Econometrics of Data Science

**Basic Statistics using R**

Author: Manasseh

2022-12-06

## Import data

df <- read.csv(file = "hw1\_speed\_data.csv", header = TRUE, sep = ",")

### Q1

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

summary(df$speed\_before)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 32.60 55.20 57.85 57.66 60.00 69.50

summary(df$speed\_after)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 45.60 57.30 59.45 60.49 63.50 72.50 192

**Findings**

The mean and median vehicular speed before the repeal of a speed limit law are 57.66 (mph) and 57.85 (mph), respectively.. The mean and median vehicular speed after the repeal of a speed limit law are 60.49 (mph) and 59.45 (mph), respectively..

* Both mean and median vehicular speed after the repeal of the speed limit were greater than before the repeal.

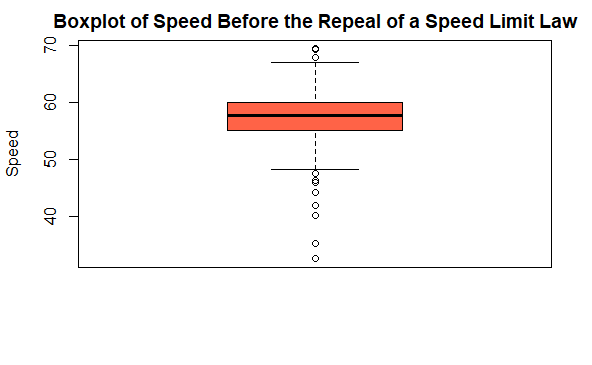
The minimum and maximum vehicular speed before the repeal of a speed limit law are 32.60 and 69.50, respectively.. The minimum and maximum vehicular speed after the repeal of a speed limit law are 45.60 and 72.50, respectively..

* The third quarter vehicular speed after (63.50 mph) the repeal of a speed limit law was slightly higher than before (60.00 mph) the repeal.
* There exists missing 192 speed data records for after the repeal of limit law.

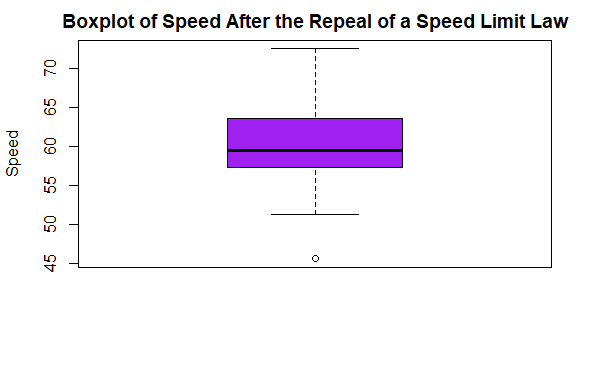
### Q2

Generate and interpret box-plots for vehicular speeds data (make them look nice).

boxplot(df$speed\_before,   
 ylab = "Speed (mph)",   
 col = "tomato",  
 main = "Boxplot of Speed Before the Repeal of a Speed Limit Law")



boxplot(df$speed\_after,   
 ylab = "Speed (mph)",  
 col = "purple",  
 main = "Boxplot of Speed After the Repeal of a Speed Limit Law")



**Findings**

The Box-plot of Speed before the repeal of a speed limit law shows there exists numerous extreme values or outliers.

The Box-plot of Speed after the repeal of a speed limit law shows there exists less outlier(s) or extreme value(s).

* Outliers have great impact on the mean value of the speed before the repeal but have little effect on the median.

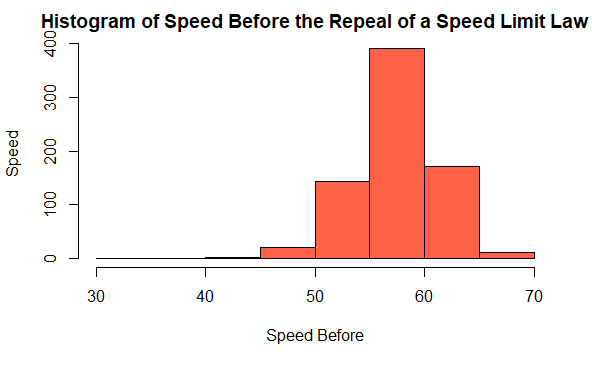
The upper whisker of Speed after the repeal of a speed limit law is longer compared to lower whisker, however it is equally same for speed before the repeal of limit law.

