**MINI PROJECT 2**

**DUE DATE : 27TH SEPTEMBER 2018**

**GROUP 0F 2 :**

**NAMES : SIDDHANT AGARWAL ,**

**KRISHAN KUMAR**

**UTD-IDS : 2021435363(SA) , 2021449376 (KK)**

**COURSE : 6313.001**

Contribution : Both the members have done the assignment individually and separately in terms of R code and problem solving apart from discussing key concepts when stuck somewhere. Reports has been collaborated by both after discussing and analyzing the results of each other’s work.

1. (12 points) Consider the dataset roadrace.csv posted on eLearning. It contains observations on 5875 runners who finished 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 identifies 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.

Ans : a)

We plot the variable Maine which identifies whether a runner is from Maine or somewhere else using a barplot function in R. The R code and its results along with the inferences drawn from it as below :

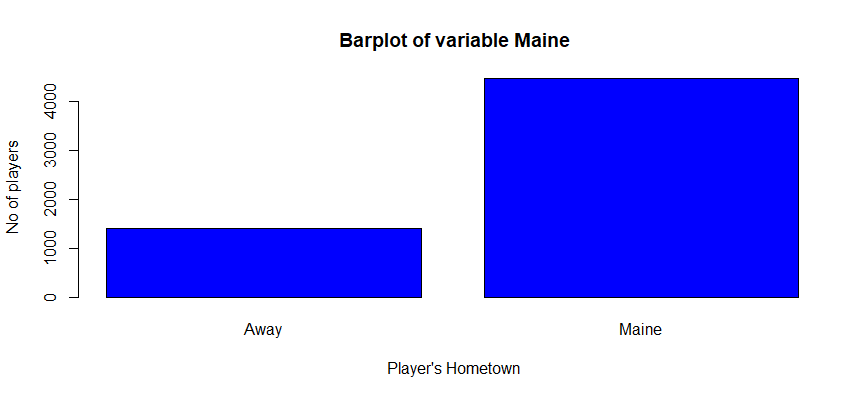
**R code**

**roadrace <- read.csv(file = "roadrace.csv", , header = TRUE , na.strings = c(" ", "\*", "NA")); #reads the csv file and puts it into the variable roadrace**

**height <- table(roadrace$Maine); #gives the total height or count of distinct values in column Maine**

**barplot(height, xlab = " Player's Hometown", ylab = " No of players", main = " Barplot of variable Maine", col = "blue");**

**#barplot function plots the bargraph of the count evaluated above from the table command**



We can conclude from the above plot that the runner’s were in majority from Maine itself. Approximately 76 % of the runner’s were from Maine and rest 24 % were away runners.

To backup the above conclusion we can see the summary statistics of the variable Maine :

> **summary( roadrace$Maine)**

* Away Maine

1417 4458 #gives the summary of the data of the variable Maine

We can see that the barplot matches exactly with the summary statistics of the variable Maine.

b) The runners' times (given in minutes) for the Maine group and the Away groups are calculated as follows :

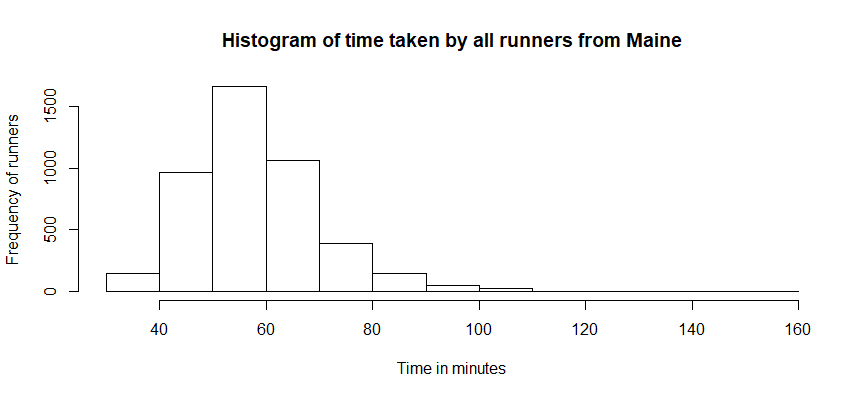
**Rcode:**

**home <- roadrace$Time..minutes.[roadrace$Maine == "Maine"];**

**#gives the time in minutes for all the runners from Maine**

**hist(home, main = "Histogram of time taken by all runners from Maine", xlab= "Time in minutes", ylab = "Frequency of runners");**

