CS544, Fundamentals of Analytics Final Project

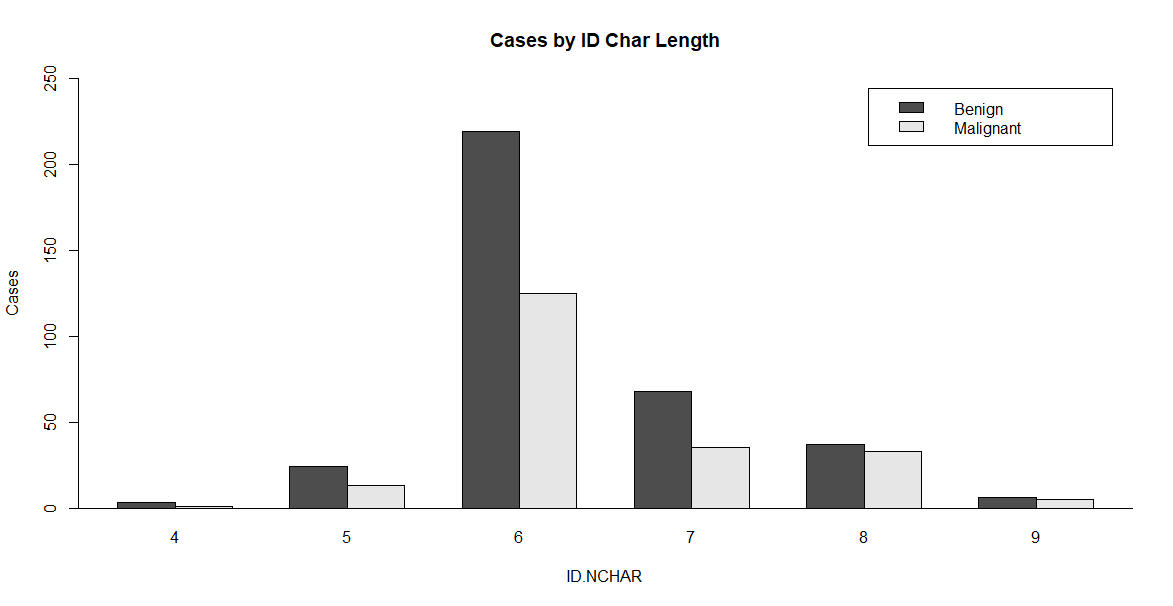
Jefferson Parker

[japarker@bu.edu](mailto:japarker@bu.edu)

For the final project I chose to analyze the Breast Cancer Wisconsin data set available from kaggle.com (<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data>). From the site, the data are derived from digitized images of fine needle aspirate breast tissue samples. Image features are computationally determined and mapped to diagnostic information of the sample as being benign or malignant. The data is comprised of 569 sample observations. The data was downloaded as a .csv file and loaded into R using the read.csv() function. The only data preprocessing step involved removing an empty value column from the data set.

