Kazi Islam

Statistics 295

Professor Benedict

Final Project

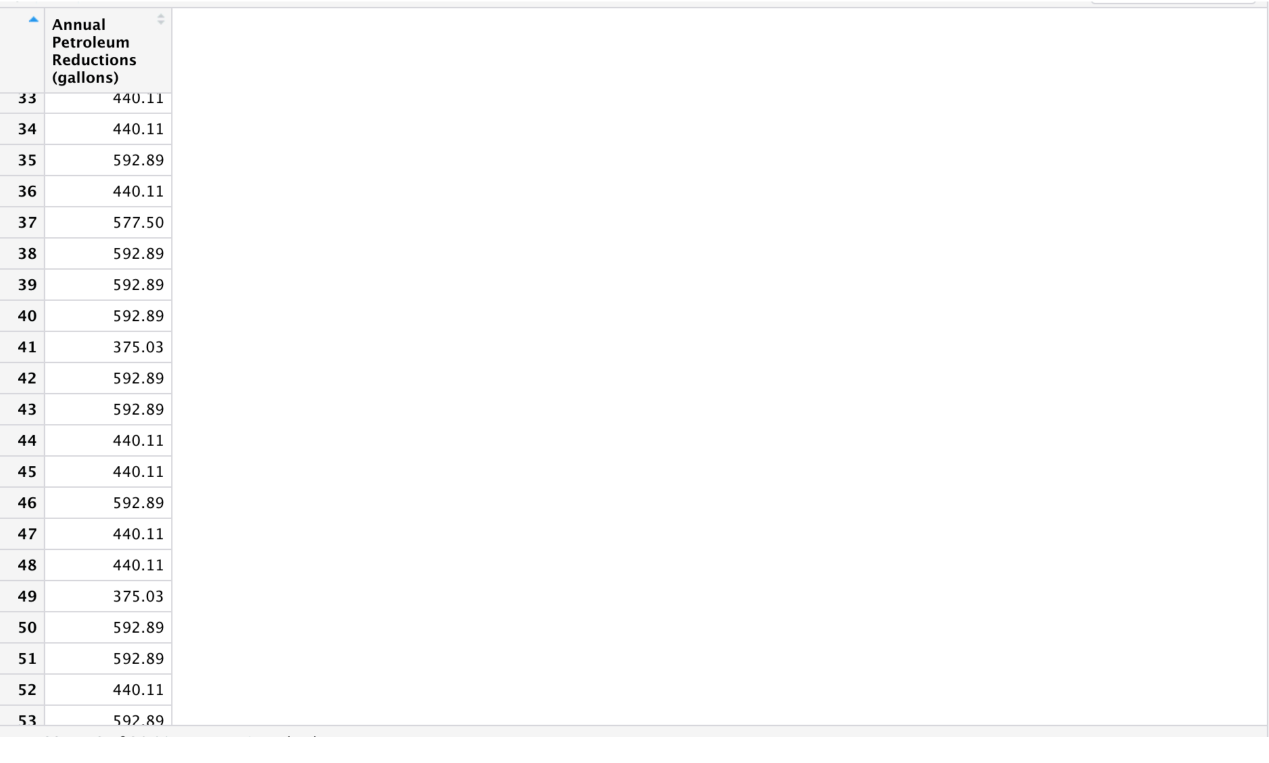
Abstract:

This study analyzes the purported annual petroleum reductions of purchased electric vehicles in New York as reported by the New York State Energy Research and Development Authority. The study first analyzes 36000+ observations of electric vehicle purchases partaking in the Drive Clean Rebate initiative. Population level statistical analyses reveals that majority of observations report annual petroleum reductions above 350 gallons with outliers laying in even the negative spectrum. From this dataset of 36000+ observations, 8 observations were chosen to run further statistical analyses. It was revealed that the 8 observations remained a good estimator for predicting the distribution of the 36000+ observations. Sampling distributions were created from the 8 observations by taking n=3 samples at a time. The resulting sampling distribution of the sample means represented a very accurate depiction of the target population. In fact, the sampling distribution of the sample means may be produce more useful estimates because it dislodges outliers. The sampling distribution of the sample means depicted a much more normal distribution than both the target and the reduced target population. Finally, the calculated variances of individual samples of the sampling distribution proved that single samples represent strong estimators of the target population. This goes to prove that statistical analyses are much better on smaller and fewer observations as opposed to running statistical analyses on costly and timely target populations.

