Project: Capstone Project 1 - Inferential Statistics

For this project, the Amazon Find Food Reviews dataset was analyzed. The goal of this project is to create a predictive model which will classify a product as having positive or non-positive feedback from the customers based on the number of stars (score). From a user's point of view, it is clear that there will definitely be a strong correlation between the reviews of a product to the score received. However, this generates the question, are there other features which might help in the classification of the products? To answer this question, a couple of features (HelpfulnessNumerator, HelpfulnessDenominator) were analyzed and tested for statistical significance.

In order to test the above scenario, a hypothesis testing was utilized. The testing was set up as follows:

h_0: There is no correlation between the two features and score variable.

h_1: There is some correlation between the two features and score variable.

Then, the observed correlation coefficient was calculated from the original datasource using the below function, where x and y are the two lists of values to find correlations between.

```
def pearson_r(x,y):
    corr_mat = np.corrcoef(x,y)
    return corr_mat[0,1]
```

From the above, it was found that HelpfulnessNumerator has a pearson correlation of -0.0325901134286 and HelpfulnessDenominator a value of -0.0979859576873 to the feature Score. It seems there is some correlation according to the value, but further testing was done to check how significant these values were. A permutation test was utilized with the assumption that correlation is not significant. The helpfulness values were permuted and pearson correlation was calculated again. This trial was ran 5000 times. Using the same function above, a 99% confidence interval of correlation were found to be:

- HelpfulnessNumerator has range [-0.00517265 0.0029612] with confidence interval 99
- HelpfulnessDenominator has range [-0.00476361 0.00298221] with confidence interval 99

Since the observed correlation is outside the bounds of the permutation samples' confidence intervals, it can be concluded that there is real correlation between the variables.