**#plots the histogram for the time taken by all runners from Maine**



We can see from the above histogram plot that the distribution is slightly **Right-Skewed distribution**

as its peak is slightly off-center and the tail stretches away to the right. The shape of histogram is asymmetrical and hence skewed.

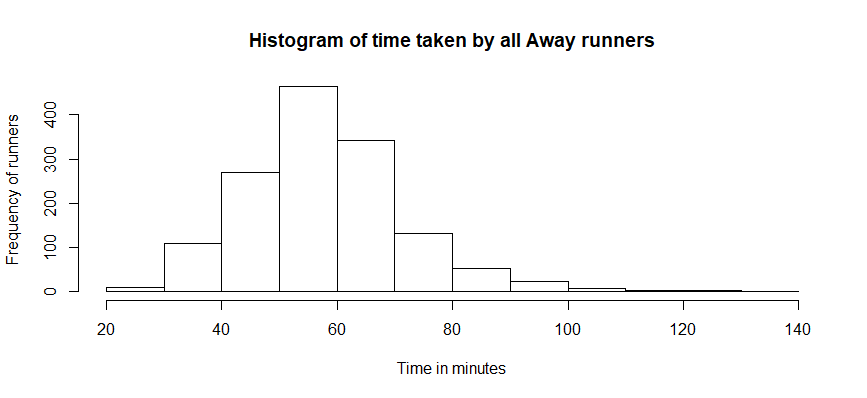
**Rcode :**

**Away <- roadrace$Time..minutes.[roadrace$Maine == "Away"];**

**#gives the time in minutes for all the runners from other places than Maine**

**hist(Away, main = "Histogram of time taken by all Away Runners", xlab= "Time in minutes", ylab = "Frequency of runners");**

**#plots the histogram for the time taken by all Away Runners**



We can see that the above plot for the time taken from away runners has a better symmetry than time taken from maine runners and forms bell shaped normal distribution curve but this also is not perfectly symmetrical

and tail stretches a bit to the right along with the peak being slightly off-centered. Hence, the distribution is

**slightly right skewed distribution.**

**Relevant Summary statistics :**

* **summary( home )**

**#gives summary statistics for the variable home containing times taken by all maine runners**

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

30.57 50.00 57.03 58.20 64.24 152.17

* **summary ( Away )**

**#gives summary statistics for the variable home containing times taken by all away runners**

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

27.78 49.15 56.92 57.82 64.83 133.71

* **sd(home)** **#standard deviation for home runners**

**>**12.18511

* **sd( Away )** **#standard deviation for away runners**

**>**13.83538

* **diff(range(home))** #gives range for home runners

>121.6

* **diff(range(away)**) #gives range for away runners

>105.928

* **IQR(home)**  #gives inter quartile range for home runners

>14.24775

* **IQR(away)** #gives inter quartile range for away runners

>15.674

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Runners** | **Min Value** | **1st quart** | **Median** | **Mean** | **3rd quart** | **Max value** | **Range** | **Standard deviation ( sd)** | **IQR** |
| Maine | 30.57 | 50.00 | 57.03 | 58.20 | 64.24 | 152.17 | 121.6 | 12.18511 | 14.24775 |
| Away | 27.78 | 49.15 | 56.92 | 57.82 | 64.83 | 133.71 | 105.928 | 13.83538 | 15.674 |

