EDUC 641 Assignment 03 Key

1. Dataset (20% point)

1.1. Open your RStudio, create a project and save it. Go to the root directory of the project and create a folder named “data”. Download the [cont.csv dataset](file:///C:\Users\clair\Documents\0_projects\educ-quant\641\assignments\data\cont.csv) and store it in the folder “data”. Create an Rmd/R file in the project root directory. Then read the data into your R environment. Make sure to check whether the variable treat is factor as well as whether coursework and vocabulary are double. Transform them into correct types if needed.

1.2. Write your own code to view the dataset and write 3-4 sentences about the structure of the data (how many variables are there, what type is each variable, how many rows/observations, etc.).

* There are four variables in the dataset: tchid, treat, coursework, and vocabulary
* Currently, all variables are double, but tchid and treat should be factor
* There are 126 teacher-level observations

#### 2. Descriptive statistics of the outcome variable (40% point)

2.1. Central tendency

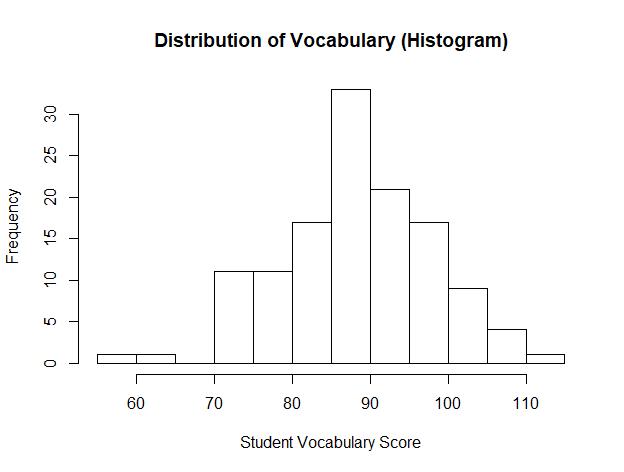
2.1.1 What are the mean and median of the outcome variable, vocabulary? Interpret the statistical meanings of these measures.

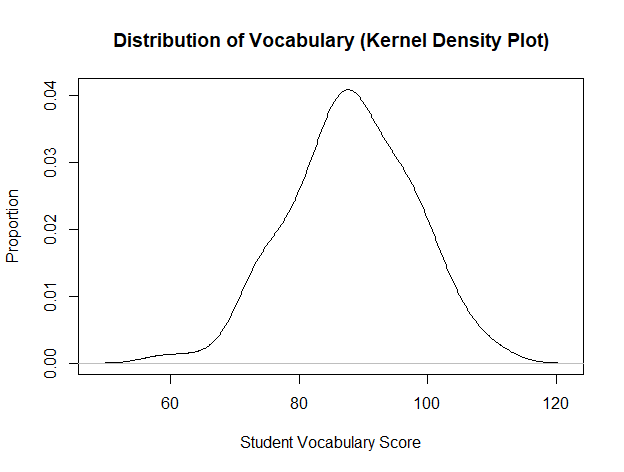
* The mean of vocabulary is 88.12, which means that the average student score for the sampled teachers is 88.12
* The median of vocabulary is 88.35, which means that half of the teachers have an average student score above or below 88.35

2.1.2. Create a binned frequency table (suggested bin width = 5) for vocabulary. Then create a plot to show the distribution of vocabulary and make sure to label the x and y axis. Is it normally distributed? If not, describe how and how much the distribution is skewed. Did you notice any outliers?

|  |  |
| --- | --- |
| (58.5,63.5] | 0 |
| (63.5,68.5] | 1 |
| (68.5,73.5] | 6 |
| (73.5,78.5] | 12 |
| (78.5,83.5] | 17 |
| (83.5,88.5] | 28 |
| (88.5,93.5] | 24 |
| (93.5,98.5] | 20 |
| (98.5,104] | 11 |
| (104,108] | 4 |

* Plot options:





* The distribution of vocabulary is approximately normal. No outliers are observed.

2.2. Variability

3.2.1. What are the variance and standard deviation of vocabulary? Interpret the statistical meanings of these measures.

* The variance of vocabulary is 92.78, the standard deviation is 9.63.
* The average squared deviation of each score from the mean (88.12) is 92.78 and its square root is 9.63.

2.2.2. List the quartiles, percentiles, interquartile range, and range of vocabulary.

* The quartiles of vocabulary are 81.51, 88.35, 94.74
* The percentiles are 74.84, 80.25, 83.76, 85.71, 88.34, 90.29, 93.34, 96.56, 100.25
* The interquartile range is 13.24
* The range is 58.50 - 111.44

