Lab4

Vergil

November 2, 2015

1.

A.

```
playboy <- read.csv("~/Desktop/playboy.csv",stringsAsFactors=FALSE)
playboy$ratio_wh <- playboy$Waist/playboy$Hips
```

B.

```
playboy$ratio_BMI <- playboy$Weight/(playboy$Height^2)
```

C.

```
mean(playboy$ratio_wh,na.rm = T)
```

```
## [1] 0.6745107
```

```
sd(playboy$ratio_wh,na.rm = T)
```

```
## [1] 0.03757309
```

```
mean(playboy$ratio_BMI,na.rm = T)
```

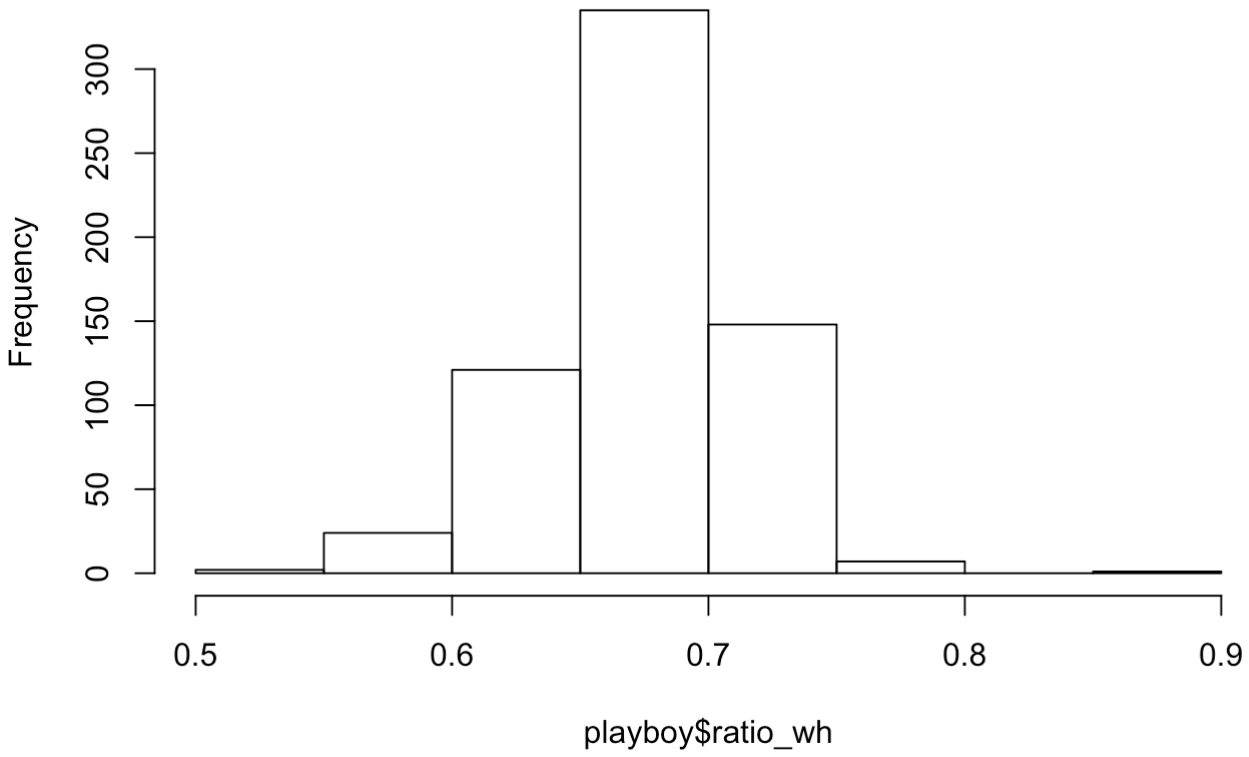
```
## [1] 0.02633712
```

```
sd(playboy$ratio_BMI,na.rm = T)
```