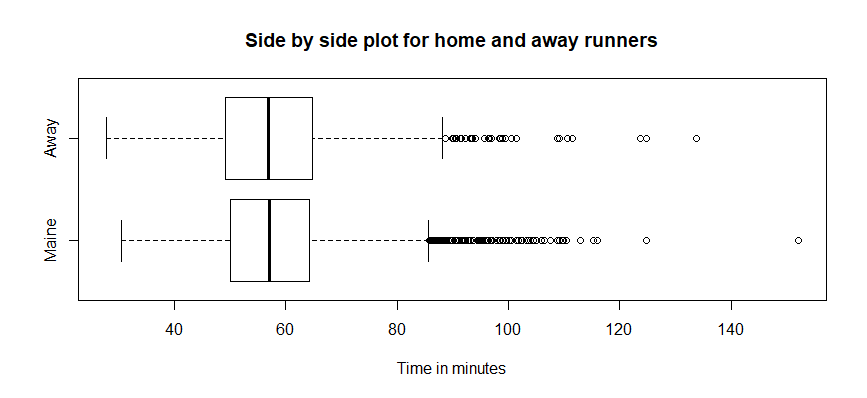
We can see that (Median – Min) < ( Max – Median ) hence the conclusion about right skewed distribution is verified by the statistics.

c) Repeating above using side by side plot :

**hometown <- c("Maine","Away");**  #makes a vector of names

**boxplot(home , Away , names= hometown, horizontal = TRUE, main = "Side by side plot for home and away runners", xlab = "Time in minutes");**

**#plots the side by side graph of the runners time taken**



We can see that the box plot matches with the summary statistics above and from the boxplot it is evident

That both the distributions are slightly skewed towards the right as (Median – Min) < ( Max – Median ) for

both the cases.

Hence, the side by side boxplot confirms the skewness of the distribution interpreted by the histogram above.

d) The runners' ages (given in years) for male and female runners are plotted as below :

{ RED BY KRISHAN:

> roadrace\_data <- read.csv(file = "roadrace.csv", , header = TRUE , na.strings = c(" ", "\*", "NA"));

> View(roadrace)

> male\_age\_tb <- roadrace\_data$Age[roadrace\_data$Sex == "M"]

> female\_age\_tb <- roadrace\_data$Age[roadrace\_data$Sex == "F"]

> head(male\_age\_tb)

[1] 25 21 28 25 21 22

>par(mfrow=c(1,2))

> boxplot(male\_age\_tb)

> boxplot(female\_age\_tb)

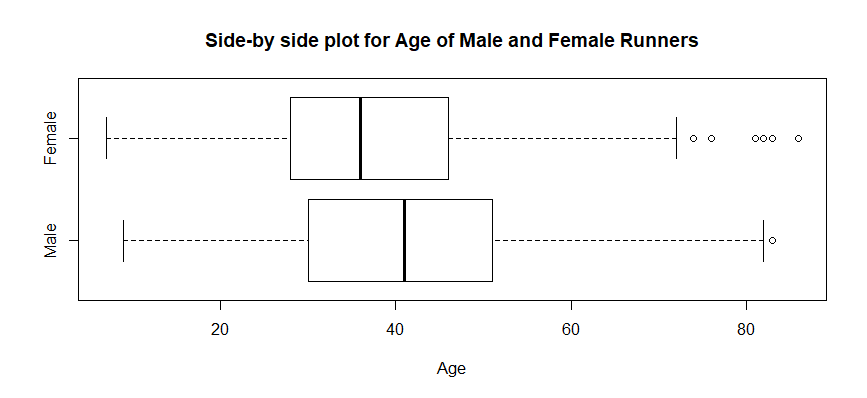
}

**MaleAge <- roadrace$Age[roadrace$Sex == "M"]; #gives the age of all male runners**

**FemaleAge <- roadrace$Age[roadrace$Sex == "F"]; #gives the age of all female runners**

**Sex <- c("Male", "Female"); #gives vector of sexes**

**boxplot(MaleAge,FemaleAge,names= Sex, horizontal = TRUE, main = "Side-by side plot for Age of Male and Female Runners", xlab = "Age"); #boxplot for the ages by sexes**



From the above side by side boxplot we can easily conclude that both the distributions are **Right Skewed Distributions** as both the distribution have **(Max – Median) > (Median – Min ).**