**ID Variable Hypothesis**

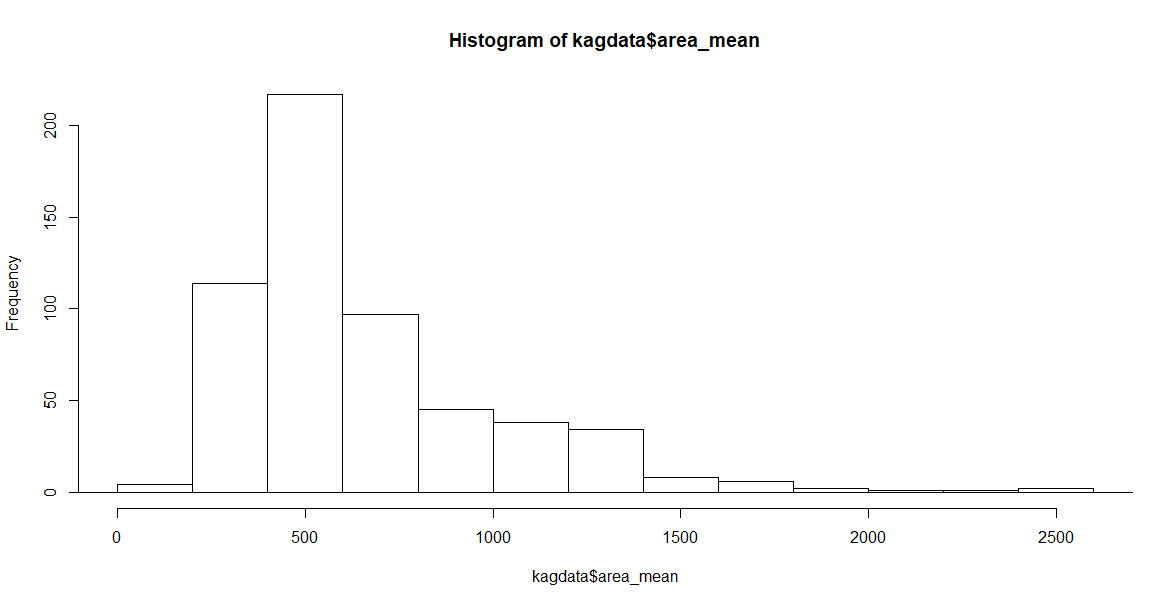
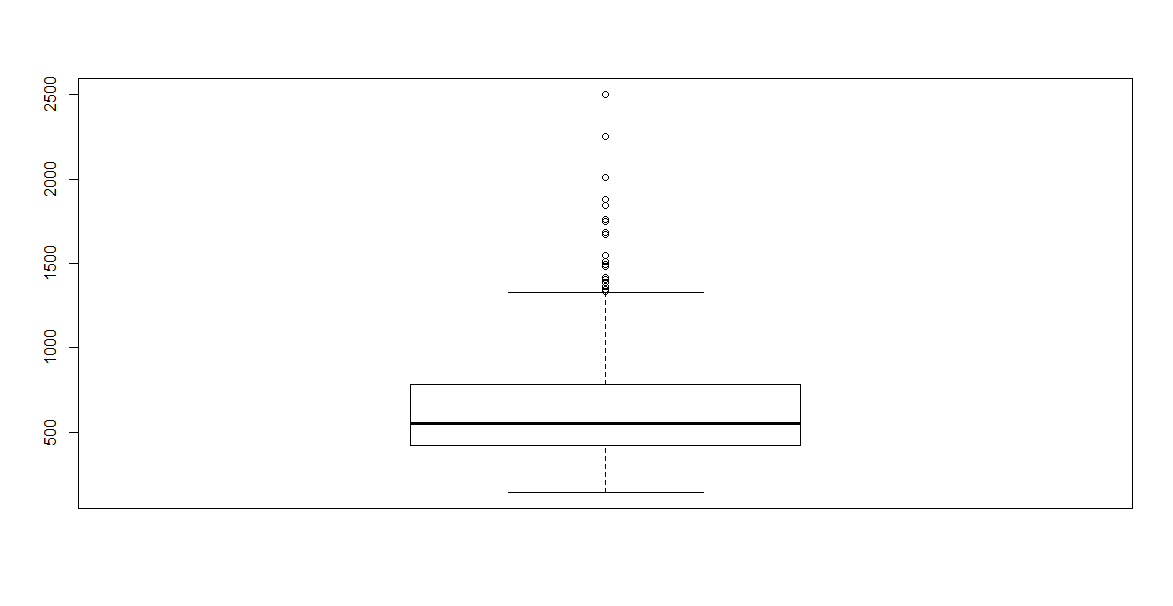
Upon examining the data, my first observation was that the ID variable (sample identifier) had an unusual distribution of values, ranging from the thousands to the millions. One hypothesis was that the different ranges of values were important, possibly representing different times of sample collection (larger numbers being later) or different sample collection sites. In order to investigate further, I created a derived categorical variable, ID.NChar, representing the length of the ID variable.



