the offense type sample mean.

The chi-square formula is the same for the Independence test and the Goodness-of-fit test. The main difference for the Independence test is what I use for the observed and expected variables in the formula. In the independence test, I used random sample distributions of my data to generate my observed data, and calculate my expected data. First, I created a random sample distribution of victim race and offense type. Second, I created a crosstab of the random distributions with race as the columns and offense type as the index. Third, I got the observed distribution from the crosstab. Fourth, I calculated the expected distribution by taking the row totals and column totals of the crosstab, performing an outer product on them with the np.outer() function and dividing by the number of observations. Fifth, I calculated the chi-square statistic using the chi-square formula. Finally, I computed the critical value and p-value using the chi-squared statistic. Based on the critical value and p-value, **I reject the null hypothesis and** conclude that no significant relationship or correlation detected between victim race sample mean is related to the offense type sample mean.

Since all of my data was categorical and non-numerical, I struggled to apply what I learned from the examples in my mini-projects and lessons to my categorical data. There may have been other questions to answer and other methods using statistical analysis on my categorical data, but I chose to use the chi-square tests on my categorical data.

The **chi-square** statistic is commonly **used for testing relationships between categorical variables**. The null hypothesis of the **Chi-Square Independence Test** is that no relationship exists on the categorical variables in the population; they are independent. Based on the p-value and critical value, there is no relationship between victim race and population. I also used the **Chi-Square Goodness-of-Fit test** to determine if the sample data matches the population. Based on the result, our victim race sample does not match the NYC 2019 census race population.