R Assignment 2

Akshita Gundavarapu

2021-01-27

```
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#library(ggplot2)
#library(RcmdrMisc)
knitr::opts_chunk$set(echo = TRUE)
SHOW_SOLUTIONS = TRUE
```

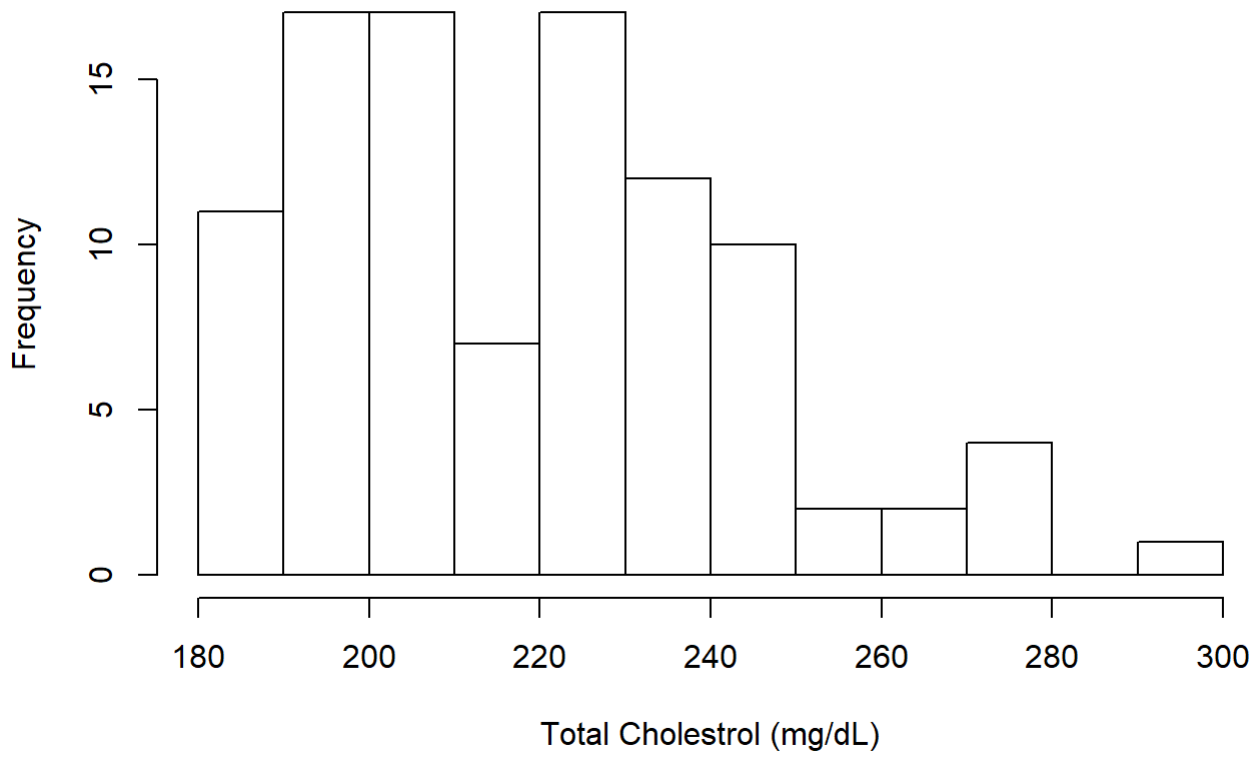
```
D.df <- read.csv("Patient_Data.csv", header=TRUE, as.is=TRUE)
D.df$Sex <- as.factor(D.df$Sex)
D.df$MaritalStat <- as.factor(D.df$MaritalStat)
```