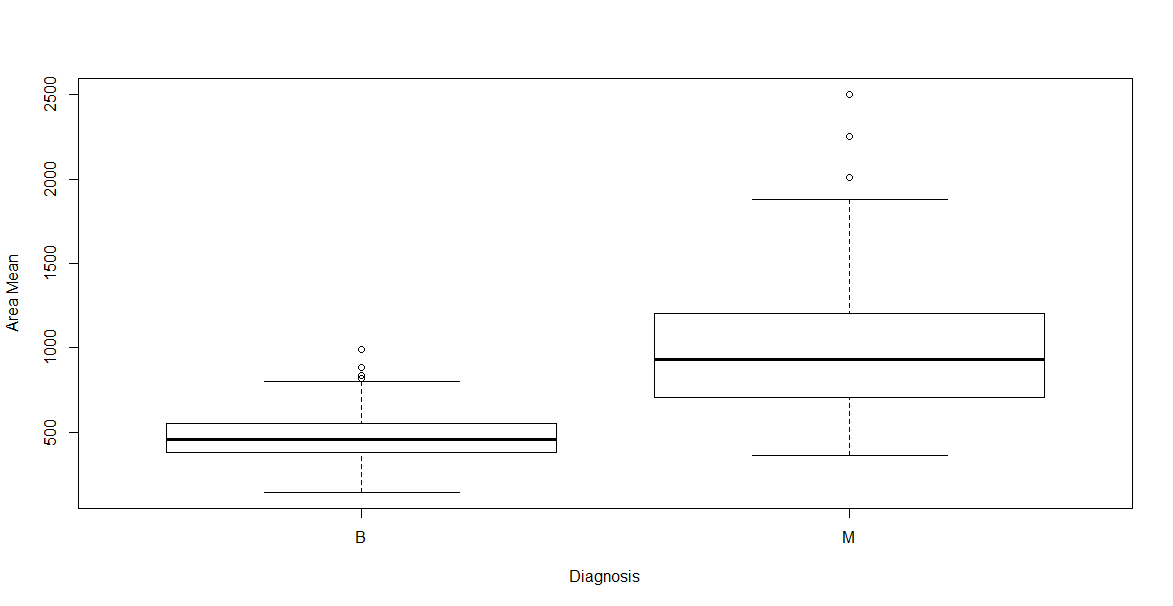
A bar plot of the sample diagnoses per ID.NChar value shows a range of samples being taken at different values of ID.NChar, consistent with the length of the sample ID being meaningful but not revealing the nature of the source.

**Area Mean**

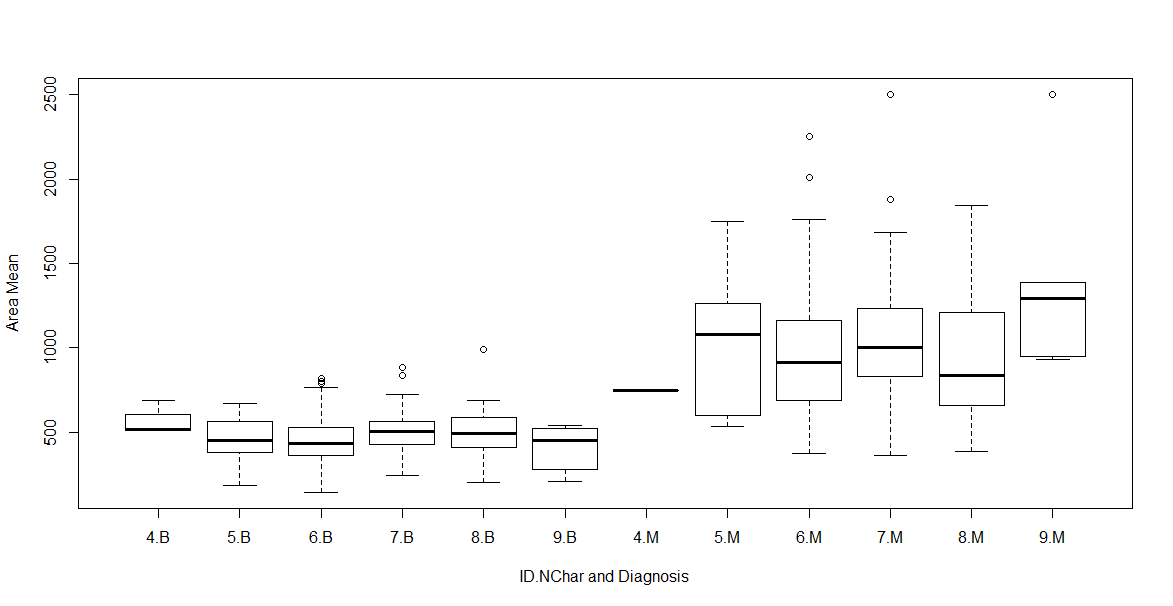
To begin the quantitative analysis of the data, I explored the output of the summary() function (data not shown) to identify any variables with extreme values or other indications that they may be interesting. One such variable was ‘area\_mean’, defined as the smoothness\_mean of the image data. I generated a histogram and a boxplot of this variable to learn more about its distribution:

Clearly there is a large range of values for area\_mean, including several outlier values. Given that this is medical data with a diagnostic parameter, the next question was whether the area mean was influenced by diagnosis (benign or malignant). Sure enough, there was a difference in the quantitative measure by diagnosis.



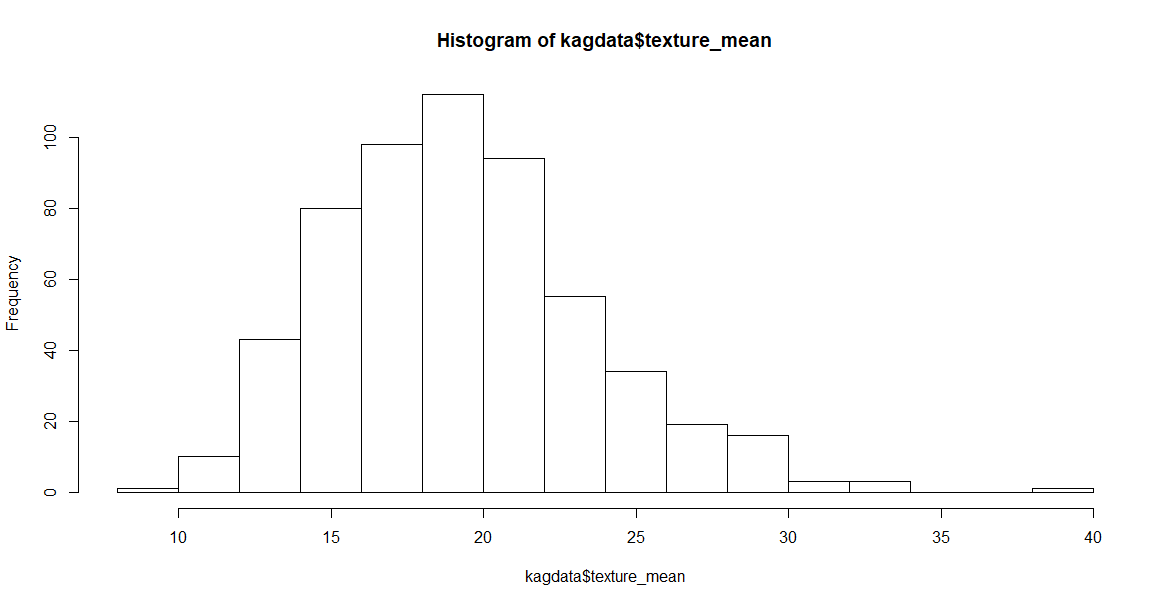
The area mean for benign (B) samples was much lower than for the malignant (M) samples with means of approximately 500 and 1000 respectively. Given the difference in the area\_mean variable by diagnosis, I further explored the possibility of any interaction between area\_mean, diagnosis and our mystery ID.NChar variable.



