# COMPSCI- 5590-0015

# ECONOMETRICS OF DATA SCIENCE
# HOMEWORK-1

**Name: Mohammad Shaik**

**Student id: 16319967**

**1.Generate summary statistics for vehicular speeds data and compare the results.**

**Ans:**

To read the data, first I need to check whether the dataset located in the current working directory. Use getwd() function which returns the current Working Directory in fetching dataset.

Note: It's not mandatory to use getwd() function, to find the dataset. Instead we can create the project folder having the dataset added manually.

Initially, I have read the "hw1_speed_data.csv " dataset into the dataframe 'data' by using read.csv() function and read the first 10 Rows from the dataframe.

```
> #Get current Working Directory to fetch dataset
> getwd()
[1] "/Users/mohammadshaik/Desktop/UMKC/Econometrics of Data Science/Hw-1"
> #Read CSV file
> data <- read.csv("hw1_speed_data.csv", header=TRUE)
> # Reading First 10 Rows from the dataframe
> data[1:10,]
   speed_before speed_after
1          32.6        63.5
2          35.3        56.9
3          40.2        54.1
4          41.9          NA
5          44.2        60.3
6          46.0        56.1
7          46.3        56.6
8          47.5        56.9
9          48.3        59.8
10         48.3        56.7
```

Basically, summary provides the statistics for each column of the entire dataset. Now to generate the summary statistics information for vehicular speeds data, I have used summary() function and then it returns the result as displayed in the console as in the below fig.

```
> summary(data)
  speed_before    speed_after
 Min.   :32.60   Min.   :45.60
 1st Qu.:55.20   1st Qu.:57.30
 Median :57.85   Median :59.45
 Mean   :57.66   Mean   :60.49
 3rd Qu.:60.00   3rd Qu.:63.50
 Max.   :69.50   Max.   :72.50
                 NA's   :192
>
```

From the above fig., Looking at the summary values, we can see that the summary stats are higher for the 'speed_after' than 'speed_before' as the vehicle speeds goes up. On the other hand, there were 192 missing values for vehicular speed column 'speed_after'.

Similarly, I found the summary statistics for the individual column by passing parameter to the summary() function and returns the stats as shown below:

```
> summary(data$speed_before)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.60   55.20   57.85   57.66   60.00   69.50
```

Also, I found individual stats i.e. mean for the selected column by passing it as a parameter for mean() function as shown below:

```
> mean(data$speed_before)
[1] 57.65565
```

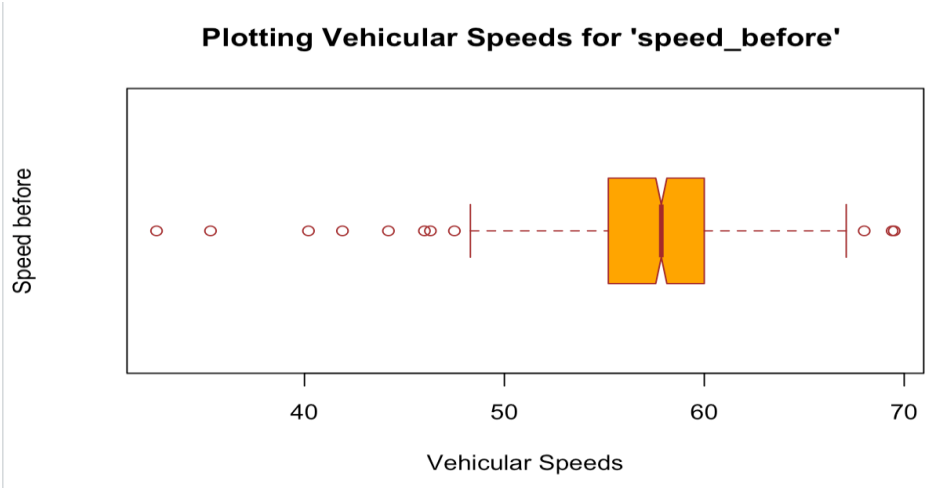
**2. Generate and interpret box plots for vehicular speeds data.**