The median line of Speed after the repeal of speed limit law is not centered, however it is centered for speed before the repeal of limit law. This shows the speed after the repeal data is skewed.

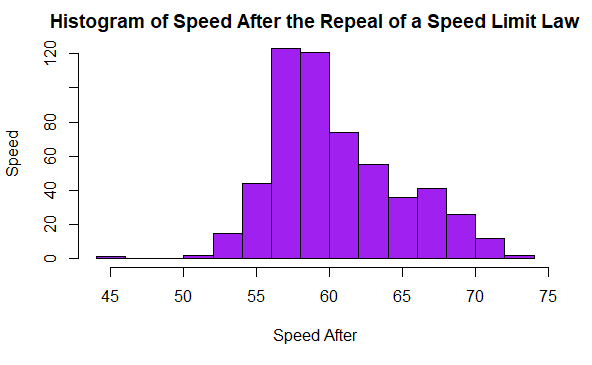
### Q3

Generate and interpret histograms for vehicular speeds data (make them look nice).

hist(df$speed\_before,  
 ylab = "Speed (mph)",  
 xlab = "Speed Before",  
 col = "tomato",  
 main = "Histogram of Speed Before the Repeal of a Speed Limit Law")



hist(df$speed\_after,  
 ylab = "Speed (mph)",  
 xlab = "Speed After",  
 col = "purple",  
 main = "Histogram of Speed After the Repeal of a Speed Limit Law")



**Findings**

The histogram plot of Speed before the repeal of a speed limit law shows the distribution appears normal.

The histogram plot of Speed after the repeal of a speed limit law shows the distribution is skewed.

### Q4

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.

library(tidyverse)  
  
df |>   
 drop\_na() |>  
 filter(speed\_before > 60) |>   
 summarise(MEAN\_SPEED = mean(speed\_after),  
 MEDIAN\_SPEED = median(speed\_after))

## MEAN\_SPEED MEDIAN\_SPEED  
## 1 61.3731 59.9

**Findings**

The mean and median values of the after-speed data for those particular vehicles whose speeds before the repeal were greater than 60 mph were 61.3731 mph and 59.9 mph, respectively.

### Q5

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

library(gtsummary)  
  
df |>  
 select(speed\_after)|>  
 mutate(speed\_after\_ = as.factor(speed\_after)) |>  
 tbl\_summary() |>  
 as\_flex\_table()

