# STATISTICAL METHODS FOR DATA SCIENCE CS6313-001 FALL 2019

**Mini Project #2**

**Participant : Reetish Chand Guntakal Patil (RXG190006)**

## Question 1:

(12 points) Consider the dataset roadrace.csv posted on eLearning. It contains

observations on 5875 runners who \_nished the 2010 Beach to Beacon 10K Road Race

in Cape Elizabeth, Maine. You can read the dataset in R using read.csv function.

(a) Create a bar graph of the variable Maine, which identi\_es whether a runner is

from Maine or from somewhere else (stated using Maine and Away). You can

use barplot function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.

(b) Create two histograms the runners' times (given in minutes) | one for the

Maine group and the second for the Away group. Make sure that the histograms

on the same scale. What can we conclude about the two distributions? Back

up your conclusions with relevant summary statistics, including mean, standard

deviation, range, median, and interquartile range.

(c) Repeat (b) but with side-by-side boxplots.

(d) Create side-by-side boxplots for the runners' ages (given in years) for male and

female runners. What can we conclude about the two distributions? Back up

your conclusions with relevant summary statistics, including mean, standard

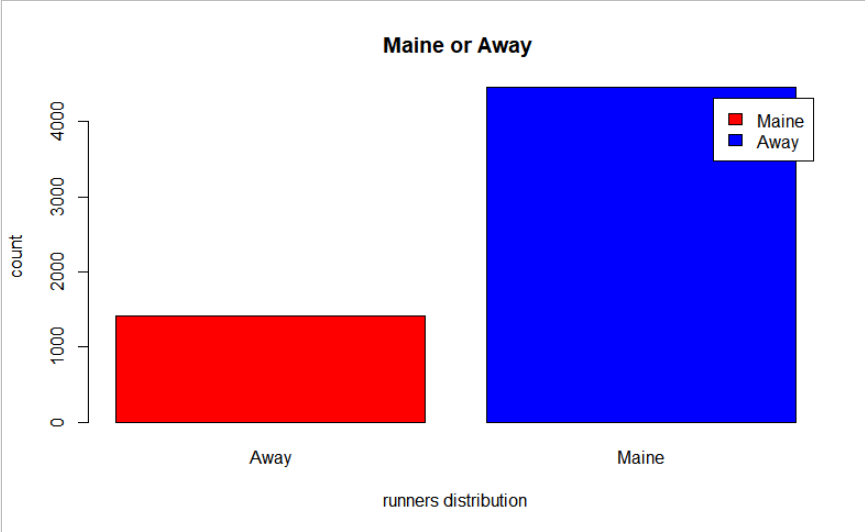
deviation, range, median, and interquartile range.

Reading data from “roadrace.csv” file:

* *data = read.csv("roadrace.csv")*

1. Creating a bar graph for variable “Maine”

* *barplot(table(data$Maine),main = " Maine or Away",xlab="runners distribution",ylab="cou nt",col=c("Red","Blue"),legend=name)*



Observations:

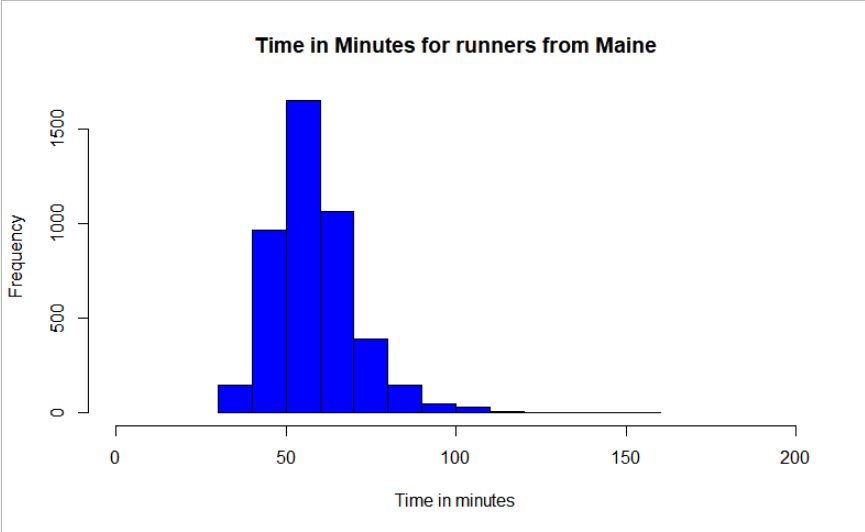
No. of runners who are away = 1400(approx.)

