Background

In this project, our client (a real estate developer) sought an application which discovered development opportunities based on market trends. My first task was to compare a summary statistic for a full data set, which I extracted as my population from www.factfinder.census.gov, against a summary statistic for one-sample which I had to compute. I discovered that my summary statistics for the Full Data Set as compared to my summary statistics for one random sample were closely related. I also learned that the statistics for one random sample was a relatively good representation of the overall population of the full data set with the average means being almost the same. The average number of bedrooms for the overall population (Full Data Set) was 3.21 and the average number of bedrooms for One Random Sample was 3.45 bedrooms. This was a difference of 0.24 which if I was to round to whole numbers the average of each, population vs. sample, I could safely round to approx. 3 average number of bedrooms for both population and sample set.

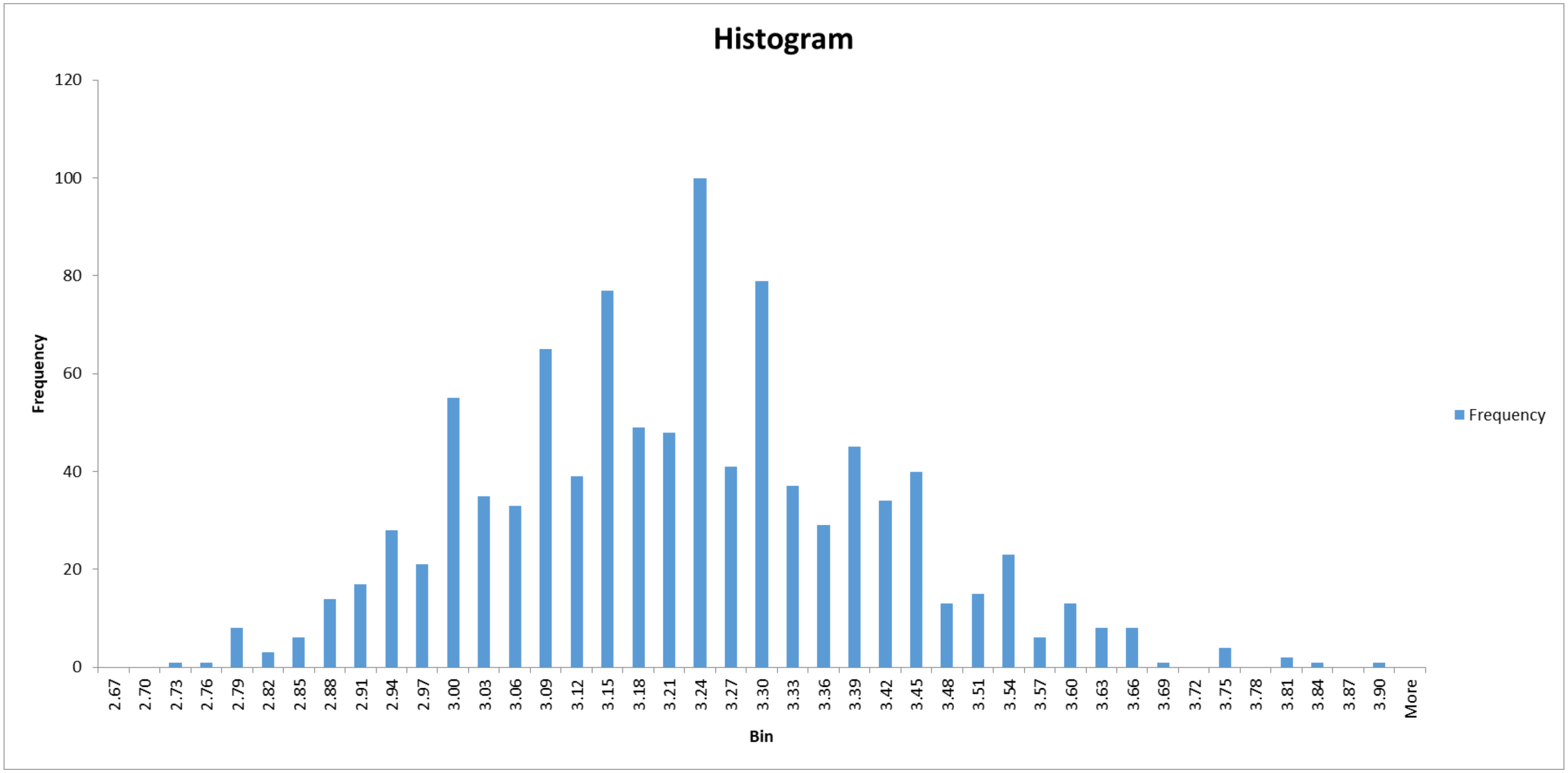
I then proceeded to compute my mean (average) of the sampling distribution obtained from 1,000 samples and compared this to the mean of my full data set and against the mean for one random sample. The mean of the sampling distribution obtained from 1,000 samples (3.21) compared to the mean of my full data set (3.21) and mean for one random sample (3.45) were all closely related.

I analyzed the one-sample mean versus the mean of the sampling distribution means to decide which of the two was a better approximation of the population's mean (the average for the full data set). After careful analysis, I learned that the Mean of the Sampling Distribution of the 1,000 mean samples, was, indeed, a better, more accurate, representation/approximation of the overall population's mean (Average of the Full Data Set) because the sample was much larger (1,000 Random Samples vs. 1 Random Sample), bringing the mean to almost the same value: Full Data Mean at 3.209360298 versus the Means of 1,000 Random Samples at 3.206319149. When rounded up, they were both approx. 3.21.

Moreover, I compared the standard deviation of my 1,000 sample means (standard error) to the standard error based on the standard deviation of my population. The standard deviation of my 1,000 sample means (standard error) of 0.19 compared to the 0.14 standard error based on the standard deviation of my population in that they were both relatively close and a good representation. That being said, my standard deviation of the 1,000-sample means was a good approximation of my standard error.

Next, I looked for any outliers within my 1,000 sample means, in other words, the amount of sample mean values that were more than 3 standard errors away from the mean of sample means. Based on my 1,000 sample means, I had at least one outlier because max. value was 3.87 > 3.77 outside my range (values more than 3 standard errors away from the mean of sample means). This was consistent with the 68-95-99.7% empirical rule because looking at my histogram chart below, it was evident that my data spreads out according to the Bell-Shaped Normal Distribution Curve, with most of my data falling within one standard deviation (68%). And when I looked at 3 standard deviations (99.7%), my data aligned according to my chart with most of my values within that 2.64 to 3.77 range.

Based on my histogram for the distribution of 1,000 sample means, I ended up with a nice bell shape curve. This shape for my histogram made it consistent with any predictions used in the Central Limit Theorem.



Histogram for the distribution of 1,000 sample means.