FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

SEMESTER 1 SESSION 2020/2021

BITI 2233 STATISTICS AND PROBABILITY

ASSIGNMENT using R Studio

(REPORT)

Group members:

KHAW TENG XIAN	B031910447
LIM WEN NI	B031910441
RECA SENG BINTI MOHD FADZIL SENG	B031910187

Courses: 2 BITI S1G1

Lecturer: TS. DR. SEK YONG WEE

Due date: 24/01/2021

Question 01

Question 1 (20 marks)

1. Use RStudio to answer the following:

    a. Display all the data in a table, then summarize it with frequency details.
    b. Plot a suitable graph for each variable. Explain the results.
    c. Compute the probability value for each total of the spending variable.

    $P = X-\mu/\sigma$

    Total sum of variable B = 27245

    Total square sum of variable B = 5490225

    Variance of variable B = 10051.98

    Standard deviation of Variable B, σ = 100.2595

    μ = 110.752

    Variable B1(x= mean of B1): z = 0.7089324

    P value = 0.7608168

    Variable B2(x= mean of B2): z = 0.5179646

    P value = 0.6977585

    Variable B3(x= mean of B3): z = -0.5499953

    P value = 0.2911613

    Variable B4(x= mean of B4): z = -0.1899541

    P value = 0.4246725

    Variable B5(x= mean of B5): z = -0.02452983

    P value = 0.490215

    Variable B6(x= mean of B6): z = -0.4624177

    P value = 0.3218909

```
#z value and P value for Variable B1
z=(mean(df2$`Variable-B1`)-meanofB)/stdB
pnorm(abs(z))
#0.7608168

#z value and P value for Variable B2
z=(mean(df2$`Variable-B2`)-meanofB)/stdB
pnorm(abs(z))
#0.6977585

#z value and P value for Variable B3
z=(mean(df2$`Variable-B3`)-meanofB)/stdB
pnorm(-abs(z))
#0.2911613
```

```
#z value and P value for Variable B4
z=(mean(df2$`Variable-B4`)-meanofB)/stdB
pnorm(-abs(z))
#0.4246725

#z value and P value for Variable B5
z=(mean(df2$`Variable-B5`)-meanofB)/stdB
pnorm(-abs(z))
#0.490215

#z value and P value for Variable B6
z=(mean(df2$`Variable-B6`)-meanofB)/stdB
pnorm(-abs(z))
#0.3218909
```

d. For all variables, compute the expected values, variances, and standard deviation accordingly. Explain the results.

Mean for Variable A: 991.4634

Variance for Variable A: 3167613

Standard Deviation for Variable A:1779.779

Mean for Variable B1: 181.8293

Variance for Variable B1: 22269.7

Standard deviation for Variable B1: 149.2303

Mean for Variable B2: 162.6829

Variance for Variable B2: 8290.122

Standard deviation for Variable B2: 91.05011

Mean for Variable B3: 55.60976

Variance for Variable B3: 2005.244

Standard deviation for Variable B3: 44.77995

Mean for Variable B4: 91.70732

Variance for Variable B4: 3984.512

Standard deviation for Variable B4: 63.12299

Mean for Variable B5: 108.2927

Variance for Variable B5: 8829.512

Standard deviation for Variable B5: 93.96548

Mean for Variable B6: 64.39024

Variance for Variable B6: 2800.244

Standard deviation for Variable B6: 52.9173

```
mean(df2$`Variable-A`)
#[1] 991.4634
var(df2$`Variable-A`)
#[1] 3167613
sd(df2$`Variable-A`)
#[1] 1779.779
mean(df2$`Variable-B1`)
#[1] 181.8293
var(df2$`Variable-B1`)
#[1] 22269.7
sd(df2$`Variable-B1`)
#[1] 149.2303
mean(df2$`Variable-B2`)
#[1] 162.6829
var(df2$`Variable-B2`)
#[1] 8290.122
sd(df2$`Variable-B2`)
#[1] 91.05011
mean(df2$`Variable-B3`)
#[1] 55.60976
var(df2$`Variable-B3`)
#[1] 2005.244
sd(df2$`Variable-B3`)
#[1] 44.77995
mean(df2$`Variable-B4`)
#[1] 91.70732
var(df2$`Variable-B4`)
#[1] 3984.512
sd(df2$`Variable-B4`)
#[1] 63.12299
```

```
mean(df2$`Variable-B5`)
#[1] 108.2927
var(df2$`Variable-B5`)
#[1] 8829.512
sd(df2$`Variable-B5`)
#[1] 93.96548
mean(df2$`Variable-B6`)
#[1] 64.39024
var(df2$`Variable-B6`)
#[1] 2800.244
sd(df2$`Variable-B6`)
#[1] 52.91733
```

According to the results above, the standard deviation of Variable A is larger than its mean. The standard deviation value widely spread, away from the mean value. Therefore, we can conclude that there is more frequency of extreme values for data of Variable A. Hence, the data in Variable A is more likely to be interpreted by using median values and interquartile ranges to get a better representation of the data. By comparing the mean value of Variable B, students tend to spend the most on house or hostel rental (Variable B1) as the mean value of value of Variable B1 is the largest. On the other hand, by comparing the mean value of Variable B, we can also conclude that students tend to spend the least on transportation (Variable B3) as the mean value of value of Variable B3 is the smallest.

Appendix (Question 1)

Filter

	Dataset	Variable-A	Variable-B1	Variable-B2	Variable-B3	Variable-B4	Variable-B5	Variable-B6
1	Set1 - Normal semester @UTeM - without MCO	600	180	120	250	50	60	60
2	Set1 - Normal semester @UTeM - without MCO	350	120	120	90	60	50	80
3	Set1 - Normal semester @UTeM - without MCO	150	90	90	80	60	70	70
4	Set1 - Normal semester @UTeM - without MCO	350	150	90	50	50	60	80
5	Set1 - Normal semester @UTeM - without MCO	2050	500	400	110	60	600	90
6	Set1 - Normal semester @UTeM - without MCO	450	130	130	50	40	80	0
7	Set1 - Normal semester @UTeM - without MCO	450	140	230	20	50	160	60
8	Set1 - Normal semester @UTeM - without MCO	450	180	150	30	130	60	50
9	Set1 - Normal semester @UTeM - without MCO	800	80	360	20	60	130	60
10	Set1 - Normal semester @UTeM - without MCO	600	110	80	50	130	130	160
11	Set1 - Normal semester @UTeM - without MCO	250	85	70	60	80	60	90
12	Set1 - Normal semester @UTeM - without MCO	1200	230	120	40	80	50	0
13	Set1 - Normal semester @UTeM - without MCO	4200	250	110	50	90	40	60
14	Set1 - Normal semester @UTeM - without MCO	3000	500	210	20	60	230	60
15	Set1 - Normal semester @UTeM - without MCO	1500	80	80	120	110	160	80
16	Set1 - Normal semester @UTeM - without MCO	450	140	130	60	230	80	90
17	Set1 - Normal semester @UTeM - without MCO	150	150	130	50	60	90	80
18	Set1 - Normal semester @UTeM - without MCO	600	250	230	20	130	160	60
19	Set1 - Normal semester @UTeM - without MCO	450	250	120	20	60	160	160
20	Set1 - Normal semester @UTeM - without MCO	1500	100	130	10	50	50	50
21	Set1 - Normal semester @UTeM - without MCO	450	250	70	50	60	60	0

Figure 1 Data for Set 1 (row 1-21)

	Dataset	Variable-A	Variable-B1	Variable-B2	Variable-B3	Variable-B4	Variable-B5	Variable-B6
22	Set1 - Normal semester @UTeM - without MCO	450	250	140	60	150	60	60
23	Set1 - Normal semester @UTeM - without MCO	11000	900	500	20	150	160	60
24	Set1 - Normal semester @UTeM - without MCO	1100	150	190	30	160	80	160
25	Set1 - Normal semester @UTeM - without MCO	450	160	130	20	60	90	60
26	Set1 - Normal semester @UTeM - without MCO	450	120	110	20	40	120	60
27	Set1 - Normal semester @UTeM - without MCO	250	230	210	120	350	250	280
28	Set1 - Normal semester @UTeM - without MCO	800	90	80	40	60	60	0
29	Set1 - Normal semester @UTeM - without MCO	450	160	210	30	220	70	60
30	Set1 - Normal semester @UTeM - without MCO	450	90	220	50	60	80	40
31	Set1 - Normal semester @UTeM - without MCO	450	90	140	50	40	80	60
32	Set1 - Normal semester @UTeM - without MCO	350	230	230	130	60	40	60
33	Set1 - Normal semester @UTeM - without MCO	350	160	80	60	80	60	80
34	Set1 - Normal semester @UTeM - without MCO	450	90	220	40	70	80	0
35	Set1 - Normal semester @UTeM - without MCO	350	160	80	40	150	150	60
36	Set1 - Normal semester @UTeM - without MCO	1100	90	130	50	60	60	50
37	Set1 - Normal semester @UTeM - without MCO	450	90	230	20	30	180	0
38	Set1 - Normal semester @UTeM - without MCO	450	120	150	20	80	60	0
39	Set1 - Normal semester @UTeM - without MCO	450	110	120	40	90	80	60
40	Set1 - Normal semester @UTeM - without MCO	250	110	220	130	120	60	50
41	Set1 - Normal semester @UTeM - without MCO	600	90	110	60	30	80	0

Showing 22 to 41 of 41 entries, 8 total columns

Figure 2 Data for set 1 (row 22-41)

Table 1 Frequency Table for Variable A

	Var1	Freq
1	150	2
2	250	3
3	350	5
4	450	16
5	600	4
6	800	2
7	1100	2
8	1200	1
9	1500	2
10	2050	1
11	3000	1
12	4200	1
13	11000	1

Showing 1 to 13 of 13 entries, 2 total columns

Table 2 Frequency Table for Variable B1

	Var1	Freq
1	80	2
2	85	1
3	90	8
4	100	1
5	110	3
6	120	3
7	130	1
8	140	2
9	150	3
10	160	4
11	180	2
12	230	3
13	250	5
14	500	2
15	900	1

Showing 1 to 15 of 15 entries, 2 total columns

Table 3 Frequency Table for B2

	Var1	Freq
1	70	2
2	80	5
3	90	2
4	110	3
5	120	5
6	130	6
7	140	2
8	150	2
9	190	1
10	210	3
11	220	3
12	230	4
13	360	1
14	400	1
15	500	1

Table 4 Frequency Table for B3

	Var1	Freq
1	10	1
2	20	10
3	30	3
4	40	5
5	50	9
6	60	5
7	80	1
8	90	1
9	110	1
10	120	2
11	130	2
12	250	1

Showing 1 to 12 of 12 entries, 2 total columns

Table 5 Frequency Table for B4

	Var1	Freq
1	30	2
2	40	3
3	50	4
4	60	13
5	70	1
6	80	4
7	90	2
8	110	1
9	120	1
10	130	3
11	150	3
12	160	1
13	220	1
14	230	1
15	350	1

Showing 1 to 15 of 15 entries, 2 total columns

Table 6 Frequency Table for B5

	Var1	Freq
1	40	2
2	50	3
3	60	11
4	70	2
5	80	8
6	90	2
7	120	1
8	130	2
9	150	1
10	160	5
11	180	1
12	230	1
13	250	1
14	600	1

Showing 1 to 14 of 14 entries, 2 total columns

Table 7 Frequency Table for Variable B6

	Var1	Freq
1	0	8
2	40	1
3	50	4
4	60	15
5	70	1
6	80	5
7	90	3
8	160	3
9	280	1

Showing 1 to 9 of 9 entries, 2 total columns

```
FreqA<-table(df2$`Variable-A`, exclude = NULL)

View(FreqA)

FreqB1<-table(df2$`Variable-B1`, exclude = NULL)
```

```
View(FreqB1)

FreqB2<-table(df2$`Variable-B2`, exclude = NULL)

View(FreqB2)

FreqB3<-table(df2$`Variable-B3`, exclude = NULL)

View(FreqB3)

FreqB4<-table(df2$`Variable-B4`, exclude = NULL)

View(FreqB4)

FreqB5<-table(df2$`Variable-B5`, exclude = NULL)

View(FreqB5)

FreqB6<-table(df2$`Variable-B6`, exclude = NULL)

View(FreqB6)
```

