**Mini Project 4**

**Name :**

**Utkarsh Gandhi ( usg170030 )**

**Dhwani Kaneria( drk170130 )**

**Contribution of team members:**

Utkarsh:

* Wrote R code for question 1
* Done analysis of the plots and output

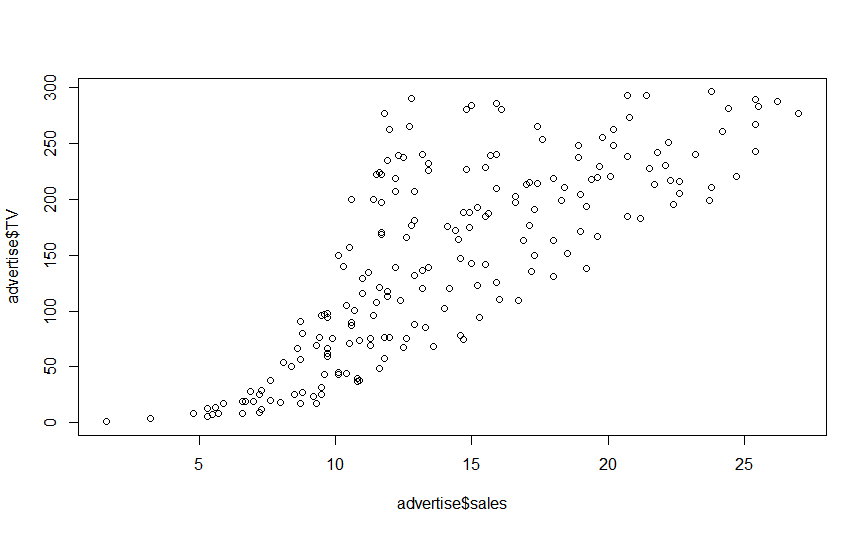
Dhwani:

* Wrote R code for question 2
* Done analysis of the plots and derived results

**Answer 1:**

**#R-code:**

1. file="Advertising.csv"; # taking file name into a object
2. advertise=read.csv(file); # reading csv file into a dataframe
3. plot(advertise$sales,advertise$TV) #scatter plot of sales vs TV

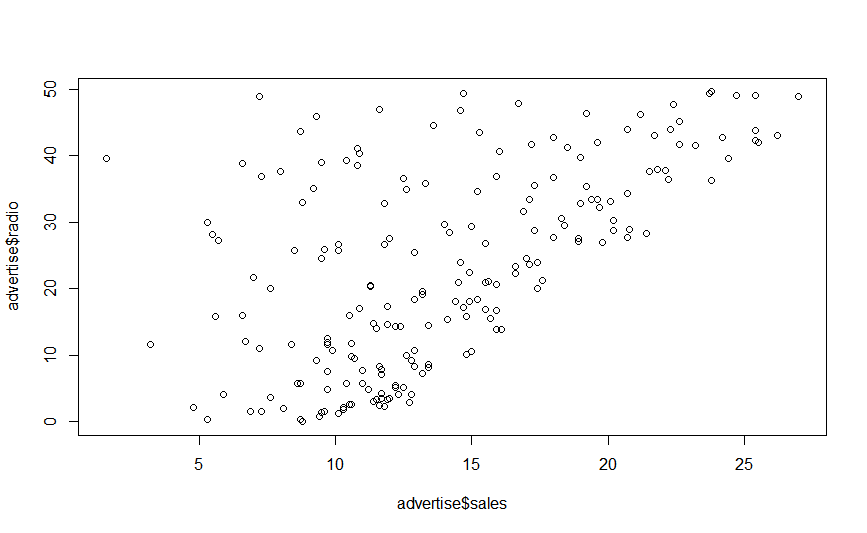


**Analysis:**

From the above scatter plot, we can see that there is kind of linear relationship between the TV and sales. As the sales value increases the TV value also increases in approximate linear manner. We can also infer that there is positive correlation between them. The correlation came out to be 0.782, calculated using cor function in R.

#scatter plot of sales vs Radio

plot(advertise$sales,advertise$radio)



**Analysis:**

From the above scatter plot of radio vs sales, we can infer that there is no proper linear correlation between them. The points are scattered a lot. They also don’t form a linear relationship. As calculated below the correlation coefficient came out to be 0.57 which is less than 1. Thus, no relation can be inferred from the above scatter plot.