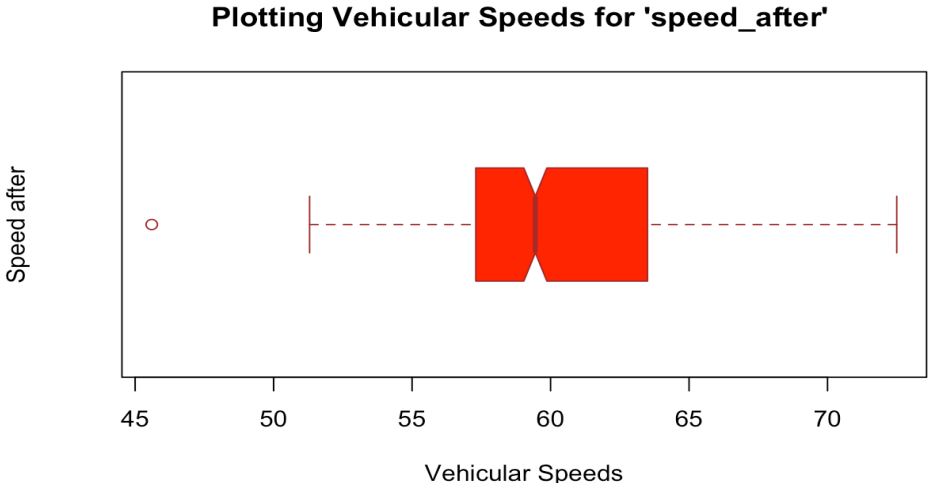
**Ans**:

Boxplot draws a plot for any of the vector, which takes the numeric vectors only. In our case Boxplots represent the comparison of vehicular speed data for both 'speed_before' and 'speed_after' and plot show the mean, spread, skewness, and outliers of the entire dataset.

Firstly, I have plotted the boxplot individually for both vehicular speed 'speed_before' as well as 'speed_after' columns. Using boxplot() with parameters x axis label name, colour, border and legend for the title, I executed the code and the result you can see as follows:

**Individual plots for both vehicular speeds before and after:**

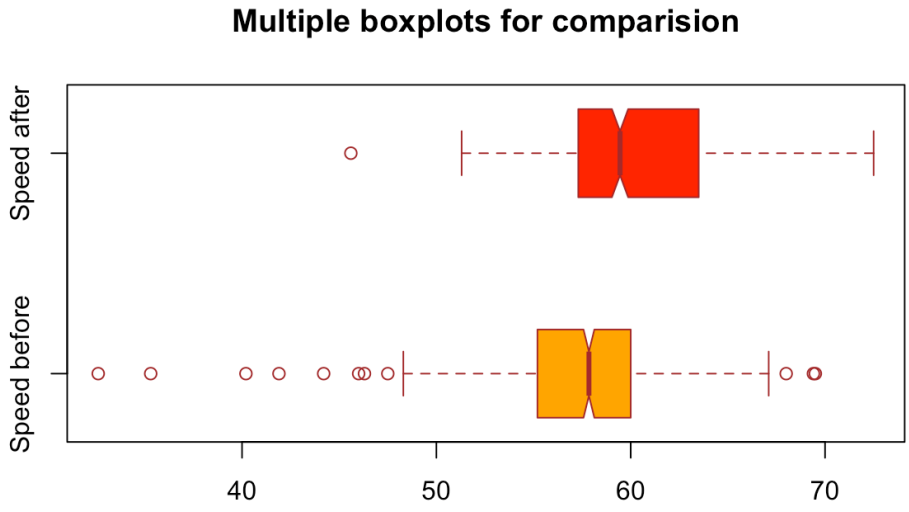



- For the ‘speed_before’ column of Vehicular speed data, we got the average value as 58 and for ‘speed_after’, it is 59. The average value is incremented just by less value for the following speed.
- For median, data is spread more evenly for the both the speeds.
- Also, you can see the least value for ‘speed_before’ is about 30 while for ‘speed_after’ data, it is about 45. Similarly, the highest value you can see for both the data is 70 and 74.
- The Vehicular speed before data boxplot shows a right-skewed distribution with a large tail on the right side, whereas the vehicular speed after data boxplot shows a symmetrical distribution with equal tails towards both ends. increase.

