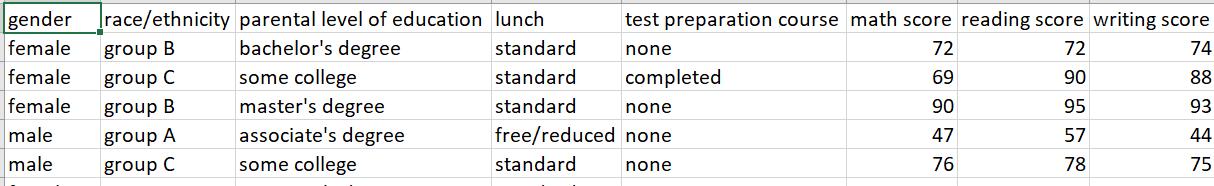
**Aim:**

To apply a linear regression model for predicting scores of students, based on various factors, such as parental level of education, whether the student has completed the test preparation course. We wish to analyze several features and study correlation between some factors and test scores.   
After building the regression model, we assess it using the adjusted R-squared, which is a measure of how much variability in the target variable can be explained by the explanatory variables. We wish to perform this assessment on models run on samples(created using various sampling techniques). During the process, we also verify the applicability of Central Limit Theorem.   
  
**Dataset:**

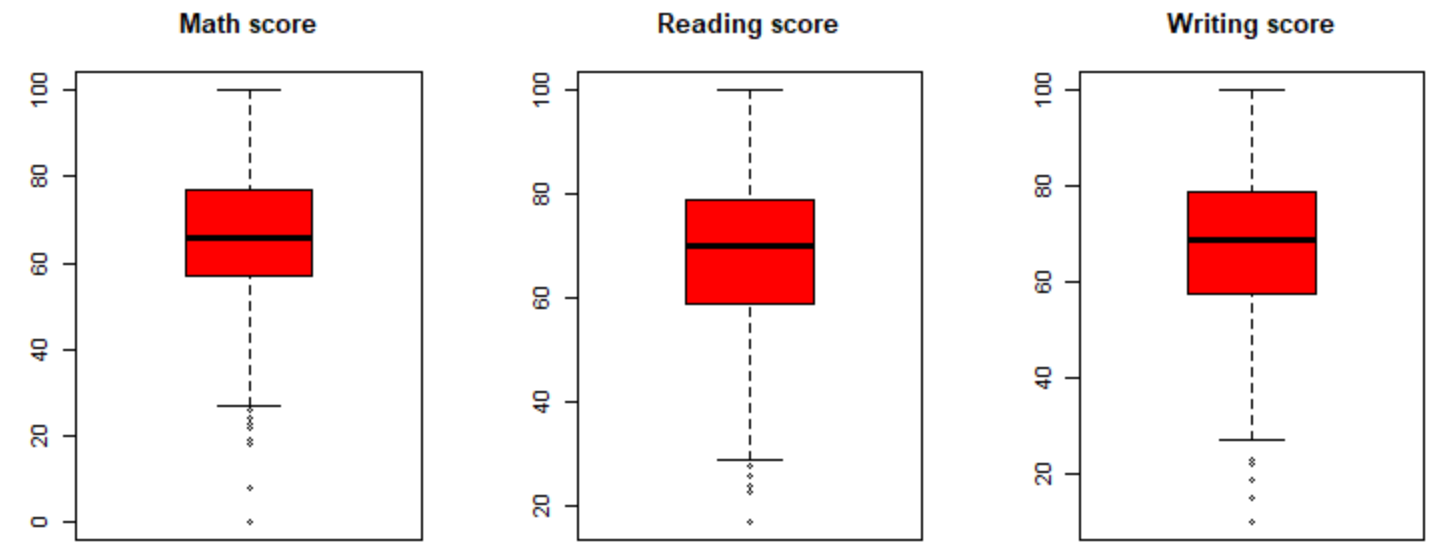
The dataset is 1000 rows long and there are 8 columns. The dataset is taken from https://www.kaggle.com/adithyabshetty100/student-performance

Here is a glimpse of the csv file.