The dataset, “NYSERDA Electric Vehicle Drive Clean Rebate Data: Beginning 2017,” passes the completeness test because the dataset is authored by the New York State Energy Research and Development Authority (NYSERDA). This dataset can be found at the New York State’s database collection. Furthermore, the dataset owner is the NYSERDA. This means that the dataset is published by a government agency which gives the dataset viewer a high degree of confidence that the dataset is complete and not missing values since government agencies are thorough with their work. Additionally, the website publishes the list of variables and how the values for those variables are collected.

The dataset passes the consistency test. Although the dataset contain 36000+ observations, I randomly selected observations and verified that the data was consistent. All the values for the randomly selected observations was consistent across all variables and did not include any spelling discrepancies.

Based on random selection and verification, the dataset is also accurate. Again, I randomly selected many



	Annual Petroleum Reductions (gallons)
33	440.11
34	440.11
35	592.89
36	440.11
37	577.50
38	592.89
39	592.89
40	592.89
41	375.03
42	592.89
43	592.89
44	440.11
45	440.11
46	592.89
47	440.11
48	440.11
49	375.03
50	592.89
51	592.89
52	440.11
53	592.89

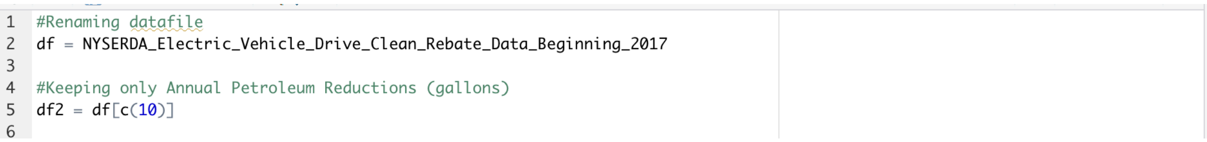
observations and looked for discrepancies. Although some observations included discrepancies, such as outliers, these values occurred more frequently than once. Thus, these values are most likely not inaccurate but represent outliers in the population.

The dataset passes the validity test because it does not contain any missing values. It is unlikely that the NYSERDA would not collect all the data for a processed observation. Thus, the dataset contains an unscrupulous amount of data.

The target population represents all the rebate applications submitted to the NYSERDA by car dealerships. Thus, each observation represents a rebate application submitted to the NYSERDA by car dealerships for eligible vehicles in the Drive Clean Rebate initiative. In other words, every observation represents the purchase of one eligible vehicle that participated in the Drive Clean Rebate program.

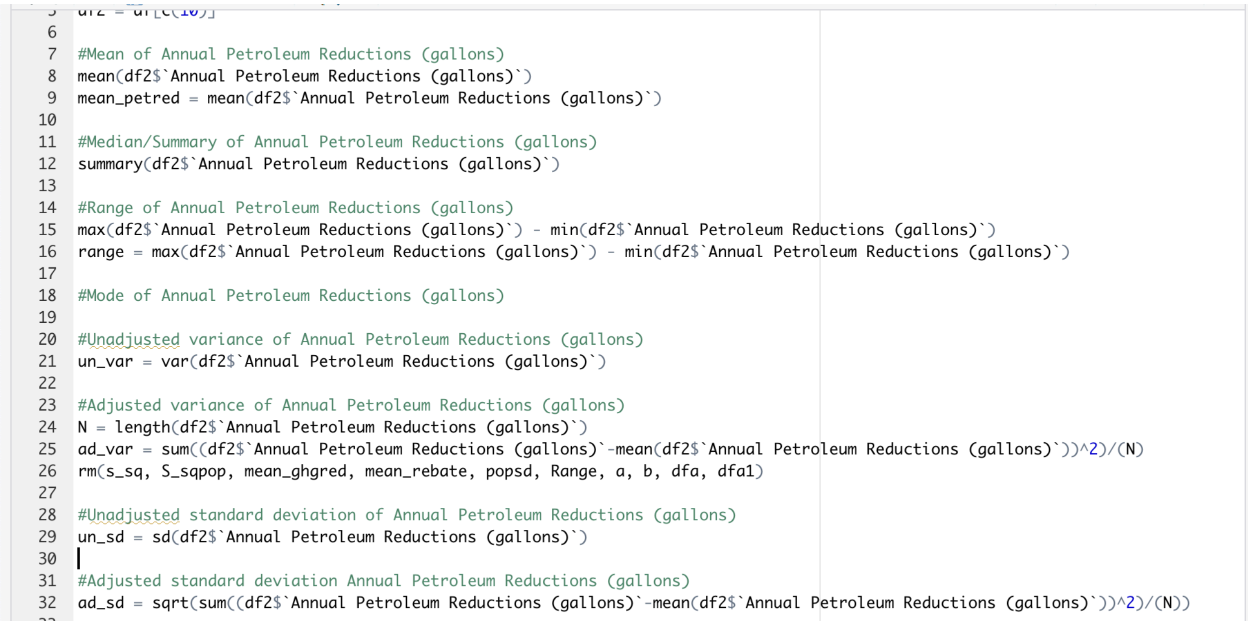
Exploratory Analysis:

The variable Annual Petroleum Reductions (gallons) was picked for analysis from the target population. The accompanying pictures illustrate the code used to isolate this variable and a section of the observations for this variable. This variable contains a varied range of values.

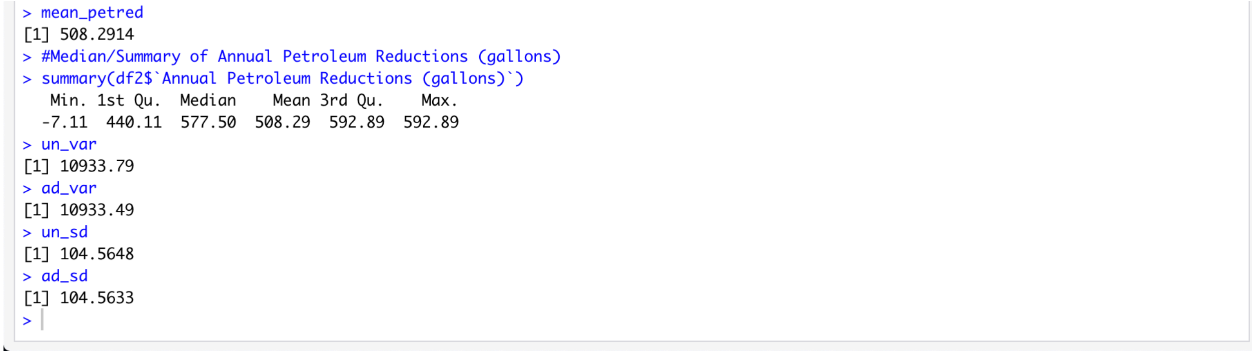


```
#Renaming datafile
df = NYSERDA_Electric_Vehicle_Drive_Clean_Rebate_Data_Beginning_2017

#Keeping only Annual Petroleum Reductions (gallons)
df2 = df[c(10)]

```

After running some analysis, the mean of this variable was found to be 508.2914 gallons. The median of this dataset was 577.50 gallons, and the range of this variable was 600 gallons. The unadjusted and adjusted variances of this variable were 10933.7876 and 10933.4913 respectively. The unadjusted and adjusted standard deviations were 104.5647 and 104.5633 respectively. The accompanying illustrations portray the code used to run these mathematical analyses and their outputs.

