Statistical Methods for Data Science

Mini Project 4

Members:

SHIVA RANGA CHAWALA	sxc167630
KRISHNA SINDHU KOTA	kxk171030

Contribution:

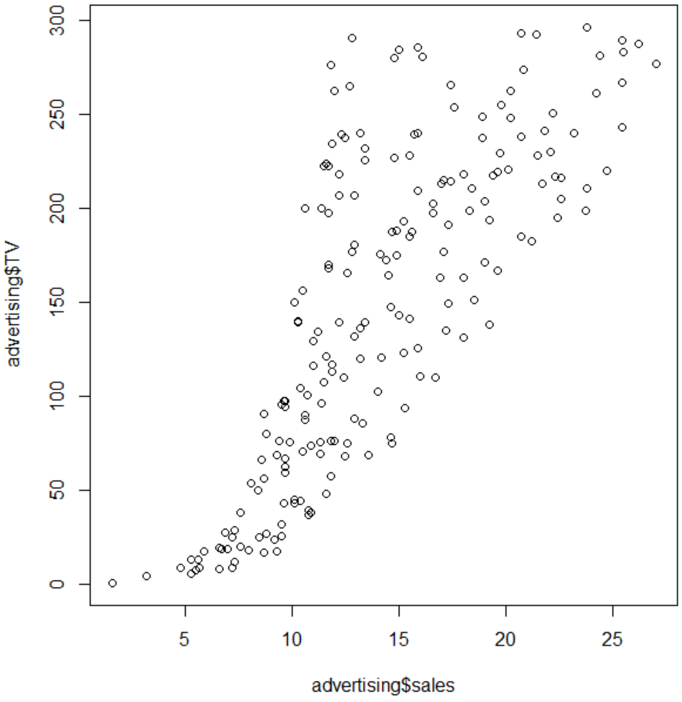
Shiva Ranga Chawala – Equally contributed

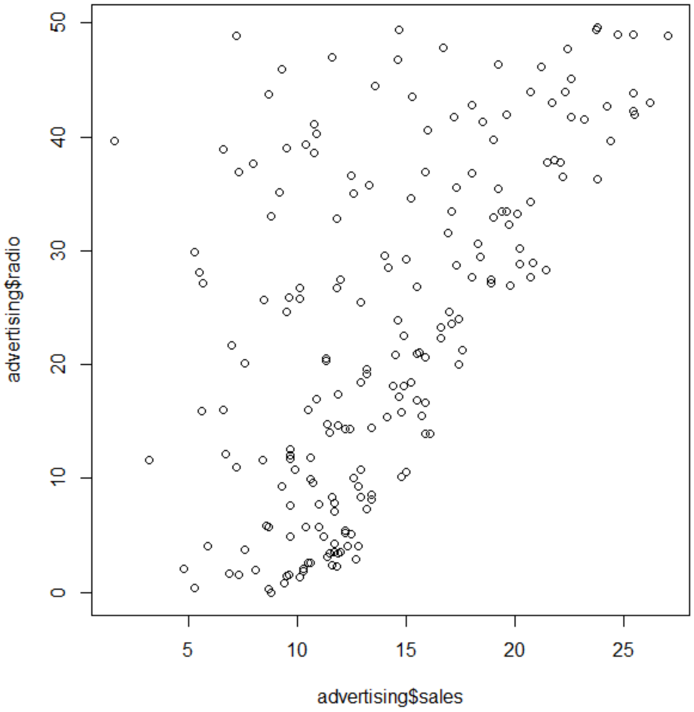
Krishna Sindhu Kota – Equally contributed