| Characteristic | N = 7441 |
| --- | --- |
| speed\_after | 59.5 (57.3, 63.5) |
| Unknown | 192 |
| speed\_after\_ |  |
| 45.6 | 1 (0.2%) |
| 51.3 | 1 (0.2%) |
| 51.7 | 1 (0.2%) |
| 52.4 | 2 (0.4%) |
| 52.5 | 1 (0.2%) |
| 52.6 | 1 (0.2%) |
| 52.7 | 1 (0.2%) |
| 52.8 | 1 (0.2%) |
| 53.1 | 1 (0.2%) |
| 53.3 | 1 (0.2%) |
| 53.4 | 2 (0.4%) |
| 53.6 | 3 (0.5%) |
| 53.7 | 1 (0.2%) |
| 53.9 | 1 (0.2%) |
| 54.1 | 3 (0.5%) |
| 54.3 | 3 (0.5%) |
| 54.4 | 3 (0.5%) |
| 54.7 | 2 (0.4%) |
| 54.8 | 1 (0.2%) |
| 54.9 | 1 (0.2%) |
| 55 | 1 (0.2%) |
| 55.1 | 2 (0.4%) |
| 55.2 | 1 (0.2%) |
| 55.3 | 3 (0.5%) |
| 55.4 | 5 (0.9%) |
| 55.5 | 2 (0.4%) |
| 55.6 | 3 (0.5%) |
| 55.7 | 2 (0.4%) |
| 55.8 | 6 (1.1%) |
| 55.9 | 3 (0.5%) |
| 56 | 3 (0.5%) |
| 56.1 | 4 (0.7%) |
| 56.2 | 7 (1.3%) |
| 56.3 | 6 (1.1%) |
| 56.4 | 3 (0.5%) |
| 56.5 | 2 (0.4%) |
| 56.6 | 8 (1.4%) |
| 56.7 | 5 (0.9%) |
| 56.8 | 7 (1.3%) |
| 56.9 | 6 (1.1%) |
| 57 | 8 (1.4%) |
| 57.1 | 9 (1.6%) |
| 57.2 | 5 (0.9%) |
| 57.3 | 8 (1.4%) |
| 57.4 | 13 (2.4%) |
| 57.5 | 4 (0.7%) |
| 57.6 | 5 (0.9%) |
| 57.7 | 3 (0.5%) |
| 57.8 | 8 (1.4%) |
| 57.9 | 7 (1.3%) |
| 58 | 5 (0.9%) |
| 58.1 | 8 (1.4%) |
| 58.2 | 2 (0.4%) |
| 58.3 | 7 (1.3%) |
| 58.4 | 9 (1.6%) |
| 58.5 | 7 (1.3%) |
| 58.6 | 3 (0.5%) |
| 58.7 | 7 (1.3%) |
| 58.8 | 9 (1.6%) |
| 58.9 | 6 (1.1%) |
| 59 | 7 (1.3%) |
| 59.1 | 6 (1.1%) |
| 59.2 | 4 (0.7%) |
| 59.3 | 10 (1.8%) |
| 59.4 | 6 (1.1%) |
| 59.5 | 9 (1.6%) |
| 59.6 | 8 (1.4%) |
| 59.7 | 6 (1.1%) |
| 59.8 | 3 (0.5%) |
| 59.9 | 3 (0.5%) |
| 60 | 1 (0.2%) |
| 60.1 | 4 (0.7%) |
| 60.2 | 10 (1.8%) |
| 60.3 | 7 (1.3%) |
| 60.4 | 3 (0.5%) |
| 60.5 | 1 (0.2%) |
| 60.6 | 6 (1.1%) |
| 60.7 | 4 (0.7%) |
| 60.8 | 5 (0.9%) |
| 60.9 | 1 (0.2%) |
| 61 | 5 (0.9%) |
| 61.4 | 5 (0.9%) |
| 61.5 | 5 (0.9%) |
| 61.6 | 3 (0.5%) |
| 61.7 | 2 (0.4%) |
| 61.8 | 6 (1.1%) |
| 61.9 | 4 (0.7%) |
| 62 | 3 (0.5%) |
| 62.1 | 1 (0.2%) |
| 62.2 | 4 (0.7%) |
| 62.3 | 3 (0.5%) |
| 62.4 | 2 (0.4%) |
| 62.5 | 4 (0.7%) |
| 62.6 | 3 (0.5%) |
| 62.7 | 2 (0.4%) |
| 62.8 | 2 (0.4%) |
| 62.9 | 1 (0.2%) |
| 63 | 5 (0.9%) |
| 63.1 | 2 (0.4%) |
| 63.2 | 2 (0.4%) |
| 63.3 | 1 (0.2%) |
| 63.4 | 1 (0.2%) |
| 63.5 | 3 (0.5%) |
| 63.6 | 3 (0.5%) |
| 63.7 | 5 (0.9%) |
| 63.8 | 7 (1.3%) |
| 63.9 | 3 (0.5%) |
| 64 | 1 (0.2%) |
| 64.1 | 2 (0.4%) |
| 64.2 | 2 (0.4%) |
| 64.3 | 2 (0.4%) |
| 64.4 | 3 (0.5%) |
| 64.5 | 1 (0.2%) |
| 64.6 | 6 (1.1%) |
| 64.7 | 1 (0.2%) |
| 65.1 | 2 (0.4%) |
| 65.2 | 1 (0.2%) |
| 65.5 | 3 (0.5%) |
| 65.6 | 1 (0.2%) |
| 65.7 | 6 (1.1%) |
| 65.9 | 2 (0.4%) |
| 66 | 4 (0.7%) |
| 66.2 | 1 (0.2%) |
| 66.3 | 2 (0.4%) |
| 66.4 | 4 (0.7%) |
| 66.5 | 3 (0.5%) |
| 66.7 | 5 (0.9%) |
| 66.8 | 2 (0.4%) |
| 66.9 | 3 (0.5%) |
| 67 | 2 (0.4%) |
| 67.1 | 1 (0.2%) |
| 67.2 | 1 (0.2%) |
| 67.3 | 5 (0.9%) |
| 67.4 | 3 (0.5%) |
| 67.5 | 1 (0.2%) |
| 67.6 | 2 (0.4%) |
| 67.7 | 3 (0.5%) |
| 67.8 | 3 (0.5%) |
| 68.1 | 2 (0.4%) |
| 68.2 | 2 (0.4%) |
| 68.3 | 2 (0.4%) |
| 68.5 | 2 (0.4%) |
| 68.6 | 2 (0.4%) |
| 68.7 | 3 (0.5%) |
| 68.8 | 3 (0.5%) |
| 68.9 | 2 (0.4%) |
| 69 | 1 (0.2%) |
| 69.1 | 1 (0.2%) |
| 69.5 | 1 (0.2%) |
| 69.6 | 1 (0.2%) |
| 69.7 | 1 (0.2%) |
| 69.8 | 2 (0.4%) |
| 69.9 | 1 (0.2%) |
| 70.1 | 1 (0.2%) |
| 70.2 | 2 (0.4%) |
| 70.3 | 1 (0.2%) |
| 70.4 | 1 (0.2%) |
| 70.6 | 2 (0.4%) |
| 70.9 | 2 (0.4%) |
| 71 | 1 (0.2%) |
| 71.5 | 1 (0.2%) |
| 71.8 | 1 (0.2%) |
| 72.2 | 1 (0.2%) |
| 72.5 | 1 (0.2%) |
| Unknown | 192 |
| 1Median (IQR); n (%) | |