**Statistics Details :**

* **summary ( MaleAge )**

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

9.00 30.00 41.00 40.45 51.00 83.00

* **summary ( FemaleAge )**

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

7.00 28.00 36.00 37.24 46.00 86.00

* **sd( MaleAge)**

>13.99289

* **sd ( FemaleAge )**

**>**12.26925

* **IQR ( MaleAge )**

>21

* **IQR ( FemaleAge )**

>18

* **Diff(Range(MaleAge))**

>74

* **Diff(Range(FemaleAge))**

>79

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Runners** | **Min Value** | **1st quart** | **Median** | **Mean** | **3rd quart** | **Max value** | **Range** | **Standard deviation ( sd)** | **IQR** |
| Male | 9 | 30 | 41 | 40.45 | 51 | 83 | 74 | 13.99289 | 21 |
| Female | 7 | 28 | 36 | 37.24 | 46 | 86 | 79 | 12.26925 | 18 |

We can see that from the statistics we know for male runners,

Median – Mean = 41 – 9 =32 and Max- Median = 83-41 = 42

Hence , **(Max – Median) > (Median – Min ) holds and the distribution is Right Skewed for Male.**

Similary for Female runners,

Median – Mean = 36 – 7 =29 and Max- Median = 86-36 = 50

Hence , **(Max – Median) > (Median – Min ) holds and the distribution is Right Skewed for Female runners.**

**2)** The dataset motorcycle.csv contains the number of fatalmotorcycle accidents that occurred in each county of South Carolina during 2009. The plots for the accident by each county is as follows :

**Rcode**

**motorcycle <- read.csv(file = "motorcycle.csv", header = TRUE , na.strings = c(" ", "\*", "NA"));**

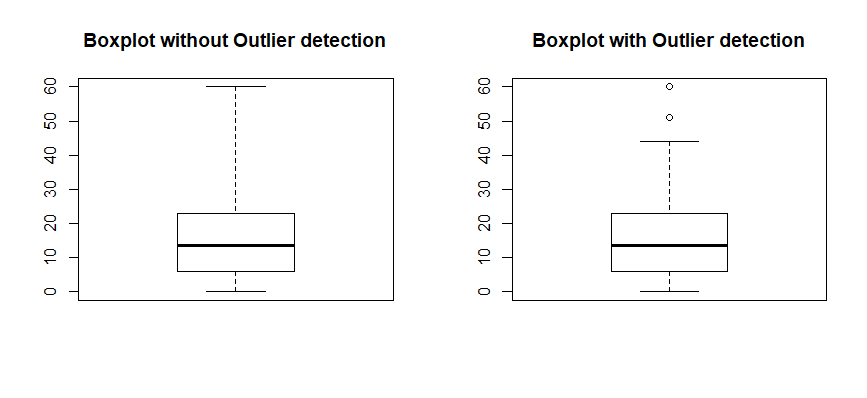
**# reads the file**

**accident <- motorcycle$Fatal.Motorcycle.Accidents #gives the accident variable data**

**par(mfrow=c(1,2)) # for side by side plot**

**boxplot(accident,main = "Boxplot with Outlier detection**") #boxplot with outlier default

**boxplot(accident,range = 0,main = "Boxplot without Outlier detection") #boxplot without default outlier detection**



We can see from the boxplot that the distribution is **right skewed**.  After applying the outlier detection, we find that the counties ‘Greenville’ and ‘Horry’ can be considered outliers as the number of fatal motorcycle accidents in these counties are at extreme (above 50). Summary statistics show that that on an average(median), there are 13.50 motorcycle fatal accidents in South Carolina.

As per the data, ‘Greenville’ and ‘Horry’ counties might have the highest number of fatalities in South Carolina because traffic rules might not be followed strictly in those counties.

Summary statistics :

* **Summary (accident)**

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

0.00 6.00 13.50 17.02 23.00 60.00

Hence, we can see that Median – Min < Max -Median from the statistics which verifies that the distribution is **Right-skewed.**