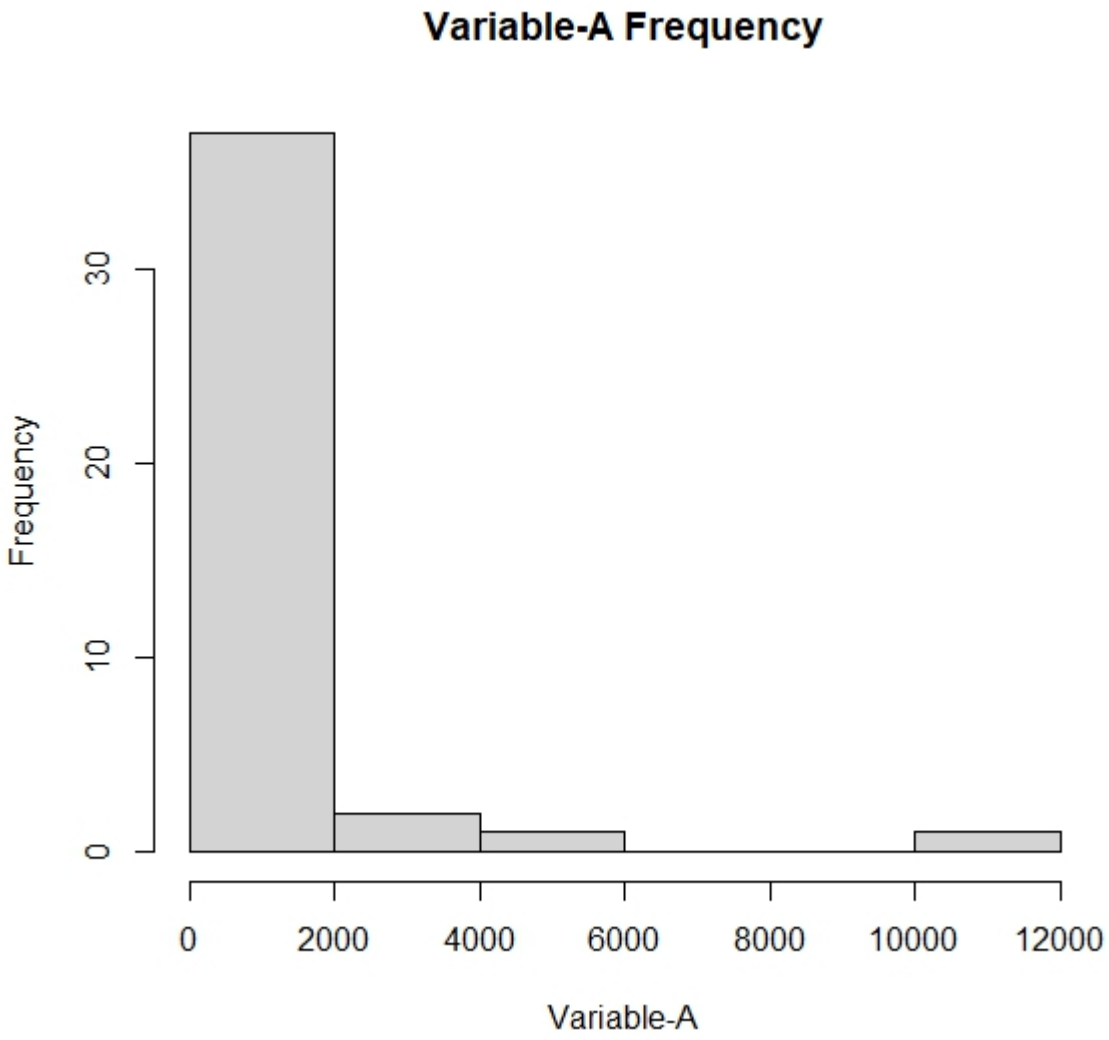

Figure 3 Histogram of Variable A

Most of the students in Dataset 1 receive pocket money around RM0-RM2000 each semester. This is because the according to the histogram in Figure 3, the frequency of the pocket money received each semester between RM0 to RM2000 is the highest. Meanwhile, least students in Dataset 1 receive pocket money within RM4000-RM6000 as well as RM10000-RM12000 each semester. This is because according to the histogram in Figure 3, the frequency of the pocket money received each semester within RM4000-6000 as well as RM10000-RM12000 is the lowest.

```
x=df2$`Variable-A`

hist(x,xlab = "Variable-A", ylab = "Frequency", main = "Variable-A Frequency")
```

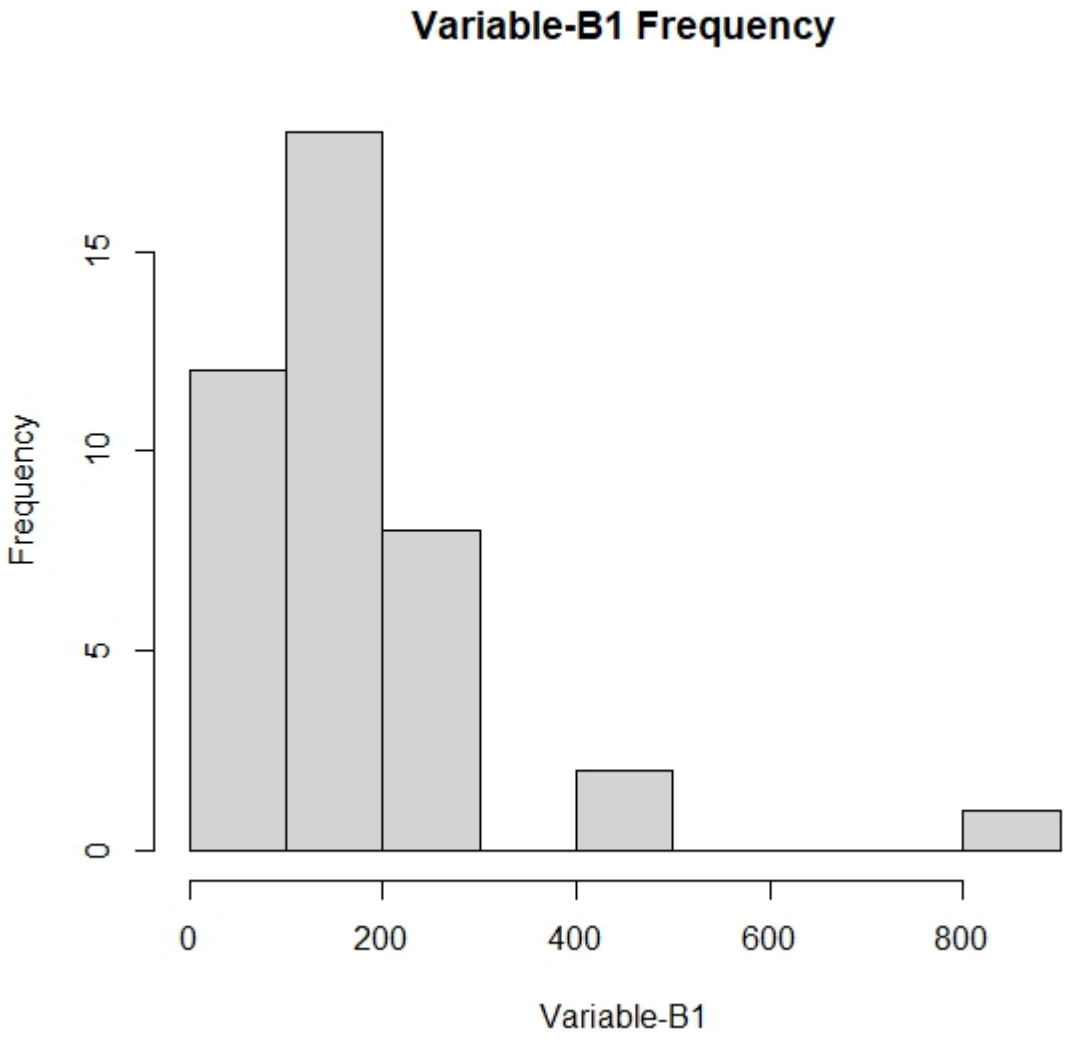

Figure 4 Histogram of Variable B1

Most of the students in Dataset 1 spend between RM100-RM200 per month on house or hostel rental. This is because the according to the histogram in Figure 4, the frequency of the spending per month on house or hostel rental between RM100-RM200 is the highest. Meanwhile, least students in Dataset 1 spend above RM800 per month on house or hostel rental. This is because the according to the histogram in Figure 4, the frequency of the spending per month on house or hostel rental above RM800 is the lowest.

```
x=df2$`Variable-B1`

hist(x,xlab = "Variable-B1", ylab = "Frequency", main = "Variable-B1 Frequency")
```

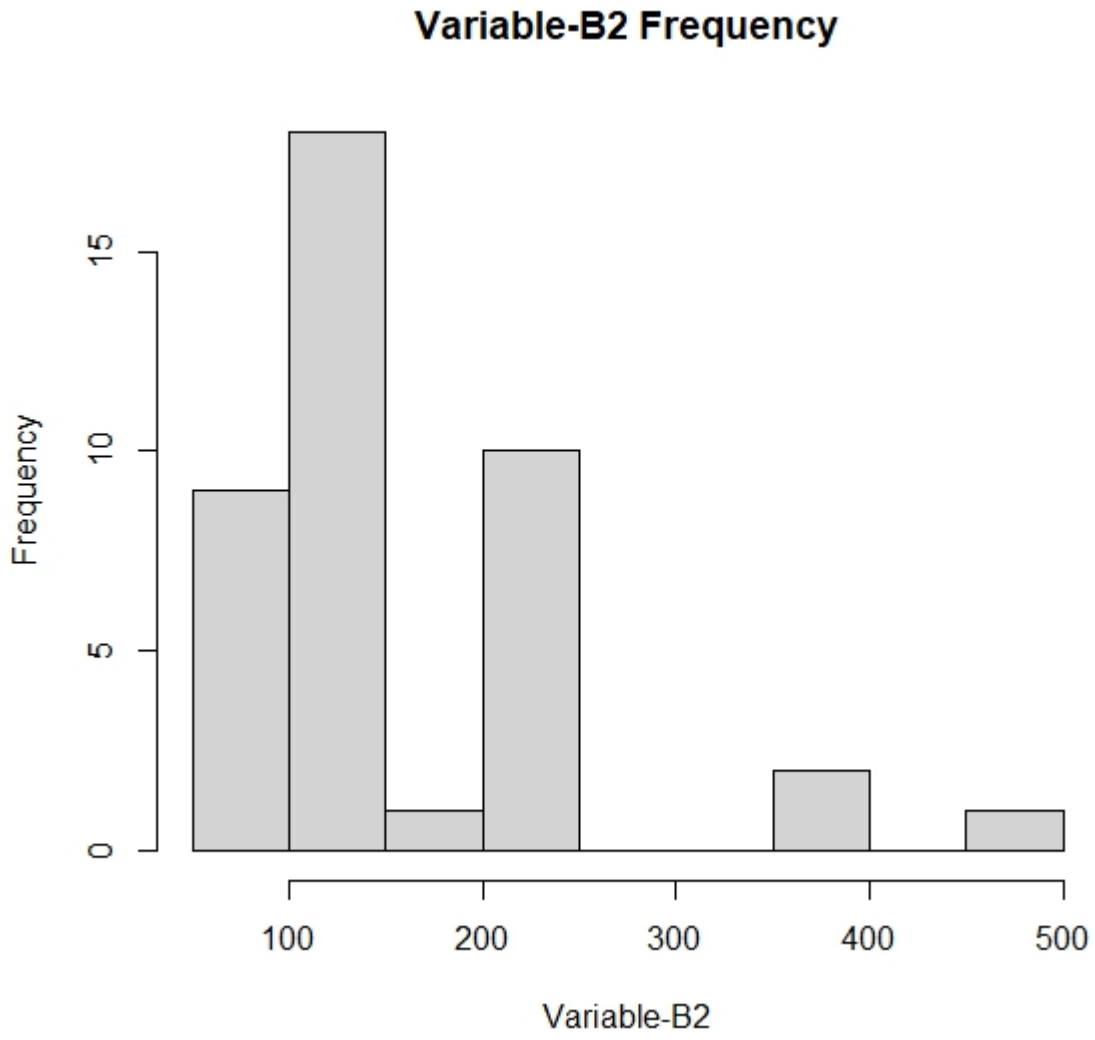

Figure 5 Histogram of Variable B2

Most of the students in Dataset 1 spend between RM100-RM150 per month on food and groceries. This is because the according to the histogram in Figure 5, the frequency of the spending per month on food and groceries between RM100-RM150 is the highest. Meanwhile, least students in Dataset 1 spend between RM150-RM200 as well as between RM450-RM500 per month on food and groceries. This is because the according to the histogram in Figure 5, the frequency of the spending per month on food and groceries between RM150-RM200 as well as between RM450-RM500 are the lowest.

```
x=df2$`Variable-B2`

hist(x,xlab = "Variable-B2", ylab = "Frequency", main = "Variable-B2 Frequency")
```

