3.3 **Income Inequality with Regions**

Regions are also one of the most important factors that may impact the population’s income level and income equality. In the dataset3, the data was related to the income level in different regions and countries around the world. The question we wanted to address is whether the income inequality existed in different regions.

**3.1.1 EDA**

We subgrouped the income shares percentages on 5 quintiles of population (Q1, Q2, Q3, Q4, Q5) based on the four regions (Asia, America, Africa, Europe). Our focus was on the Q1 and Q5 population because each of them represents how much income shares the 20% of the population with lowest income and highest income in those regions occupied. Based on those simple grouping and we did some exploratory data analysis.

Firstly, we did the summary analysis on the quintile 5 population in different region. From the summary table below, we can easily find out that America has the highest mean, which is 54.54, which can be interpreted as the top 20% population with highest income occupied about 54.54% of total income shares. From this significant number, we can first speculate that the income in America region is highly unequally distributed because the 20% of population occupied about half of the income shares. Comparably, Asia and Europe has lower income inequality than Africa and America.

A screenshot of a computer

Description automatically generated

Next, we made two box plots related to the income shares of Q1 and Q5 population with different regions. As shown in the first box plot, which is plot the lower quintile population’s income shares percentage, we can see America population has the lowest median percentage of income shares, around 4. This means that the 20% of population with lower income occupies only about 4% of total income. The upper and bottom quartiles are approximately 3 and 5, respectively. There is one outlier that is close to 1, and the range goes from about 2 to 6. This shows that there is a lower number that differs significantly from the rest of the data, even though the central tendency is larger than in Asia and Europe. Other useful observation is that the existence of outliers in the plots for Asia and Africa indicates the presence of extreme values, which may be the result of anomalies, excessive wealth or poverty in certain regions.

The second Boxplot is plotting the upper quintile population’s income shares percentage. From the plot, we can see America population has the highest median percentage of income shares, around 55. This means that the 20% of population with upper income occupies more than 50% of total income, which is a significant number. A consistent distribution of upper-income levels is suggested by the income distribution's mild symmetry around the median. Comparing Asia and Africa to Europe and the Americas, more variation can be seen in top income levels.

A diagram of a graph

Description automatically generated with medium confidenceA diagram of a bar chart

Description automatically generated with medium confidence

Based on our initial observation, we can first get a hypothesis that America are more income inequal than the other three regions. To prove this hypothesis, we will next run a hypothesis test.

**3.1.2 Hypothesis Testing and Methods**

Based on the dataset, we picked two regions to represent the developed region, America, and developing region, Asia, and made the null hypothesis and alternative hypothesis as following:

Frist, we used T-test method identify the hypothesis at 95% significance level. The code and result are shown as below. From the result, we can have the p-value is less than 2.2e-16 <0.05, which indicates that the observed difference in means is highly statistically significant. This significant small p-value also indicate that we should reject the null hypothesis. So, we can conclude that at 95% of significance level, the mean of "americas\_q5" is significantly larger than the mean of "asia\_q5."

A screenshot of a computer program

Description automatically generated

We also used the bootstrap method to identify the result and double check the result before. The code and result are shown as below. From this result, the variance of the bootstrap means is 0.000378, which is a significantly small number, and it suggest a relatively stable estimate. Since 1 is not in the 95% bootstrap confidence bound, we should reject the null hypothesis. Thus, we have strong evidence that the mean of "americas\_q5" is significantly larger the mean of "asia\_q5."

A screenshot of a computer

Description automatically generated

**3.1.3 Results**

Based on both results from the T-test and Bootstrap, we should reject the null hypothesis, meaning that at 95% significance level, we can conclude that America's upper quintile population of income shares is larger than Asia's upper quintile population of income shares. The result also shows that, the 20% upper income population in America earn more income than the 20% upper income population in Asia in their own regions. As each represented as developed region and developing region, we can also conclude that upper-income population in developed region control comparably more money than population in developing region. So, there is more income inequality in developed region, like America, than developing region, like Asia.