Statistical Methods for Data Science
CS 6313.001: Mini Project #2

Due on Tuesday February 26, 2019 at 10am

Instructor: Prof. Min Chen

Shyam Patharla (sxp178231)

Contents

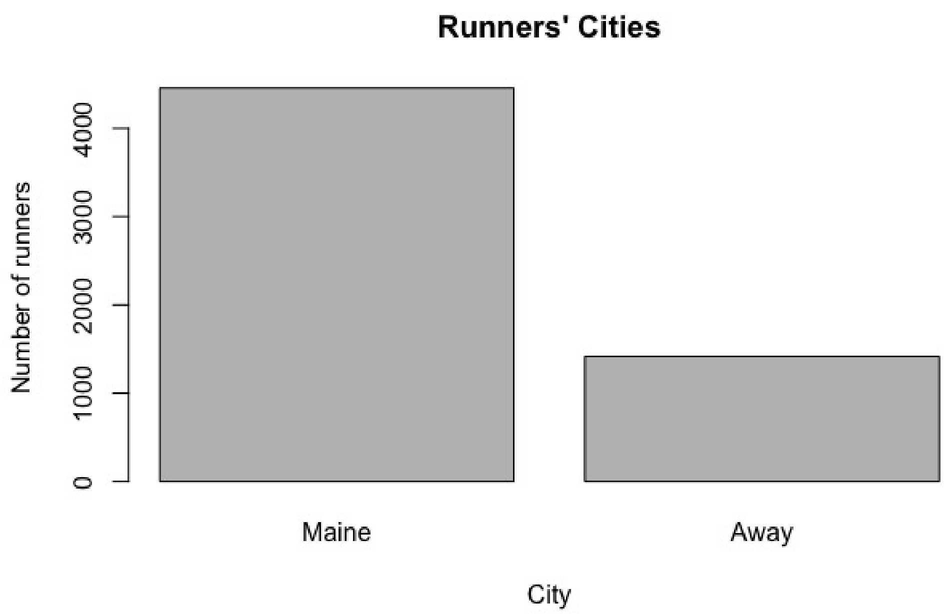
Section 1 Answers

Problem 1.1

(a) Create a bar graph for the variable Maine

1. Read the *roadrace.csv* file.
2. Get the tuples of runners who are from Maine. Store them in the vector **from.maine**.
3. Get the tuples of runners who are not from Maine(Away). Store them in the vector **not.from.maine**.
4. Get the number of rows in from.maine and not.from.maine and store them in a vector.
5. Barplot the vector.

We get the following barplot:



Inferences:

1. The number of runners from Maine is quite high compared to the number of Runners not form Maine.
2. The number of runners from Maine (4458) is cloose to thrice the number of runners not from Maine (1417).
3. This is reasonable since the number of participants from the city in which marathon is conducted is of often more.

(b) Histograms for Runner's Times - Maine and Away

1. We plot the histogram of the values in the 12th column of the **from.maine** vector i.e. the runners' time (minutes) for the runners from Maine.

2. We plot the histogram of the values in the 12th column of the **not.from.maine** vector i.e. the runners' time (minutes) for the runners not from Maine.

We get the following two histograms:

We can infer the follwing:

1. The histogram in the case of Maine runners looks right-skewed. It has 2 bars to the left of its highest bar and 5 bars to the right of its highest bar.

2. The histogram in the case of Maine runners, though it looks like a normal distribution, is also right-skewed. It has 3 bars to the left of its highest bar and 5 bars to the right of its highest bar.