1. **Consider the advertising data stored in the Advertising.csv file available on eLearning. Make scatterplots of sales against TV and radio, and comment on the strength of linear relationship between sales and TV and sales and radio. Let p1 and p2 respectively denote the population correlation between sales and TV and between sales and radio. For each of the two correlations, provide a point estimate, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results.**

```
#Installing boot packages
install.packages("boot")
# loading the boot packages
library(boot)

#Reading the csv file
advertising = read.csv("Advertising.csv", header = TRUE, sep = ",")

#Scatter plots for Sales vs TV and Sales vs Radio
plot(advertising$sales, advertising$TV)
plot(advertising$sales, advertising$radio)

#Point Estimate of Correlation between Sales and TV
p1 = cor(advertising$sales, advertising$TV)
print(paste("Point Estimate of correlation between Sales and TV is:", p1))
# Parameter of interest: Correlation between Sales and TV
corr.tv = function(x, indices)
{
  result = cor(advertising$sales[indices], advertising$TV[indices])
  return(result)
}

#Non-parametric boot call
corr.tv.boot = boot(advertising, corr.tv, R = 999, sim = "ordinary", stype = "i")
corr.tv.boot

#95% confidence interval for correlation based on 999 bootstrap replicates
boot.ci(corr.tv.boot)

#Calculating the bias
bias.tv = mean(corr.tv.boot$t) - p1
print(paste("Bias is:", bias.tv))

#OR we can use the below code to get 95% confidence interval using percentile bootstrap method
sort(corr.tv.boot$t)[c(25, 975)]
```

Results:


advertising$TV
0
50
100
150
200
250
300
5
10
15
20
25
advertising$sales


advertising$radio
0
10
20
30
40
50
5
10
15
20
25
advertising$sales

```
"Point Estimate of correlation between Sales and TV is: 0.782224424861606 "

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = advertising, statistic = corr.tv, R = 999, sim = "ordinary",stype = "i")

Bootstrap Statistics :
original      bias    std. error
t1* 0.7822244 -0.0001073533  0.02806657


BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = corr.tv.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.7273,  0.8373 )   ( 0.7299,  0.8409 )

Level     Percentile            BCa
95%   ( 0.7235,  0.8345 )   ( 0.7163,  0.8326 )
Calculations and Intervals on Original Scale

"Bias is: -0.000107353255132225"

> sort(corr.tv.boot$t)[c(25, 975)]
[1] 0.7235084 0.8345154
```

```
#Point Estimate of Correlation between Sales and Radio
p2 = cor(advertising$sales, advertising$radio)
print(paste("Point Estimate of correlation between Sales and Radio is:", p2))

# Parameter of interest: Correlation between Sales and Radio
corr.radio = function(x, indices)
{
  result = cor(advertising$sales[indices], advertising$radio[indices])
  return(result)
}

#Non-parametric boot call
corr.radio.boot = boot(advertising, corr.radio, R = 999, sim = "ordinary", stype = "i")
corr.radio.boot
```

```
#95% confidence interval for correlation based on 999 bootstrap replicates
boot.ci(corr.radio.boot)

#Calculating the bias
bias.radio = mean(corr.radio.boot$t) - p2
print(paste("Bias is:", bias.radio))

#OR we can use the below code to get 95% confidence interval using percentile bootstrap method
sort(corr.radio.boot$t)[c(25, 975)]
```

Results:

```
 "Point Estimate of correlation between Sales and Radio is: 0.576222574571055"

 ORDINARY NONPARAMETRIC BOOTSTRAP

 Call:
 boot(data = advertising, statistic = corr.radio, R = 999, sim = "ordinary", stype = "i")

 Bootstrap Statistics :
     original       bias    std. error
 t1* 0.5762226 -1.608647e-05  0.05435051

 BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
 CALL :
 boot.ci(boot.out = corr.radio.boot)

 Intervals :
 Level      Normal              Basic
 95%   ( 0.4697,  0.6828 )   ( 0.4750,  0.6829 )

 Level     Percentile            BCa
 95%   ( 0.4695,  0.6775 )   ( 0.4619,  0.6755 )
 Calculations and Intervals on Original Scale

 "Bias is: -1.60864733380617e-05"

 > sort(corr.radio.boot$t)[c(25, 975)]
 [1] 0.4695111 0.6774818
```

Interpretations:

The value of the correlation coefficient tells us about the strength of the linear relationship.
If we see our point estimate of correlation of Tv and Sales (0.782224424861606) which is greater than 0.7 hence, we can say that they have a Strong Positive Linear Correlation. We can also observe the above characteristic by using the scatter plot of the Tv vs Sales which has got a positive slope.

If we see our point estimate of correlation of Radio and Sales (0.576222574571055) which is in the range of 0.3 to 0.7 hence, we can say that they have a Moderate Positive Linear Correlation. We can also observe the above characteristic by using the scatter plot of the Radio vs Sales which has got a positive slope and we can also observe that as the magnitude of the correlation coefficient decreases the points on the plot are more scattered.