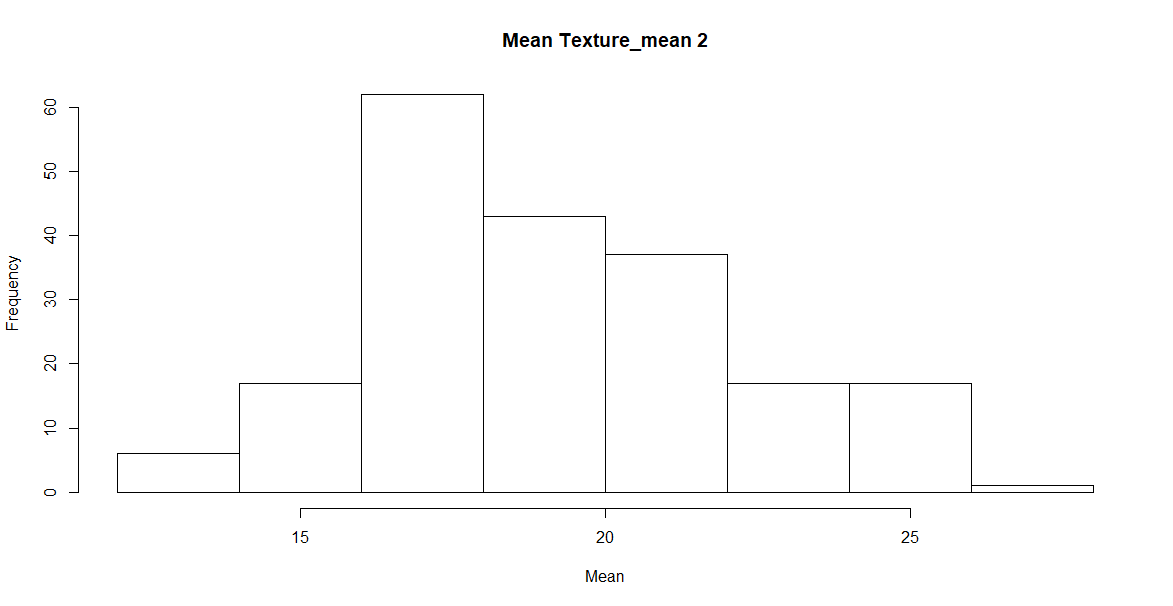
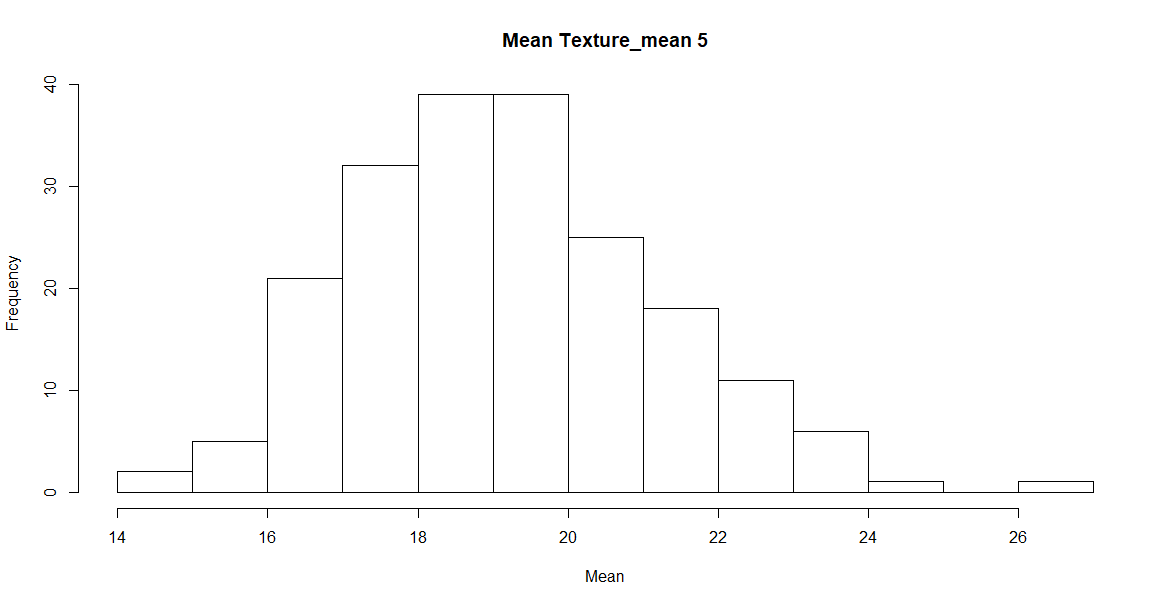
While there is no clear trend, the Benign samples have approximately the same range of values across all instances of ID.NChar, but the malignant values have clear differences in both inter-quartile range and median from one to the next. Still, the meaning of the ID.NChar variable is not clear.