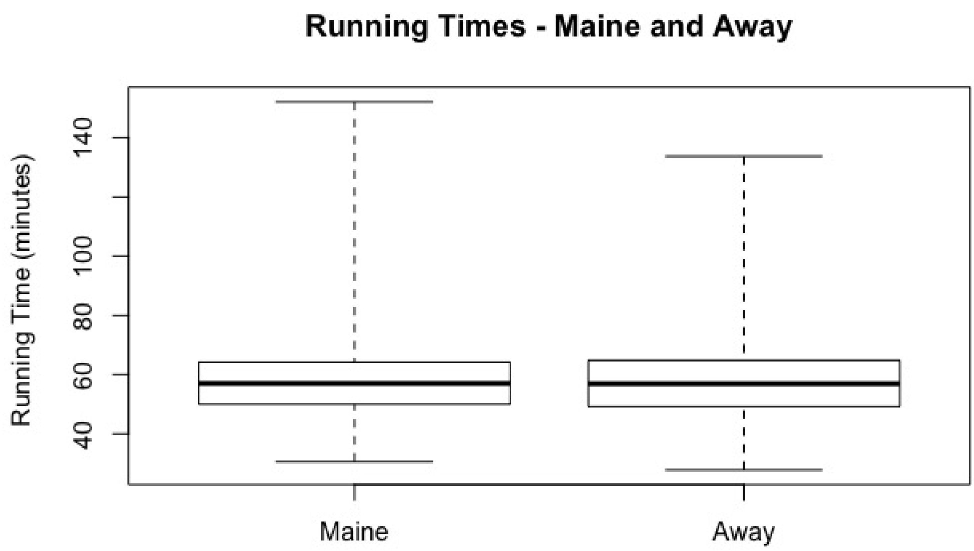
3. We can make more inferences after drawing the boxplots for the 2 cases as is shown in the next question.

(c) Boxplots for Runners' Times - Maine and Away

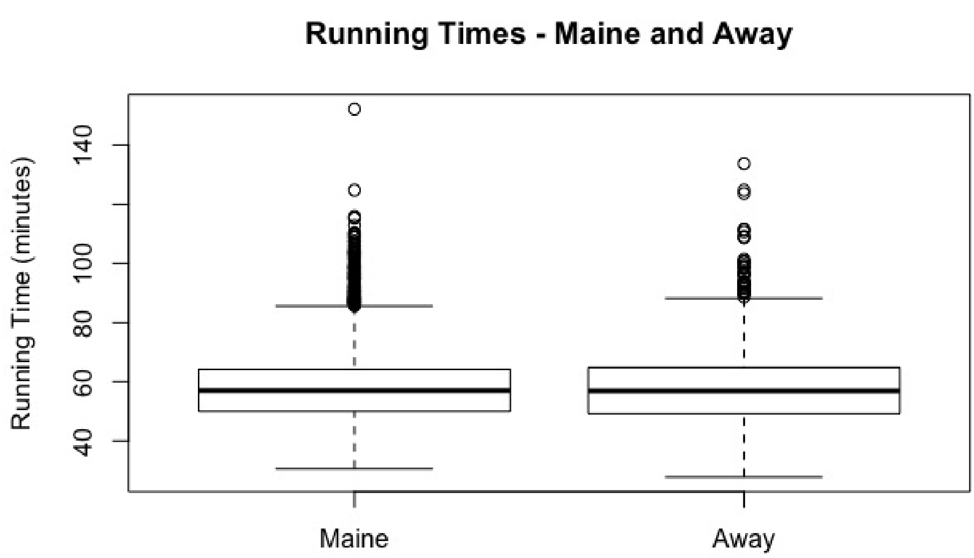
1. We boxplot the values in the 12th column of the **from.maine** vector i.e. the runners' time (minutes) for the runners from Maine.

2. We boxplot the values in the 12th column of the **not.from.maine** vector i.e. the runners' time (minutes) for the runners not from Maine.