No. of runners who are from Maine = 4400(approx.) Total no. of runners = 5800(approx.)

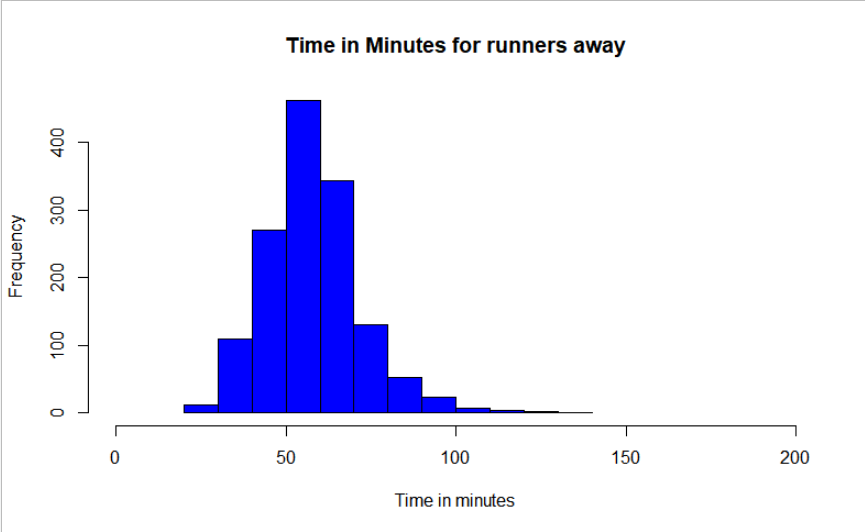
The no. of runners in the two categories stated above are in 1:3 ratio

1. Creating two histograms the runners’ times (given in minutes)

* *d1 = filter(data,data$Maine=="Maine")*
* *d2 = filter(data,data$Maine=="Away")*
* *hist(d1$Time..minutes.,main="Time in Minutes for runners from Maine", xlab="Time in minutes",ylab = "Frequency", xlim = c(0,200), col="Blue")*

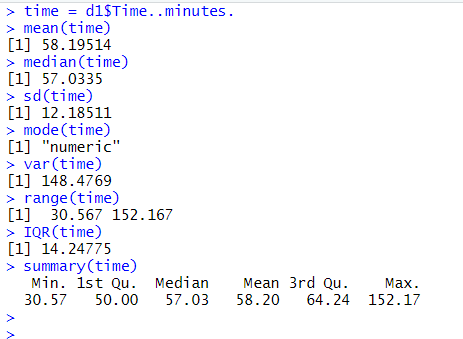


* *hist(d2$Time..minutes.,main="Time in Minutes for runners away", xlab="Time in minutes", ylab = "Frequency", xlim = c(0,200),col="Blue")*

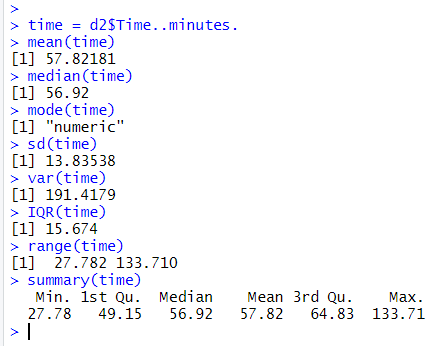


Observations :

The runners away from Maine took less time in minutes comparatively. Statistics for runners form Maine:

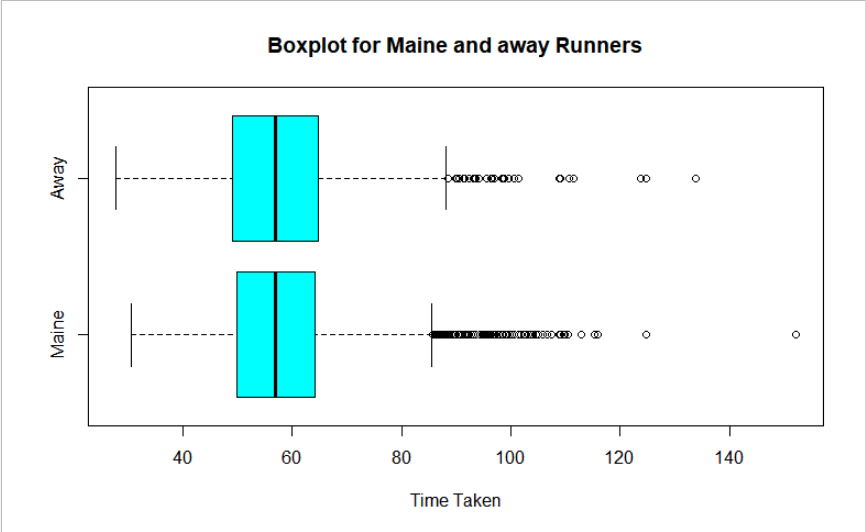


Statistics for runners away from Maine:



1. Repeating (b) but with side-by-side boxplots.

* *name=c("Maine","Away")*
* boxplot(d1$Time..minutes.,d2$Time..minutes.,horizontal = TRUE,names=name,main = "Bo xplot for Maine and away Runners",xlab="Time Taken",col="cyan")

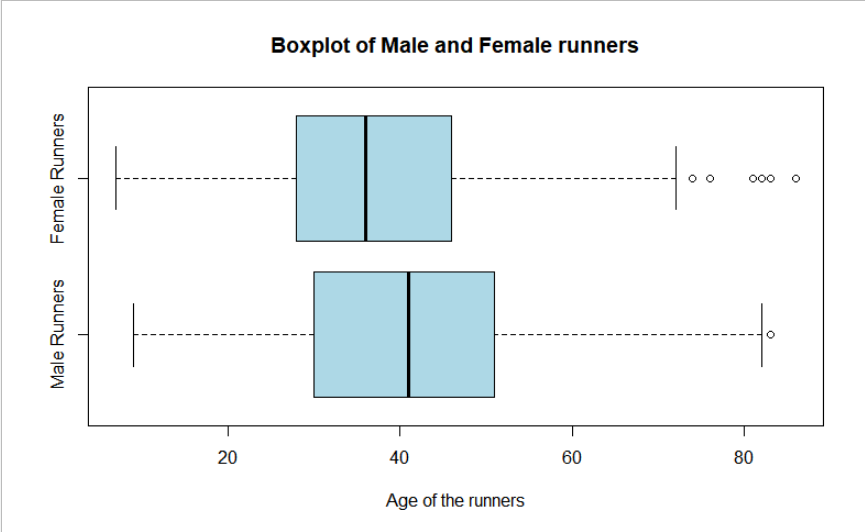


Observation:

No. of outliers in the Maine plot are greater than those away from Maine.

1. Creating side-by-side boxplots for the runners’ ages for male and female runners.

* *d3 = filter(data,data$Sex=="M")*
* *d4 = filter(data,data$Sex=="F")*
* *boxplot(d3$Age,d4$Age,horizontal = TRUE,names=gender,main = "Age of Male and Femal e runners",xlab="Age of runners",col="lightblue")*

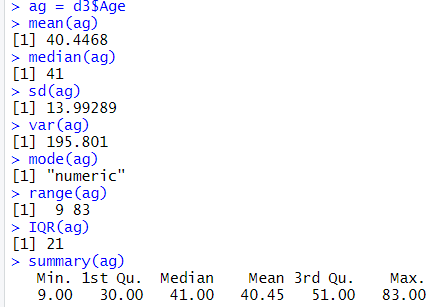


Observation :