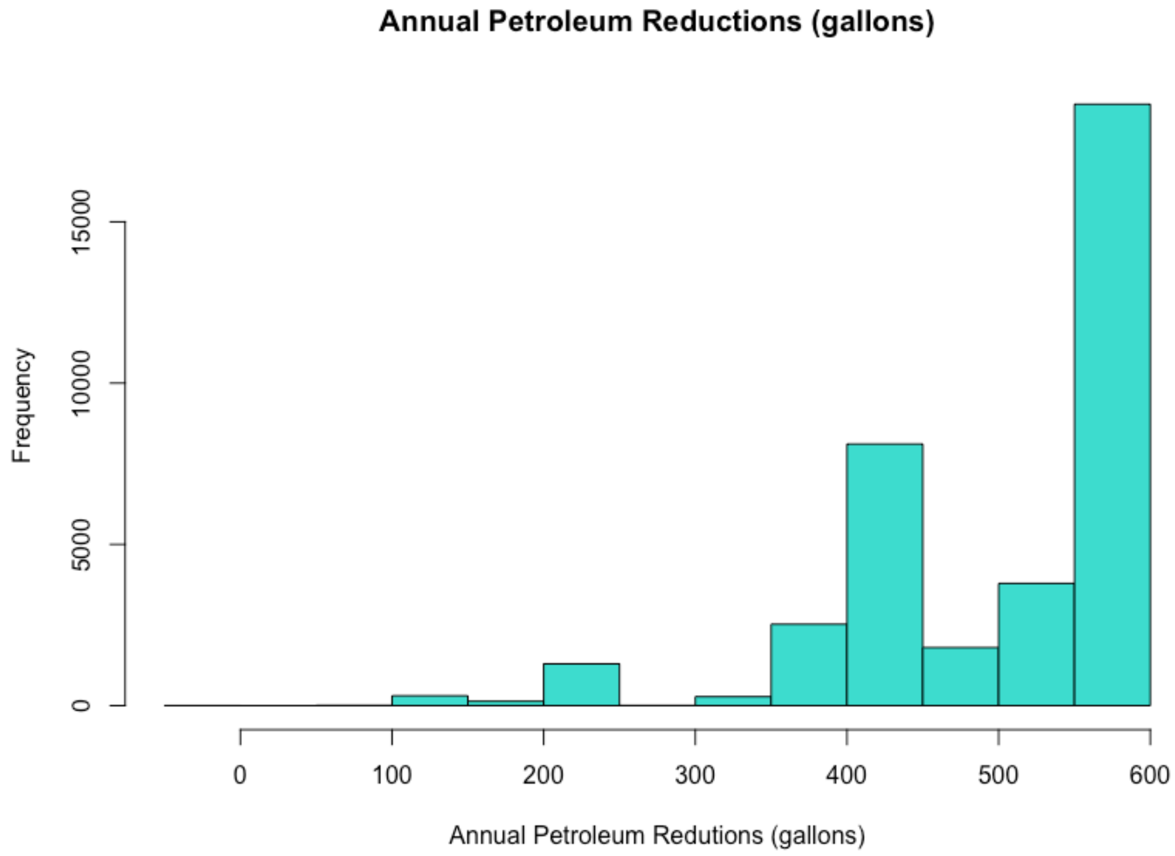


```
df2 = df[c(10)]

#Mean of Annual Petroleum Reductions (gallons)
mean(df2$`Annual Petroleum Reductions (gallons)`)
mean_petred = mean(df2$`Annual Petroleum Reductions (gallons)`)

#Median/Summary of Annual Petroleum Reductions (gallons)
summary(df2$`Annual Petroleum Reductions (gallons)`)

#Range of Annual Petroleum Reductions (gallons)
max(df2$`Annual Petroleum Reductions (gallons)`) - min(df2$`Annual Petroleum Reductions (gallons)`)
range = max(df2$`Annual Petroleum Reductions (gallons)`) - min(df2$`Annual Petroleum Reductions (gallons)`)

#Mode of Annual Petroleum Reductions (gallons)

#Unadjusted variance of Annual Petroleum Reductions (gallons)
un_var = var(df2$`Annual Petroleum Reductions (gallons)`)

#Adjusted variance of Annual Petroleum Reductions (gallons)
N = length(df2$`Annual Petroleum Reductions (gallons)`)
ad_var = sum((df2$`Annual Petroleum Reductions (gallons)`-mean(df2$`Annual Petroleum Reductions (gallons)`))^2)/(N)
rm(s_sq, S_sqpop, mean_ghgred, mean_rebate, popsd, Range, a, b, dfa, dfa1)

#Unadjusted standard deviation of Annual Petroleum Reductions (gallons)
un_sd = sd(df2$`Annual Petroleum Reductions (gallons)`)

#Adjusted standard deviation Annual Petroleum Reductions (gallons)
ad_sd = sqrt(sum((df2$`Annual Petroleum Reductions (gallons)`-mean(df2$`Annual Petroleum Reductions (gallons)`))^2)/(N))
```



```
> mean_petred
[1] 508.2914
> #Median/Summary of Annual Petroleum Reductions (gallons)
> summary(df2$`Annual Petroleum Reductions (gallons)`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -7.11  440.11  577.50  508.29  592.89  592.89 
> un_var
[1] 10933.79
> ad_var
[1] 10933.49
> un_sd
[1] 104.5648
> ad_sd
[1] 104.5633
> 
```