Problem 1

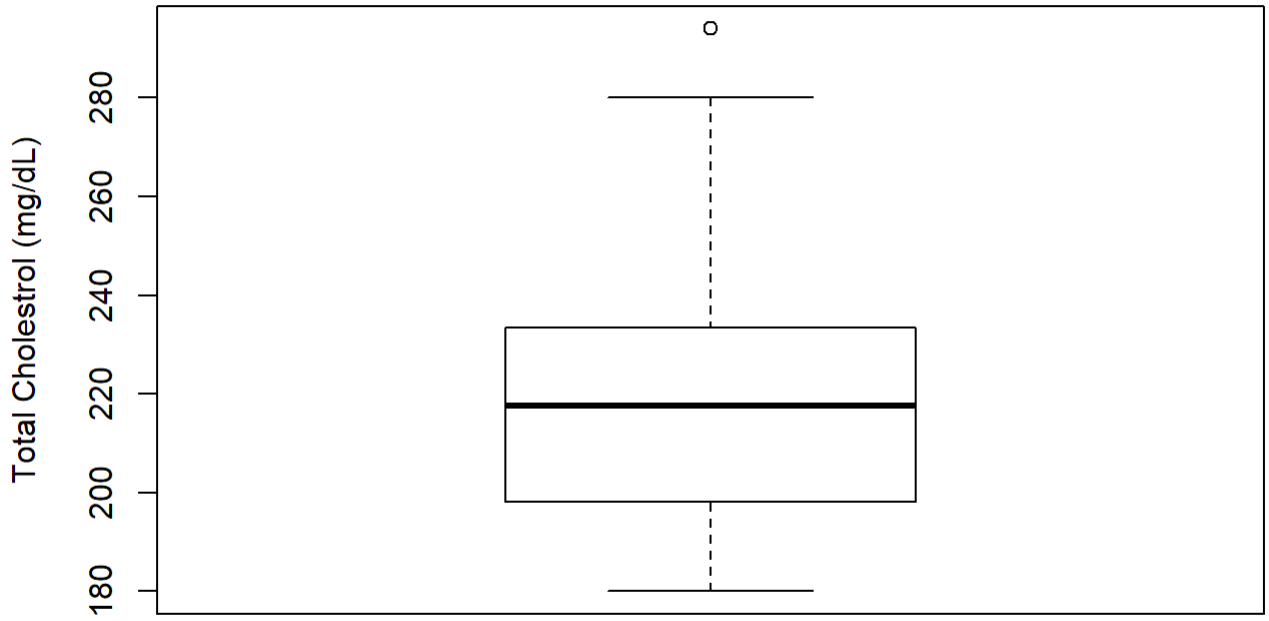
```
(sum.weight <- D.df %>% summarize(Min = min(TotChol, na.rm=TRUE),
                                  Q1 = quantile(TotChol, 0.25, na.rm=TRUE),
                                  Median = quantile(TotChol, 0.50,
                                                    na.rm=TRUE),
                                  Q3 = quantile(TotChol, 0.75, na.rm=TRUE),
                                  Max = max(TotChol, na.rm=TRUE),
                                  Mean = mean(TotChol, na.rm=TRUE),
                                  Standard_deviation = sd(TotChol, na.rm=TRUE)))
```

```
##   Min  Q1 Median     Q3 Max   Mean Standard_deviation
## 1 180 198  217.5 233.25 294 219.27           25.10497
```

```
hist(D.df$TotChol, main="", xlab="Total Cholestrol (mg/dL)", breaks = 10)
```