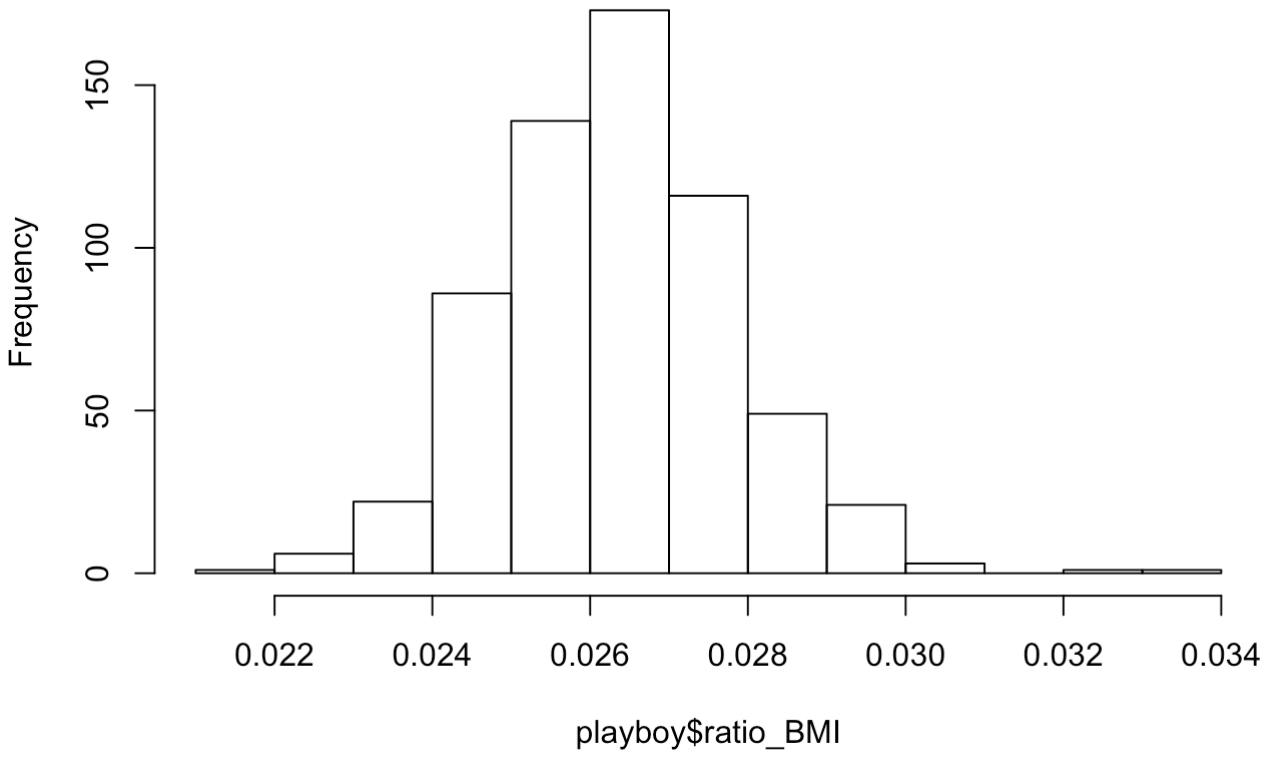
```
## [1] 0.001481924
```

D.

```
hist(playboy$ratio_wh)
```




```
hist(playboy$ratio_BMI)
```



The shape of the histogram for both hip ratio and BMI are pretty normal. Altough there might be a vague sign of right-skewed in the histogram of BMI ratio, but overall, it is still pretty normal. Both of them are centered around the mean