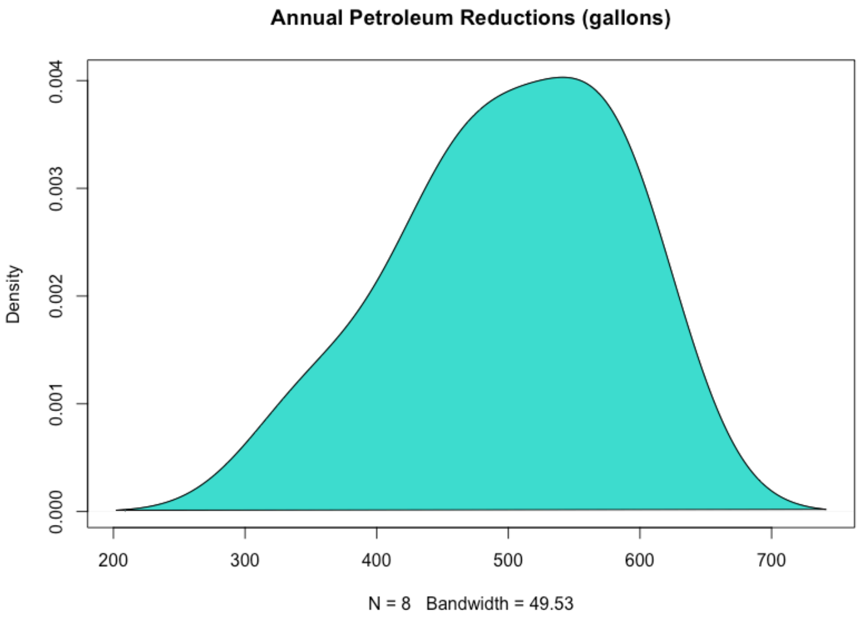
Portraying the distribution of the variable, Annual Petroleum Reduction (gallons), the histogram shows the frequency of each the observations with a bin equal to 50. The x-axis contains the value of the observations grouped according to their respective bins, while the y-axis shows the frequencies of the bins. A perfunctory glance at the histogram reveals that the data distribution is strongly skewed to the left. Thus, the median is the best indicator of central tendency of this dataset. The accompanying graph reveals this distribution.



The reduced target population of this variable was composed by taking N=8 observations. These observations were chosen from the summary of the dataset, such as the first quartile, median, third quartile and maximum. Interestingly, the minimum value of -7.11 was not used in the reduced target population because this observation represented an outlier. Instead, the minimum was replaced by identifying all of the unique values of the variable and then picking an observation from the 350-400 range because the histogram proves that all values below that bin represent outliers which would greatly skew the mathematical analyses in the reduced target population. The other 3 observations were randomly chosen from the highest frequency bins.