#### 3. Inferential statistics of the outcome variable (40% point)

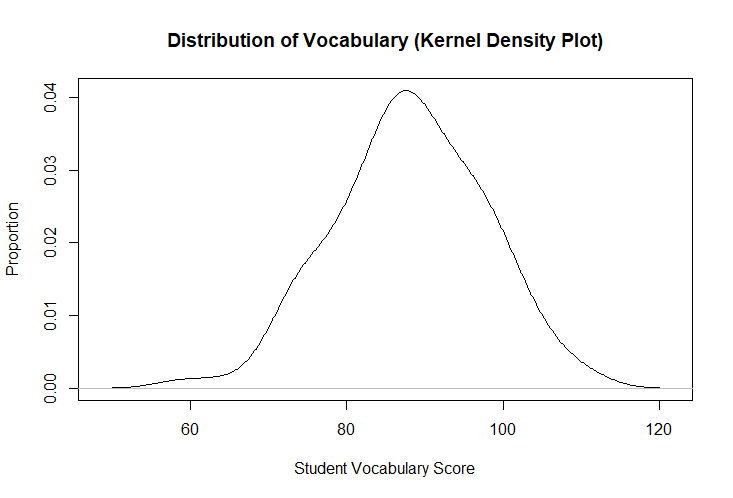
3.1 Random sample from the population

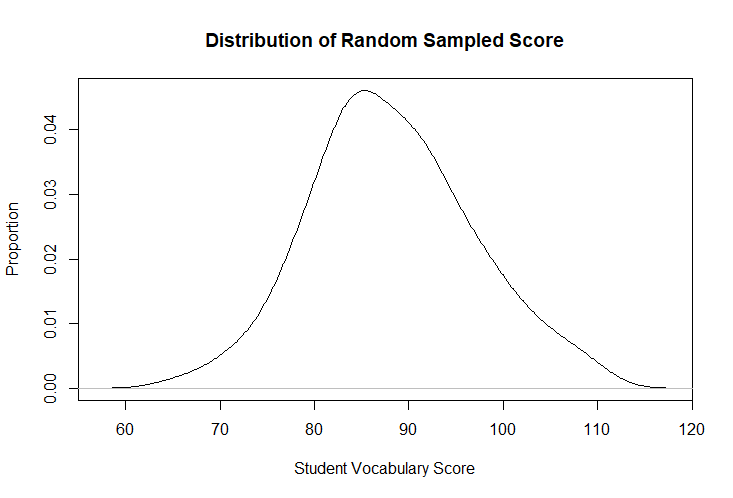
3.1.1. Using the mean and standard deviation of *vocabulary* to generate a normally distributed variable and name it *vocabulary\_random* (hint: use the rnorm function that randomly select a vector of numbers based on the number of observations, mean, and standard deviation you define).

IMPORTANT NOTE: before simulation, make sure to put set.seed(123) in your code so you can get the same sample and same results with those in the key.

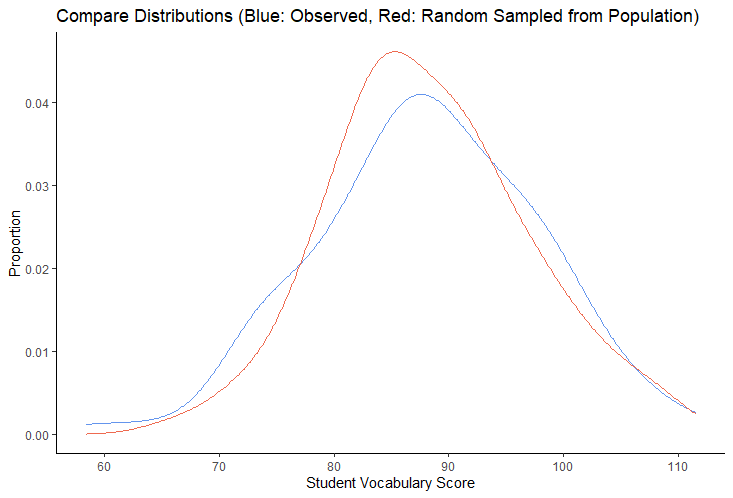
3.1.2. Using the same type of plot you used for vocabulary, plot the distribution of vocabulary\_random and compare the two plots. You may choose one of the two ways to do this: (a) use base R code to put the two plots side-by-side, or (b) use {ggplot2} package to put the two into one plot.

* Option (a)





* Option (b)



3.1.3. What are the statistical meanings of this new plot? Compare these two distributions and write one sentence to describe what you have observed. Circle back to question 2.1.2, do you think now whether vocabulary is normally distributed?

* It represents the distribution of a randomly generated sample from the population that share the same mean and standard deviation with the variable vocabulary
* The distributions of vocabulary (the observed sample in our dataset) and vocabulary\_random (the random sample from the population) are very similar.
* We can conclude now that visually, the observed variable, vocabulary, is normally distributed.

3.2 One-sample t-test

3.2.1 Despite whether *vocabulary* is normally distributed, you want to assume normal distribution in your future analysis. At this point, please specify your null hypothesis for the one-sample t-test (which you’ll conduct in the following questions to see if your assumption holds).

* The mean value of vocabulary is not significantly different from the population mean.

3.2.2. What is the mean of vocabulary\_random? Is it equal to the mean of \*vocabulary\*? Explain why they are identical/not identical.

* The mean of vocabulary\_random is 87.77 while the mean of vocabulary is 88.12.

3.2.3. We consider the population mean is equal to the mean of vocabulary\_random. To test whether your null hypothesis in 3.2.1 is true, run a one-sample t-test and write 2-3 sentences to interpret your result.

* The *t* statistic is 0.405 with a *p*-value above 0.5, indicating that the probability of the value of t differs from zero is only by chance. As a result, we keep the null hypothesis and conclude that the mean of vocabulary is not significantly different from the population mean.