Additionally, we can represent to both the plots with in one boxplot as follows:

Here few more parameters were passed to the function to compare the plots.

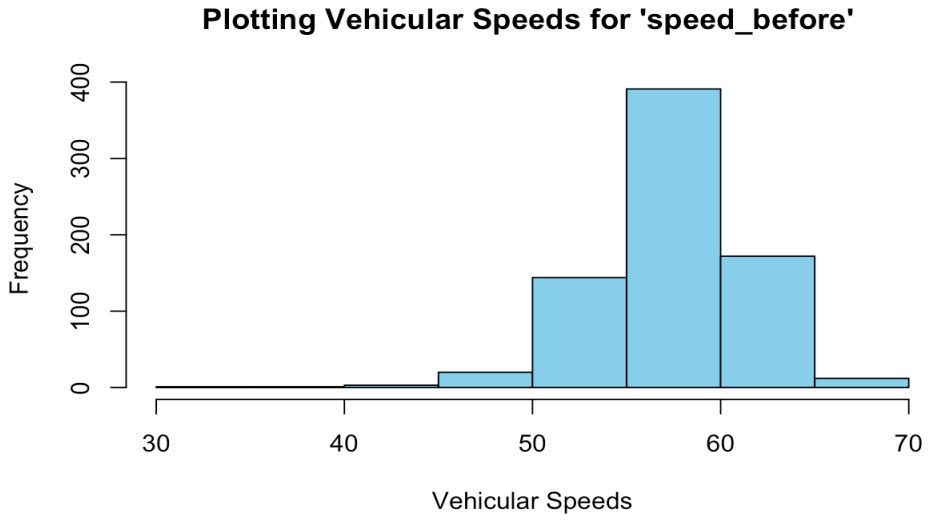



**3. Generate and interpret histograms for vehicular speeds data.**

**Ans**:

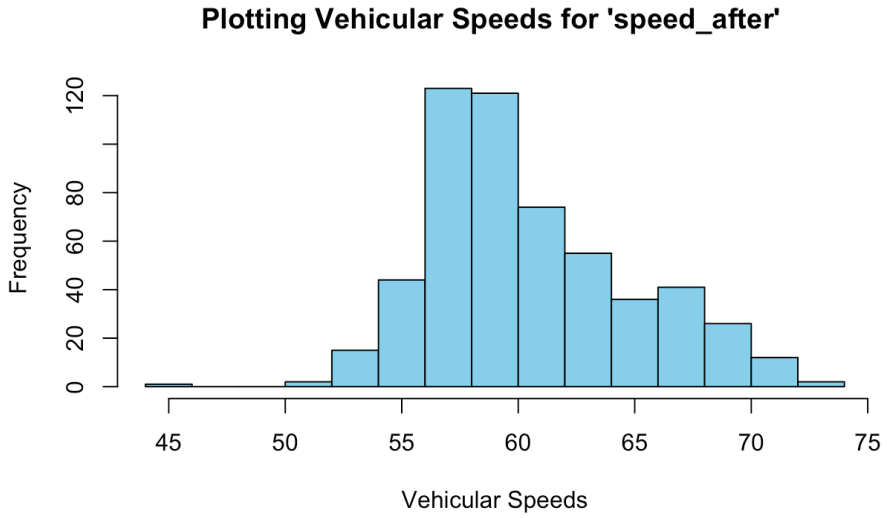
A histogram is a bar chart representation dividing the data into a set of frequencies (or) classes as columns along the horizontal x-axis whereas y-axis represents the count or percentage within the data for each column. Histogram is plotted by using hist() which takes a vector of values.

In our case, passing parameters like x axis label name, legend for the title, also add colour and border to the hist() function I executed the code and the result you can see as follows:



Generally, histograms help in finding the distribution of the data and their skewness.

For the above plot, it represents the distribution is not symmetrical and there is a left skewed distribution where the outliers located to the left side of the plot.




Similarly, the histogram with added parameters for "speed_after" data represents the distribution is a right skewed distribution.

Hence, both the plots are normally distributed but the ‘speed_before’ plot has a higher peak and the "speed_after" skews more to the left w.r.t other plots.

