Capstone Project 1 - Statistical Data Analysis

Factors Influencing ACT Scores

The statistical analysis on the data consists of three sections.

1. Checking the Normality of the data

For the ACT data, my assumption was the average scores were normally distributed and this stood correct by looking at the distribution plot of the data. In this section, I checked if this was also true statistically by comparing the ECDF of the data for average scores to the theoretical CDF of the Normal Distribution. To compute the theoretical CDF, I used np.random.normal with the mean and std of the data for average scores with a sample size 10000. And looking at the graphical results, I concluded that the average scores of ACT data were approximately Normally distributed.

But in this analysis, to generate the normal distribution I used the mean and standard deviation of the original data. Then I investigated if these parameters were appropriate values for the normal parameters? To answer this question, I generated a bootstrap sample and repeated the process 10000 times. Using this bootstrap sample which is created by replacement technique, I also generated bootstrap replicates 10000 times to compute the mean. Then I computed the standard deviation of bootstrap replicates and created a histogram showing the distribution of bootstrap estimate of the mean. The original mean of the data was 21.3474. In the graph, it showed that 95% of the sample means would lie within the 95% confidence interval [21.24857604 21.44879386].

2. Investigating the Correlation between the variables

In this section, I showed the correlation matrix for some variables using df.corr() and sns.heatmap() techniques. The observed correlation between some variables seemed very strong. But this condition may just be by chance, those variables may actually be totally independent of each other. So I tested this hypothesis in this section. I investigated the correlation between the mean Act scores and the percentage of Hispanic students, as well as for the percentage of White non-Hispanic students and also the mean Act scores and the percentage of low income students.

To do this, I used pairs bootstrap for linear regression for highly correlated variables. I generated replicates for slope and intercept and showed the confidence interval for the replicates. Here are my findings for this section:

The bootstrap replicates with a 95 % confidence interval indicate that the slope and intercept between the average scores and the percentage of Hispanic students have a 95 % chance of lying within [-0.09720368 -0.09239305] and [26.20255784 26.51325565]. The original slope and the intercept was - 0.0949256254 and 26.36371850.

The bootstrap replicates with a 95 % confidence interval indicate that the slope and intercept between the average scores and the percentage of White non-Hispanic students have a 95 % chance of lying within [0.09900614 0.10519674] and [18.76642025 18.98418596]. The original slope and the intercept was 0.1018791975 and 18.877755183.

The bootstrap replicates with a 95 % confidence interval indicate that the slope and intercept between the average scores and the percentage of low income students have a 95 % chance of lying within [-0.10996054 -0.10618294] and [27.21568846 27.44642057]. The original slope and the intercept was - 0.1080966003627 and 27.33169245579.

3. Hypothesis test for mean difference

In the exploratory data analysis section, I observed that the mean scores were different for different years in four subject tests. There were some trends for some test scores. So in this section, I found answers for these questions: Is it possible that this effect is just due to random chance? In other words, what is the probability that we would get the observed difference in mean scores if the means were the same? The hypothesis is that the means are equal.

To perform this hypothesis test, I shifted the two data sets so that they had the same mean and then used bootstrap sampling to compute the difference of means. Then I estimated the difference between the means of the samples from 2014 and 2018 and reported a 95% confidence interval.

For all of the test subjects and the overall mean scores, p-value was less than 0.05. Based on the p-value, I rejected the null hypothesis and concluded that the mean scores for 2014 and 2018 were not equal to each other.