Capstone1 Statistical Analysis

For the statistical analysis of my dataset of the first capstone project, I used several different techniques to perform hypothesis testing. I compared distributions of the mean between independent variables and between the target variable and one of the predictors. I applied hypothesis testing to evaluate a correlation coefficient, as well.

The first hypothesis test was to examine the difference between the distribution of ages with respect to male and female actors. After separating the data by gender, I generated 10,000 permutation samples of each set. I plotted the Empirical Cumulative Distribution Functions (ECDF) for each and compared them to those of the observed data. The curves showed a distinct gap between the observed data, while the curves of the samples overlapped. This indicated that the distributions between the genders were not the same. Next, I compared the difference of the mean age between the two genders with that of 10,000 permutation replicates. The p-value was given as zero. This meant that the probability was extremely low of observing a difference of the mean age as extreme as the one observed between the two gender sets.

For the second hypothesis test, I examined the difference between the mean age of the two genders, without considering their distribution. I created 10,000 bootstrap replicates of the difference of the mean by shifting the means of the two sets to coincide, generating bootstrap replicates from each of those arrays, and taking the difference of them, elementwise. The p-value was given as zero, once again. This confirmed that regardless of whether or not the two sets came from the same distribution, it was highly unlikely that the difference in mean ages was due to chance.

The next hypothesis test focused on the target variable, actor value. The goal of this test was to determine if there was a significant difference between genders with respect to the target values. Again, I drew 10,000 permutation samples for each gender and plotted all four ECDFs. The curves lined up tightly, with a small gap showing toward the larger positive values. After drawing the permutation replicants of the difference of the mean actor value, the p-value was not zero, but it was still below 1%. This meant that we could be more than 99% confident that there was a real difference between the distribution of the mean actor values between the genders.

Finally, I applied hypothesis testing to examine the correlation coefficient between the target variable and the predictor variable, gender. The correlation coefficient I computed was very small, but they were small for all other predictors, as well. I permuted the values of each gender array and generated 10,000 permutation replicates of the correlation coefficients from them. The p-value for this test was just over 0.07. This meant that there were 725 replicates out of 10,000 whose correlation coefficients were greater than the ones calculated from the observed data. This meant that we could not be more than 92% confident that the two variables were correlated.