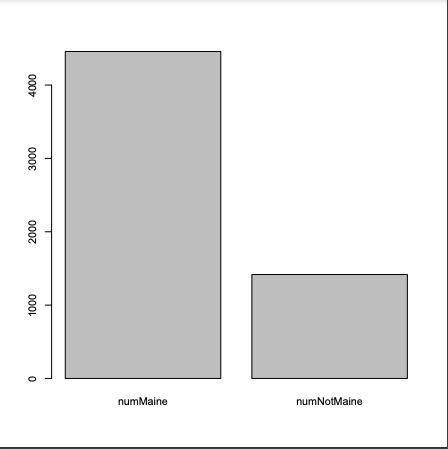
Mini Project #2

Sudarshana Jagadeeshi

Contribution of each group member: I completed the project in full

**Section 1**

1.



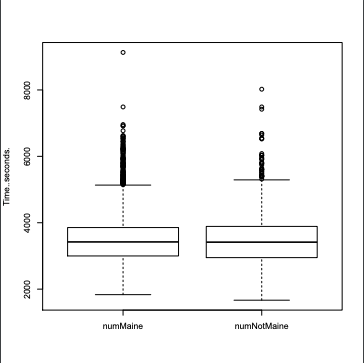
From this barplot, we can see that the number of Maine racers outnumber the number of non-Maine racers by about 3:1. We can verify this because when we print the raw count in R, we get 4458 for numMaine and 1417 for numNotMaine.

**b)**

|  |  |
| --- | --- |

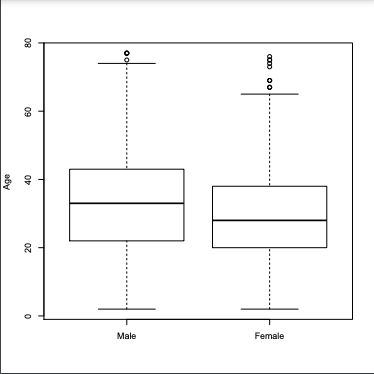
We can conclude that the distributions are very similar. They both are right-skewed, but timesNotMaine peters out slower, which is confirmed by its SD of 830, compared to Maines 731. They both have their median around 3.5k seconds. The bulk of the racers finish between 2.5k and 4.5k seconds in both distributions. A close look at the quartiles will show that notMaine has a slightly higher IQR- around 3% more.

**c)**



We can see some of what we discussed above. Notice that the numNotMaine box is fatter.

**d)**



**Summary Statistics**

Male

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 22.00 33.00 32.56 43.00 77.00

Female

Min. 1st Qu. Median Mean 3rd Qu. Max.

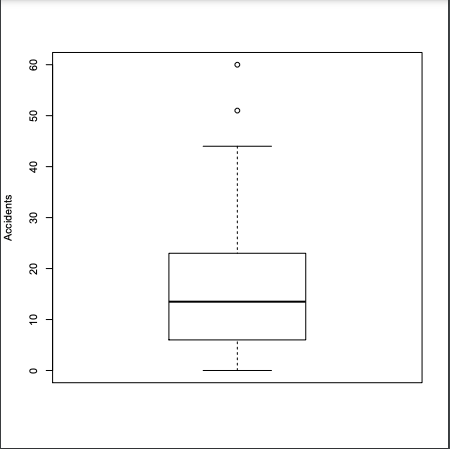
2.00 20.00 28.00 29.26 38.00 76.00

**Analysis**

We can see that the female ages are on the whole lower, as all of the quartiles are lower than the males. They are also lower in SD than the males(12.2 v 14.1) and have a lower range of ages. It is considered statistically unlikely to compete as a male over 75, and as a female over 65.

**2.**

**Boxplot:**



**Summary Statistics:**

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 6.00 13.50 17.02 23.00 60.00

**Feature Discussion:**

We see a mean of about 17, and an IQR of about 17 as well. That large IQR means this distribution probably has a larger SD than a typical normal distribution. We notice that the median is less than the mean, which indicates a right-skewed distribution. There are two high outliers, and no lower outliers.

**Outlier Counties and why:**

Horry with 60 and Greenville with 51 are the two outliers indicated on the above boxplot. The reasons for this may be varied- it could just be that the two counties have higher population and therefore more motorcycles. We would need to further analyze the rate of accidents/population to know if factors other than population are contributing to accidents, such as poor roads, poor enforcement, etc.

**Section 2**

**Question 1**

**rrdata <- read.csv("roadrace.csv")**

**#===============PART A=================**

**notMaine <- subset(rrdata, Maine=="Away") #keep all tuples with Maine=="Away"**

**numNotMaine <- nrow(notMaine) #how many non-mainers?**

**print(numNotMaine)**

**Maine <- subset(rrdata, Maine=="Maine")**

**numMaine <- nrow(Maine)**

**print(numMaine)**

**barplotvec <- c(numMaine, numNotMaine) #a vector of two values**

**mynames <- c("numMaine", "numNotMaine") #so the plot is labeled**

**barplot(barplotvec, names=mynames)**

**#print(colnames(Maine))**

**#=================PART B=================**

**timesMaine <- Maine[ ,"Time..seconds."] #keep all rows, take only time column**

**histMaine <- hist(timesMaine, xlim= c(1500, 10000))**

**timesNotMaine <- notMaine[, "Time..seconds."]**

**histNotMaine <- hist(timesNotMaine, xlim= c(1500, 10000)) #ensuring same range**

**#=================PART C=================**

**bp <- boxplot(timesMaine, timesNotMaine, names= mynames, ylab="Time..seconds.")**

**#=================PART D=================**

**male <- subset(rrdata, Sex=="M")**

**female <- subset(rrdata, Sex=="F")**

**male\_ages <- as.numeric(male[, "Age"])#as.numeric b/c R throws error otherwise**

**female\_ages <- as.numeric(female[, "Age"])**

**agehist <- boxplot(male\_ages, female\_ages, names=c("Male", "Female"), ylab= "Age")**

**print(summary(male\_ages)) #print 5-num summary, sd(vec) returns the sd**

**print(summary(female\_ages))**

**Question 2**

**motodata <- read.csv("motorcycle.csv") #stores in R dataframe**

**print(head(motodata)) #first few entries**

**accidents <- motodata[, "Fatal.Motorcycle.Accidents"] #all rows, only accidents column**

**print(summary(accidents)) #5-num summary**

**boxplot(accidents, ylab="Accidents") #default orientation vertical**

**#hist(accidents)**