**4. Find out the mean and median values of the after-speed data for those particular vehicles whose speeds before the repeal were greater than 60 mph.**

Ans:

We can find the mean and median for the dataset by using mean() and median() functions passing dataset as a parameter. Also add ‘na.rm = TRUE’ as aparameter which removes null values.

Now, filtered original dataset having the ‘speed_before’ greater than 60mph for the ‘speed_after’ data and then removing null values, using mean() and median() functions the result is as follows:

```
> #mean for after-speed data by filtering speeds before > 60
> mean(data[data$speed_before>60,'speed_after'],na.rm = TRUE)
[1] 61.3731
> #median for after-speed data by filtering speeds before > 60
> median(data[data$speed_before>60,'speed_after'],na.rm = TRUE)
[1] 59.9
```

Hence, mean and median for after speed data for particular speeds before repeal greater than 60mph are 61.37 and 59.9

**5. Find out the frequency distribution of vehicular after-speed data and interpret results.**

**Ans:**

Create a list of class boundaries then Group the data into bins. Finally, find the Frequency distribution of Vehicular after speed data as follows:

Here the original was in a bad format:

freqdistdata

(45,50] (50,55] (55,60] (60,65] (65,70] (70,75]

1 31 274 146 86 14

Now the same table was shown in a better format:

| | freqdistdata | Freq |
|---|---|---|
| 1 | (45,50] | 1 |
| 2 | (50,55] | 31 |
| 3 | (55,60] | 274 |
| 4 | (60,65] | 146 |
| 5 | (65,70] | 86 |
| 6 | (70,75] | 14 |

**6. Generate 99% confidence intervals for mean vehicular after-speed data assuming the population variance is unknown. Explain each step and interpret the results.**

**Ans:**

We can solve this problem by uing 3 methods.

**1st Method:**

The confidence interval for the means can be calculated mathematically by using the below formula:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

Where CI represents confidence interval,

$\bar{x}$ represents population mean,

z represents critical value,

s represents standard deviation,

n represents length of sample.

Based upon the formula, we will be calculating all the values for the terms as follows:

1. Calculate the mean of the sample data
2. Compute the size
3. Find the standard deviation
4. Find the standard error which helps to find margin error
5. Find the critical value is calculated by using qt() function provided by R studio

Finally, by passing all the values to actual formula we will be calculating the lower bound and upper bound which means the confidence interval.

```
> print(c(lower_bound,upper_bound))
[1] 60.07497 60.90257
```

**2nd Method:**

Using One Sample t-test function.

In our case, with t.test() function with confidence interval as 0.99, I got the mean value and Confidence interval as following:

```
> t.test(data$speed_after, conf.level = 0.99)

        One Sample t-test

data:  data$speed_after
t = 325.16, df = 551, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 60.00793 60.96961
sample estimates:
mean of x
 60.48877
```

Thereby using p value, we can negotiate null hypothesis

**3rd Method:**

Using confint() function which builds the linear model on the data.

```
> # Calculate the mean and standard error
> model <- lm(speed_after ~ 1, data)
> # Find the confidence interval
> confint(model, level=0.99)
                0.5 %   99.5 %
(Intercept) 60.00793 60.96961
>
```

**7. Generate 95% confidence intervals for the variance of before-speed data. Explain each step and interpret the results.**

**Ans**:

The confidence interval for Population Variance can be calculated mathematically by using the below formula:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ and are obtained from the chi-squared distribution table for α/2 and 1- α/2,

n-1 degrees of freedom,

s2 is a variance.

Based upon the formula, we will be calculating all the values for the terms as follows:

1. Calculate the variance of the sample data by using var() function
2. Compute the size
3. Find the α value
4. Find chi-squared distribution table

Finally, by passing all the values to actual formula we will be calculating the lower bound and upper bound which means the confidence interval.

```
> print(c(lower_bound,upper_bound))
[1] 14.87124 18.22765
```

**8. Test whether the mean speed is 55 mph before and 60 mph after at the α=5% significance level. Explain each step and interpret the results.**

**Ans:**

Basically, there are 2 methods to find the solution as shown in 6th question.

