Homework Exercise 1

W203 K Iwasaki

January 9, 2016

W203 Statistics for Data Science

Unit 1 Homework

Exercise

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

```
setwd("C:/Users/K/Desktop/Berkeley/00_Academics/01_2017 Summer/W203 STATS/hw01")
data <- read.csv('cars.csv')

# check the data
head(data)
```

```
##    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
nrow(data)
```

```
## [1] 25
```

```
?cars
```

```
## starting httpd help server ...
```

```
##  done
```

1. What are the variables in the file?

```
str(data)
```

```
## 'data.frame':    25 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : int  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : int  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : int  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: int  4 4 1 1 2 1 4 2 2 4 ...
```

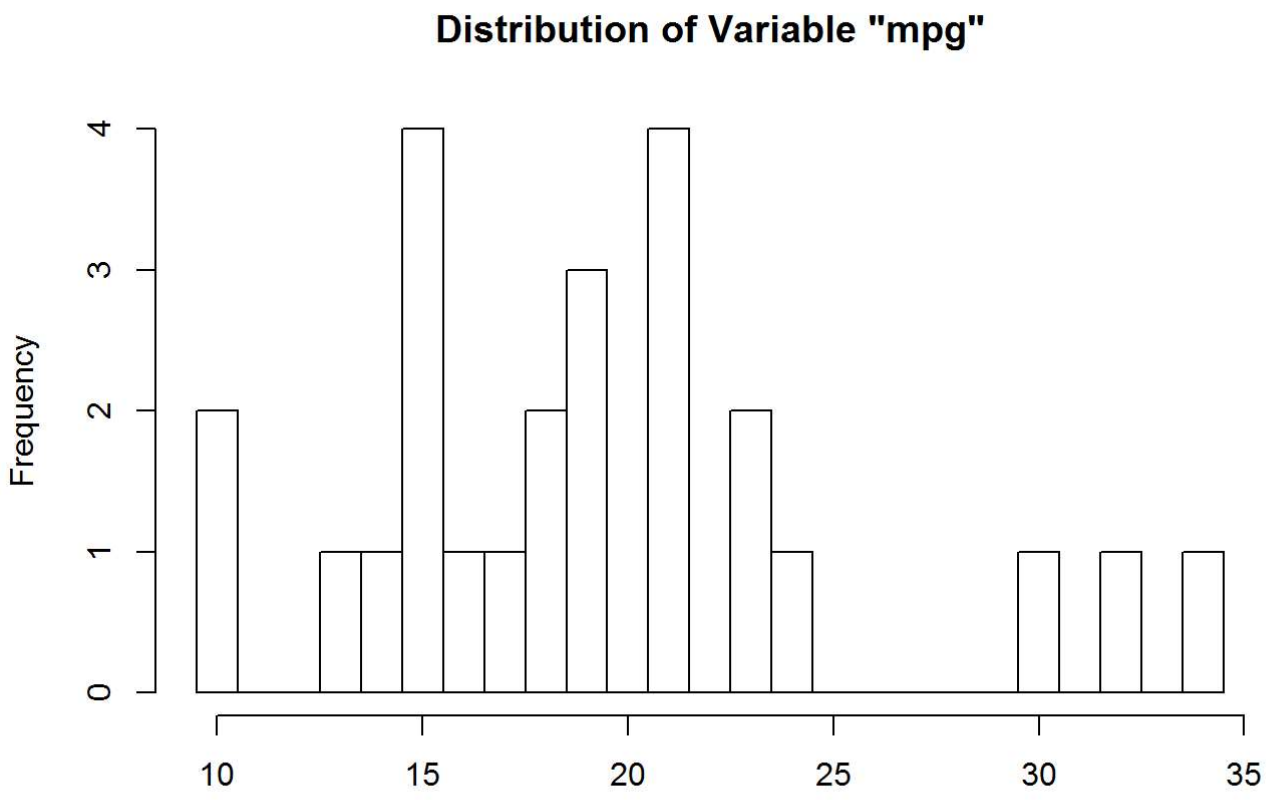
2. Find the mean, median, minimum, maximum, 1st quartile and 3rd quartile for the mpg variable.

```
summary(data$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(data$mpg, breaks = 10:35 - 0.5,
     main = 'Distribution of Variable "mpg"', xlab = NULL)
```