**Point Estimate of both correlations:**

**#correlation of Sales and TV**

cor(advertise$sales, advertise$TV, use = "everything",method = c("pearson", "kendall", "spearman"))

[1] **0.7822244**

**#correlation of sales and radio**

> cor(advertise$sales, advertise$radio, use = "everything",method = c("pearson", "kendall", "spearman"))

[1] **0.5762226**

**Bootstrap Estimate (Correlation between sales and TV)**

1. file="Advertising.csv"; # taking file name into a object
2. advertise=read.csv(file); # reading csv file into a dataframe
3. #The statistic of interest here is the correlation coefficient of sales and TV.
4. cor.npar<-function(advertise,indices)
5. {
6. data<-advertise[indices,] #taking all rows of data
7. result<-cor(data$sales,data$**TV**)

#using cor function to get correlation coefficient

1. return (result)
2. }
3. (cor.npar.boot<-boot(advertise,cor.npar,R=999,sim="ordinary", stype = "i"))
4. #using boot function and number of bootstrap samples as 999

**Output:**

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

boot(data = advertise, statistic = cor.npar, R = 999, sim = "ordinary", stype = "i")

**Bootstrap Statistics :**

**original bias std. error**

**t1\* 0.7822244 0.0001987172 0.0262113**

Here the bootstrap estimates of bias and standard error of the point estimate is

bias = 0.0001987172

std. error = 0.0262113

**Getting 95% Confidence Interval using Percentile Bootstrap method**

**#R-code:**

1. boot.ci(cor.npar.boot)

**Output:**

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 999 bootstrap replicates

CALL :

boot.ci(boot.out = cor.npar.boot)

Intervals :

Level Normal Basic

95% ( 0.7270, 0.8338 ) ( 0.7280, 0.8367 )

Level Percentile BCa

95% ( 0.7278, 0.8365 ) ( 0.7170, 0.8284 )

**#Using percentile bootstrap method**

1. s\

[1] 0.7277753 0.8364844

Hence the Confidence Interval is **[ 0.7277753, 0.8364844 ]** for **Sales vs TV**

**Interpretation:**

* As we can see that the correlation coefficient (point estimate) for sales vs Tv that we calculated correctly falls in the confidence interval which states the plausible values. The higher value here indicates that sales has some degree of positive linear relationship with TV.
* 95% confidence interval (CI) for the correlation coefficient indicates that this is the range of values that contains with a 95% confidence the 'true' correlation coefficient.

**Bootstrap Estimate (Correlation between sales and radio)**

1. file="Advertising.csv"; # taking file name into a object
2. advertise=read.csv(file); # reading csv file into a dataframe
3. #The statistic of interest here is the correlation coefficient of sales and TV.
4. cor.npar<-function(advertise,indices)
5. {
6. data<-advertise[indices,] #taking all rows of data
7. result<-cor(data$sales,data$**radio**)

#using cor function to get correlation coefficient

1. return (result)
2. }
3. (cor.npar.boot<-boot(advertise,cor.npar,R=999,sim="ordinary", stype = "i"))
4. #using boot function and number of bootstrap samples as 999

**Output:**

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

boot(data = advertise, statistic = cor.npar, R = 999, sim = "ordinary",

stype = "i")

Bootstrap Statistics :

original bias std. error

t1\* 0.5762226 -0.001863566 0.05620538

Here the bootstrap estimates of bias and standard error of the point estimate is

bias = -0.001863566

std. error = 0.05620538

**Getting 95% Confidence Interval using Percentile Bootstrap method**

**#R-code:**

1. boot.ci(cor.npar.boot)

**Output:**

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 999 bootstrap replicates

CALL :

boot.ci(boot.out = cor.npar.boot)

Intervals :

Level Normal Basic

95% ( 0.4679, 0.6882 ) ( 0.4716, 0.6949 )

Level Percentile BCa

95% ( 0.4575, 0.6809 ) ( 0.4544, 0.6788 )

Calculations and Intervals on Original Scale

**#Using percentile bootstrap method**

1. sort(cor.npar.boot$t)[c(25,975)]

[1] 0.4575466 0.6808763

Hence the Confidence Interval is [ 0.4575466, 0.6808763 ] for Sales vs radio

**Interpretation:**

As we can see that the correlation coefficient (point estimate) for sales vs radio that we calculated correctly falls in the confidence interval which states the plausible values.