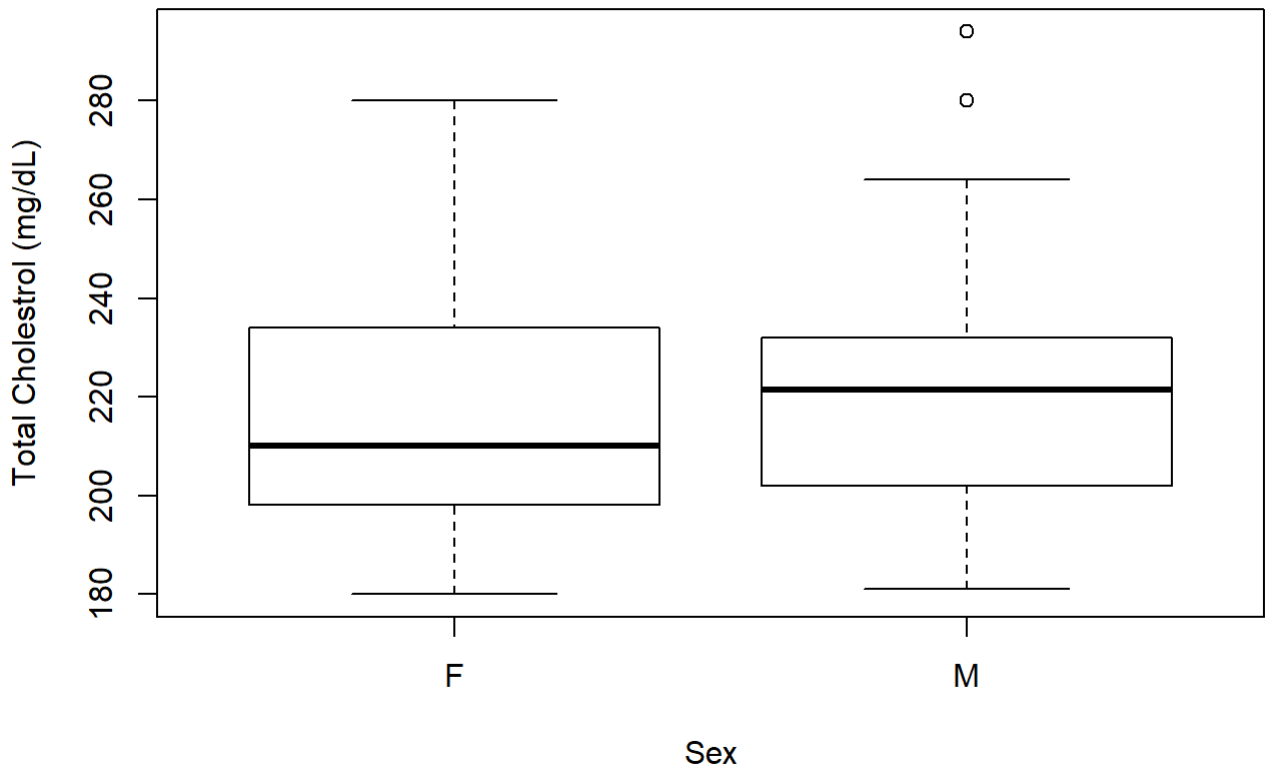

```
boxplot(D.df$TotChol, ylab="Total Cholestrol (mg/dL)")
```



From the 7 summary table, histogram and boxplot the distribution seems to be bimodal and right skewed. We can tell that it is right skewed because the mean (219.27 mg/dL) is greater than the median (217.5). Moreover, the data ranges from 184 mg/dL to 294mg/dL, so it has a range of 110. The data is also spread with a standard deviation of 25.1 mg/dL is and we also have one upper boundry outlier which is the data value of 294 mg/dL.

Problem 2

```
boxplot(D.df$TotChol ~ D.df$Sex, ylab="Total Cholestrol (mg/dL)",
        xlab = "Sex")
```



The comparative box plot shows that the median total cholestrol of females is much lower than the median total cholestrol of the males. The total cholestrol of females is also more spread out than the total cholestrol of the males. Moreover, the males data shows two ouliers which fall over the upper boundry, whereas the females data has no outliers.

Problem 3

```
(tab1 <- table(D.df$Sex, D.df$MaritalStat))
```

```
##
##      D  M  S  W
##   F 11 20  9 10
##   M 13 14 17  6
```

```
prop.table(tab1)*100
```

```
##
##      D  M  S  W
##   F 11 20  9 10
##   M 13 14 17  6
```

Table of counts and table of joint percentages have the same values. This is because the total number of datapoints is 100, so the percentages and counts are esentially the same.