Without using the 1.5 * IQR rule, we get the following boxplot:



Using the 1.5 * IQR rule, we get the following boxplot:



From the above boxplots, we can infer the following:

1. The Maine and Away distributions both have a **median** value close to 60 as the boxplots indicate. This is confirmed by:

```
> median(from.maine[,12])
## [1] 57.0335

> median(not.from.maine[,12])
## [1] 56.92
```

2. The **quartiles** of the Maine Runners are close to their counterparts in the Away distribution. This is true since:

```
> quantile(from.maine[,12])
##        0%       25%       50%       75%      100%
## 30.56700  49.99550  57.03350  64.24325 152.16700

> quantile(not.from.maine[,12])
##     0%     25%     50%     75%    100%
## 27.782  49.153  56.920  64.827 133.710
```

3. The distribution for the Away runners has a slightly higher **inter-quartile range** (Q3 - Q1) in comparison to the distribution of the Maine runners. This is confirmed by the following:

```
> quantile(from.maine[,12])[4]-quantile(from.maine[,12])[2]
##     75%
## 14.24775

> quantile(not.from.maine[,12])[4]-quantile(not.from.maine[,12])[2]
##   75%
## 15.674
```

4. The distribution for the Maine runners has a slightly higher **range** (highest value - lowest value) in comparison to the Away runners when the *1.5 * IQR* rule is **not applied**.

```
> quantile(from.maine[,12])[5]-quantile(from.maine[,12])[1]
##  100%
##  121.6
```

```
> quantile(not.from.maine[,12])[5]-quantile(not.from.maine[,12])[1]
##  100%
##  105.928
```

5. The distribution for the Away runners has a slightly higher **range** (highest value - lowest value) in comparison to the Maine runners when the *1.5 * IQR rule* is **applied**.

```
# Getting Q3 + 1.5 * IQR for Maine
> high <- 1.5 * (quantile(from.maine[,12])[4] -
quantile(from.maine[,12])[2]) + quantile(from.maine[,12])[4]
##      75%
## 85.61487

# Getting Q1 - 1.5 * IQR for Maine
> low <- 1.5 *( quantile(from.maine[,12])[4] -
quantile(from.maine[,12])[2])*(-1)+ quantile(from.maine[,12])[2]
##     75%
## 28.62388

# Range for Running times of Maine runners
> high - low
##  56.99099



# Getting Q3 + 1.5 * IQR for Away
> high <- 1.5 * (quantile(from.maine[,12])[4] -
quantile(not.from.maine[,12])[2]) + quantile(not.from.maine[,12])[4]
##      75%
##  88.338

# Getting Q1 - 1.5 * IQR for Away
> low <- 1.5 *( quantile(from.maine[,12])[4] -
quantile(not.from.maine[,12])[2])*(-1)+ quantile(not.from.maine[,12])[2]
##     75%
##   25.642

# Range for Running times of Away runners
> high-low
```

```
##   75%
## 62.696
```

6. The distribution for Maine Runners has far more **outliers** than in the case of Away Runners.

7. Most of the outliers in both boxplots are close to the (Q3 + 1.5 * IQR) value.

8. The outliers are only on the **higher end** of the distribution and not the lower end in both cases.

9. The **standard deviation** of the Away Runners' distribution is slightly higher than that of the Maine Runners' distribution:

```
> sd(from.maine[,12])
## [1] 12.18511

> sd(not.from.maine[,12])
## [1] 13.83538
```

10. The Maine distribution has a **mean** value slightly larger than that of the Away distribution. This is true since:

```
> mean(from.maine[,12])
## [1] 58.19514

> mean(not.from.maine[,12])
## [1] 57.82181
```