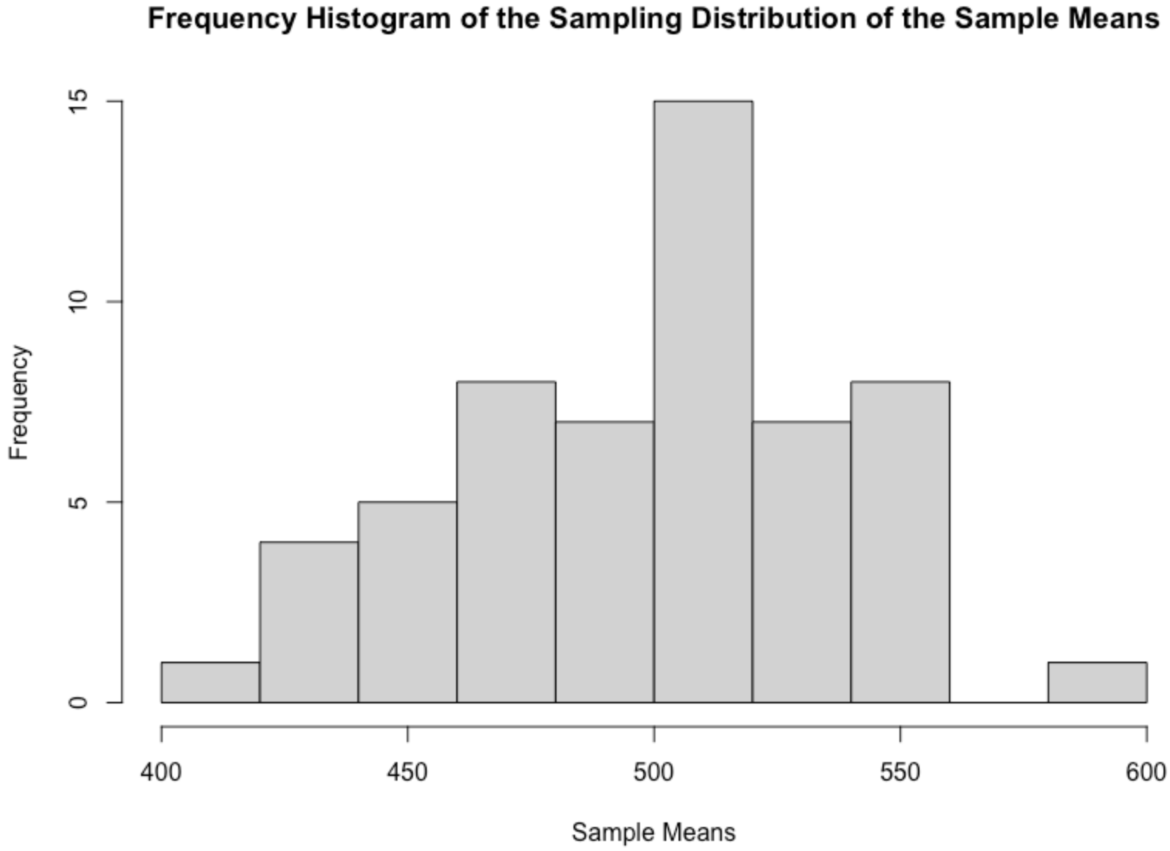
The mathematical analyses on the reduced target population occurred in much of the same way as it did on the target population. Namely, the mean of the reduced target population was 500.03125. The unadjusted and adjusted variances of the reduced target population were 6959 and 6089 respectively. The unadjusted and adjusted standard deviations of the reduced target population were 83.4 and 78.0 respectively.

The reduced target population's distribution is depicted using a density plot. The shift from histogram to density plot took place because unlike the target population which contains 36000+ observations, the reduced target population contains only 8 observations. Thus, it made sense to portray the reduced target population using a

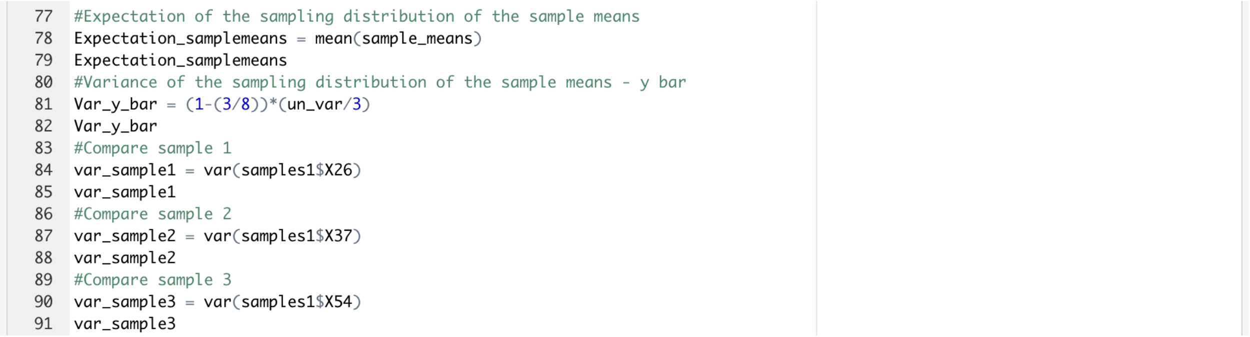


density plot. Just as the histogram of the target population portrayed a left skew, the density plots similarly portrays a left skew. The following illustration presents the distribution of the reduced target population.