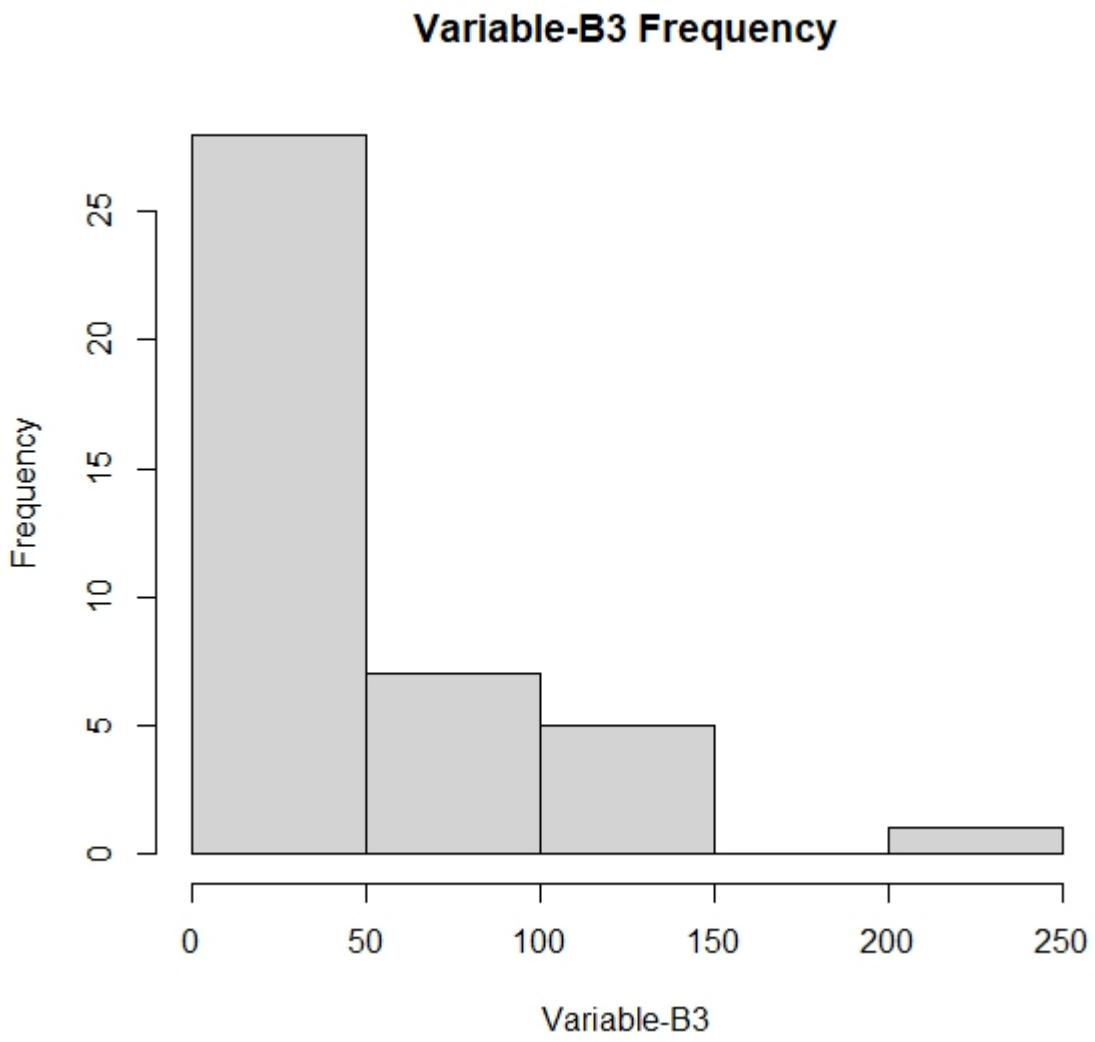

Figure 6 Histogram of Variable B3

Most of the students in Dataset 1 spend between RM0-RM50 per month on transportation. This is because the according to the histogram in Figure 6, the frequency of the spending per month on food and groceries between RM0-RM50 is the highest. Meanwhile, least students in Dataset 1 spend between RM200-RM250 per month on transportation. This is because the according to the histogram in Figure 6, the frequency of the spending per month per month on transportation between RM200-RM250 is the lowest.

```
x=df2$`Variable-B3`

hist(x,xlab = "Variable-B3", ylab = "Frequency", main = "Variable-B3 Frequency")
```

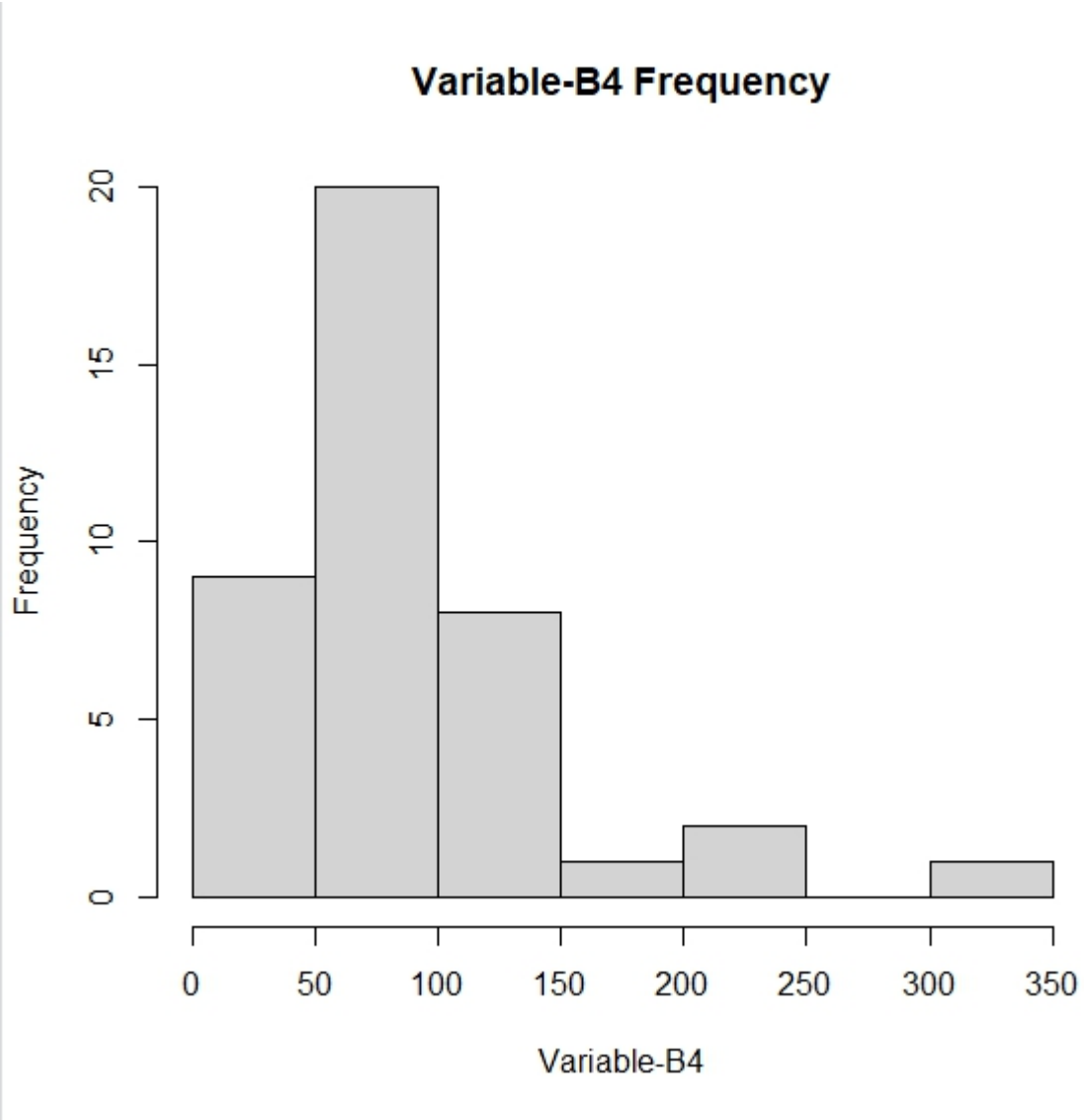

Figure 7 Histogram of Variable B4

Most of the students in Dataset 1 spend between RM50-RM100 per month on bill and utilities. This is because the according to the histogram in Figure 7, the frequency of the spending per month on bill and utilities between RM50-RM100 is the highest. Meanwhile, least students in Dataset 1 spend between RM150-RM200 as well as between RM300-RM350 per month on bill and utilities. This is because the according to the histogram in Figure 7, the frequency of the spending per month per month on bill and utilities between RM150-RM200 as well as between RM300-RM350 are the lowest.

```
x=df2$`Variable-B4`

hist(x,xlab = "Variable-B4", ylab = "Frequency", main = "Variable-B4 Frequency")
```