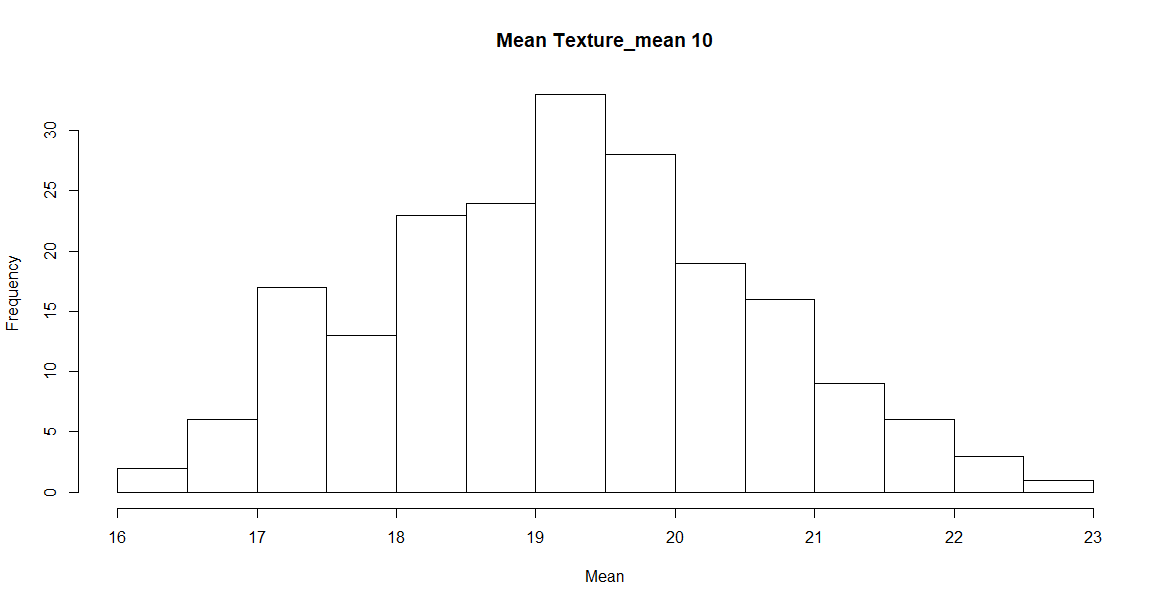
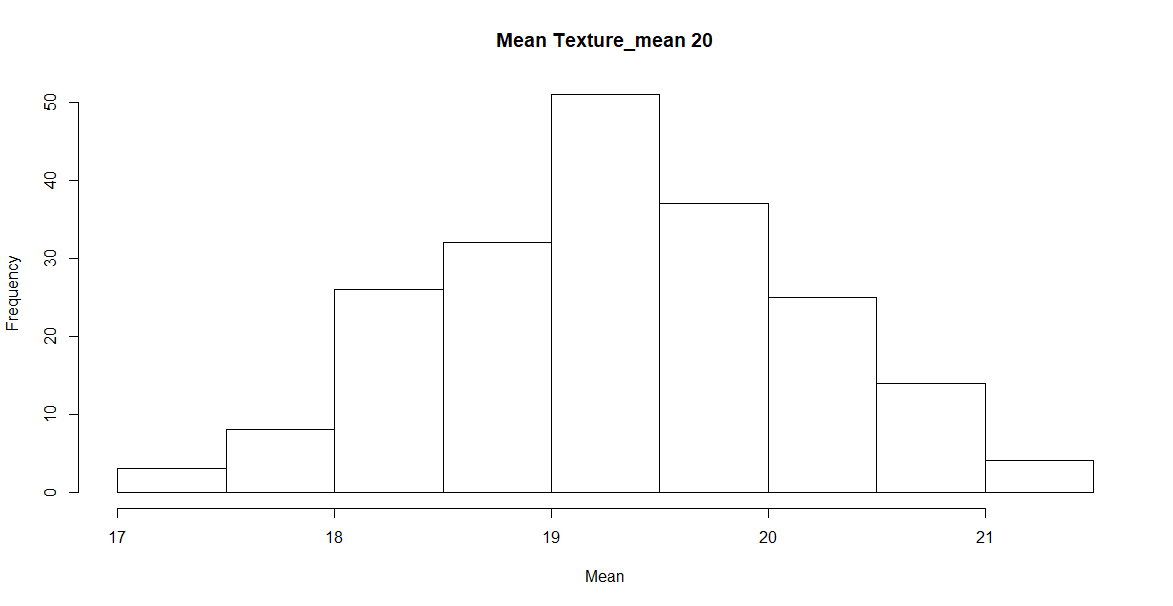
**Data Distribution and the Central Limit Theorem**