It's easy with the t.test() method.

To find the mean for the given before-speed and after speed data at α=5% significance level, firstly we need to consider the confidence level.

Confidence level is calculated based upon the significance level i.e. confidence level is denoted by (1-α)100% = 0.95

**Mean speed is 55 mph for 'speed_ before':**

By using test() function and then adding parameters mu = 55 and conf.level = 0.95 for t-test, where mu represents the number indicates true value of mean and confidence level as obtained above. This function returns One Sample t-test which helps in getting the p value in negotiating the null hypothesis.

```
> t.test(data$speed_before, mu = 55,conf.level=0.95)

        One Sample t-test

data:  data$speed_before
t = 17.875, df = 743, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 55
95 percent confidence interval:
 57.36399 57.94730
sample estimates:
mean of x
 57.65565
```

Here, we got the mean is 57.65565 which is more than the actual mean given at confidence level as 0.95.

Hence, it safely rejects the null hypothesis as because the p value obtained is less than the significance level i.e. p value < α=0.05

**Mean speed is 60 mph for 'speed_ after':**

Similarly, we pass mu = 60 and conf.level = 0.05 to the test() and calculate the mean for speed_after data as follows:

```
> t.test(data$speed_after, mu = 60,conf.level=0.95)

        One Sample t-test

data:  data$speed_after
t = 2.6274, df = 551, p-value = 0.008843
alternative hypothesis: true mean is not equal to 60
95 percent confidence interval:
 60.12336 60.85418
sample estimates:
mean of x
 60.48877
```

Here, we got the mean is 60.48877 which is more than the actual mean given at confidence level as 0.95.

Hence, it safely rejects the null hypothesis as because the p value obtained is less than the significance level i.e. p value < α=0.05

**9. Test whether the variance of after-speed data is less than 19 mph2 at the α=5% significance level. Explain each step and interpret the results**

**Ans:**

To find the variance for the given after-speed data for α=5% significance level, firstly we need to consider the confidence level.

Confidence level is calculated based upon the significance level i.e. confidence level is denoted by (1-α)100% = 0.95

Firstly, we need to install packages "EnvStats" and then import the library and then finding the variance by using varTest() function which helps in getting the p value and Chi-squared value which helps in negotiating the null hypothesis.

We Considered sigma.squared = 19 as because variance of 'speed_after' is less than 19mph2

```
> varTest(data$speed_after, alternative = "less", conf.level = 0.95, sigma.squared = 19)
$statistic
Chi-Squared
    553.969

$parameters
 df
551

$p.value
[1] 0.5435382

$estimate
variance
19.10238

$null.value
variance
      19

$alternative
[1] "less"

$method
[1] "Chi-Squared Test on Variance"
```

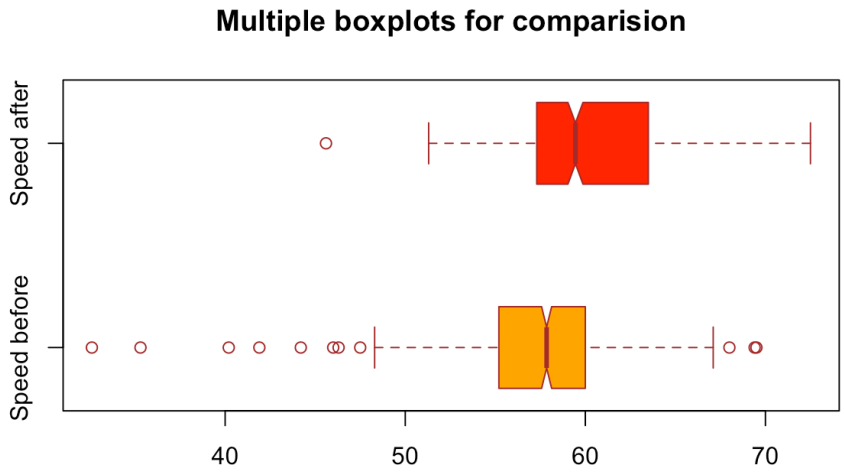
Here, we got the variance is 19.10238 which is more than the actual variance given at confidence level as 0.95.