The larger the standard error, the wider the confidence interval about the statistic.
In our case:
SE for Radio vs Sales: 0.05435051 CI: (0.4695, 0.6775)
SE for TV vs Sales: 0.02806657 CI: (0.7235, 0.8345)

Standard Error of Radio vs Sales is greater than that of TV vs Sales so the CI for Radio vs Sales is wider than TV vs Sales.

As the correlation value of TV vs Sales is greater than Radio vs Sales, we can observe that the standard error of Radio vs Sales is greater than TV vs Sales because if the correlation is nearly perfect, then the data points are close to the line and therefore the standard error of the values around the line is near zero.

The point estimate of correlation(p1) between TV and Sales matches with the Bootstrap estimate of the correlation. We can also see that the Bias value is very small which says that bootstrap estimate is very close to the actual value. Also, the bootstrap estimate of bias and calculated bias matches.

The point estimate of correlation also falls in 95% CI calculated using percentile bootstrap method from which we can see that bootstrap estimates of confidence intervals are also correct.

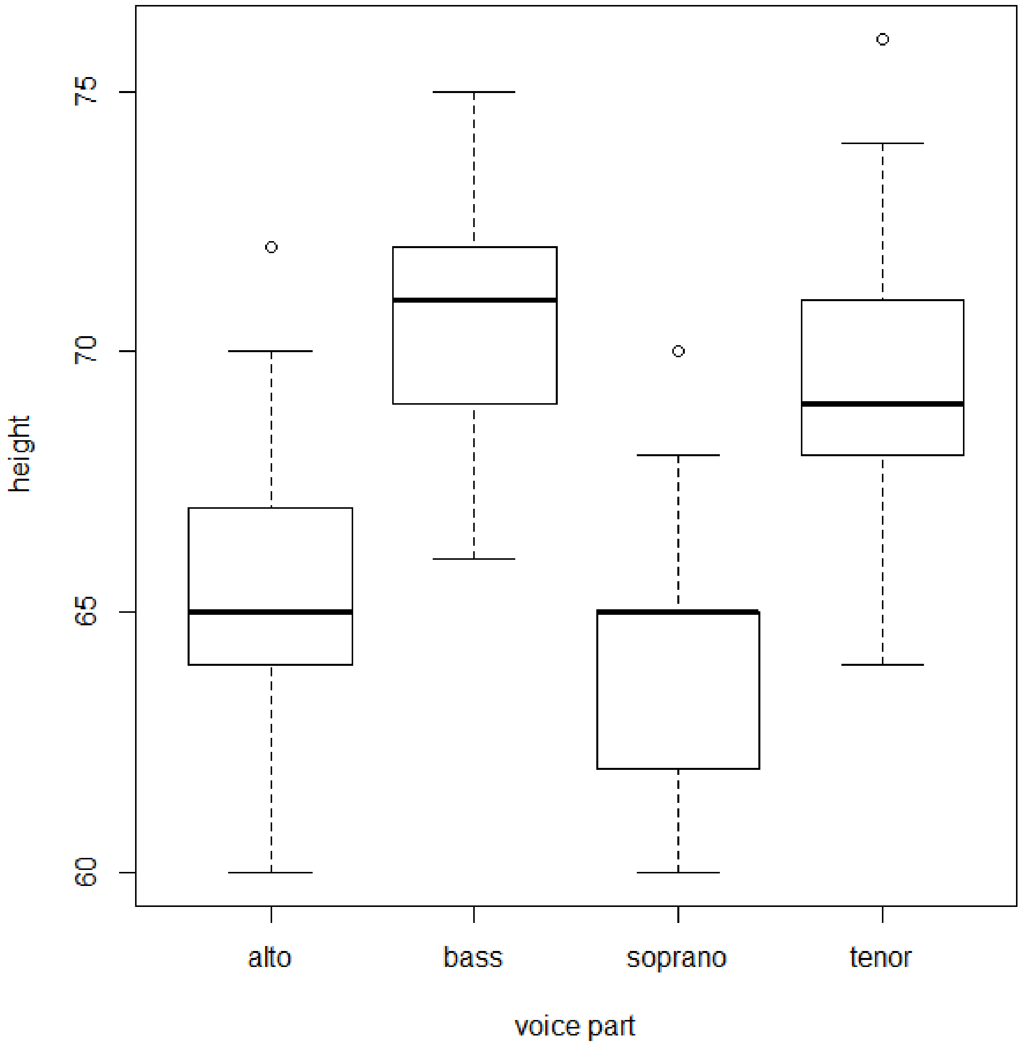
We can observe that point estimates matches the bootstrap estimates.

