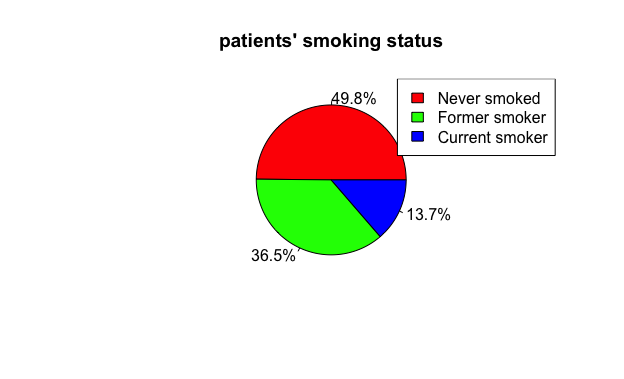
MATH-3042 HW1

Student Number: A01017545

Name: Jihyo Kim

Set: 3M

**1.**

> t = table(NutritionStudy$PriorSmoke)

> percent\_labels = round(100\*t/sum(t), 1)

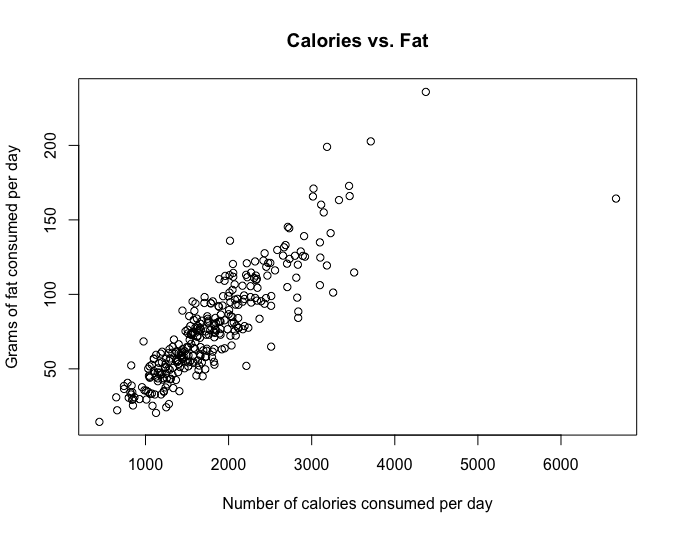
> pielabels<- paste(percent\_labels, "%", sep=“")

> cols = c('red', 'green', 'blue')

> pie(t, labels=pielabels, main="patients' smoking status", col=cols)

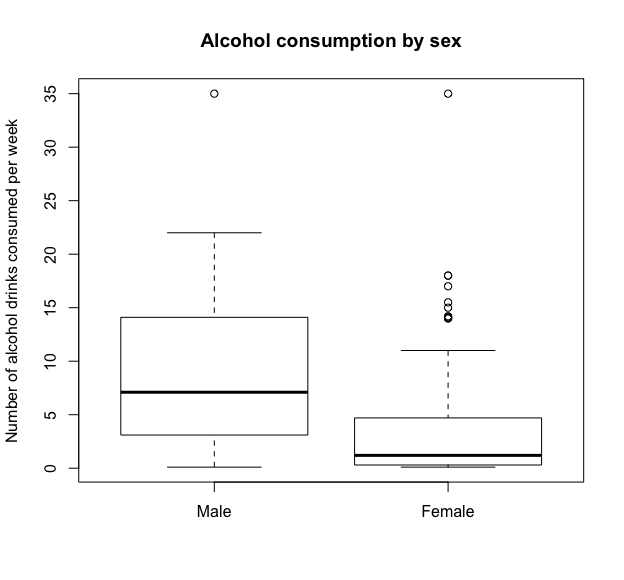
> legend("topright", c('Never smoked', 'Former smoker', 'Current smoker'), fill = cols)

**2.**



> plot(NutritionStudy$Calories, NutritionStudy$Fat, ylab="Grams of fat consumed per day", xlab="Number of calories consumed per day”)

1. There are two noticeable outliers. The one has around 4300 calories consumed per day and the other one has around 7000 calories consumed per day.
2. To answer it with a single number, it’s 80g of fat, and it seems that a person who consume 2000 calories per day consumes between 50g and 100g of fat per day. This is because according to the scatter plot above, people who consume 2000 calories are likely to appear between 50g and 100g of fat consumption.