**Findings**

The most frequent vehicular after-speed was 57.4 mph, with 13 counts, followed by 59.3 mph and 60.2 mph, with 10 counts each, respectively.

### Q6

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

# create model  
after.model <- lm(speed\_after ~ 1, df)  
  
# Calculate the confidence interval  
confint(after.model, level=0.99)

## 0.5 % 99.5 %  
## (Intercept) 60.00793 60.96961

# OR  
sample.mean <- mean(df$speed\_after, na.rm = TRUE)  
print(sample.mean)

## [1] 60.48877

sample.n <- length(df$speed\_after)  
sample.sd <- sd(df$speed\_after, na.rm = TRUE)  
sample.se <- sample.sd/sqrt(sample.n)  
print(sample.se)

## [1] 0.160235

alpha = 0.01  
degrees.freedom = sample.n - 1  
t.score = qt(p=alpha/2, df=degrees.freedom,lower.tail=F)  
print(t.score)

## [1] 2.582462

margin.error <- t.score \* sample.se  
print(margin.error)

## [1] 0.4138008

lower.bound <- sample.mean - margin.error  
upper.bound <- sample.mean + margin.error  
print(c(lower.bound,upper.bound))

## [1] 60.07497 60.90257

**Findings**

Step 1: Calculate the mean - Mean = 60.48877

Step 2: Calculate the standard error of the mean - se = 0.160235

Step 3: Find the t-score that corresponds to the confidence level (99%) - t-score = 2.582462

Step 4. Calculate the margin of error and construct the confidence interval - moe = 0.4138008; Lower = 60.07497; Upper = 60.90257

The 99% confidence intervals for mean vehicular after-speed is (60.07497 kph, 60.90257 kph)

### Q7

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

d.f <- length(df$speed\_before) - 1  
var\_speed\_before <- var(df$speed\_before, na.rm = TRUE)  
print(var\_speed\_before)

## [1] 16.42102

lower = var\_speed\_before \* d.f / qchisq(0.05/2, d.f, lower.tail = FALSE)  
upper = var\_speed\_before \* d.f / qchisq(1 - 0.05/2, d.f, lower.tail = FALSE)  
c(lower = lower, variance = var\_speed\_before, upper = upper)

## lower variance upper   
## 14.87124 16.42102 18.22765

**Findings**

Step 1: Calculate the degrees of freedom - df = 743

Step 2: Calculate the variance - var = 16.42102

Step 3: Find the Chi-Square quantile function that corresponds to the confidence level (95%) - 669.3578 and 820.4301

Step 4. Calculate and construct the confidence interval - Lower = 14.87124; Upper = 18.22765

The 95% confidence intervals for the variance of before-speed is (14.87124 kph, 18.22765 kph)

### Q8

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.