**Preparing the data**

We explore the dataset and find that there are no NA values. Since, we wish to explore what features are important, we choose to keep all the columns.  
  
We explore the boxplots of math score, reading score and writing score, and find that there are some outliers:

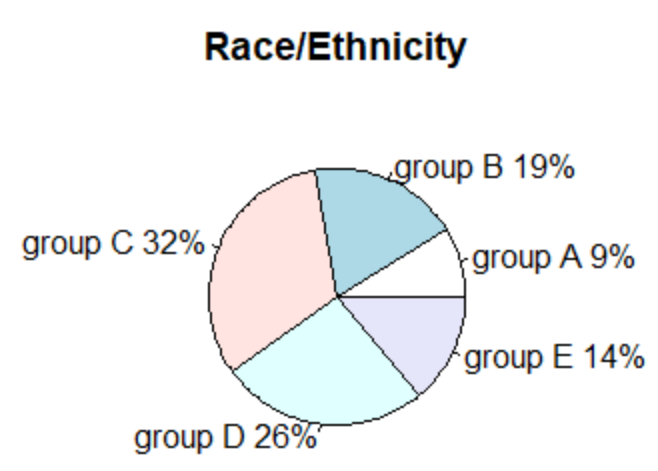
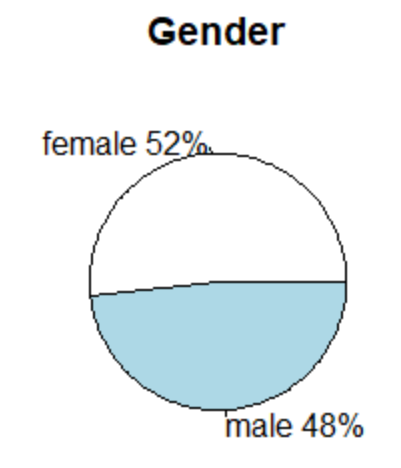


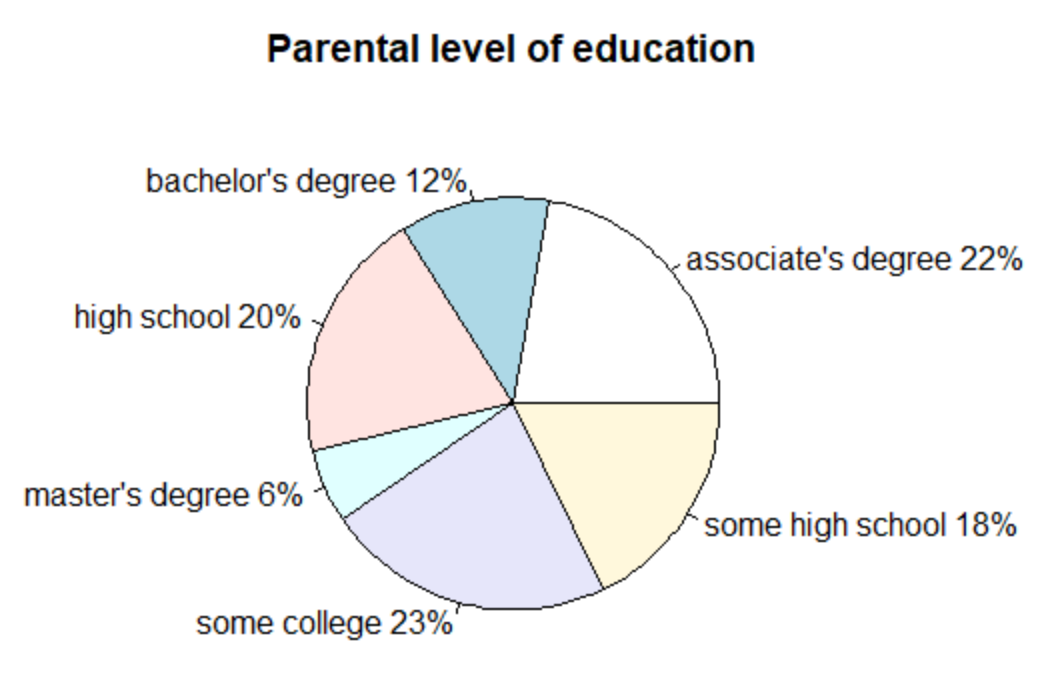
The observations (rows) corresponding to outliers were removed.  
  