(d) Boxplots for Runners' Ages - Male and Female

1. Read the *roadrace.csv* file.

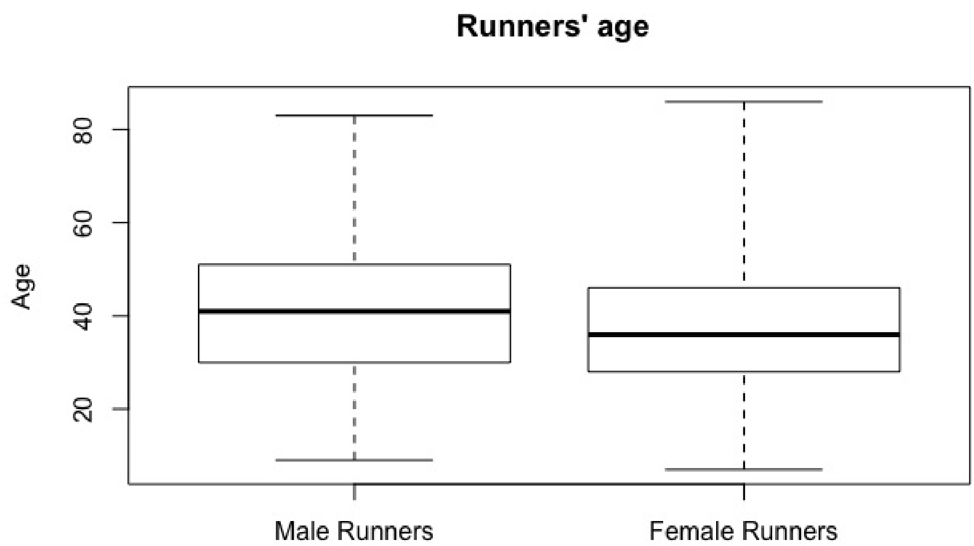
2. Get the tuples of runners who are males. Store them in the vector **male.runners**.

3. Get the tuples of runners who are females. Store them in the vector **female.runners**.

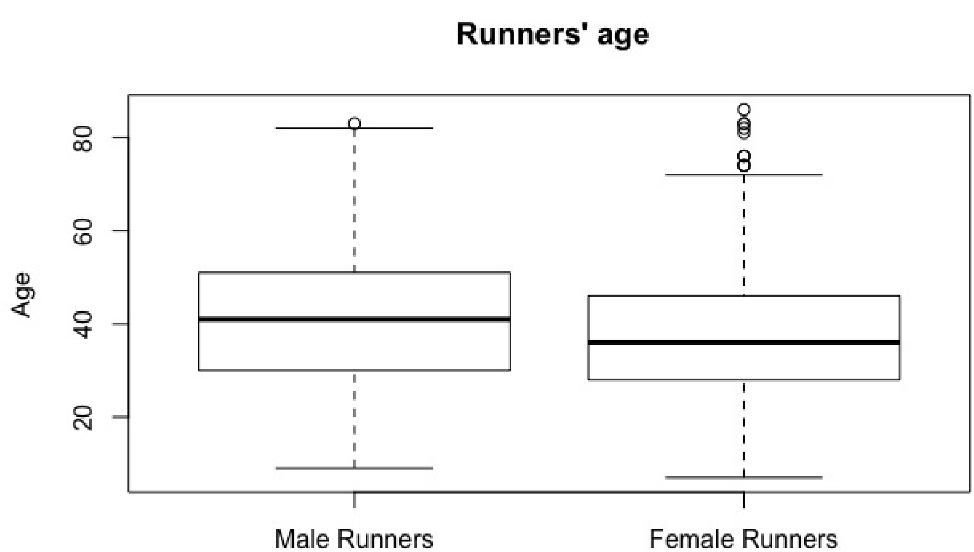
4. We boxplot the values in the 12th column of the **male.runners** vector.

5. We boxplot the values in the 12th column of the **female.runners** vector.

Without using the 1.5 * IQR rule, we get the following boxplot:



Using the 1.5 * IQR rule, we get the following boxplot:



From the above boxplots, we can infer the following:

1. The male runners' age distribution has a higher **median** in comparison to that of the female runners'. This is confirmed by:

```
> median(male.runners[,5])
## [1] 41

> median(female.runners[,5])
## [1] 36
```

2. The **quartiles** of the male runners' age distribution are close to their counterparts in the female runners' age distribution. This is true since:

```
> quantile(male.runners[,5])
##  0%  25%  50%  75% 100%
##  9   30   41   51   83

> quantile(female.runners[,5])
##    0%  25%  50%  75% 100%
##    7   28   36   46   86
```