# One-sample t-test  
speed\_before <- t.test(df$speed\_before, mu = 55)  
# Printing the results  
print(speed\_before)

##   
## One Sample t-test  
##   
## data: df$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

# One-sample t-test  
speed\_after <- t.test(df$speed\_after, mu = 60)  
# Printing the results  
print(speed\_after)

##   
## One Sample t-test  
##   
## data: df$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

**Findings ~ part 1**

Step 1: State the hypothesis Ho: mean speed before = 55 Ha: mean speed before is not equal to 55

Step 2: Calculate the degrees of freedom and sample mean - df = 743; sample mean = 57.65565

Step 3: Calculate the test statistic - t-statistic = 17.875

Step 4: Find the p-value - p-value < 2.2e-16

Step 5. Decision-making - The p-value is less than the alpha (0.05), hence we reject the null hypothesis and conclude that the mean speed is not equal to 55 mph before the repeal.

**Findings ~ part 2**

Step 1: State the hypothesis Ho: mean speed after = 60 Ha: mean speed after is not equal to 60

Step 2: Calculate the degrees of freedom and sample mean - df = 551; sample mean = 60.48877

Step 3: Calculate the test statistic - t-statistic = 2.6274

Step 4: Find the p-value - p-value = 0.008843

Step 5. Decision-making - The p-value is less than the alpha (0.05), hence we reject the null hypothesis and conclude that the mean speed is not equal to 60 mph after the repeal.

### Q9

Test whether the variance of after-speed data is less than 19 mph^2 at the α=5% significance level. Explain each step and interpret the results.

library(DescTools)  
  
VarTest(df$speed\_after, sigma.squared = 19, alternative = "less")

##   
## One Sample Chi-Square test on variance  
##   
## data: df$speed\_after  
## X-squared = 553.97, df = 551, p-value = 0.5435  
## alternative hypothesis: true variance is less than 19  
## 95 percent confidence interval:  
## 0.00000 21.15417  
## sample estimates:  
## variance of x   
## 19.10238

**Findings**

Step 1: State the hypothesis Ho: The variance of after-speed data is greater than or equal to 19 mph^2 Ha: The variance of after-speed data is less than 19 mph^2

Step 2: Calculate the degrees of freedom and sample estimate - df = 743; sample variance = 19.10238

Step 3: Calculate the test statistic - Chi square-statistic = 553.97

Step 4: Find the p-value - p-value = 0.5435

Step 5. Decision-making - The p-value is greater than the alpha (0.05), hence we fail to reject the null hypothesis and conclude that The variance of after-speed data is greater than or equal to 19 mph^2.

### Q10

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

t.test(df$speed\_before, df$speed\_after, var.equal = TRUE, conf.level = 0.90)

##   
## Two Sample t-test  
##   
## data: df$speed\_before and df$speed\_after  
## t = -12.034, df = 1294, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 90 percent confidence interval:  
## -3.220632 -2.445614  
## sample estimates:  
## mean of x mean of y   
## 57.65565 60.48877

**Findings**

Step 1: State the hypothesis Ho: The mean vehicular speeds before and after are equal Ha: The mean vehicular speeds before and after are NOT equal

Step 2: Calculate the degrees of freedom and sample mean - df = 1294; sample means (57.65565 and 60.48877)

Step 3: Calculate the test statistic - t-statistic = -12.034

Step 4: Find the p-value - p-value < 2.2e-16

Step 5. Decision-making - The p-value is less than the alpha (0.05), hence we reject the null hypothesis and conclude that the mean vehicular speeds before and after repeal are different.