The cleaned data set is 986 rows long.

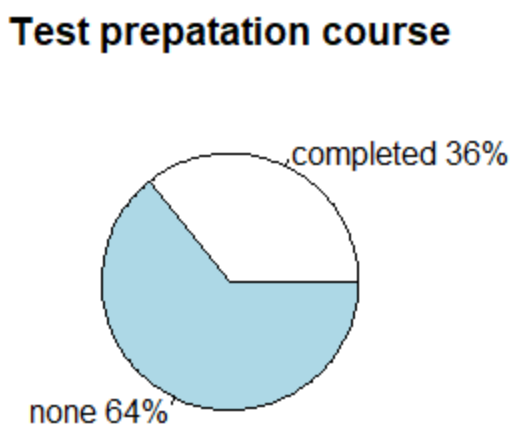
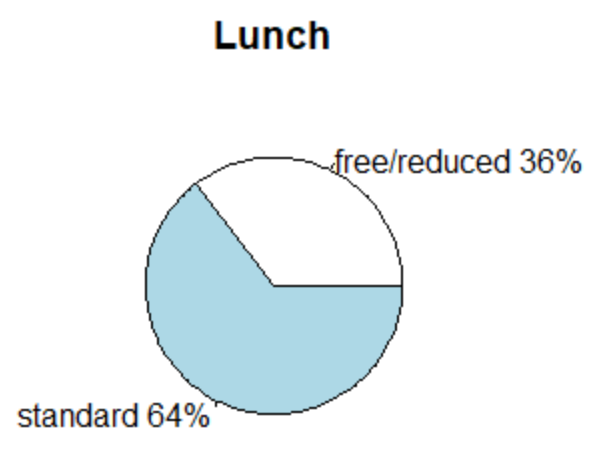
**Analyzing the data**

**Analysis for categorical variables:**

A function was defined to plot a pie chart for a given column, showing the percentages of various levels. Such pie charts were prepared for all categorical variables, to get an understanding of how many observations fall in a particular category:

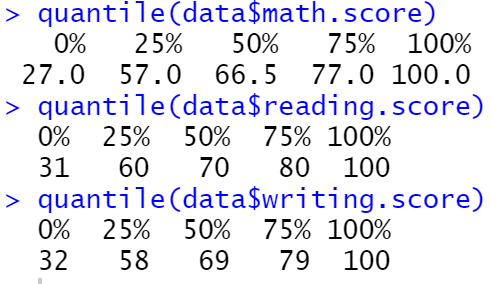


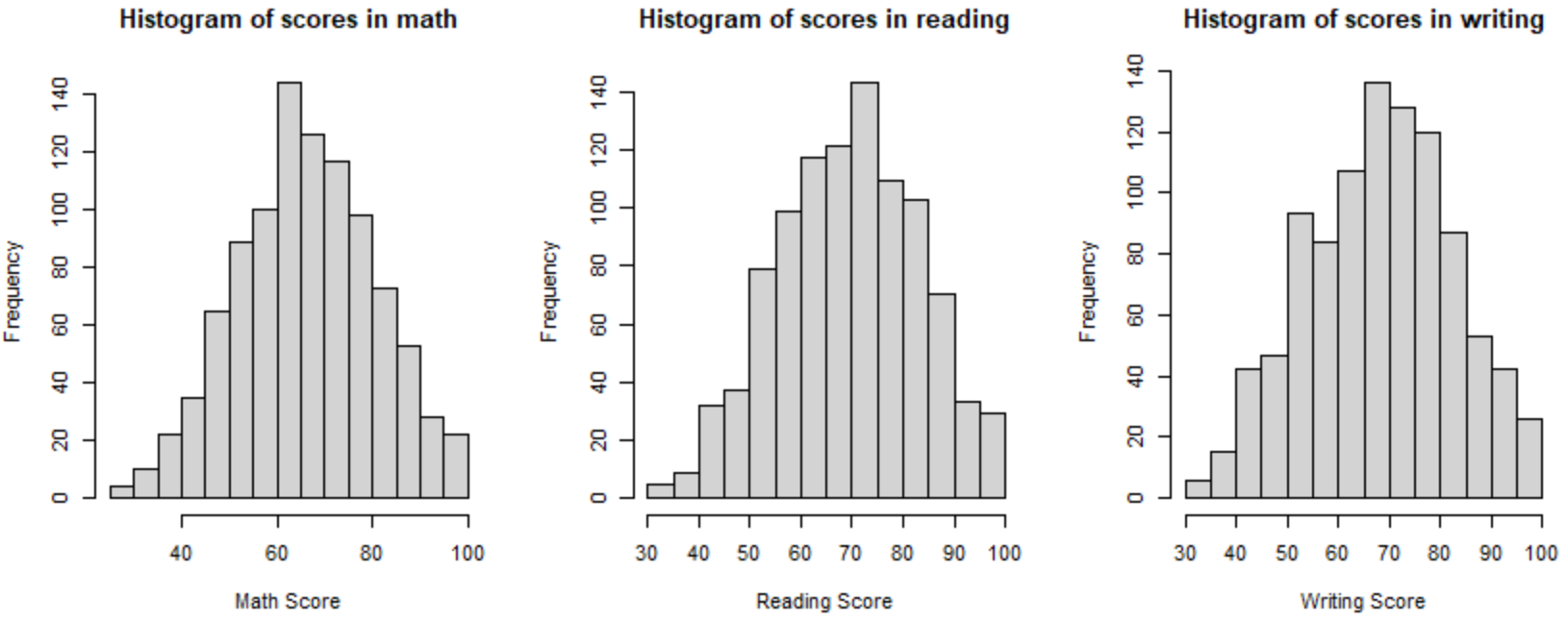




**Analysis for numerical variables:**

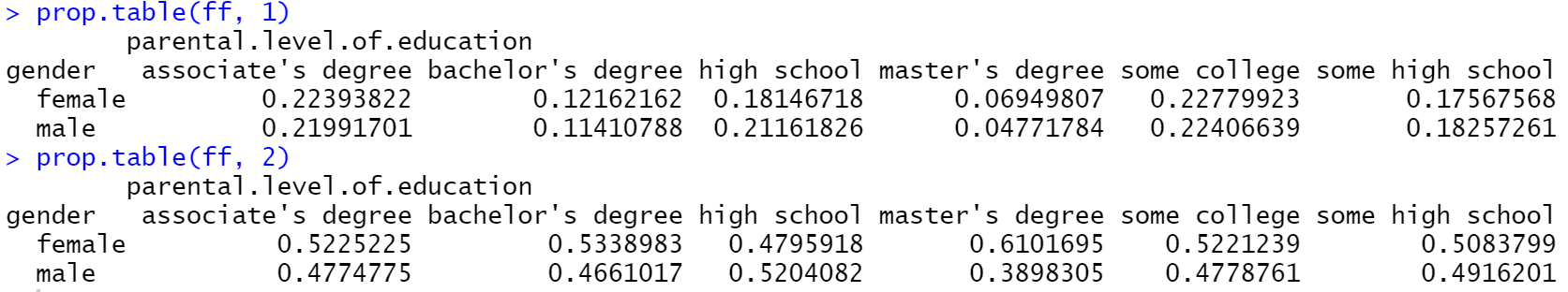
Quantiles of the three numerical variables were calculated and their histograms were created to get a better idea about the distribution of values:





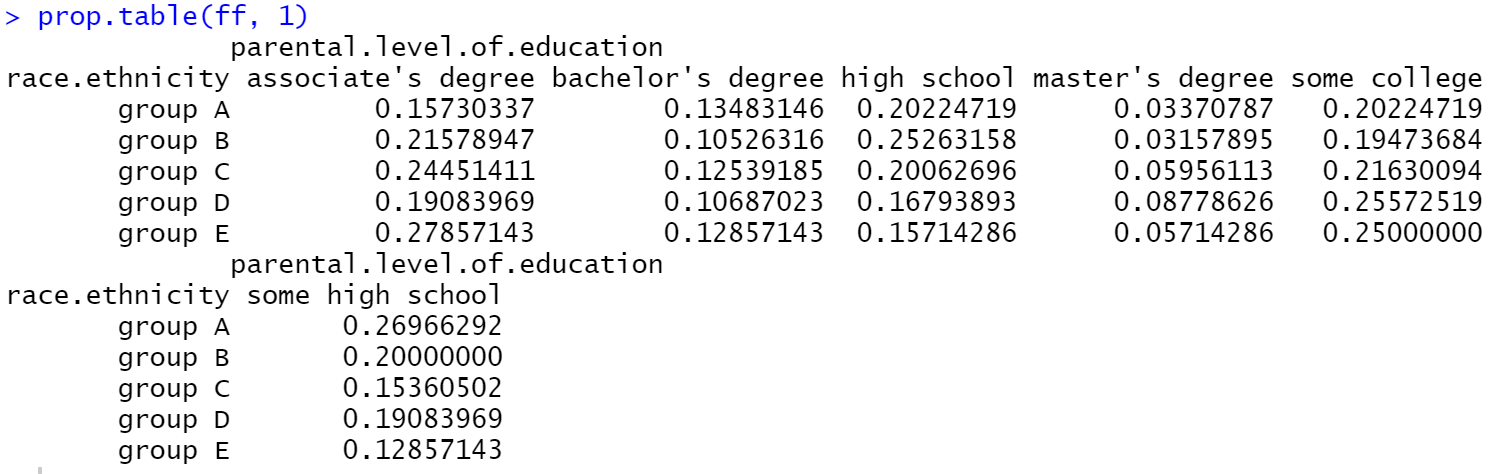
**Analysis for set of two variables:**

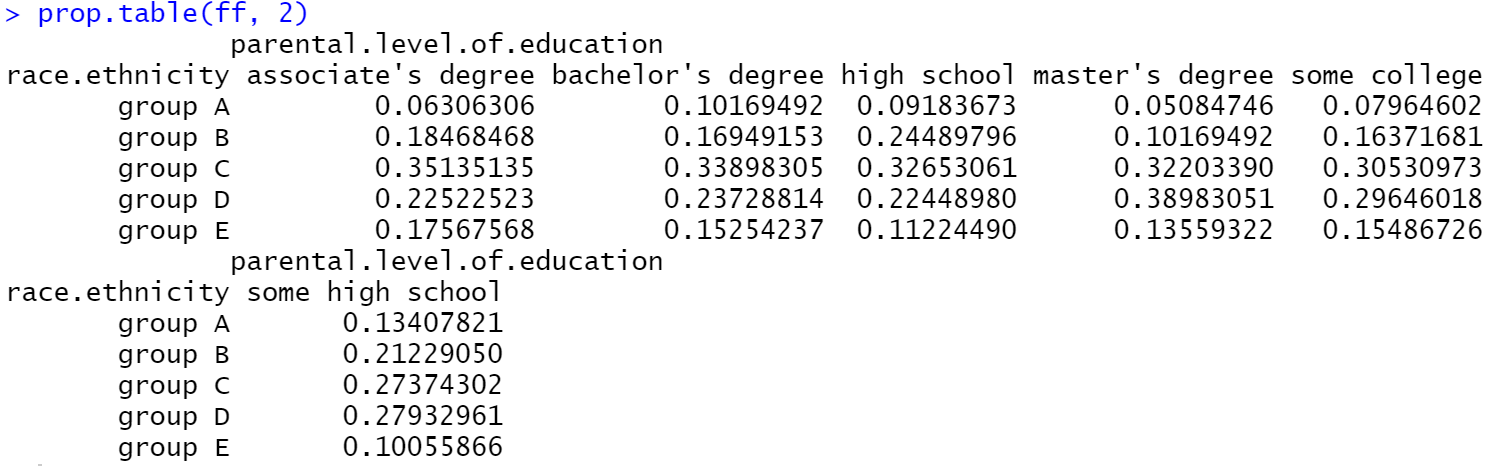
Then, for variables Gender and Parental level of Education, two way tables were created and their proportions displayed:



As we can see, the proportions are evenly distributed among the variables

Above was repeated for Race/Ethnicity and Parental level of Education:

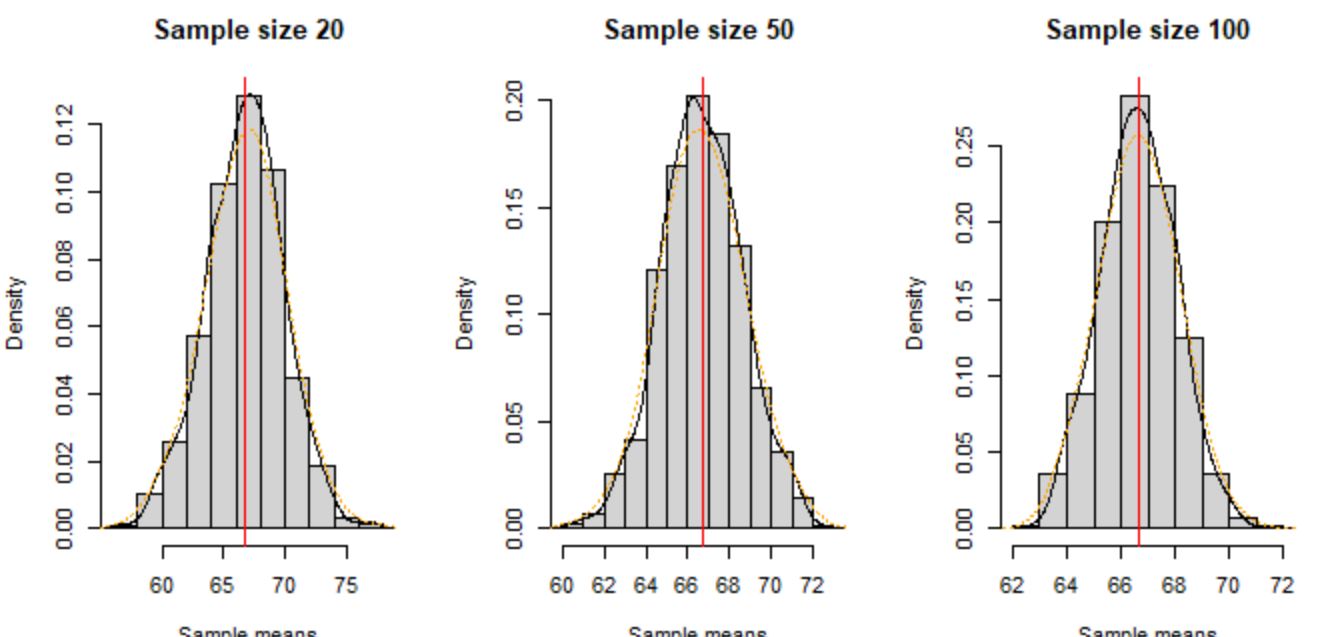


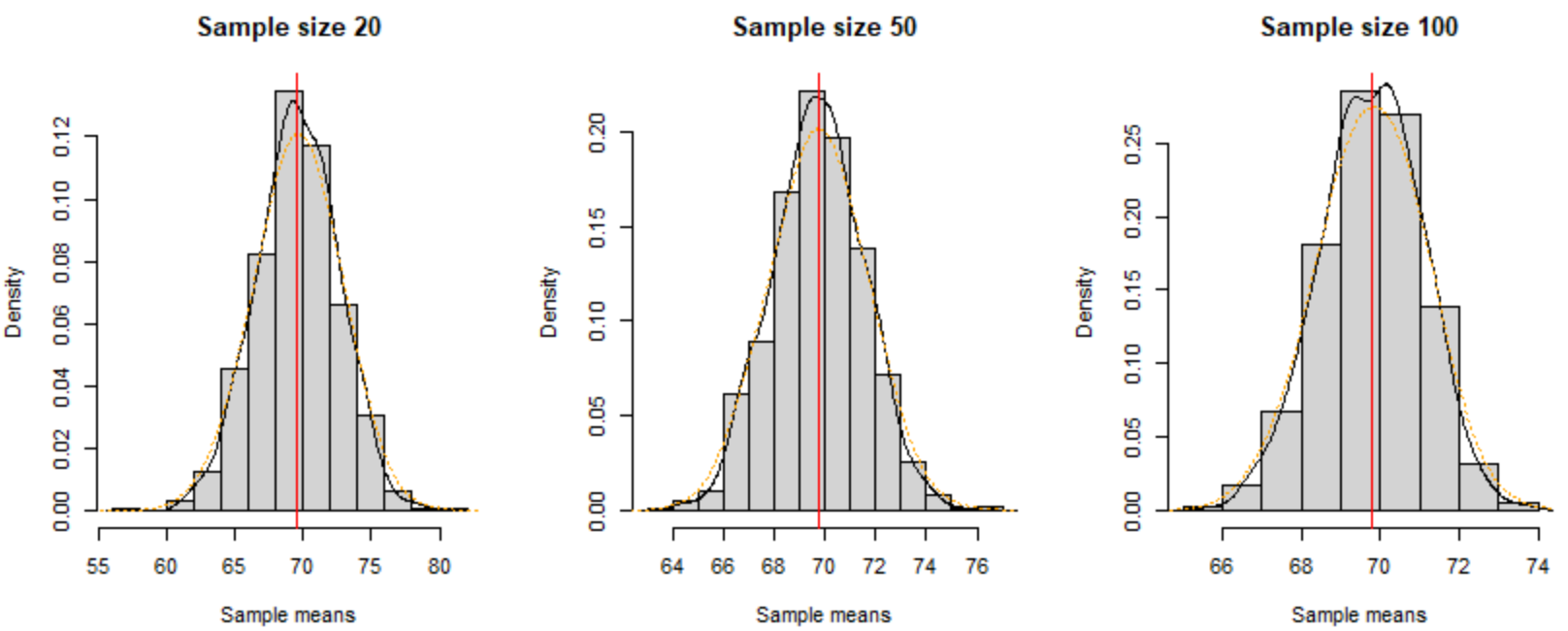


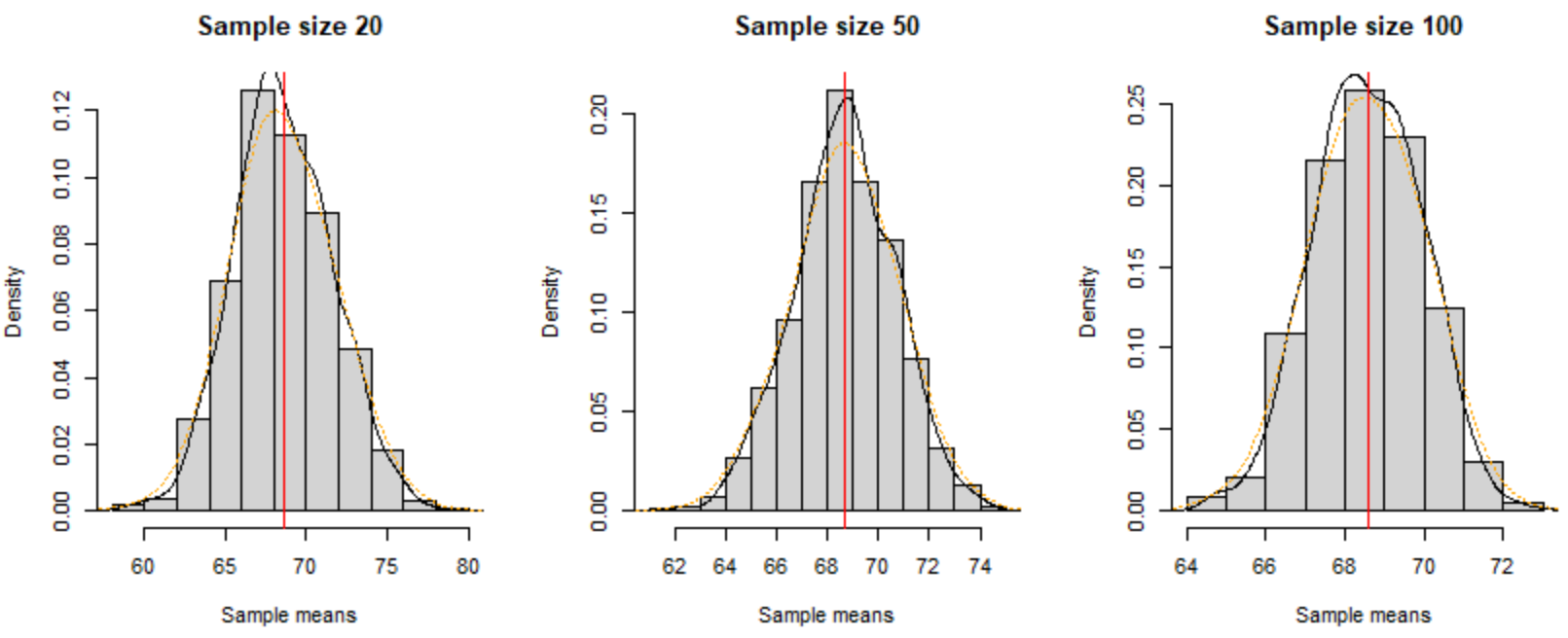
Again, the proportions are fairly evenly distributed.

**Applicability of Central Limit Theorem:**

Applicability of Central Limit theorem is verified by drawing 1000 samples of sizes 20, 50 and 100 from math.score, reading.score, writing.score. The means of these 1000 samples were plotted on a histogram, to verify whether the distribution is tending towards a normal distribution(as size of samples increases)