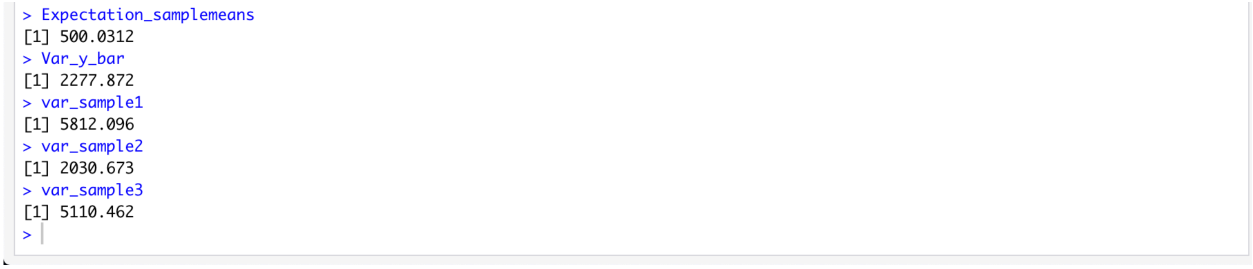
Analysis on R produced the sampling distribution with n=3 from the reduced target population. Further analyses revealed the sampling distribution of the sample means for all 56 samples. This output was plotted on a histogram to present the distribution of the sample means



R analyses were used to calculate the expected value and variance of the sampling distribution of the sample means. The expectation of the sampling distribution of the sample means was 500 and the variance was 2277. Further analyses of three random samples from n=3 revealed that the variances for each samples were 5812, 2030 and 5110 respectively. The accompanying illustrations depict the code and output of the code.



```
#Expectation of the sampling distribution of the sample means
Expectation_samplemeans = mean(sample_means)
Expectation_samplemeans
#Variance of the sampling distribution of the sample means - y bar
Var_y_bar = (1-(3/8))*(un_var/3)
Var_y_bar
#Compare sample 1
var_sample1 = var(samples1$X26)
var_sample1
#Compare sample 2
var_sample2 = var(samples1$X37)
var_sample2
#Compare sample 3
var_sample3 = var(samples1$X54)
var_sample3
```



```
> Expectation_samplemeans
[1] 500.0312
> Var_y_bar
[1] 2277.872
> var_sample1
[1] 5812.096
> var_sample2
[1] 2030.673
> var_sample3
[1] 5110.462
> 
```

Results and Conclusions

As seen in the first histogram, the distribution of the target population is heavily skewed to the left. This occurs because the most frequent observations lie on the 550 to 600 bin. However, a significant amount of observations lie on 100 to 300 range. Still the majority of observations are above 350. This means that the distribution is skewed highly to the left. Surprisingly, an insignificant number of observations lie on the 250-300 bin. Overall, the distribution is left skewed with many outliers at the lower end of the measurement and majority of the observations above 300. Although the outliers would significantly impact the mean of this dataset, the dataset remains so larger, above 36000 observations, that the mean is not very much affected. This can be seen when comparing the mean to the target population's mean, which does not contain any of the outliers in question.

The target population still carries a left skew although it is much less pronounced. The reduced target population contains observations greater than 350, which is the cutoff below which there are only outliers in the dataset. If the minimum value of the dataset was included into the reduced target population, then the distribution of observations of the reduced target population would resemble the high left skew of the target population. By not including the minimum value, which is also an outlier, the distribution of the reduced target population resembles more of a normal bell curve. In the reduced target population, majority of the observations come from the above 400 range because the target population's mean, median, mode, and quartiles come from the above 400 range.

The histogram portraying the sampling distribution of the sample means portrays a much more normal distribution. This histogram resembles.a bell curve with a maximum in the middle and then flattening of both sides. Surprisingly, there is a gap in sample means from 550-600 but aside from that the sampling distribution of the sample means resembles more of a bell curve then either of the two previous distributions. Had there not been any missing values between 550-600, the sampling distribution of the sample means would have effectively removed the left skew of the target and reduced target populations. Still, this distribution reduces the left skew substantially from the previous two distributions.

After calculating the variance of a single sample from the sampling distribution, it became clear that a single sample remains a good estimator for the variance of the sampling distribution. R analyses revealed that the variance of the sampling distribution was 2277. After analyzing the variances of a few samples from the sampling distribution, it became apparent that single samples were effective in calculating the variance of the sampling distribution. For example, the variance for the second selected sample was 2030. This is a very effective estimator for the variance of the sample distribution. Although some of the samples overestimated the variance of the sampling distribution, each sample remains a good estimator for the sampling distribution. This reveals that my sampling distribution is not very biased.