Problem 4

```
marginal_percentage_D <- ((11+13)/100)*100
marginal_percentage_M <- ((20+14)/100)*100
marginal_percentage_S <- ((9+17)/100)*100
marginal_percentage_W <- ((10+6)/100)*100

marginal_percentages <- c(marginal_percentage_D, marginal_percentage_M, marginal_percentage_S, m
arginal_percentage_W)

marginal_percentages
```

```
## [1] 24 34 26 16
```

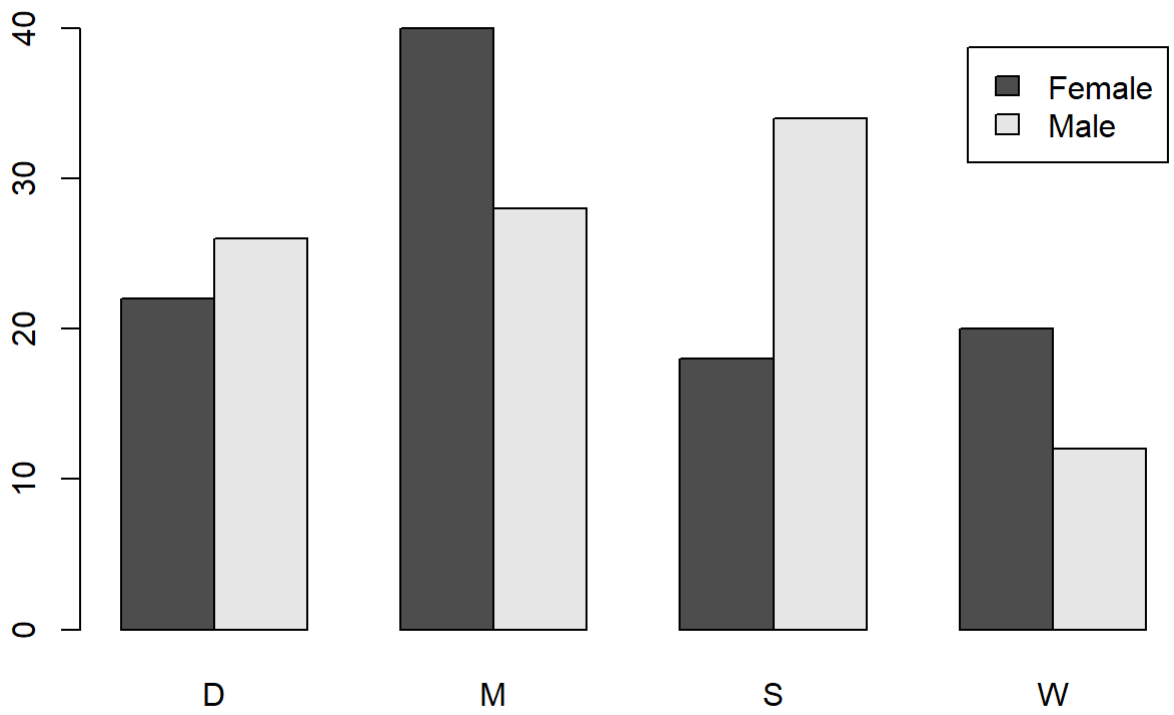
Problem 5

```
tab2 <- (Condrow.pct <- round(prop.table(tab1, margin=1) * 100, 1))
tab2[1,]
```

```
##  D  M  S  W
## 22 40 18 20
```