### Q11

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

var.test(df$speed\_before, df$speed\_after, alternative = "two.sided")

##   
## F test to compare two variances  
##   
## data: df$speed\_before and df$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

**Findings**

Step 1: State the hypothesis Ho: The variance vehicular speeds before and after are equal Ha: The variance vehicular speeds before and after are NOT equal

Step 2: Calculate the degrees of freedoms and ratio of variances - numerator df = 743; denominator df = 551; ratio of variances (0.859)

Step 3: Calculate the test statistic - F-statistic = 0.85963

Step 4: Find the p-value - p-value = 0.05591

Step 5. Decision-making - The p-value of F-test is p = 0.05591 which is greater than the significance level 0.05. In conclusion, we fail to reject null hypothesis and conclude there is no significant difference between the two variances.

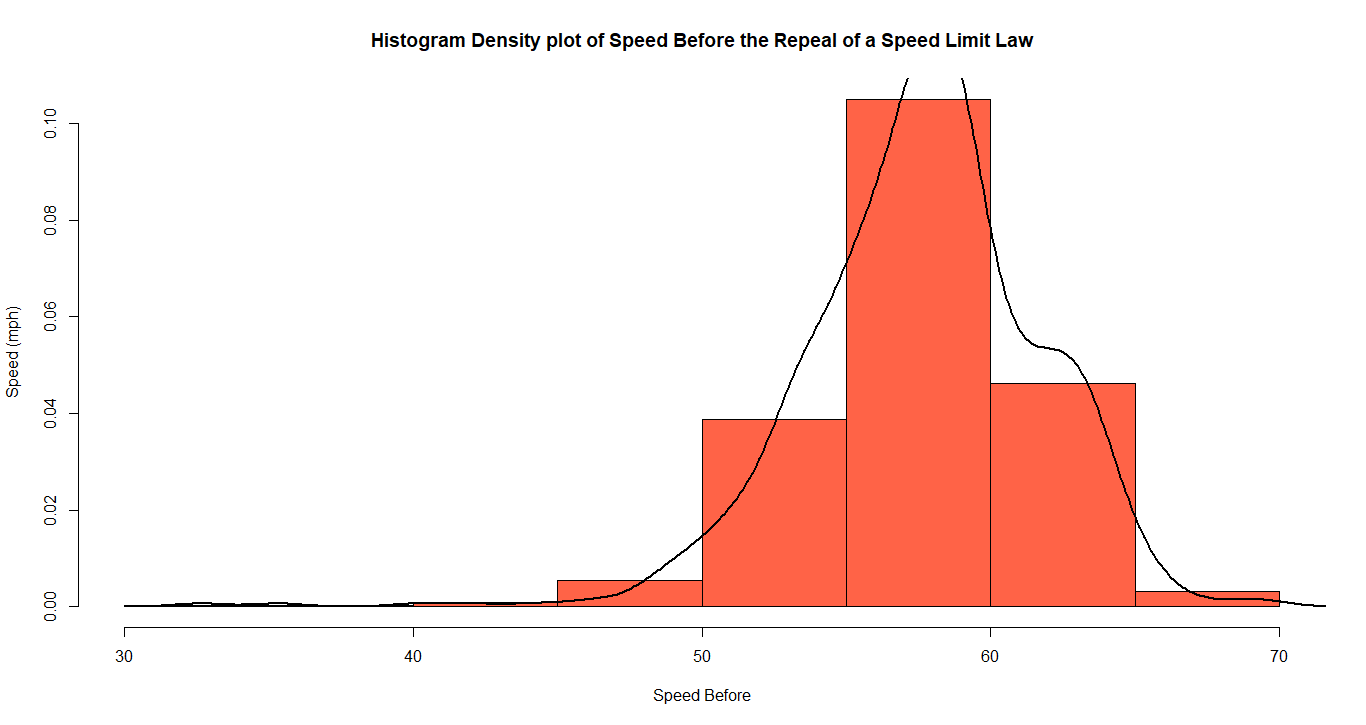
### Q12

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.