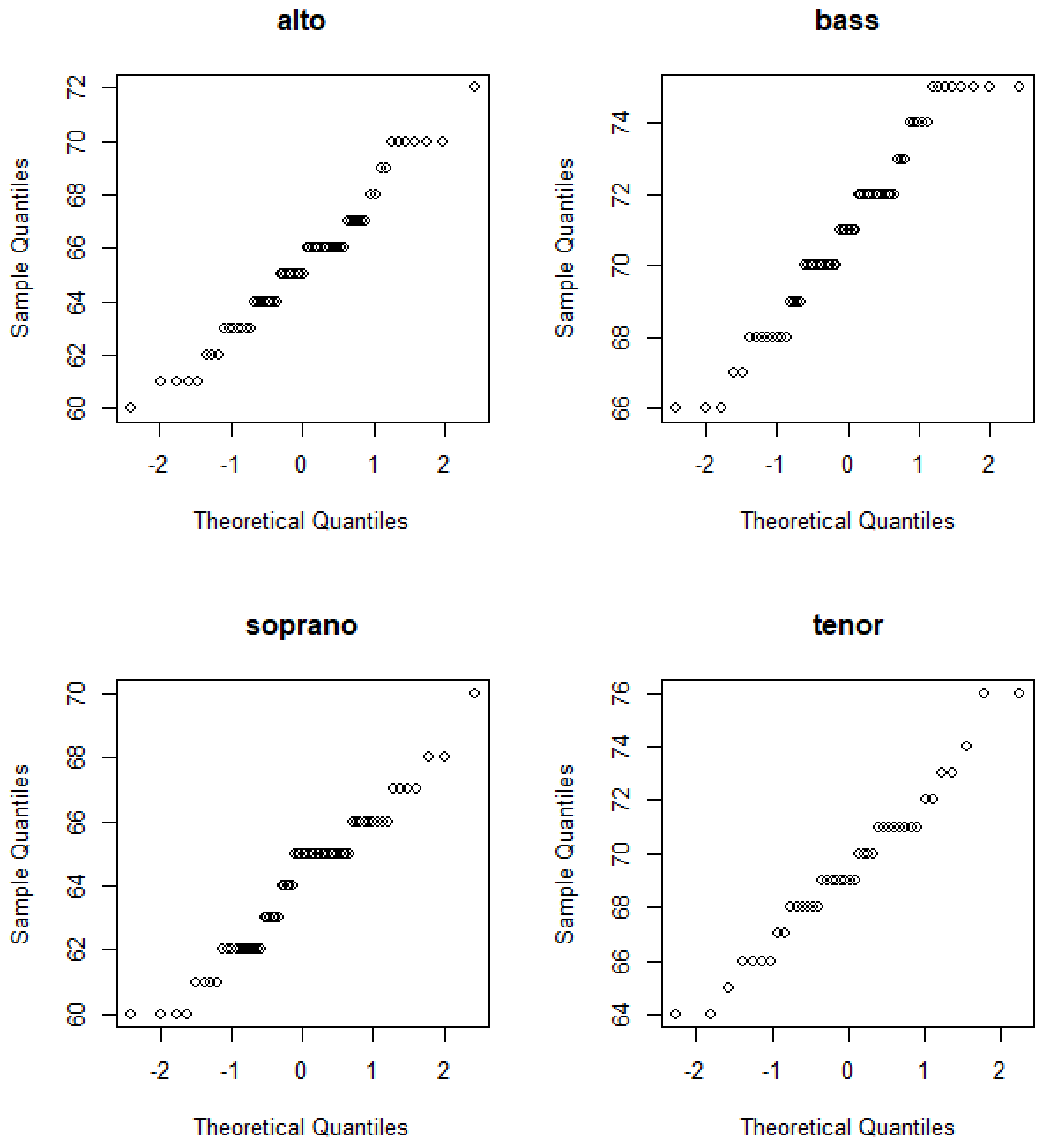
As the correlation value for Tv vs Sales is more than that of Radio vs Sales, we can infer that the sales with advertising on TV is higher than advertising on Radio.