E.

```
x <- subset(playboy,Year<1980)
y <- subset(playboy,Year>=1980)
mean(x$ratio_BMI,na.rm = T)
```

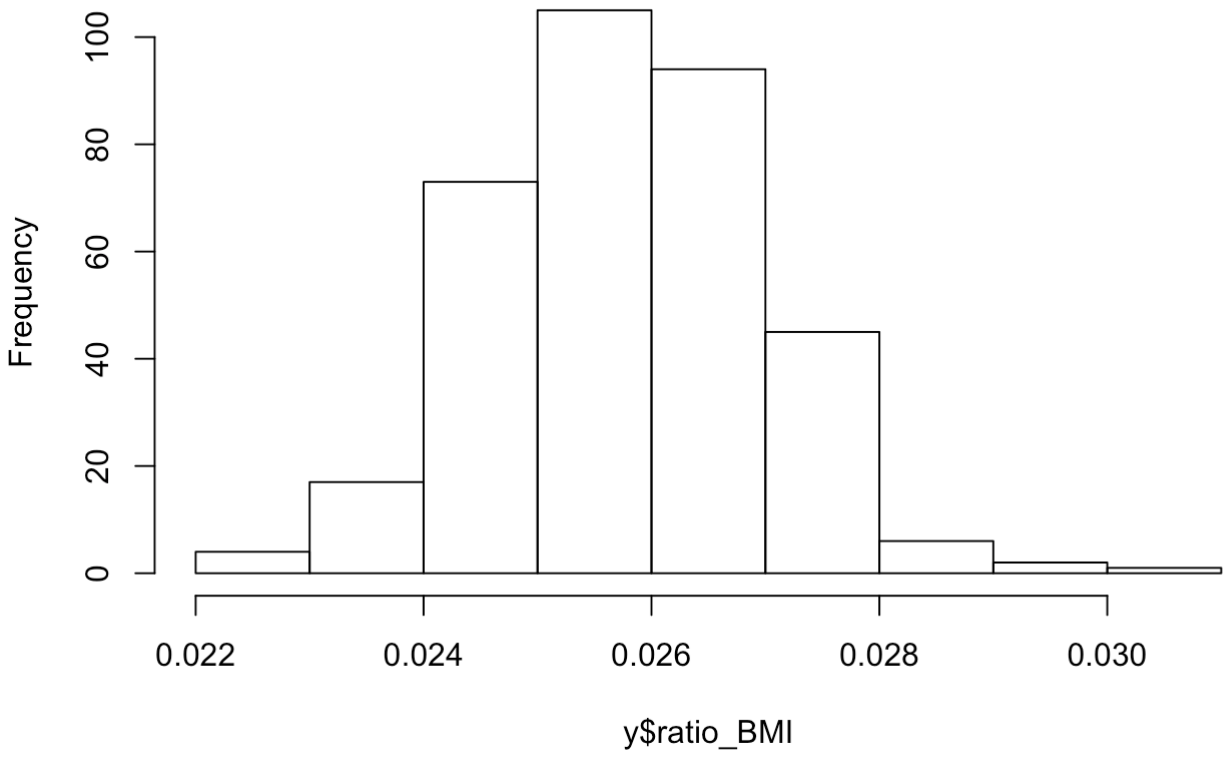
```
## [1] 0.02704221
```

```
mean(y$ratio_BMI,na.rm = T)
```

```
## [1] 0.02578647
```

```
hist(x$ratio_BMI, main="Histogram 1979-1960")
```


```
hist(y$ratio_BMI, main="Histogram 2006-1980")
```



Both the means and the plots suggest that BMI becomes smaller after 1980

2.

```
rent <- read.csv("~/Desktop/offcampus.csv",stringsAsFactors=FALSE)
char1 <- as.character(rent$Size)
char1 <- (strsplit(char1, split="br"))
char2 <- unlist(char1)

rent$rent <- gsub("\\$","",rent$Price)
rent$bedrooms <- gsub("/ ","",char2[seq(from=1, to=831, by=2)])
examp1 <- gsub(" - ", "", char2[seq(from=2,to=832,by=2)])
examp1 <- gsub(pattern = "ft2", replacement="", examp1 )
rent$sqft <- as.numeric(examp1)

rent <- rent[-c(3,4)]
```

A.

```
library(lubridate)
z <- ymd_hm(rent$Date)

sort(table(hour(z)),decreasing = T)
```

```
##
##  9 11  8 12 10 13 14 18 15 17 16  2  7 19 21 20  3  6 22  5 23  4  0  1
## 42 42 36 34 30 29 29 23 22 21 19 13 12 10 10  9  7  7  7  5  5  2  1  1
```

```
sort(table(wday(z)),decreasing = T)
```

```
##
##   2   7   1   5   6   4   3
## 126  78  72  46  40  28  26
```

9 and 11 is the most frequent hour. Monday(which in lubridate, is 2) is the most frequent day of week.

B.

```
junk <- unlist(strsplit(rent$Description, " "))
junk1 <- toupper(junk)
length(grep("UCLA",junk))
```

```
## [1] 41
```

```
length(grep("UCLA",junk,ignore.case = T))
```

```
## [1] 42
```

“UCLA” was mentioned 41 times. “UCLA”, ignoring its case, was mentioned 42 times.

C.

```
rent$Description <- gsub("Westwood Village","Westwood",rent$Description,ignore.case = T)
```

3.

A.

```
f1 <- c(rep("Wednesday", 9), rep("Friday", 91))
f1 <- factor(f1)
sort(table(f1))
```

```
## f1
## Wednesday    Friday
##         9        91
```

B.

```
library(XML)
urlname <- "http://www.boxofficemojo.com/yearly/chart/?yr=2015&p=.htm"
tables <- readHTMLTable(urlname, stringsAsFactors = FALSE )
df1 <- tables[[7]]
df1 <- df1[ -c(1,102:105),]
tablenames <- c("Rank", "Movie Title", "Studio", "TotalGross", "Theaters", "Opening", "T
heaters2", "Open", "Close" )
names(df1) <- tablenames

split1 <- gsub('\\$', '', df1$TotalGross)
split2 <- gsub('\\,', '', split1)
df1$TotalGross <- as.numeric(split2)

mean(df1$TotalGross)
```

```
## [1] 77387387
```

```
median(df1$TotalGross)
```

```
## [1] 42764323
```

```
sum(df1$TotalGross)
```

```
## [1] 7738738701
```