Problem 6

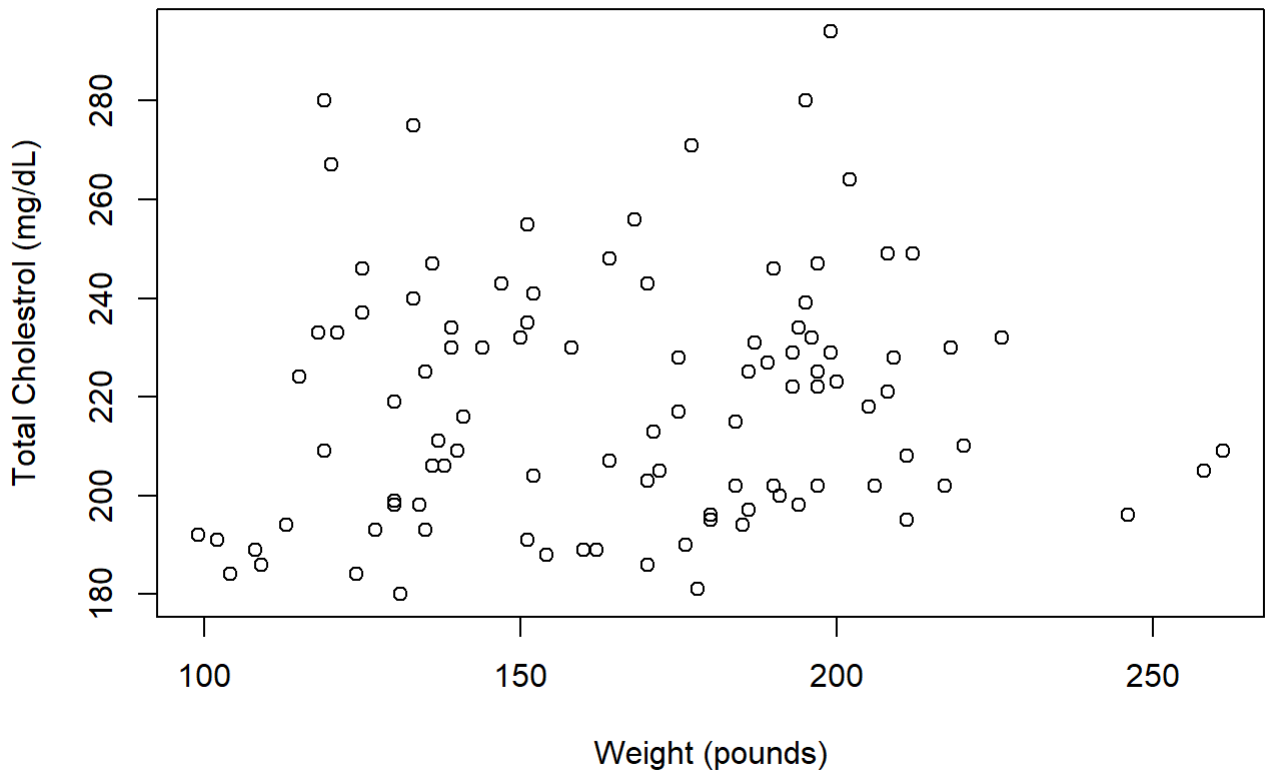
```
barplot(tab2, legend.text=c("Female","Male"), beside=TRUE)
```



From the bar plot it is evident that male data and female data do not show the same patterns in all 4 marital statuses. For example, there are more males who are divorced compared to females, however there are more females who are married than male. There are also more single males than females, however more widowed females than males. This shows that the patterns in males and males is not consistent in all 4 marital status categories, So we can say that marital status and sex appear to be statistically dependent.

Problem 7

```
#a)
plot(D.df$Weight, D.df$TotChol, xlab="Weight (pounds)", ylab="Total Cholestrol (mg/dL)")
```




```
#b)
cor(D.df$TotChol, D.df$Weight)
```

```
## [1] 0.09613236
```

```
#c)
lm.out <- lm(D.df$Weight ~ D.df$TotChol)
summary(lm.out)
```

```
##
## Call:
## lm(formula = D.df$Weight ~ D.df$TotChol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.235 -30.072   3.807  25.720  95.399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  136.5157    32.1222   4.250 4.88e-05 ***
## D.df$TotChol   0.1392     0.1456   0.956    0.341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.36 on 98 degrees of freedom
## Multiple R-squared:  0.009241,   Adjusted R-squared:  -0.0008684
## F-statistic: 0.9141 on 1 and 98 DF,  p-value: 0.3414
```

The scatter plot shows appears to show no correlation

The correcation coeeficient was found to be 0.09 which is very close to zero. This can be interpretted as no correcation or a very slight positive correlation.

The regression parameter estimates for total cholesterol on weight are b0 = 136.52 and b1 = 0.139

d. regression equation Total Cholestrol = 136.52 + 0.139(Weight)

e. The regression slope parameter is estimated to be 0.139. This would mean that for every pound increase in Weight, there is a 0.139 mg/dL increase in Total Cholestrol.

f. I don’t think this regression model is very accurate. From the scatterplot we can see that the data has no correlation so a linear regression model may not be a right fit for this data instead a quadratic or cubic model may make more sense.