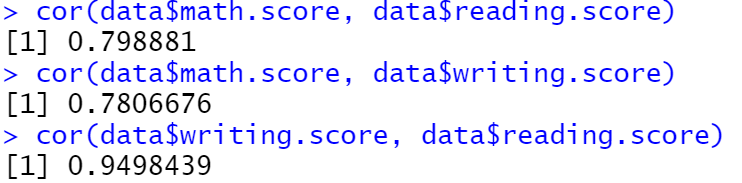




**Linear Regression:**

Multivariable linear regression was applied taking all the categorical variables as features. adjusted R2 came out to be 0.22 suggesting that model is not a good one.

Correlation between the three scores was studied and it was found that correlations are high:



So, a multi variable linear regression model was built, predicting 1 score, based on the other 2 scores. This was repeated for **samples of size 50** created using **simple random sampling, systematic sampling and stratified sampling**(two strata – Parental level of education and test preparation course). The adjusted R2 for all these models are summarized in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target variable | Whole dataset | Simple random sampling | Systematic sampling | Stratified sampling |
| Math score | 0.6423 | 0.5269 | 0.7442 | 0.5834 |
| Reading score | 0.9104 | 0.8773 | 0.9324 | 0.9391 |
| Writing score | 0.9033 | 0.8661 | 0.9175 | 0.9313 |

**Observations:**

for simple random sampling, worse models(as assessed by adjusted R2) were observed. However, this was not the case for Systematic and stratified sampling, since the model accuracy matched with the model created using the whole dataset.