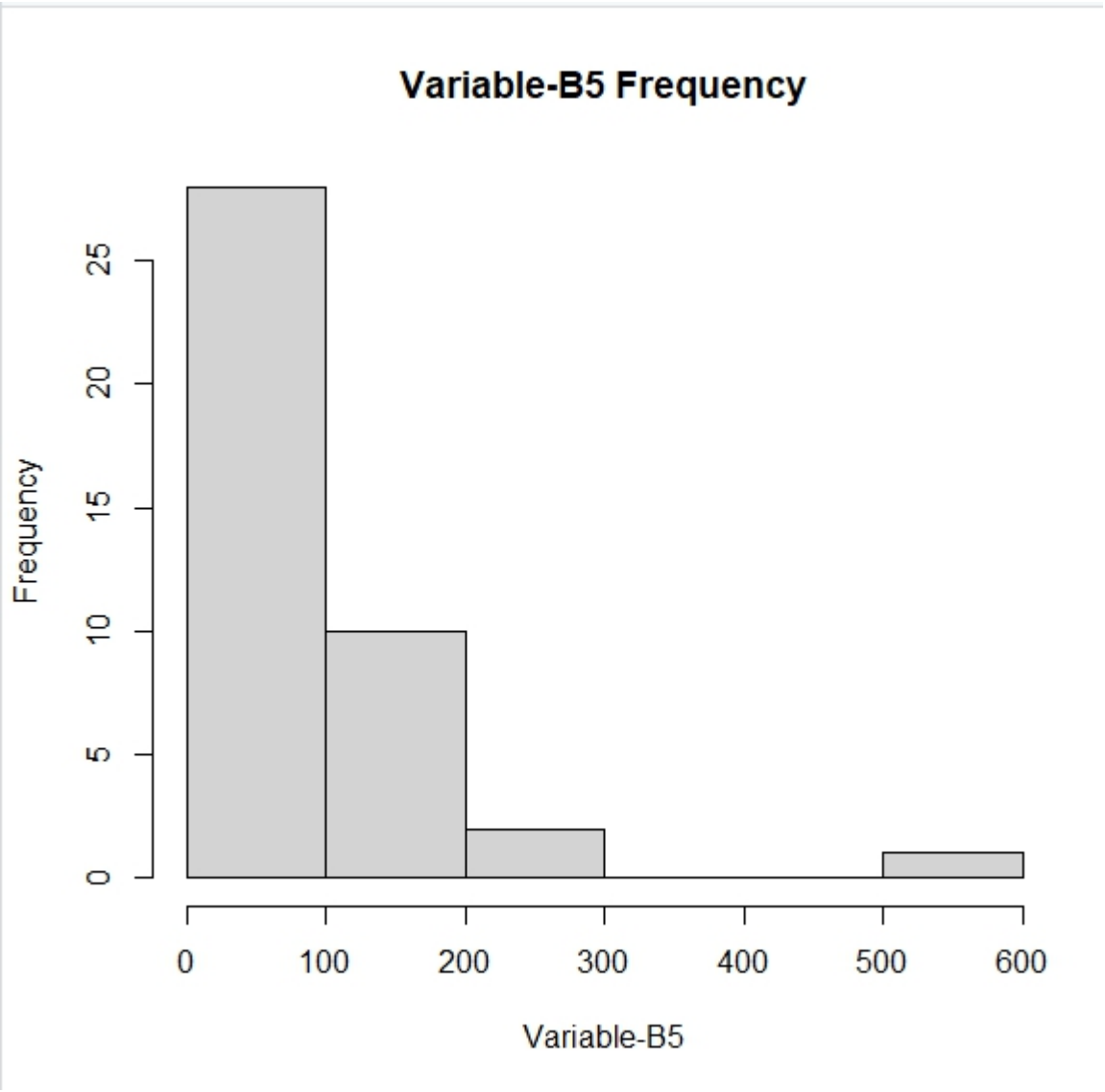

Figure 8 Histogram of Variable B5

Most of the students in Dataset 1 spend between RM0-RM100 per month on saving an emergency fund. This is because the according to the histogram in Figure 8, the frequency of the spending per month on saving an emergency fund between RM0-RM100 is the highest. Meanwhile, least students in Dataset 1 spend between RM500-RM600 per month on saving an emergency fund. This is because the according to the histogram in Figure 8, the frequency of the spending per month per month on saving an emergency fund between RM500-RM600 is the lowest.

```
x=df2$`Variable-B5`

hist(x,xlab = "Variable-B5", ylab = "Frequency", main = "Variable-B5 Frequency")
```

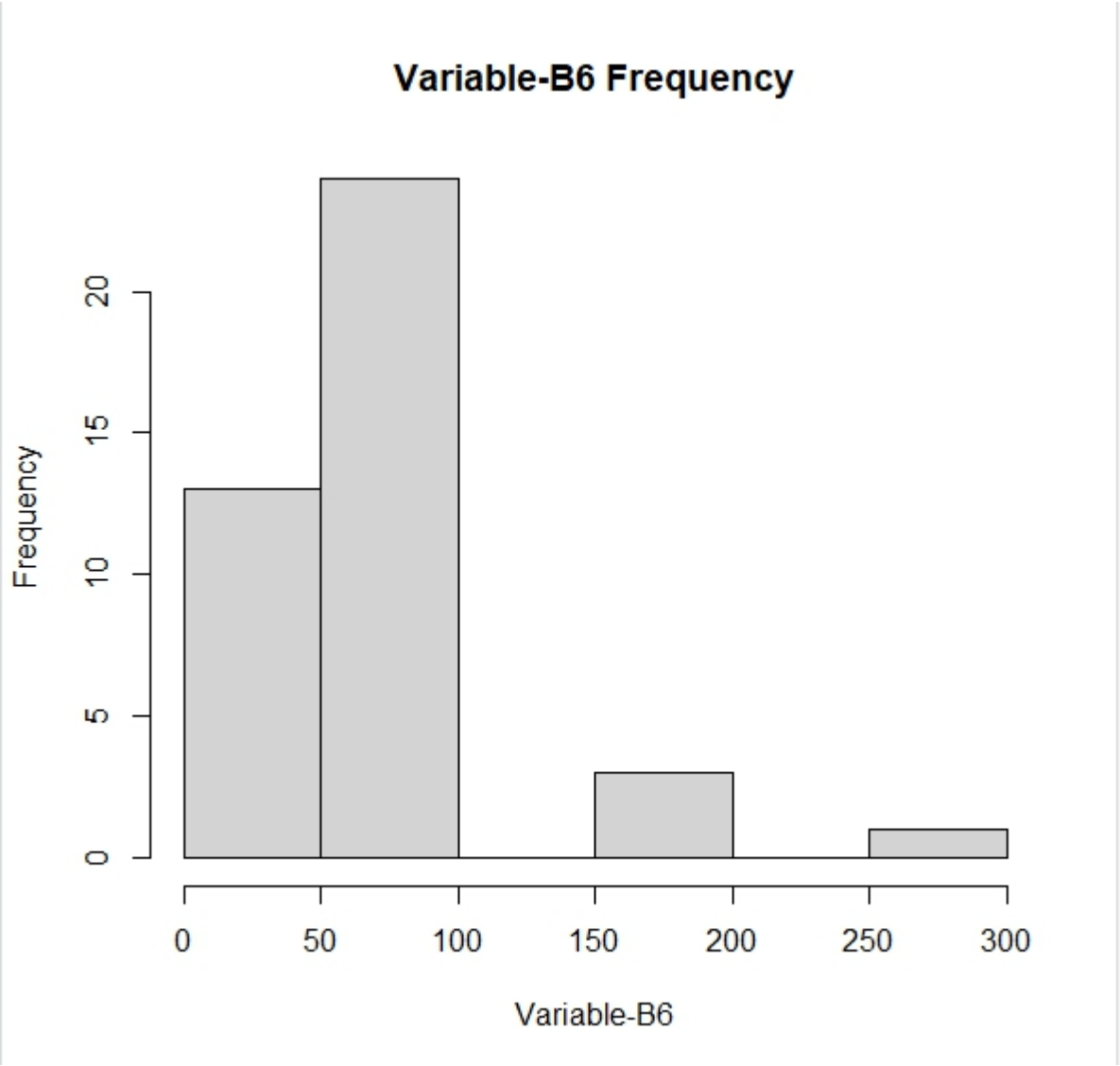

Figure 9 Histogram of Variable B6

Most of the students in Dataset 1 spend between RM50-RM100 per month on loan repayment. This is because the according to the histogram in Figure 9, the frequency of the spending per month on loan repayment between RM50-RM100 is the highest. Meanwhile, least students in Dataset 1 spend between RM250-RM300 per month on loan repayment. This is because the according to the histogram in Figure 9, the frequency of the spending per month per month on loan repayment between RM250-RM300 is the lowest.

```
x=df2$`Variable-B6`
hist(x,xlab = "Variable-B6", ylab = "Frequency", main = "Variable-B6 Frequency")
```

Question 02

Commands in R studio:

```r
#Question_02
#Question_02_(a)
#a
X~binom(n=31,p=0.447)
#b
plot(c(1:31),dbinom(c(1:31),size=31,prob=0.447),type="h",ylab="Probability Mass",xlab="Number of Success")
#c
plot(c(1:31),pbinom(c(1:31),size=31,prob=0.447),type="h",ylab="Cumulative Probability",xlab="Number of Success")
#d
dbinom( 17, size=31, prob=0.447)
#e
pbinom( 13, size= 31, prob= 0.447)
#f
pbinom( 11, size=31, prob= 0.447, lower.tail = FALSE )
#g
pbinom( 14, size = 31, prob = 0.447, lower.tail = FALSE)
#h
diff(pbinom(c(19,15), size=31, prob=0.447,lower.tail= FALSE))
#i
library(distrEx)
X = Binom(size = 31, prob =0.447)
E(X)
#j
var(X)
#k
sd(X)
#l
E(4* X + 51.324)

```

```r
#Question_02_(b)
#probability for all are chicken
(choose(8,3)*choose(7,0))/choose(15,3)

#probability for all are shrimp
(choose(8,0)*choose(7,3))/choose(15,3)

#probability for all have the same filling
((choose(8,3)*choose(7,0))/choose(15,3))+((choose(8,0)*choose(7,3))/choose(15,3))
```

Answer in console:

```
Console  Terminal  Jobs
C:/Users/User/Desktop/Task/z_STATISTIC/Statistic_ass/
> X~binom(n=31,p=0.447)
X ~ binom(n = 31, p = 0.447)
> plot(c(1:31),dbinom(c(1:31),size=31,prob=0.447),type="h",ylab="Probability Mass",xlab="Number of Success")
> plot(c(1:31),pbinom(c(1:31),size=31,prob=0.447),type="h",ylab="Cumulative Probability",xlab="Number of Success")
> dbinom( 17, size=31, prob=0.447)
[1] 0.07532248
> pbinom( 13, size= 31, prob= 0.447)
[1] 0.451357
> pbinom( 11, size=31, prob= 0.447, lower.tail = FALSE )
[1] 0.8020339
> pbinom( 14, size = 31, prob = 0.447, lower.tail = FALSE)
[1] 0.406024
> diff(pbinom(c(19,15), size=31, prob=0.447,lower.tail= FALSE))
[1] 0.2544758
> library(distrEx)
```

```
> X = Binom(size = 31, prob =0.447)
> E(X)
[1] 13.857
> var(X)
[1] 7.662921
> sd(X)
[1] 2.768198
> E(4* X + 51.324)
[1] 106.752
> (choose(8,3)*choose(7,0))/choose(15,3)
[1] 0.1230769
> (choose(8,0)*choose(7,3))/choose(15,3)
[1] 0.07692308
> ((choose(8,3)*choose(7,0))/choose(15,3))+((choose(8,0)*choose(7,3))/choose(15,3))
[1] 0.2
>
```

Graph for Question 2(a) #b:

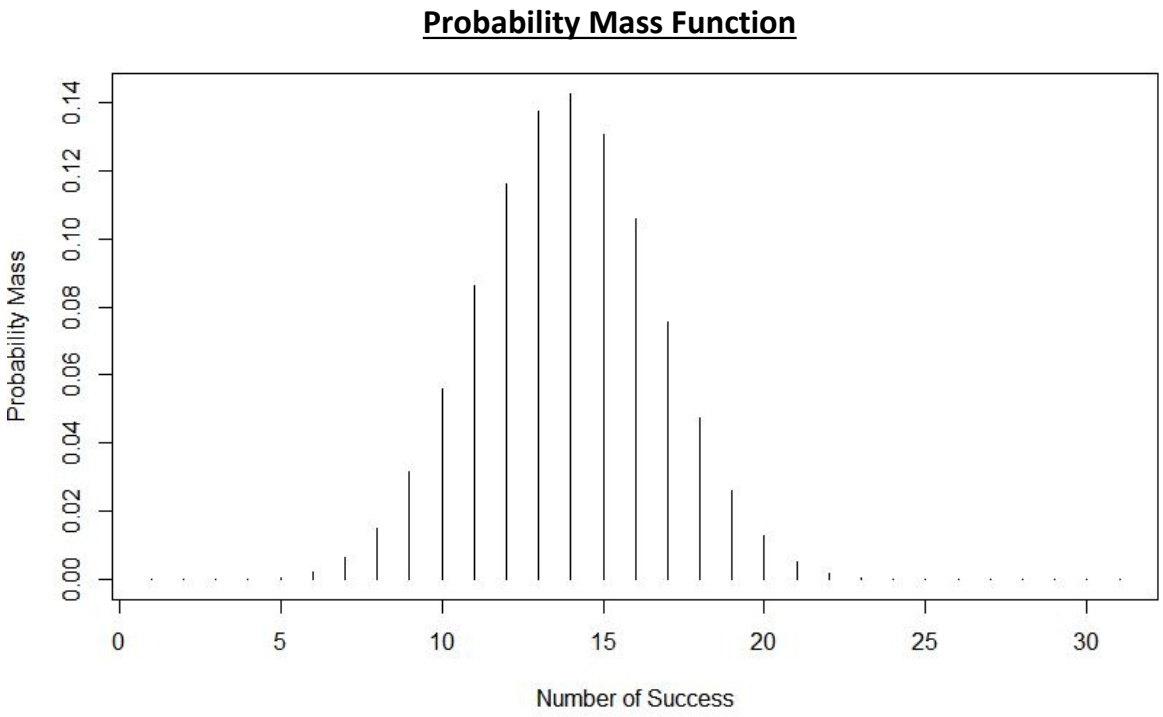

Figure 1 Probability Mass Function

Graph for Question 2(a) #c:

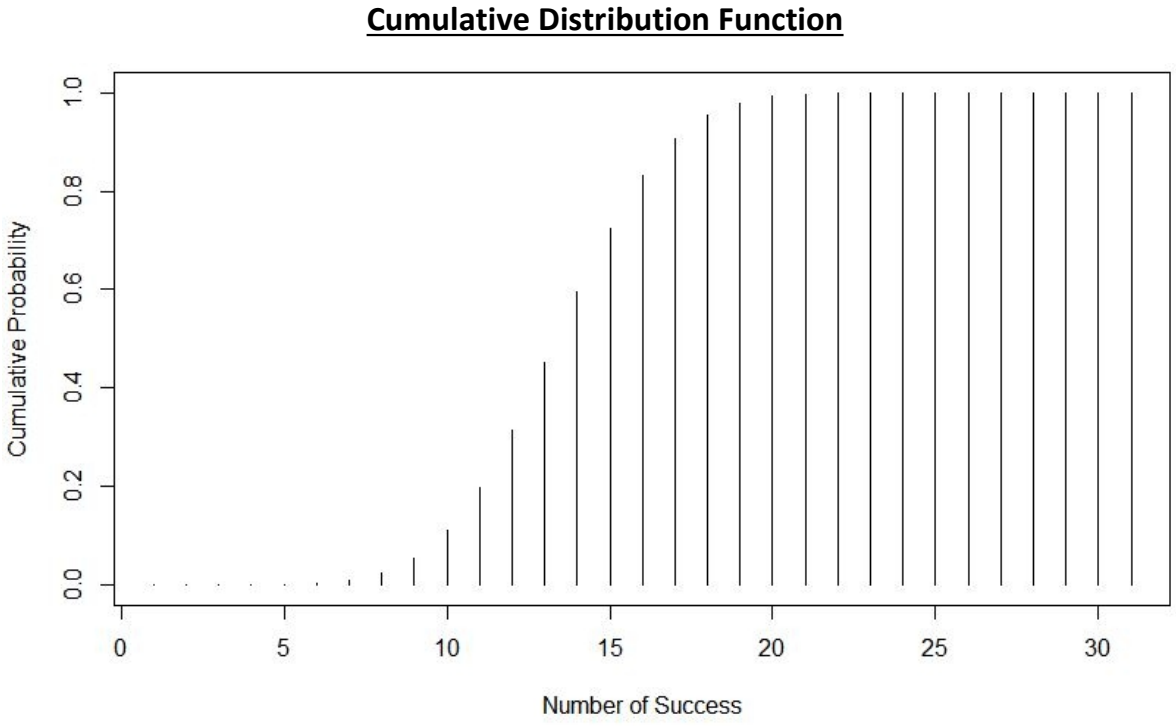

Figure 2 Cumulative Distribution Function

Question 03 (20 marks)

Student's height analysis You are interested to investigate the heights of students in the provided data set. The students data set consists of 8239 rows, each of them representing a particular student, and 16 columns, each of them corresponding to a variable/feature related to that particular student. These self-explaining variables are stud.id, name, gender, age, height, weight, religion, nc.score, semester, major, minor, score1, score2, online.tutorial, graduated, salary. Based on this data set, answer the following question by using R:

1. Create a histogram of the height of students. (you may use ggplot2 package to make a variety of histogram).

```
ggplot()+geom_histogram(data = Q3,aes(x=height),fill="white",col =
"red")+ggtitle("Histogram of height")+theme(plot.title = element_text(hjust = 0.5))
```

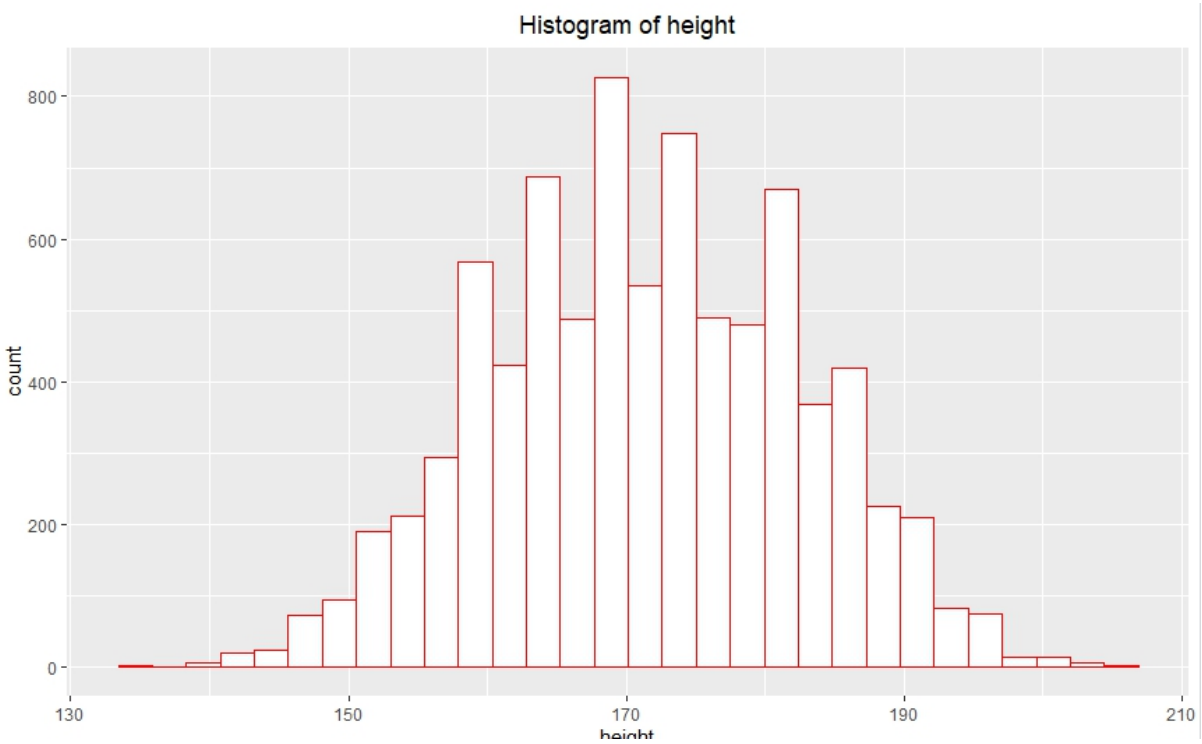

Figure 3.1: Histogram of height of students

2. Create a histogram of the height of males and females.

```
ggplot()+geom_histogram(data =
Q3,aes(x=height,color=gender),fill="white")+ggtitle("Histogram of height")+theme(plot.title
```

```
= element_text(hjust = 0.5))
```

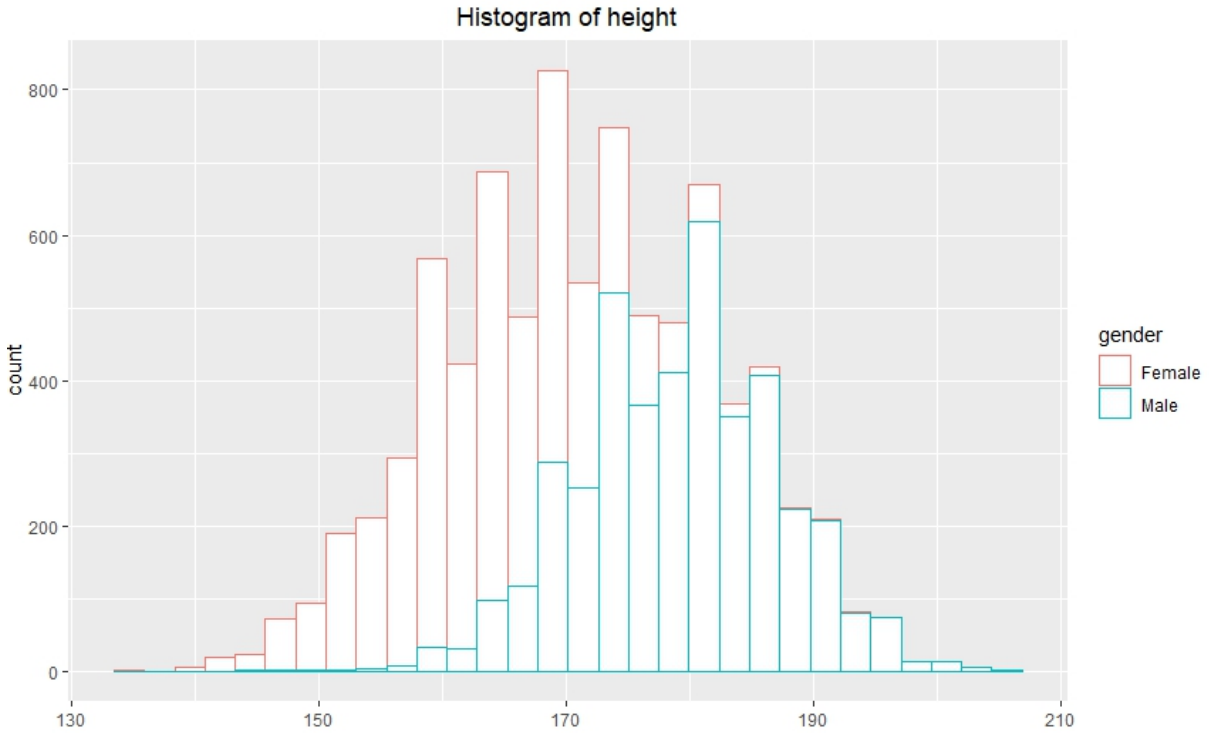

Figure 3.2: Histogram of height of males and females

3. Calculate the mean and standard deviation of the height of males and females.

```
maleMean = mean(Q3[Q3$height & Q3$gender == "Male","height"])
maleSd = sd(Q3[Q3$height & Q3$gender == "Male","height"])
```

For male:

Mean= 179.0727

Standard deviation = 7.988852

```
femaleMean = mean(Q3[Q3$height & Q3$gender == "Female","height"])
femaleSd = sd(Q3[Q3$height & Q3$gender == "Female","height"])
```

For female:

Mean =163.6533

Standard deviation = 7.919726

4. Calculate the probability of a randomly picked female student from the student data set with a height less or equal to 161 cm.

```
pnorm(161,femaleMean,femaleSd)
```

0.3688041

5. Calculate the probability of a randomly picked female student from the students data set with a height higher or equal to 170 cm

```
(1-pnorm(170,femaleMean,femaleSd))
```

0.2114556

6. Calculate the probability of a randomly picked female student from the students data set with a height between 150 and 160.

```
(pnorm(160,femaleMean,femaleSd)-pnorm(150,femaleMean,femaleSd))
```

0.2799379

7. Calculate the probability of a randomly picked female student from the students data set with a height between 170 and 180.

```
(pnorm(180,femaleMean,femaleSd)-pnorm(170,femaleMean,femaleSd))
```

0.1919492

8. What can you conclude both probabilities in questions 6 and 7?

Probability of a randomly picked female student from the students data set with a height between 150 and 160 is higher than probability of a randomly picked female student from the students data set with a height between 170 and 180. We more probably picked female with height between 150 and 160 than picked female with height between 170 and 180

9. We want to know the height of female students in our students data set that corresponds to a probability of 0.25.

```
qnorm(0.25,femaleMean,femaleSd)
```

158.3115

P(X<158.3115) = 0.25

10. We want to know the height of male students in our students data set that corresponds to a probability of 0.25.

```
qnorm(0.25,maleMean,maleSd)
```

173.6843

$P(X < 173.6843) = 0.25$

-Question 4a (10 marks)

Research about drawbacks of using social media among students have been carried out and the students will answer their degree of agreement about the statements on the drawbacks of using social media based on a 5-point Likert scale:

1 2 3 4 5

Strongly disagree ⟶ Strongly agree

The results of the previous research carried in one of the universities in Malaysia are as in Table 1.

Table 1: Results from previous research

Statements on drawbacks of using social media among students	**The average degree of agreement**
Statement 1: Using social media is stressful.	2.38
Statement 2: Social media damages one's reputation because of a lack of privacy.	3.02
Statement 3: The use of social media consumes both time and effort.	3.89
Statement 4: Social media can have a negative influence on student productivity.	4.05
Statement 5: The use of social media detaches me from direct contact with others.	4.19
Statement 6: There is a risk of negative comments on social media.	4.72

By using dataset **Dataset Q4**, select randomly 100 observations which consist of the results of the survey from FTMK students. The dataset contains gender, course, and answers for each question.