4. What is the standard deviation of mpg variable?

```
std = sd(data$mpg)
std
```

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

5. What is the variance of mpg variable?

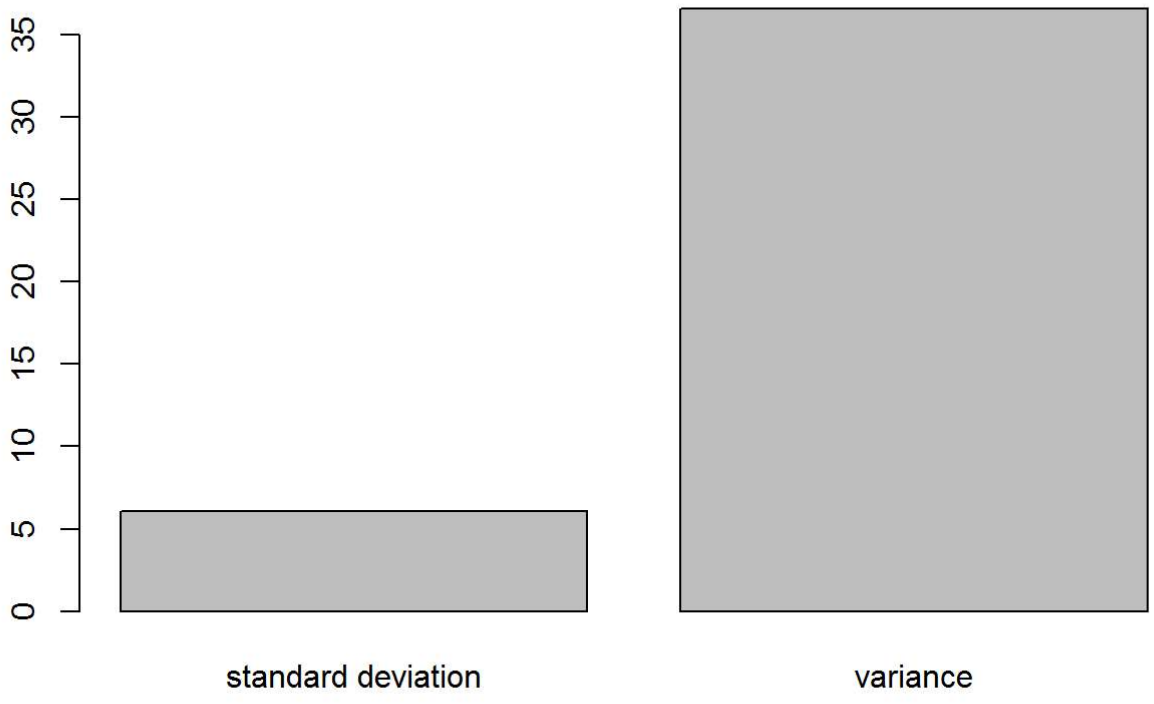
```
variance = var(data$mpg)
variance
```

```
## [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?

```
# create a table and compare the values visually
mpg_df <- data.frame(variable=c("standard deviation", "variance"), val = c(std, variance))

barplot(mpg_df$val, names.arg = mpg_df$variable)
```



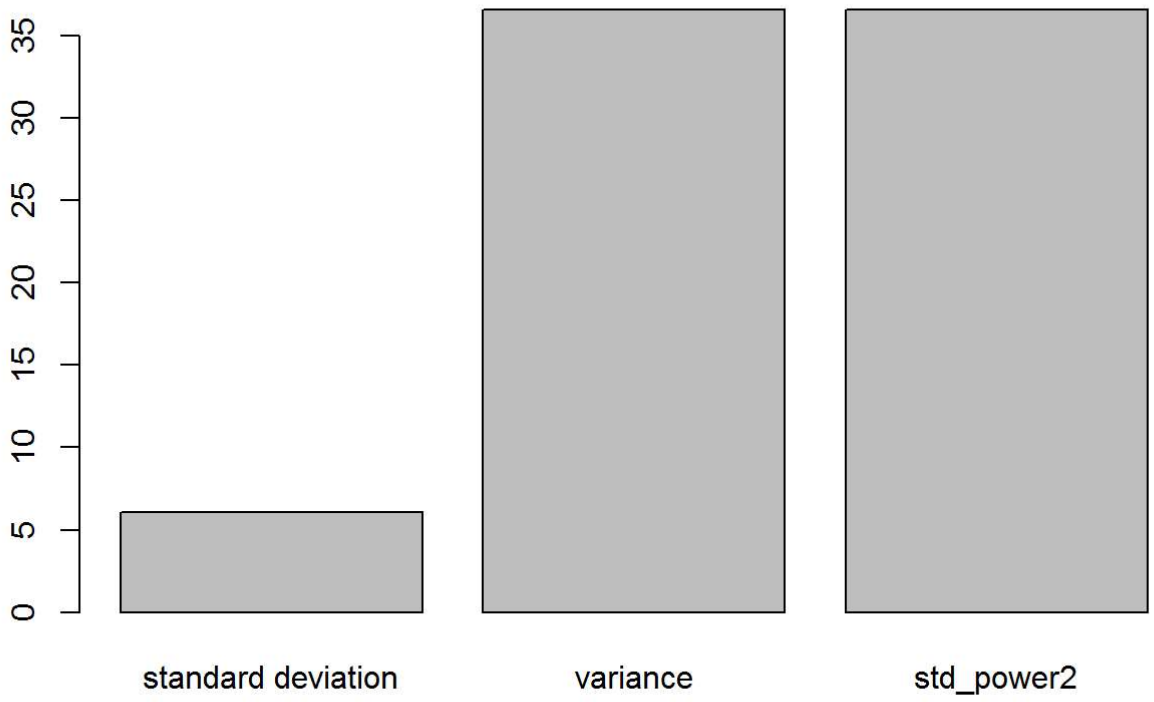

```
# The standard deviation is the square root of the variance by definition.
# Let's check if this is true here.

std_power2 <- std **2
check <- std_power2 == variance
check
```

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

```
# viualize this
new.row1 <- data.frame(variable = "std_power2", val = std_power2)
mpg_df <- rbind(mpg_df, new.row1)

barplot(mpg_df$val, names.arg = mpg_df$variable)
```



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

```
length(data$cyl)
```

```
## [1] 25
```

```
summary(data$cyl) # note there is two NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   4.000   4.000   6.000   6.261   8.000   8.000       2
```

8. What is the mean of the cyl variable?

```
mean(data$cyl, na.rm=T)
```

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