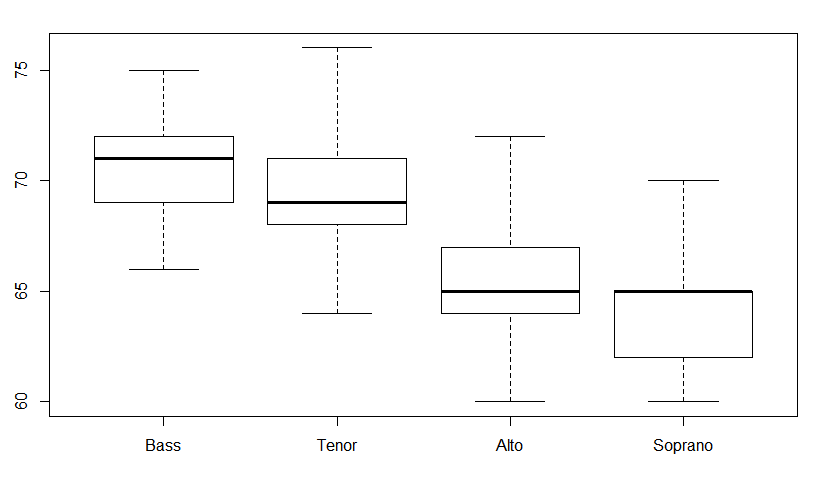
**Answer 2 :**

**a)**

**#R-code:**

1. #Read file voice.txt and making 1at row as a header
2. df <- read.table("voice.txt",header = TRUE,sep = ",")
3. #create different data frames for each category
4. df\_Bass = data.frame()
5. df\_Tenor= data.frame()
6. df\_Alto= data.frame()
7. df\_Soprano= data.frame()
8. # separate height as per category and store it in appropriate data frame
9. for(i in 1:nrow(df))
10. {
11. if(df[i,2]=="Bass")
12. df\_Bass <- rbind(df\_Bass,df[i,1])
14. if(df[i,2]=="Tenor")
15. df\_Tenor <- rbind(df\_Tenor,df[i,1])
17. if(df[i,2]=="Alto")
18. df\_Alto <- rbind(df\_Alto,df[i,1])
20. if(df[i,2]=="Soprano")
21. df\_Soprano <- rbind(df\_Soprano,df[i,1])
23. }
24. #give name to column of dataframe
25. colnames(df\_Bass)<- c("height")
26. colnames(df\_Tenor)<- c("height")
27. colnames(df\_Alto)<- c("height")
28. colnames(df\_Soprano)<- c("height")
29. #side by side boxplot of each data frame
30. boxplot(df\_Bass$height,df\_Tenor$height,df\_Alto$height,df\_Soprano$height,names = c("Bass", "Tenor","Alto","Soprano"))

**Output :**



**Analysis:**

From the above box plots of height distribution of four voice parts, we can infer that height is not normally distributed for any of the voice part. For Soprano and Bass, distribution of heights is left skewed and for Alto and Tenor, distribution is right skewed. Even we can infer from the quartiles, that increasing order of heights is Soprano, Alto, Tenor and Bass.

**b)**

**#R-code:**

1. # generating 95% confidence interval
2. n <- length(df\_Alto$height)
3. m <- length(df\_Soprano$height)
4. Alto\_height\_mean <- mean(df\_Alto$height)
5. Soprano\_height\_mean<- mean(df\_Soprano$height)
6. Alto\_height\_var<- var(df\_Alto$height)
7. Soprano\_height\_var <- var(df\_Soprano$height)
8. Alto\_height\_mean - Soprano\_height\_mean+ c(-1,1) \* qnorm(1-(1-0.95)/2) \*sqrt(((Alto\_height\_var/n) + (Soprano\_height\_var /m)))

**output :**

**[1] 0.4219834 2.1097859**

mean of x mean of y

**65.38710 64.12121**

**Analysis:**

After generating confidence interval for Alto and Soprano, We get interval as  **[0.4219834, 2.1097859**]. That concludes, there is difference in the mean heights of Alto and Soprano singers. We get mean of heights as 65.38 for Alto and 64.121 for Soprano. So difference is 1.259. Since the entire interval is above zero, we can conclude that mean of Alto > mean of Soprano, and plausible values for mean difference is in between **[0.4219834,2.1097859**]. Here we assume that number of sample size is large and observed that sample distribution is not normal as per box plots and als population variance is not same for Alto and Soprano.

**c)**

conclusion in B compared to analysis of A , confirmed that mean height of Alto is higher than mean height of Soprano and plausible values for mean difference is in between **[0.4219834,2.1097859**].