2. **Consider the dataset stored in the file singer.txt on eLearning. This dataset contains heights in inches of the singers from a choral society. The data are grouped according to voice part. There are four voice parts, namely, Bass, Tenor, Alto, and Soprano. The vocal range for each voice part increases in pitch from Bass to Soprano.**

   i. **Perform an exploratory analysis of the data by examining the distributions of the heights of the singers in the four groups. Comment on what you see. Do the four distributions seem similar? Justify your answer.**

```
#Reading the text file into a table
singer = read.table(file = "singer.txt", header = T, sep = ",")

#Storing different types of voice.part in different data frames
par(mfrow = c(1, 1))
alto = subset(singer, voice.part == "Alto")
bass = subset(singer, voice.part == "Bass")
soprano = subset(singer, voice.part == "Soprano")
tenor = subset(singer, voice.part == "Tenor")

#boxplot for Alto, Bass, Soprano and Tenor
boxplot(alto$height, bass$height, soprano$height, tenor$height, names =
c("alto", "bass", "soprano", "tenor"), xlab = "voice part", ylab = "height")

# summary statistics
summary(alto$height)
summary(bass$height)
summary(soprano$height)
summary(tenor$height)

# draw normal QQ plots
par(mfrow = c(2, 2))
qqnorm(alto$height, main = "alto")
qqnorm(bass$height, main = "bass")
qqnorm(soprano$height, main = "soprano")
qqnorm(tenor$height, main = "tenor")
```


75
70
65
60
height
alto
bass
soprano
tenor
voice part



```
> summary(alto$height)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 60.00   64.00   65.00   65.39   67.00   72.00
> summary(bass$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 66.00   69.00   71.00   70.98   72.00   75.00
> summary(soprano$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 60.00   62.00   65.00   64.12   65.00   70.00
> summary(tenor$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   64.0    68.0    69.0    69.4    71.0    76.0
```

Interpretations:

No, the four distributions doesn’t seem similar. Based on mean, Q1 and Q3 the bass singers are the tallest followed by tenors, altos and sopranos. The four distributions have same variability because they have same IQR value (Q3 – Q1 =3) in all cases.

ii. **Is there any difference in the mean heights of Alto and Soprano singers? If yes, how much is the difference? Answer these questions by constructing an appropriate confidence interval. Clearly state the assumptions, if any, and be sure to verify the assumptions.**

```
#Confidence Intervals
alpha = 1-0.95
alto_mean = mean(alto$height)
soprano_mean = mean(soprano$height)

alto_var = var(alto$height)
soprano_var = var(soprano$height)

print(paste("Variance for Alto singers:", alto_var))
print(paste("Variance for Soprano singers:", soprano_var))

n1 = nrow(alto)
n2 = nrow(soprano)

print(paste("Number of Alto singers:", n1))
print(paste("Number of Soprano singers:", n2))

diff = alto_mean - soprano_mean + c(-1,1) * qnorm(1-(alpha/2)) *
sqrt((alto_var/n1) + (soprano_var/n2))
diff
```

Output:

```
> diff
[1] 0.4219834       2.1097859

[1] "Number of Alto singers: 62"
[1] "Number of Soprano singers: 66"

"Variance for Alto singers: 7.02802749867795"
"Variance for Soprano singers: 4.75431235431235"
```

Interpretation:

As 0 doesn’t fall in the 95% confidence interval so we can say that there is difference in the mean heights of Alto and Soprano singers. The difference lies in the interval [0.4219834, 2.1097859].

From the qq-plots we can see that the data is not normal and we can also see that n1 and n2 values are large and sample variances are not equal. So, we have made no assumptions here.

iii. How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?

From (a),
The mean height of Alto singers is : 65.39
The mean height of Soprano singers is: 64.12

The difference in the mean heights of Alto and Soprano singers is 1.27 which lies in the 95% CI calculated in (b), which is [0.4219834, 2.1097859]

As we can observe that 0 is not present in the CI and the CI values are positive, hence we can interpret that the mean height of Alto singers is greater than the mean height of Soprano singers.