Based on the results from the previous research and the selected observations from **Dataset Q4**, answer the following questions. Assume that all assumptions are true for constructing confidence interval and hypothesis testing.
Report the hypothesis testing process step by step and conclude the findings.

a. Construct a 70% confidence interval for the average degree of agreement of statement 3 for all students.

   From the sample data,

   $\alpha = 0.30$

Sample standard deviation = 1.06837

n=100

MoE = ±0.1107294

Average = 2.90

μ = (2.789271, 3.010729)

We are 70% confident that the interval from 2.789271 to 3.010729 has the average degree of agreement of statement 3 of all students.

```
> #Question 4a.a
> stdS3<-sd(sampleData$stmnt3); stdS3
[1] 1.06837
> n=100
> error <- qnorm(0.85)*stdS3/sqrt(n); error
[1] 0.1107294
> avgS3<-mean(sampleData$stmnt3); avgS3
[1] 2.9
> left <- avgS3-error; left
[1] 2.789271
> right<- avgS3+error; right
[1] 3.010729
>
```

b. Construct an 80% confidence interval for the average degree of agreement of statement 3 for all students.

From the sample data,

α= 0.20

Sample standard deviation = 1.06837

n=100

MoE = ± 0.1369171

Average = 2.90

μ = (2.763083, 3.036917)

We are 80% confident that the interval from 2.763083 to 3.036917 has the average degree of agreement of statement 3 of all students.

```
> #Question 4a.b
> stdS3<-sd(sampleData$stmnt3); stdS3
[1] 1.06837
> n=100
> error <- qnorm(0.90)*stdS3/sqrt(n); error
[1] 0.1369171
> avgS3<-mean(sampleData$stmnt3); avgS3
[1] 2.9
> left <- avgS3-error; left
[1] 2.763083
> right<- avgS3+error; right
[1] 3.036917
>
```

c. Construct a 90% confidence interval for the average degree of agreement of statement 3 for all students.

From the sample data,

$\alpha = 0.10$

Sample standard deviation = 1.06837

n=100

MoE = ± 0.1757312

Average = 2.90

μ = (2.724269, 3.075731)

We are 90% confident that the interval from 2.724269 to 3.075731 has the average degree of agreement of statement 3 of all students.

```
> #Question 4a.c
> stdS3<-sd(sampleData$stmnt3); stdS3
[1] 1.06837
> n=100
> error <- qnorm(0.95)*stdS3/sqrt(n); error
[1] 0.1757312
> avgS3<-mean(sampleData$stmnt3); avgS3
[1] 2.9
> left <- avgS3-error; left
[1] 2.724269
> right<- avgS3+error; right
[1] 3.075731
>
```

d. * What can you comment about the width of the confidence interval obtained from (a), (b), and (c)?

As the significance level increases, the width of the confidence interval also increases.

e. Test at the 3.5% significance level if the average degree of agreement for statement 3 has been decreased from previous research by using the p-value approach.

Step 1: H_0: $\mu \geq 3.89$, H_1: $\mu < 3.89$

Step 2: $\mu = 3.89$

xbar = 2.90

n=100

Z= -9.266454

Step 3: P-value = 9.621871e-21

Step 4: $\alpha = 0.35$ -> P-value<0.35

Reject H_0 at $\alpha = 0.35$.

Step 5: The average degree of agreement for statement 3 has been decreased from previous research.

```
> #Question 4a.e
> meansample<-mean(sampleData$stmnt3); meansample
[1] 2.9
> averageS3 = 3.89
> n=100
> z <- (meansample-averageS3)/(sd(sampleData$stmnt3)/sqrt(n)); z
[1] -9.266454
> pnorm(-abs(z))
[1] 9.621871e-21
>
```

f. Test at the 1% significance level if the average degree of agreement for statement 4 has been increased from previous research by using the p-value approach.

Step 1: $H_0 : \mu \leq 4.05$, $H_1 : \mu > 4.05$

Step 2: : $\mu = 4.05$

xbar = 3.27

n=100

Z= -7.214841

Step 3: P-value = 2.699855e-13

Step 4: $\alpha = 0.01$ -> P-value<0.01

Reject H_0 at $\alpha = 0.01$.

Step 5: The average degree of agreement for statement 4 has been increased from previous research.

```
> #Question 4a.f
> meansample<-mean(sampleData$stmnt4); meansample
[1] 3.27
> averageS4 = 4.05
> n=100
> z <- (meansample-averageS4)/(sd(sampleData$stmnt4)/sqrt(n)); z
[1] -7.214841
> pnorm(-abs(z))
[1] 2.699855e-13
>
```

g. Test at the 5% significance level if the average degree of agreement for statement 5 is the same as previous research by using the p-value approach.

Step 1: $H_0 : \mu = 4.19$, $H_1 : \mu \neq 4.19$

Step 2: : μ=4.19

xbar=2.96

n=100

Z = -9.867067

Step 3: P-value= 5.783092e-23

Step 4: $\alpha = 0.05$ -> P-value<0.05

Reject H_0 at $\alpha = 0.05$.

Step 5: The average degree of agreement for statement 5 is not the same as previous research.

```
> #Question 4a.g
> meansample<-mean(sampleData$stmnt5); meansample
[1] 2.96
> averageS5 = 4.19
> n=100
> z <- (meansample-averageS5)/(sd(sampleData$stmnt5)/sqrt(n)); z
[1] -9.867067
> 2*pnorm(-abs(z))
[1] 5.783092e-23
>
```

h. * Based on your literature research on the internet, in 1-page, give your own opinions and comments on the drawbacks of using social media among students. Relate your opinions and comments with the results in Table 1.

Based on my literature research on the internet, I agree with the results obtained in Table 1 regarding the drawbacks of using social media among students. Anyone around the world can now be 'Friends' on social media. 'Friends' can be made by sending a request and getting approval from the social media user. People who are 'Friends' on social media might not be actual friends in real life. A random stranger can be 'Friends' too. They can be anyone who is not close to you and not knowing you personally. On social media, people have the freedom to give feedback on any posts that are uploaded to the platform. Some strangers can just use social media to vent pressure and anger by commenting using curses and inappropriate words on someone. By looking at the comments section, it leads to social anxiety, depression, and exposure to content that is not appropriate which can adversely affect mood. Chronic social media users are prone to report having poor mental health, including symptoms of anxiety and depression. Social bullying cases are also increasing as a result. As there are no ways in restricting people to give negative comments on social media, I agree with Statement 6 for getting the highest average degree of agreement and being the top drawback of using social media among students.

Browsing social media can be stressful. Nowadays, people are focusing more on social media than building friendships and interact with family members. As the fear of missing out on the latest trend and updates, people tend to spend a lot of time on social media to reply to comments, posting photos, videos, share interesting social media posts, playing games, etc. Heavy attention is put on every post to upload to get good feedback and reviews from their friends and followers. However, this issue can be easily overcome by restricting the necessities and duration of using social media. For example, people can set a timer to track how long they spend on social media and manage their time to be away from it. Besides, we can choose to turn off notifications so we will not easily get distracted while trying to concentrate on completing the daily tasks. As there are many ways of dealing with stress while using social media, I agree with Statement 1 for getting the lowest average degree of agreement and being the bottom drawback of using social media among students.

In conclusion, the existence of social media comes with its pros and cons. Students need to acknowledge the importance of discipline while using social media. Students should continuously filter the information available on social media so that they will not be distracted. Advice and guidance from family members are necessary so that social media platforms can be utilized in a meaningful way.

-Question 4b (10 marks)

Collect the prices of 100 houses to be sold with the same characteristics (for eg 3 bedrooms, single-storey, built-in area = min square feet is 2000) from any online advertisement (for eg mudah.com). Find it's the mean and standard deviation. Assume that the mean and standard deviation you've obtained is the mean and standard deviation for the population. Then, select any prices from a particular state (from 100 data you've selected i.e for eg Melaka with sample size = frequency of data from Melaka) and find its mean and standard deviation. Assume that the mean and standard deviation is the sample mean and sample standard deviation respectively.
Assume that all assumptions are true for constructing confidence interval and hypothesis testing.

```
pmean<-mean(Dataset_Q4b_Price_of_houses_in_Malaysia$`Price (RM)`); pmean
```

Population mean = 572003.1

```
Pstd<-sd(Dataset_Q4b_Price_of_houses_in_Malaysia$`Price (RM)`); Pstd
```

Population standard deviation = 221829.6

```
johormean<-mean(johorhouses$`Price (RM)`); johormean
```

Sample mean (State: Johor) = 639727.3

```
johorstd<-sd(johorhouses$`Price (RM)`); johorstd
```

Sample standard deviation (State: Johor) = 92284.44

a. Compute the 80% confidence intervals for the prices from sample data.

```
Pstd<-sd(Dataset_Q4b_Price_of_houses_in_Malaysia$`Price (RM)`)
n=100
error <- qnorm(0.90)*Pstd/sqrt(n); error
johormean<-mean(johorhouses$`Price (RM)`)
left <- johormean-error; left
```

```
right<- johormean+error; right
```

Significance level = 0.20

Population standard deviation = 221829.6

n=100

MoE = ± 28428.61

Average = 639727.3

μ = (611298.7, 668155.9)

We are 80% confident that the interval from 611298.7 to 668155.9 for the prices from sample data.

b. Conduct a test of hypothesis to determine whether the price of houses for the chosen state is less than the population mean price. Report the hypothesis testing process step by step and conclude the findings.

```
johormean<-mean(johorhouses$`Price (RM)`); johormean
pmean<-mean(Dataset_Q4b_Price_of_houses_in_Malaysia$`Price (RM)`); pmean
n=100
z <- (johormean-pmean)/(sd(Dataset_Q4b_Price_of_houses_in_Malaysia$`Price (RM)`)/sqrt(n)); z
```

Step 1: $H_0 : \mu \geq 572003.1$, $H_1 : \mu < 572003.1$

Step 2: $\mu = 572003.1$

xbar = 639727.3

n=100

Population standard deviation = 221829.6

Z= 3.052982

Step 3: At $\alpha = 0.05$ -> $Z_\alpha = -1.65$

Step 4: Accept H_0 at $\alpha = 0.05$.

Step 5: From the sample, there is not enough evidence to conclude that the price of houses for the chosen state is less than the population mean price.

c. From the population and sample data, calculate the proportion of houses that have a price less than RM *X* (your own choice for eg: RM200000). State *p* and $\hat{p}$.

```
n = length(johorhouses$`Price (RM)`); k # valid responses count
k = sum(johorhouses$`Price (RM)`<600000); k
pbar = k/n; pbar
nforMalaysiahouseprice = length(Dataset_Q4b_Price_of_houses_in_Malaysia$`Price
(RM)`); nforMalaysiahouseprice
kforMalaysiahouseprice = sum(Dataset_Q4b_Price_of_houses_in_Malaysia$`Price
(RM)`<600000); kforMalaysiahouseprice
p = kforMalaysiahouseprice/nforMalaysiahouseprice; p
```

X = RM600000

$p = 0.59$

$\hat{p} = 0.3636364$

d. Compute the 95% confidence interval for the proportion of houses that have a price of less than RM *X* from sample data.

```
SE = sqrt(pbar*(1-pbar)/n); SE     # standard error
E = qnorm(.975)*SE; E              # margin of error
pbar + c(-E, E)
```

$\alpha = 0.05$

$\sigma p' = \sqrt{p'(1-p')/n}$

$p = p' \pm [Z(a/2)\ \sqrt{p'(1-p')/n}]$

Standard error = 0.1450407

margin of error = 0.2842746

p = (0.07936175, 0.64791098)

We are 95% confident that the interval from 0.07936175 to 0.64791098 for the proportion of houses that have a price of less than RM X from sample data.

e. Conduct a test of hypothesis to determine whether the proportion of houses that have a price of less than RM *X* is less than 40%. Report the hypothesis testing process step by step and conclude the findings.

```
p=0.4
n = length(johorhouses$`Price (RM)`); n
pbar = k/n; pbar
z= pbar-p/ sqrt((p*(1-p))/n);z
```

p= 0.4

$\hat{p}$ = 0.3636364

Step 1: $H_0 : p \geq 0.4$, $H_1 : p < 0.4$

Step 2: $Z=(\hat{p}-p)/\sqrt{(pq/n)}$

Z = -2.344376

Step 3: At $\alpha = 0.05$ -> $Z_\alpha = -1.65$

Step 4: Reject H_0 at $\alpha = 0.05$.