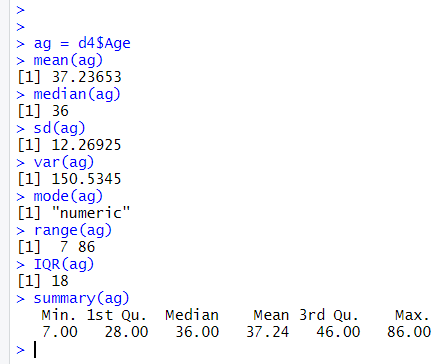
Median of female runners is less than male runners Range of female runners is more than male runners

No of outliers for female runners are more than male runners

Statistics for male runners:



Statistics for female runners:



## Question 2

(8 points) Consider the dataset motorcycle.csv posted on eLearning. It contains

the number of fatal motorcycle accidents that occurred in each county of South Car-

olina during 2009. Create a boxplot of data and provide relevant summary statistics.

Discuss the features of the data distribution. Identify which counties may be con-

sidered outliers. Why might these counties have the highest numbers of motorcycle

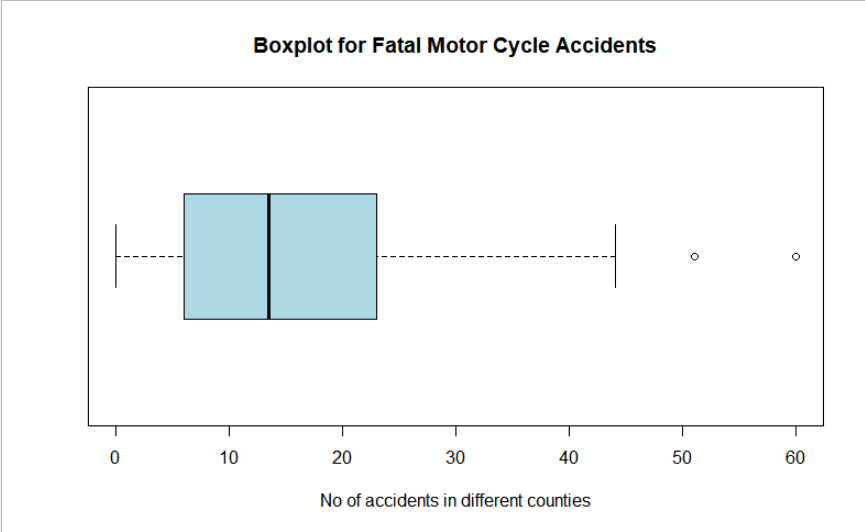
## fatalities in South Carolina?

Reading the csv file

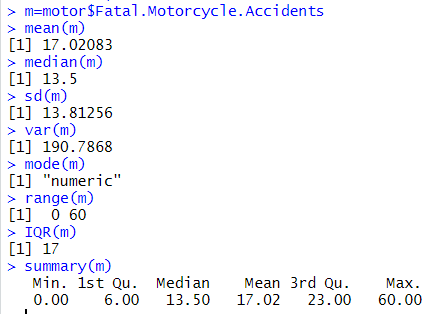
* *motor = read.csv("motorcycle.csv")*

Creating the boxplot :

* *boxplot(motor$Fatal.Motorcycle.Accidents,horizontal = TRUE,col="lightblue",main = "Boxp lot for Fatal Motor Cycle Accidents",xlab="No of accidents in different counties")*



Observation: Statistics



* + As the mean is greater than the mode the histogram generated must be right- skewed
  + There are two outliers with number of accidents 51in Greenville and 60 in Horry
  + From the statistics above the most traffic deaths occurred in Greenville followed by Horry followed by Richland and Charleston
  + Uneducated drivers and lenient traffic regulations might be one of the reasons the traffic deaths in these countries. B road conditions might also be a reason for such deaths.
* *hist(motor$Fatal.Motorcycle.Accidents,col = "lightblue",main="Histogram for Motorcycle a ccidents", xlab="Number of accidents", ylab = "Frequency", xlim = c(0,100),border="Black")*

