**STAT 40001/50001 Statistical Computing Fall 2024**

**Test 1- Solution**

**Name: Answer key**

**PUID:012345678**

*This exam consists of 4 questions. Please provide the R codes that have been used to perform all the calculations and graphics along with the interpretation of the output. It is important that you distinguish between the R code and the result description. You may use different fonts or different colors.*

**Q.N. 1)** For a given number *x*, the R code *as.numeric(strsplit(as.character(x),"")[[1]])* splits the data into its digits. Use x=[Your PUID, no leading zero ]

**a)** Find the sum of the digits in your PUID

> x=012345678

> ID=as.numeric(strsplit(as.character(x),"")[[1]])

> ID

[1] 1 2 3 4 5 6 7 8

> sum(ID)

[1] 36

**b)** Find the mean and the variance of the digits of your PUID.

> ID=c(1,2,3,4,5,6,7,8)

> mean(ID)

[1] 4.5

> var(ID)

[1] 6

**c)** Generate 500 random numbers from a normal distribution whose mean and standard deviation are the mean and the standard deviations of the digits of your PUID. Make sure that the random number you generated are reproducible. Please print first 5 numbers.

> set.seed(1)

> R=rnorm(500, mean=mean(ID), sd=sd(ID))

> head(R,5)

[1] 2.965508 4.949832 2.453136 8.407624 5.307126

>

**d)** Calculate the five-number summary of the data that you generated in part (c).

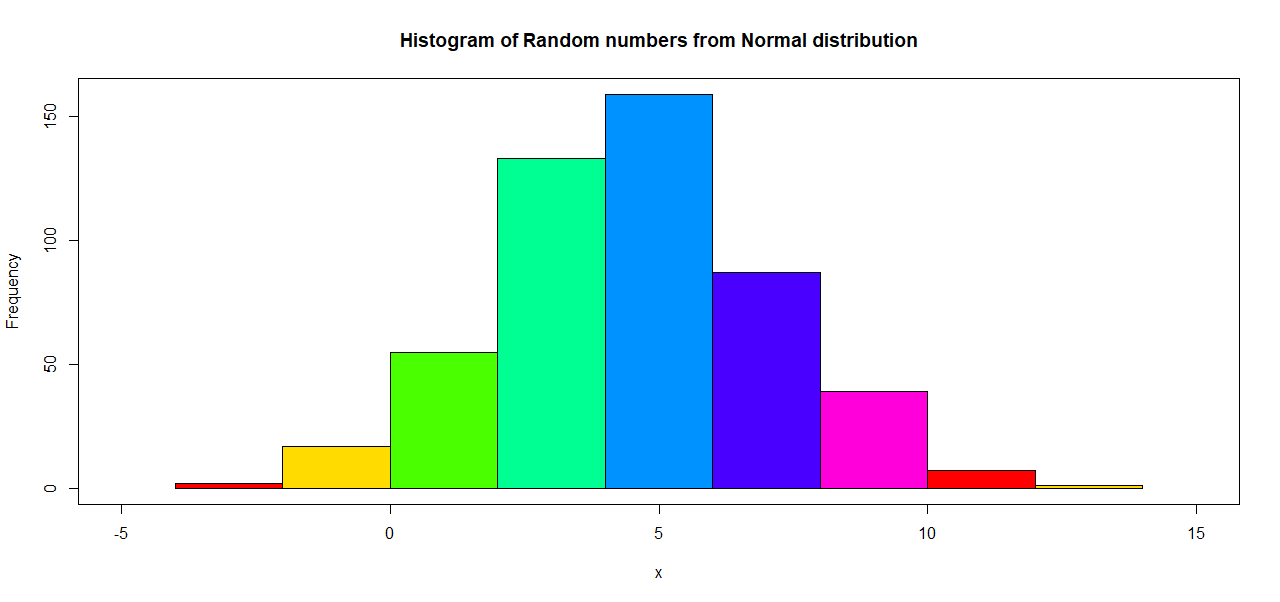
> fivenum(R)

[1] -2.868184 2.978246 4.409912 6.162678 13.833234

**e)** Draw a histogram of the data you generated in part(c). Please update the graph using different colors and title, labels etc.