Step 5: From the sample, there is enough evidence to conclude that the proportion of houses that have a price of less than RM X is less than 40%.

Question 5 (20 marks)

Question 5a (10 marks) DAILY RAINFALL BY STATE FOR PENINSULAR MALAYSIA 2020

1. Download the Daily Rainfall by State for Peninsular Malaysia data for the year 2020. Understand the data.

2. Calculate the average and standard deviation for API of November and December for the FOUR (4) states of your choice. Visualize and explain the data:

i. Dataset 1: State 1

ii. Dataset 2: State 2

iii. Dataset 3: State 3

iv. Dataset 4: State 4

v. Dataset 5: State 5

Dataset1= NSembilan

```
m1 <- mean(Dataset1_Nov_Dec$`Rainfall (mm)`)

sd1 <- sd(Dataset1_Nov_Dec$`Rainfall (mm)`)

boxplot(Dataset1_Nov_Dec$`Rainfall (mm)`, main="NSembilan rainfall",xlab
="NSembilan",ylab="Rainfall (mm)")
```

Average dataset1 : 8.896359

Sd datasetset1 : 9.857991

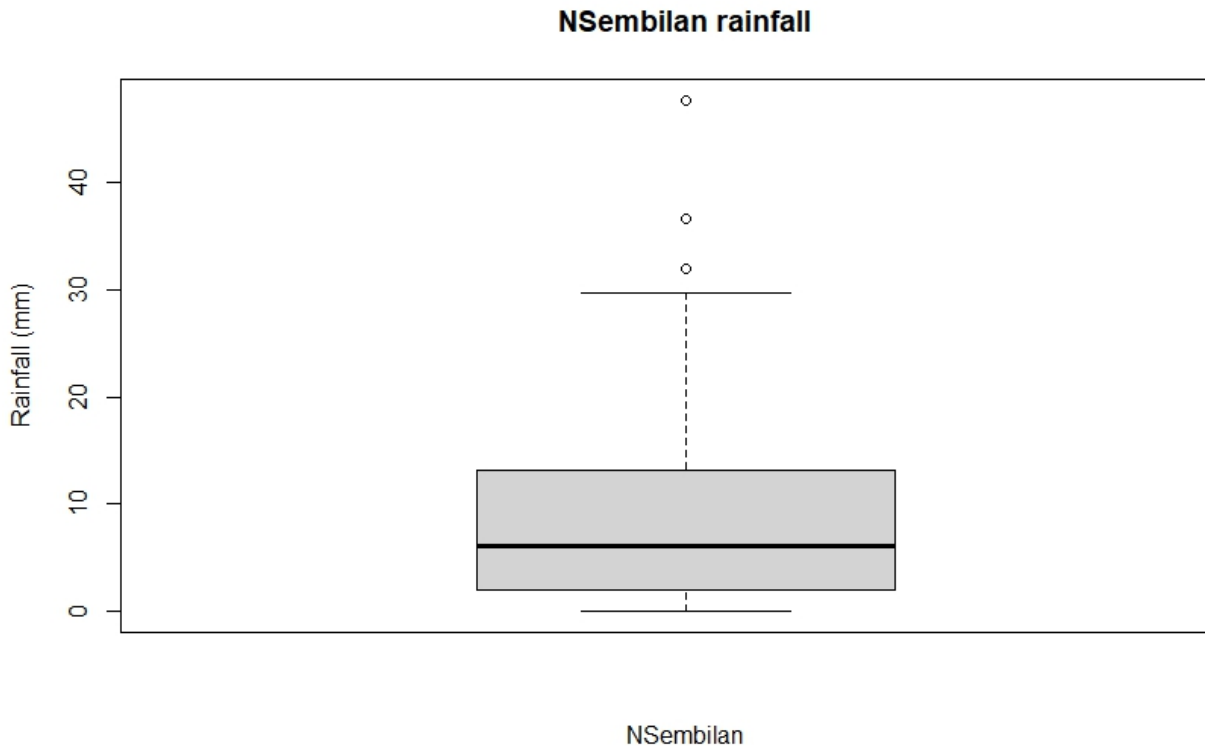

Figure 5.1 NSembilan rainfall boxplot

Lowest observation near to 0. Lower quartile around 2. Upper quartile is around 13. Median is around 6. Highest observation excluding outlier is near to 30. There are 3 outliers.

Dataset2 = Pahang

```
m2 <- mean(Dataset2_Nov_Dec $`Rainfall (mm)`)

sd2 <- sd(Dataset2_Nov_Dec $`Rainfall (mm)`)

boxplot(Dataset2_Nov_Dec $`Rainfall (mm)`,main="Pahang
rainfall",xlab="Pahang",ylab="Rainfall (mm)")
```

Average dataset2: 13.5811

Sd datasetset2: 11.52974

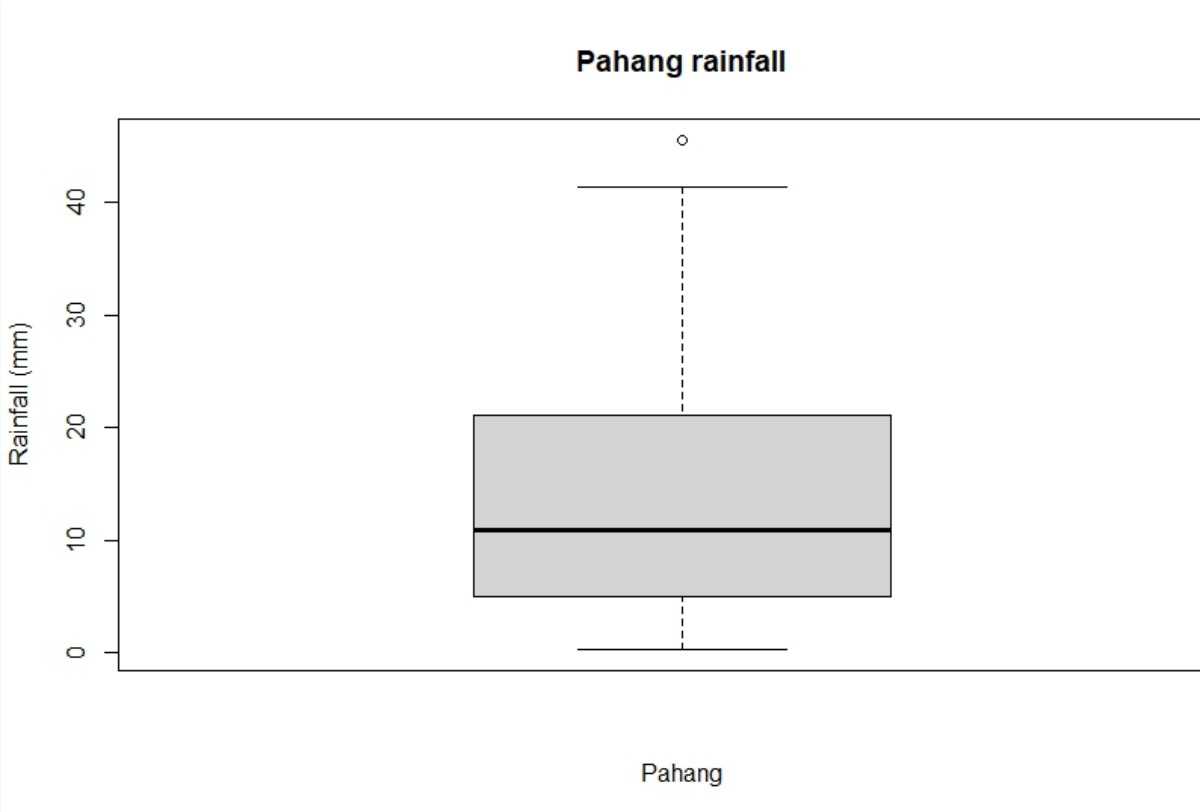

Figure 5.2 Pahang rainfall boxplot

Lowest observation is near to 0. Lower quartile is near to 5. Median is around 11. Upper quartile is around 21. Highest observation excluding outlier is inside 41 to 42. There are 3 outliers.

Dataset3 = Penang

```
m3 <- mean(Dataset3_Nov_Dec $`Rainfall (mm)`)

sd3 <- sd(Dataset3_Nov_Dec$`Rainfall (mm)`)

boxplot(Dataset3_Nov_Dec $`Rainfall (mm)`,main="Penang
rainfall",xlab="Penang",ylab="Rainfall (mm)")
```

Average dataset3: 10.87891

Sd dataset3: 14.92963

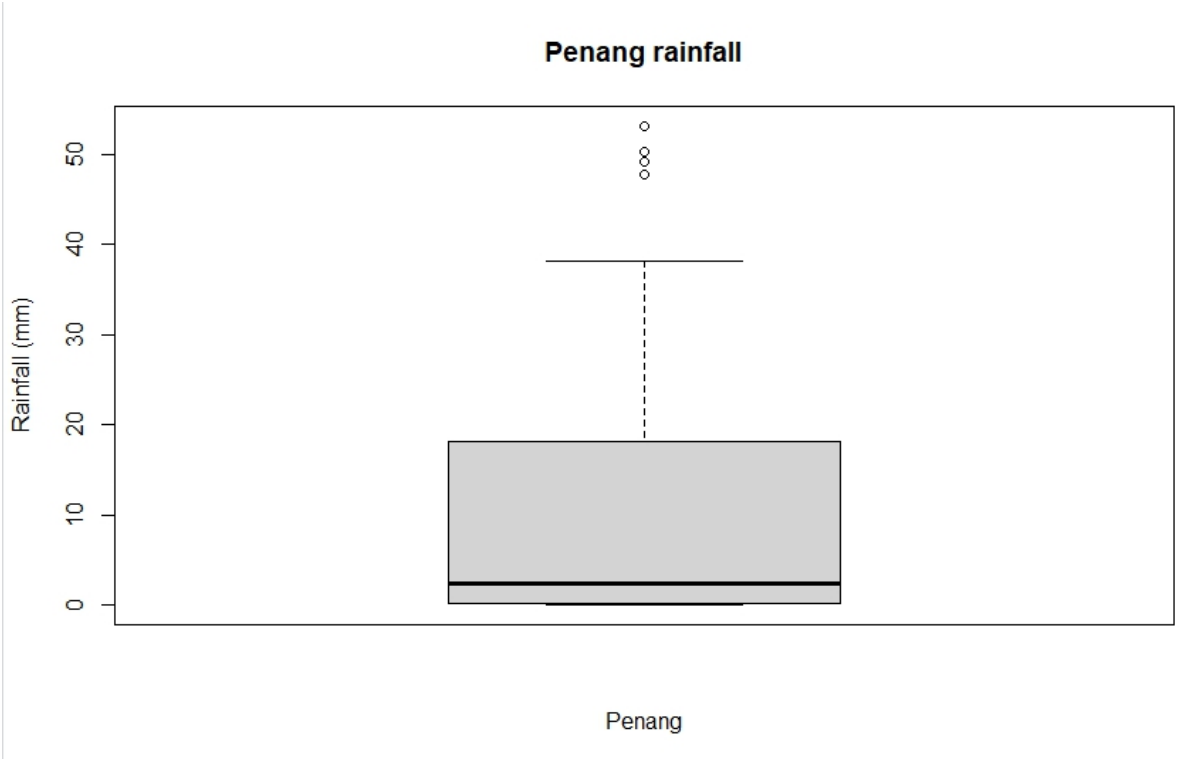

Figure 5.3: Penang rainfall boxplot

Lowest observation is 0. The lower quartile is near to 0.Median is around 2. The upper quartile is around 18. The highest observation excluding outlier is near to 39. There are 4 outliers.

