‘*The Application of Statistical Inference*’ - Adrian Halarewicz

The DataFrame containing the cost of living index for cities throughout the world can be treated and processed as a sample for determining the true mean of the cost of living. The statistical analysis for this project will focus on the cost of living data and obtaining a sense of the true mean for the cost of living throughout the United States.

First, we can examine the distribution of the data points representing the cost of living index for individual cities. Because the scope of this project is built around student loan data in the United States, the data for the rest of the word will not play much of a significant role in this statistical analysis. It would be beneficial for the user to understand how the cost of living index of their particular city compares to the average cost of living index in the US.

Assuming there are cities in the US that are not accounted for in the cost of living index data, a challenge appears when trying to determine the average cost of living index throughout all cities in the US. To solve this problem, different methods of statistical inference can be applied to our sample. This project utilizes a Frequentist, Bootstrap, and Bayesian approach to statistical inference to determine the mean cost of living index in the US.

Before any analysis or inference is performed, a few functions are defined to support the analysis. Functions to compute ECDFs and plot histograms of different bin sizes aid in additional Exploratory Data Analysis to asses the distribution of the sample. After plotting the ECDF and multiple histograms, it is clear that the sample data is skewed right and appears to be not normally distributed.

The first attempt at computing the mean utilized a frequentist approach. The application of frequentist inference yields an interval with a 95% probability of containing the true mean within the interval. The function takes the sample as a single parameter and begins by computing the observed mean of the sample. The standard deviation for the sample is computed while accounting for 1 degree of freedom. The standard deviation of the sample is only an approximation of the population mean, and utilizing (n-1) instead of (n) in the denominator provides a better approximation for sigma (the true population mean). Utilizing the standard deviation and the central limit theorem the standard error is computed. The standard error, in combination with the critical z-value, is used to compute the margin of error and produce a 95% confidence interval for the mean.

The frequentist approach provides an interval that described the mean with a specified uncertainty. Bootstrapping can provide the same results while also describing the likelihood of obtaining the recorded sample. Through bootstrap statistics, we can simulate running the experiment many times to obtain a distribution of our desired characteristic. The bootstrap function creates 10,000 replicates of our sample and stores the mean in an array. Functions were defined to generate bootstrap replicates.

The confidence interval was computed and a visualization was constructed for the distribution.

Finally, Bayesian inference was the third method of inference applied to the US cost of the living sample to estimate the mean. Using appropriate estimates for the exponential priors alpha and beta, a gamma distribution is constructed. 10,000 draws are performed and the 95% credible intervals are computed and plotted. The results of the Bayesian inference do not agree with the results of the other two methods of inference. Therefore, the next steps would be to review the Bayesian model’s algorithm to see where corrections or improvements could be made.

For the three individual methods of inference, the results are as follows:

Frequentist	[69.26 , 72.55]	95% Confidence Interval
Bootstrap	70.9	Most probable mean
Bayesian	68.86	Most probable mean