3. The distribution for the male runners' ages has a higher **inter-quartile range** (Q3 - Q1) in comparison to the age distribution of the female runners. This is confirmed by the following:

```
> quantile(male.runners[,5])[4] - quantile(male.runners[,5])[2]
## 75%
##  21

> quantile(female.runners[,5])[4] - quantile(female.runners[,5])[2]
## 75%
## 18
```

4. The distribution for the male runners has a slightly lower **range** (highest value - lowest value) in comparison to female runners when the *1.5 * IQR* rule is **not applied**.

```
# Range for male runners' ages
> quantile(male.runners[,5])[5] - quantile(male.runners[,5])[1]
## 100%
##  74

# Range for female runners' ages
> quantile(female.runners[,5])[5] - quantile(female.runners[,5])[1]
## 100%
##  79
```

5. The age distribution for male runners has a slightly higher **range** (highest value - lowest value) in comparison to that of female runners when the *1.5 * IQR rule* is **applied**.

```
# Getting Q3 + 1.5 * IQR for males
> high <- 1.5 * (quantile(male.runners[,5])[4] -
quantile(male.runners[,5])[2]) + quantile(male.runners[,5])[4]
##     75%
## 82.5

# Getting Q1 - 1.5 * IQR for males
> low <- 1.5 * (quantile(male.runners[,5])[4] -
quantile(male.runners[,5])[2])*(-1) + quantile(male.runners[,5])[2]
##    75%
## -1.5

# Range for Running times of male runners
> high - low
##  84


# Getting Q3 + 1.5 * IQR for females
> high <- 1.5 * (quantile(female.runners[,5])[4] -
quantile(female.runners[,5])[2]) + quantile(female.runners[,5])[4]
##   75%
##  73

# Getting Q1 - 1.5 * IQR for females
> low <- 1.5 * (quantile(female.runners[,5])[4] -
quantile(female.runners[,5])[2])*(-1) + quantile(female.runners[,5])[2]
##  75%
##   1

# Range for Running times of female runners
> high-low
##  75%
## 72
```

6. The age distribution for female runners has far more **outliers** than in the case of male runners.

7. The outliers are only on the **higher end** of the distribution and not the lower end.

8. The **standard deviation** of the male runners' ages is slightly higher than that of the female runners' distribution.

```
> sd(male.runners[,5])
## [1] 13.99289

> sd(female.runners[,5])
## [1] 12.26925
```

9. The male runners' age distribution has a **mean** value slightly larger than that of the female runners's age distribution. This is true since:

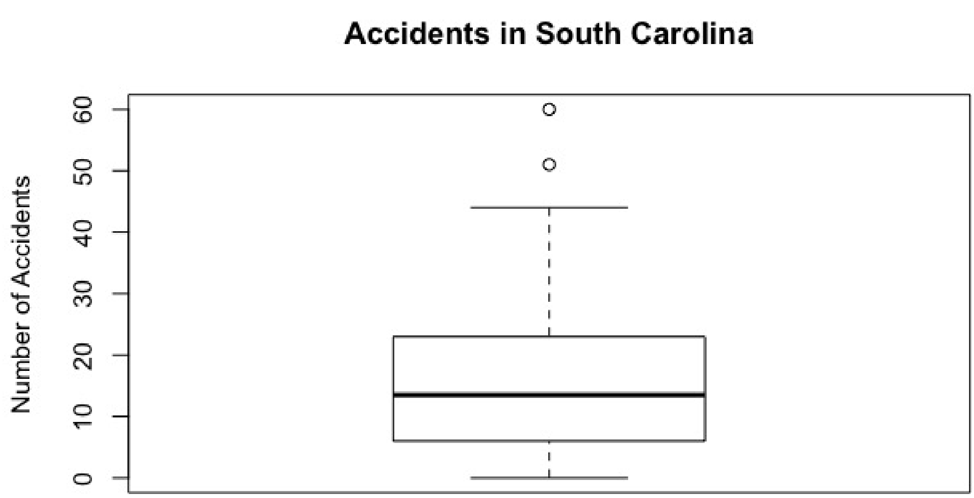
```
> mean(male.runners[,5])
## [1] 40.4468
> mean(female.runners[,5])
## [1] 37.23653
```