Next, I turned to exploration of the ‘texture\_mean’ variable, which measures the standard deviation of gray-scale values. As shown, the texture mean data is somewhat left skewed with a mean of 19.3 and Standard Deviation of 4.3.



Using a uniform distribution, 200 samples of size 2, 5, 10 or 20 were drawn from the texture\_mean data. The density of the sample means was plotted for each:

The mean and standard deviation values for each sampling show that the sample mean density all have about the same mean value while the standard deviations have an inversely proportional relationship with the sample size, becoming smaller with increased sample size.

|  |  |  |
| --- | --- | --- |
| **Size** | **Mean** | **StdDev** |
| All | 19.3 | 4.30 |
| 2 | 19.1 | 3.00 |
| 5 | 19.2 | 1.97 |
| 10 | 19.2 | 1.32 |
| 20 | 19.3 | 0.84 |

**Data Sampling**

One last exploration was of the effect of different sampling methods on the representation of ID.NChar groups. I looked briefly at Simple Random Sampling Without Replacement and Systematic Sampling. As shown in the table below, the representation frequency of the different ID.NChar values is shown in the top row (ALL).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID.Nchar** | **4** | **5** | **6** | **7** | **8** | **9** |
| **All** | 0.01 | 0.07 | 0.60 | 0.18 | 0.12 | 0.02 |
| **SRSWOR** |  | 0.10 | 0.75 | 0.15 |  |  |
| **Systematic** |  | 0.1 | 0.7 | 0.1 | 0.05 | 0.05 |

The sampling frequency using SRSWOR resulted in all of the three smallest ID.NChar groups being lost from the sample population, while Systematic Sampling resulted in only the smallest ID.NChar group being lost completely. Keep in mind that ID.NChar group = 4 only had four samples total.

**Conclusion**

While our not quite rigorous analysis would suggest that the length of the sample ID variable, captured as ID.NChar, likely has some meaningful explanation it is still unclear just what it means. However, we have been able to demonstrate that the data can be plotted and analyzed in several useful ways and the Central Limit Theorem still holds true.