Hence, it fails to reject the null hypothesis as because the p value obtained is not less than the significance level i.e. $p = 0.5435 > \alpha = 0.05$

**10. Test that the mean vehicular speeds before and after are equal at the α=10%significance level. Explain each step and interpret the results.**

**Ans:**

Firstly, we will be plotting Boxplots to find the whether the mean vehicular speeds for both the speeds before and after are equal.




From the actual Boxplots, the means are not equal. So we will be finding whether the means are equal at α=10% significance level

To find the means for the given vehicular speeds before and after for α=10% significance level, firstly we need to find the confidence level.

Confidence level is calculated based upon the significance level i.e. confidence level is denoted by (1-α)100% = 0.90

By using help() method, we find the parameters required for t-test, assuming variances are uneaqual and alternative with two sided test. Then finding the means by using t.test() function, a Welch Two Sample t-test which helps in getting the p value in negotiating the null hypothesis.

```
> t.test(data$speed_before,data$speed_after,alternative="two.sided",mu=0,var.equal = FALSE,paired=F,conf=
0.90)

        Welch Two Sample t-test

data:  data$speed_before and data$speed_after
t = -11.9, df = 1135.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -3.225032 -2.441214
sample estimates:
mean of x mean of y
 57.65565  60.48877
```

mean of x 57.65565,

mean of y 60.48877

Here the mean for before and after speed data are not equal.

Hence, it rejects the null hypothesis as because the p value obtained is less than the significance level considering no differences in the mean.

**11. Test that the vehicular speed variances before and after are equal at the α=5% significance level. Explain each step and interpret the results.**

**Ans:**

To find the variances for the given vehicular speeds before and after for α=5% significance level, firstly we need to find the confidence level.

Confidence level is calculated based upon the significance level i.e. confidence level is denoted by (1-α)100% = 0.95.

To find the variances, we can use var.test() function to compare two variances.

By using help() method, we find the parameters required for f test, assuming alternative with two sided test. Then finding the means by using var.test() function with added parameters, a F test to compare two variances which helps in getting the p value in negotiating the null hypothesis.

```
> var.test(data$speed_before,data$speed_after, alternative = 'two.sided',conf.level = 0.95)

        F test to compare two variances

data:  data$speed_before and data$speed_after
F = 0.85963, num df = 743, denom df = 551, p-value = 0.05591
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7348115 1.0038179
sample estimates:
ratio of variances
         0.8596321
```

Here, we got the variance is 0.8596321 and the p value 0.05591.

Hence, it fails to reject the null hypothesis as because the p value obtained is not less than the significance level i.e. p = 0.05591> α=0.05 considering that there is no significant difference between the variances.

**12. Use a Mann-Whitney-Wilcoxon test to assess whether the distributions of speeds before and after are equal. Also draw density plots using before and after speeds data. Interpret the results based on the test and drawing**

**Ans:**

To find the distribution for the given vehicular speeds before and after are equal, let's consider significance level as α=5%, now we need to find the confidence level.

Confidence level is calculated based upon the significance level i.e. confidence level is denoted by (1-α)100% = 0.95.

As given by using Mann-Whitney-Wilcoxon test, we will be checking distribution of speeds. By using help() method, to get more info on this test and then add parameters required for wilcox test, assuming alternative with two sided test.

Finally, wilcox.test() gives the p value which helps in considering the null hypothesis or not.

```
> wilcox.test(data$speed_before,data$speed_after,alternative = "two.sided",conf.int = T,conf.level = 0.9
5,,correct = T)

        Wilcoxon rank sum test with continuity correction

data:  data$speed_before and data$speed_after
W = 137860, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -2.799986 -1.899980
sample estimates:
difference in location
             -2.300024
```

Hence, it rejects the null hypothesis as because the p value obtained is not less than the significance level considering that the mean is significant different and there is negative difference in location.

**Density plots forvehicular speeds data before and after:**

Let take d1 as speed_before data, d2 as speed_after data by removing null values

Then plot the density diagrams by using the plot() function and provide the legends as follows:

From the below density plots, we can see the almost similar plots.