Problem 1.2 Boxplot for motorycle accidents in South Carolina

1. Read the *motorcycle.csv* file.

2. Boxplot the values the second column.

We get the following boxplot:



We can infer the following:

1. The data clearly has a **right skewed** distribution since there are more values at the right end.

2. The **mean** is higher than the **median**, as implied by the right-skewedness of the distribution.

```
> mean (motorcycle[,2])
## [1] 17.02083

> median(motorcyle[,2])
## [1] 13.5
```

3. The **quartiles** also show the right-skewedness of the data since the first 3 quartiles are close to each other and there is a huge gap between the 3rd and 4th quartiles.

```
> quantile(motorcycle[,2])
##  0%  25%  50%  75% 100%
## 0.0  6.0 13.5 23.0 60.0
```

4. The data has an **inter-quartile range** (Q3 - Q1) of 19.

```
> quantile(motorcycle[,2])[4] - quantile(motorcycle[,2])[2]
##  75%
## 19
```

5. The distribution has a **range** (highest value - lowest value) of 60 when the *1.5 * IQR* rule is **not applied**.

```
# Range for  number of accidents
> quantile(motorcycle[,2])[5] - quantile(motorcycle[,2])[1]
## 100%
##  60
```

6. The distribution has a slightly higher **range** (highest value - lowest value) of 68 when the *1.5 * IQR* rule is **applied**.

```
> high <- 1.5*(quantile(motorcycle[,2])[4]-
quantile(motorcycle[,2])[2])+quantile(motorcycle[,2])[4]
## 75%
## 48.5

> low <- -1.5*(quantile(motorcycle[,2])[4]-
quantile(motorcycle[,2])[2])+quantile(motorcycle[,2])[2]
## 75%
## -19.5

> high - low
## 75%
## 68
```

7. The distribution has a **standard deviation** of around 13.

```
> sd(motorcycle[,2])
## 100%
## [1] 13.81256
```

8. There are 2 counties which are outliers:
   (a) Greenville, with 51 accidents
   (b) Horry, with 60 accidents

Section 2 R Code

```
####################################
# R code for question 2
####################################
# Reading the .csv File #
motorcycle <- read.csv(file="/users/psprao/downloads/stats/datasets/motorcycle
   .csv")

# Two plots in a row
par(mfrow=c(1,1))

# boxplot
boxplot(motorcycle[,2],range=1.5,ylab="Number of Accidents",main="Accidents in
    South Carolina")
```

```
library(sqldf)

data <- read.csv.sql(file="/users/psprao/downloads/stats/datasets/voltage.csv"
   )

voltages.remote<-sqldf("select * from data where location=0")[,2]
voltages.local<-sqldf("select * from data where location=1")[,2]

par(mfrow=c(1,1))
boxplot(voltages.remote,voltages.local,range=1.5,main="Voltage readings",
        ylab="Voltage", names = c("Remote locations", "Local locations"))
```

```
library(sqldf)

data <- read.csv.sql(file="/users/psprao/downloads/stats/datasets/voltage.csv"
   )

voltages.remote<-sqldf("select * from data where location=0")[,2]
voltages.local<-sqldf("select * from data where location=1")[,2]

remote.mean<-mean(voltages.remote)
remote.var=var(voltages.remote)
n=NROW(voltages.remote)

# Getting the statistics for the local voltages
local.mean<-mean(voltages.local)
local.var=var(voltages.local)
m=NROW(voltages.local)

# Estimator for difference in means
theta.hat<- remote.mean-local.mean

# Standard error for mean difference estimator
pooled.var<-((n-1)*remote.var + (m-1)*local.var)/(n+m-2)

# COnfidence interval
mean.diff.ci<-theta.hat + c(-1,1)*qt(1-(1-0.95)/2,n+m-2)*sqrt((pooled.var/n) +
     (pooled.var/m))

print(mean.diff.ci)
```