Homework Exercise 1

Shan He

Sep 7, 2017

W203 Statistics for Data Science

Unit 1 Homework

Exercise

Load the dataset found in the file, cars.csv.

```
cars = read.csv("cars.csv")
```

1. What are the variables in the file?

```
names(cars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

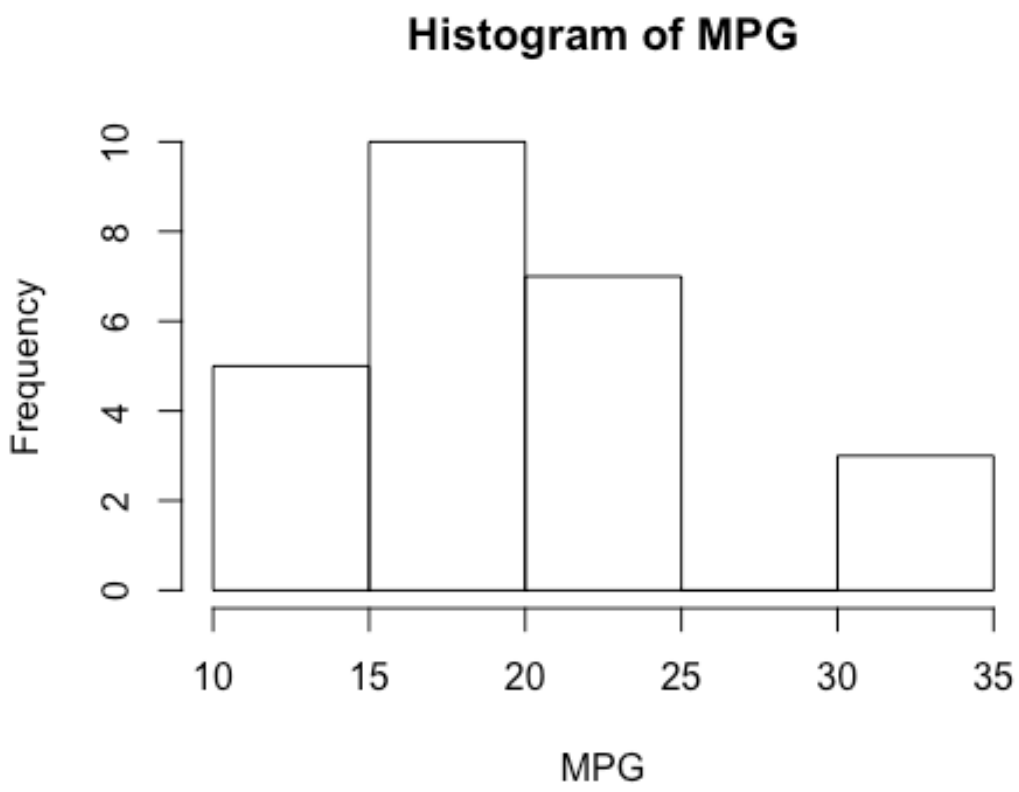
2. Find the mean, median, minimum, maximum, 1st quartile and 3rd quartile for the mpg variable.

```
summary(cars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.20   18.70   19.49   21.50   33.90
```

3. Create a histogram of the mpg variable.

```
hist(cars$mpg, main = "Histogram of MPG", xlab =  "MPG")
```



4. What is the standard deviation of mpg variable?

```
sd(cars$mpg)
```

```
## [1] 6.047446
```

5. What is the variance of mpg variable?

```
var(cars$mpg)
```

```
## [1] 36.5716
```

6. What is the relationship of the standard deviation to the variance? Why does the standard deviation and variance of the mpg variable differ?

```
sd(cars$mpg) == sqrt(var(cars$mpg))
```

```
## [1] TRUE
```

The standard deviation is the square root of the variance.

7. How many data points are there for the cyl variable?

```
sum(!is.na(cars$cyl))
```

```
## [1] 23
```

8. What is the mean of the cyl variable?'

Case 1: When the mean should be 'NA' if there is any empty entry

```
mean(cars$cyl)
```

```
## [1] NA
```

Case 2: When the mean should be the mean of all non 'NA' values

```
mean(cars$cyl, na.rm = TRUE)
```

```
## [1] 6.26087
```