Dataset4 =Perak

```
m4 <- mean(Dataset4_Nov_Dec$`Rainfall (mm)`)

sd4 <- sd(Dataset4_Nov_Dec$`Rainfall (mm)`)

boxplot(Dataset4_Nov_Dec$`Rainfall (mm)`,main="Perak
rainfall",xlab="Perak",ylab="Rainfall (mm)")
```

Average dataset4: 12.35873

Sd dataset4: 11.52715

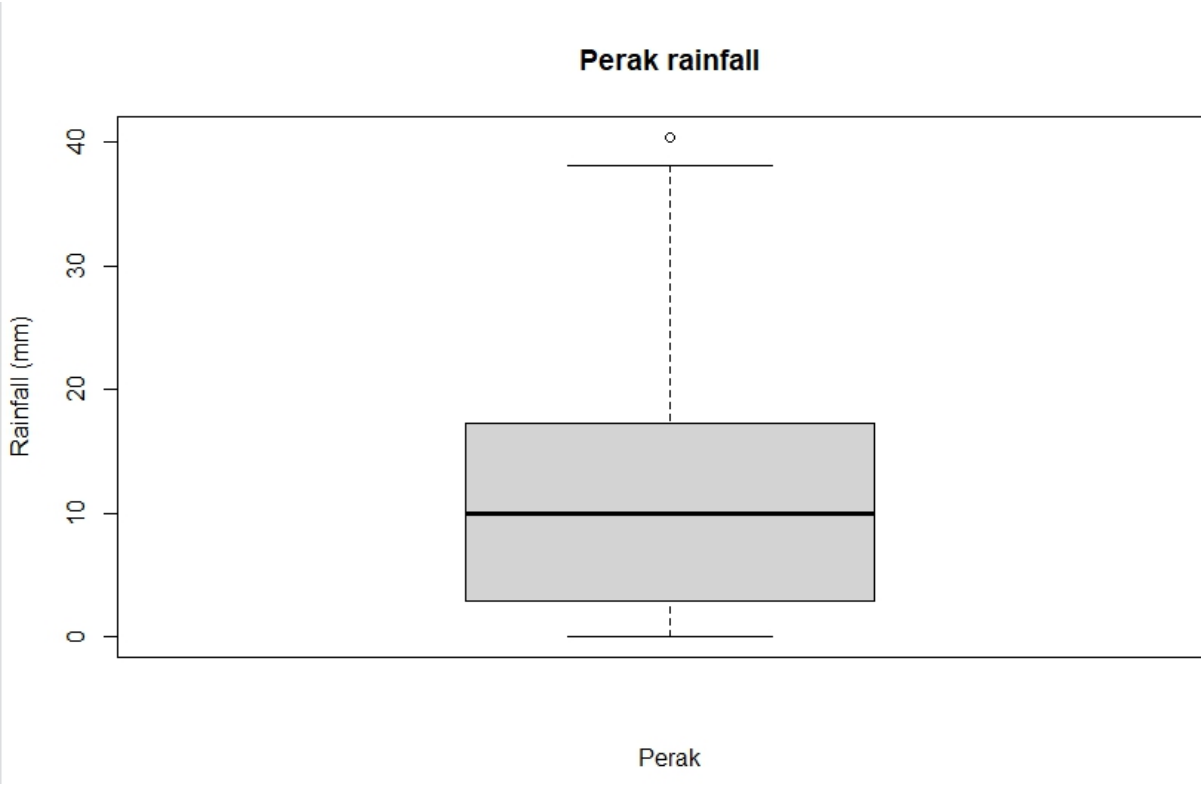

Figure 5.4: Perak rainfall boxplot

The lowest observation is near to 0. The lower quartile is near to 3. Median is near to 10.The upper quartile is around 17. The highest observation excluding outlier is around 38. There is 1 outlier.

Dataset5= Perlis

```
m5 <- mean(Dataset5_Nov_Dec$`Rainfall (mm)`)

sd5 <- sd(Dataset5_Nov_Dec$`Rainfall (mm)`)

boxplot(Dataset5_Nov_Dec$`Rainfall (mm)`,main="Perlis
rainfall",xlab="Perlis",ylab="Rainfall (mm)")
```

Average dataset5: 4.604918

Sd dataset5: 9.868862

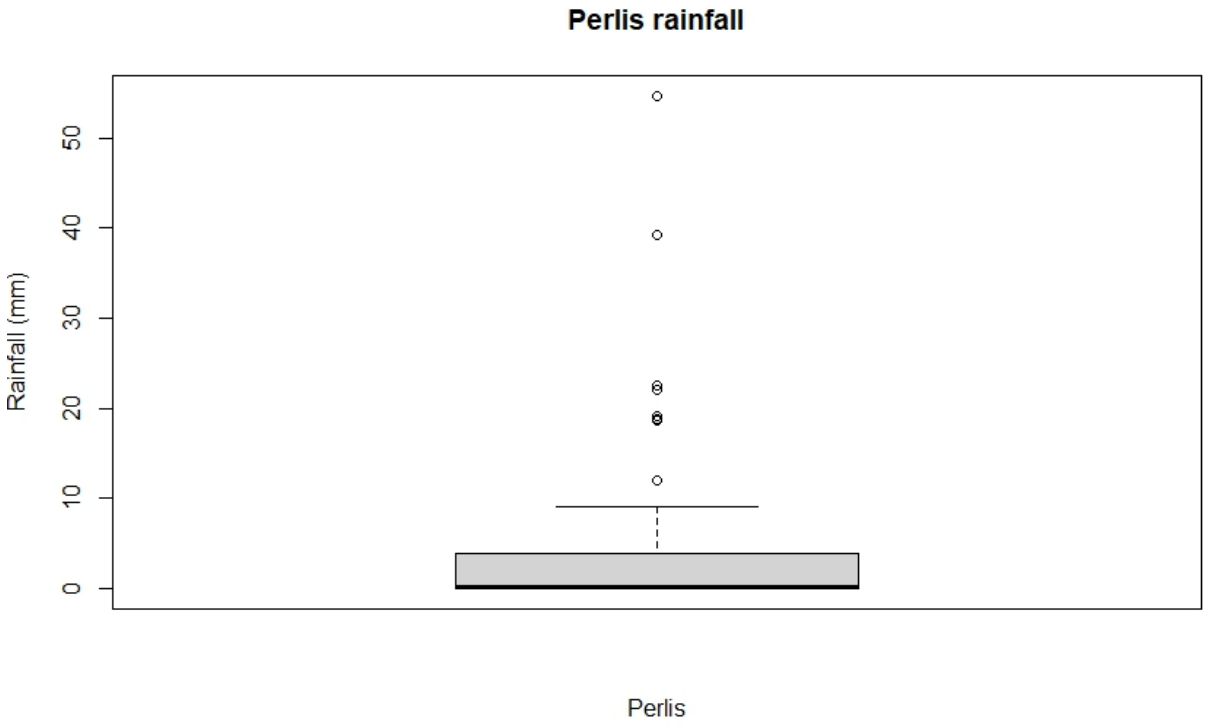

Figure 5.5 Perlis rainfall boxplot

The lowest observation is 0. The lower quartile is 0. Median is near to 0. The upper quartile is near to 4. The highest observation excluding outliers is around 9. There are around 8 outliers.

5i

Construct a 95% confidence interval for the average rainfalls of Dataset 1 for November and December 2020.

```
error <- qnorm(0.975)*sd1/sqrt(61)

m1-error

m1+error
```

95% confidence interval for the average rainfalls of Dataset 1 for November and December 2020 is (6.422519,11.3702)

5ii

Construct a 98% confidence interval for the average rainfalls of Dataset 1 for November and December 2020.

```
error <- qnorm(0.99)*sd1/sqrt(61)
```

```
m1-error
m1+error
```

98% confidence interval for the average rainfalls of Dataset 1 for November and December 2020 is (5.960074, 11.83264)

5iii

By using the p-value approach, test at the 5% significance level if the average rainfalls are greater for Dataset 2 than Dataset 3. Assume that degrees of the agreement are both normally distributed with equal standard deviations.

$H_0 = \mu_2 \leq \mu_3$

$H_1 = \mu_2 > \mu_3$

```
x = (Dataset2_Nov_Dec $`Rainfall (mm)`)
y = (Dataset3_Nov_Dec $`Rainfall (mm)`)
t.test(x,y,alternative="greater",conf.level = 0.95,var.equal = TRUE)
```

```
Two Sample t-test

data:  x and y
t = 1.1188, df = 120, p-value = 0.1327
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.301389       Inf
sample estimates:
mean of x mean of y
 13.58110  10.87891
```

At significance level of 0.05, p-value = 0.1327 > 0.05

Accept H_0 .Average of dataset2 is not larger than average of dataset3

5iv

By using the p-value approach, test at the 2.5% significance level if the average rainfalls are different for Dataset 4 and Dataset 5. Assume that degrees of the agreement are both normally distributed with unequal standard deviations.

$H_0 : \mu_4 = \mu_5$

$H_1 : \mu_4 \neq \mu_5$

```
x = (Dataset4_Nov_Dec$`Rainfall (mm)`)
y = (Dataset5_Nov_Dec$`Rainfall (mm)`)
t.test(x,y,alternative="two.sided",conf.level = 0.975)
```

```
Welch Two Sample t-test

data:  x and y
t = 3.9908, df = 117.22, p-value = 0.0001151
alternative hypothesis: true difference in means is not equal to 0
97.5 percent confidence interval:
  3.342314 12.165314
sample estimates:
mean of x mean of y
12.358732  4.604918
```

At significance level of 0.025, p-value = 0.0001151 < 0.025

Reject H_0. Average of dataset4 is different with average of dataset5.

5bi) Is the percentage of Plan 1 has been decreased from previous research at a 0.05 significance level?

$H_0 : p \geq 0.429$

$H_1 : p < 0.429$

```
prop.test(x= p_1$n ,n = 60,p = 0.429, correct = FALSE, alternative = "less",conf.level = 0.95)
```

```
1-sample proportions test without continuity correction


data:  p_1$n out of 60, null probability 0.429
X-squared = 54.338, df = 1, p-value = 1
alternative hypothesis: true p is less than 0.429
95 percent confidence interval:
 0.0000000 0.9474025
sample estimates:
  p
0.9
```

At significance level of 0.05, p-value=1 > 0.05.

Accept H_0 . The percentage of Plan 1 has not been decreased from previous research at a 0.05 significance level.

5bii) Is the percentage of Plan 2 has been increased from previous research at the 0.01 significance level?

$H_0 : p \leq 0.156$

$H_1 : p > 0.156$

```
prop.test(x=p_2$n, n=60, p=0.156,correct =  FALSE,alternative = "greater",conf.level = 0.99)
```

```
1-sample proportions test without continuity correction


data:  p_2$n out of 60, null probability 0.156
X-squared = 9.4495, df = 1, p-value = 0.001056
alternative hypothesis: true p is greater than 0.156
99 percent confidence interval:
 0.1837003 1.0000000
```

```
sample estimates:
  p
0.3
```

At significance level of 0.01, p-value = 0.001056 <0.01

Reject H_0 . The percentage of Plan 2 has been increased from previous research at the 0.01 significance level

5biii)Is the percentage of Plan 3 is as same as previous research at a 0.05 significance level?

$H_0 : p = 0.338$

$H_1 : p \neq 0.338$

```
prop.test(x=p_3$n, n=60, p=0.338,correct =  FALSE,alternative = "two.sided",conf.level = 0.95)
```

```
1-sample proportions test without continuity correction


data:  p_3$n out of 60, null probability 0.338
X-squared = 2.4371, df = 1, p-value = 0.1185
alternative hypothesis: true p is not equal to 0.338
95 percent confidence interval:
 0.3157240 0.5589656
sample estimates:
      p
0.4333333
```

At significance level of 0.05, p-value = 0.1185 >0.05

Accept H_0 .The percentage of Plan 3 is same as previous research at a 0.05 significance level

5biv). Is the percentage of Plan 4 has been decreased from previous research at the 0.1 significance level?

$H_0 : p \geq 0.044$

H_1 : p < 0.044

```
prop.test(x=p_4$n, n=60, p=0.044,correct =  FALSE,alternative = "less",conf.level = 0.9)
```

```
1-sample proportions test without continuity correction


data:  p_4$n out of 60, null probability 0.044
X-squared = 60.531, df = 1, p-value = 1
alternative hypothesis: true p is less than 0.044
90 percent confidence interval:
 0.0000000 0.3276541
sample estimates:
   p
0.25
```

At significance level of 0.1, p-value = 1 >0.1

Accept H_0 .The percentage of Plan 4 has not been decreased from previous research at the 0.1 significance level