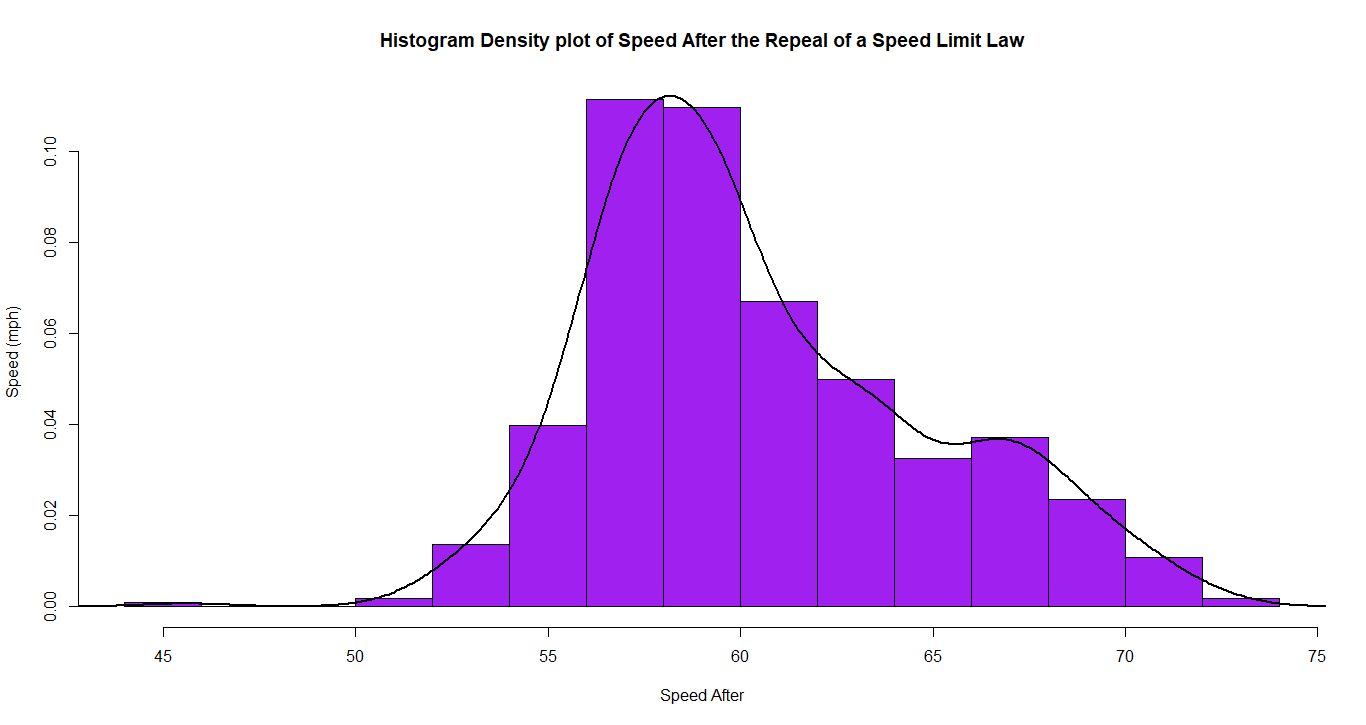
wilcox.test(df$speed\_before, df$speed\_after, data=df)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: df$speed\_before and df$speed\_after  
## W = 137860, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0

hist(df$speed\_before,  
 prob = TRUE,  
 ylab = "Speed (mph)",  
 xlab = "Speed Before",  
 col = "tomato",  
 main = "Histogram Density plot of Speed Before the Repeal of a Speed Limit Law")  
lines(density(df$speed\_before),  
 lwd = 2,  
 col = "black")



hist(df$speed\_after,  
 prob = TRUE,  
 ylab = "Speed (mph)",  
 xlab = "Speed After",  
 col = "purple",  
 main = "Histogram Density plot of Speed After the Repeal of a Speed Limit Law")  
lines(density(df$speed\_after, na.rm = TRUE),  
 lwd = 2,  
 col = "black")



**Findings**

Step 1: State the hypothesis Ho: The vehicular speeds before and after are sampled from populations with identical distributions. Ha: The vehicular speeds before and after are sampled from populations with different distributions

Step 2: Rank and Calculate the Wilcoxon Value - W = 137860

Step 3: Find the p-value - p-value < 2.2e-16

Step 4. Decision-making - At .05 significance level, we conclude that the vehicular speeds before and after are nonidentical populations.

* The density plot shows skewed distribution for the vehicular speeds after the repeal of limit law.
* The density plot shows approximately normal distribution for the vehicular speeds before the repeal of limit law.