**3.**

> males = filter(NutritionStudy, Sex == 'Male')

> females = filter(NutritionStudy, Sex == "Female")

> males = filter(males, Alcohol != 0)

> females = filter(females, Alcohol != 0)

> males = filter(males, Alcohol < 200)

> boxplot(males$Alcohol, females$Alcohol, names=c("Male", "Female"), ylab="Number of alcohol drinks consumed per week", main="Alcohol consumption by sex”)

According to the box plot, around 50% of the number of alcohol drinks consumed by men per week is ranged between 3.5 and 14, while the one for women is only ranged between 1 and 4. However, it is noticeable that there are more outliers for women than men.

**4.**

> favstats(men$Calories)

min Q1 median Q3 max mean sd n missing

827.9 1721.325 2023.6 2604.55 6662.2 2155.786 916.5653 42 0

> favstats(women$Calories)

min Q1 median Q3 max mean sd n missing

445.2 1305.4 1628.5 2052.4 4373.6 1741.404 620.2701 273 0

> favstats(men$Fiber)

min Q1 median Q3 max mean sd n missing

4.7 10.025 12.1 16.725 26.3 13.41429 4.84348 42 0

> favstats(women$Fiber)

min Q1 median Q3 max mean sd n missing

3.1 8.9 12.1 15.1 36.8 12.69231 5.402966 273 0

1. Overall, male patients consumed more calories per day compared to female patients. as the mean of daily calories consumption of male patients is 2122.064 and the one for female patients is 1793.569. When it comes to the variation, the consumptions of the male patients are less variable than the ones of the female patents, as the coefficient of variation of male patients is 0.3014991 and the one of the female patients is 0.36764.
2. Overall, male patients consumed more fibre per day compared to the female patients, as the mean of daily fibre consumption of the male patients is 13.5, and the one of female patients is 13.02079. When it comes to the variation, the finer consumptions of the male patients are less variable than the ones of the female patients, as the coefficient of variation of male patients is 0.2878424 and the one of the female patients is 0.4272725

**5.**

There are two methods to calculate outliers: using IQR and using z-scores.

Using IQR:

> iqrbmi = IQR(NutritionStudy$BMI)

> q1bmi = quantile(NutritionStudy$BMI, 0.25)

> q3bmi = quantile(NutritionStudy$BMI, 0.75)

> low\_outliers = filter(NutritionStudy, BMI < q1bmi - 1.5 \* iqrbmi)

> high\_outliers = filter(NutritionStudy, BMI > q3bmi + 1.5 \* iqrbmi)

Using z-scores:

> mean\_bmi = mean(NutritionStudy$BMI)

> sd\_bmi = sd(NutritionStudy$BMI)

> zscores = abs(NutritionStudy$BMI - mean\_bmi) / sd\_bmi

> outliers\_zscores = zscores[zscores > 2]

Result:

Using IQR, there are 15 outliers, whereas, using z-scores, there are 18 outliers.

**6.**

To compare the two distributions, I calculated the coefficient of variation and the Pearsonian coefficient of Skewness. The commands I used to calculate the two values are following:

> favstats\_males\_fat = favstats(males$Fat)

> favstats\_females\_fat = favstats(females$Fat)

> cod\_males\_fat = favstats\_males\_fat$sd / favstats\_males\_fat$mean \* 100

> cod\_females\_fat = favstats\_females\_fat$sd / favstats\_females\_fat$mean \* 100

> sk\_males\_fat = 3 \* (favstats\_males\_fat$mean - favstats\_males\_fat$median) / favstats\_males\_fat$sd

> sk\_females\_fat = 3 \* (favstats\_females\_fat$mean - favstats\_females\_fat$median) / favstats\_females\_fat$sd

cod\_males\_fat: 35.8447

cod\_females\_fat: 44.55007

sk\_males\_fat: -0.0498762

sk\_females\_fat: 0.5463848

As a result, the fat consumption for the male patients is almost symmetric (but very slightly skewed to the left), as the Pearsonian coefficient of skewness is -0.0498, while the fat consumption for the female students is seemingly symmetric (but more skewed than the male patients, and skewed to the right), as the Pearsonian coefficient of skewness is 0.546.