> hist(R, main="Histogram of Random numbers from Normal distribution", col=rainbow(7), xlim=c(-5,15), xlab="x")

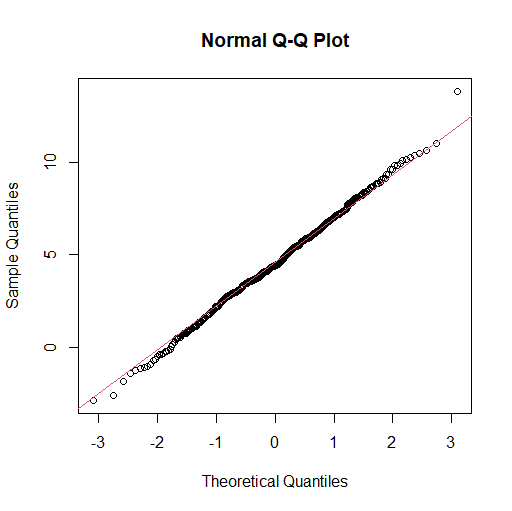
> box()



**f)** Draw a normal QQ-plot of the data you generated in part (c).

> qqnorm(R)

> qqline(R, col=2)



**2)** Use the sample function to choose three letters from a, b, c, d and e and solve those three questions

> sample(letters[1:5],3)

Answer:

> sample(letters[1:5],3)

[1] "d" "e" "c"

You were requested to answer above three questions. However, I will solve all five questions

1. The GPA of students is provided in the link below: <https://advstats.psychstat.org/data/gpa.txt> . In the dataset the missing values of GPA are given as 999 or 99. Please clean the data by removing the missing values.

*Solution: R code below can be used to import the data and create a Clean dataset.*

>a2=read.table("https://advstats.psychstat.org/data/gpa.txt", skip=2, header=T, na.string=c(99, 999))

> head(a2,3)

id gender college gpa weight

1 1 f yes 3.6 110

2 2 m yes 3.5 170

3 3 m no NA 165

> Clean=na.omit(a2)

> head(Clean,3)

id gender college gpa weight

1 1 f yes 3.6 110

2 2 m yes 3.5 170

6 6 m yes 3.7 200

1. Daily air quality measurements in New York, May to September 1973 are provided in the base package under the dataset “airquality”. How many observations have missing values?

*Answer: Base on the R output below 153-111=42 observations have at least one missing value*

> b2=airquality

> dim(airquality)

[1] 153 6

> newb2=na.omit(b2)

> dim(newb2)

[1] 111 6

1. Use functions rep and seq to create the sequence 5, 10, 10, 15, 15, 15, 20, 20, 20, 20.

*Answer: R code below produces the desired sequence of numbers*

> rep(seq(5,20,5), seq(1,4,1))

[1] 5 10 10 15 15 15 20 20 20 20

**d)** Solve the system of equations:

*Answer: Based on the R out put below, we have x=1, y=-1 and z=0*

> data=c(3,-1,2,-4,2,-3,7,-5,7)

> M=matrix(data, ncol=3,byrow=T)

> M

[,1] [,2] [,3]

[1,] 3 -1 2

[2,] -4 2 -3

[3,] 7 -5 7

> Y=c(4,-6,12)

> Y=matrix(Y, ncol=1)

> solve(M)%\*%Y

[,1]

[1,] 1

[2,] -1

[3,] 0

1. Girth, Height and Volume for Black Cherry trees are provided in the base package trees dataset. Import the dataset and determine its dimension. Please display the height of the trees using a histogram.

*Solution: R code below can be used to access the data and display the height of the trees using a histogram.*

> data(trees)

> head(trees)

Girth Height Volume

1 8.3 70 10.3

2 8.6 65 10.3

3 8.8 63 10.2

4 10.5 72 16.4

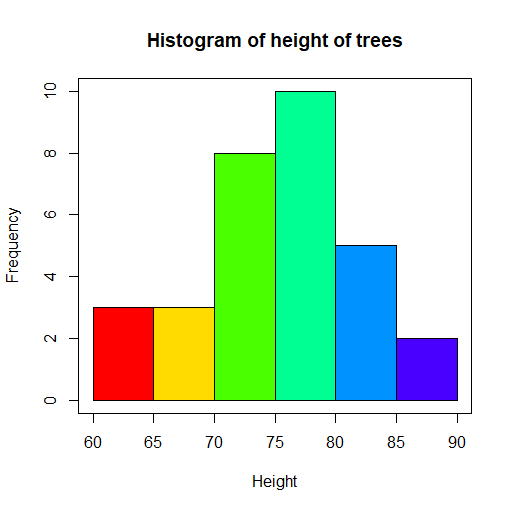
5 10.7 81 18.8

6 10.8 83 19.7

> attach(trees)

> hist(Height, col=rainbow(7), main="Histogram of height of trees")

> box()



**Q.N. 3)** Chicago’s Department of Business Affairs and Consumer Protection provides monthly reports of all taxi trips in Chicago, tagged with trip distances, trip durations, fare amounts, and tip amounts. The dataset Chicago\_taxi holds data on taxi trips for which the duration exceeded 1 minute and for which payment was made either in cash or with a credit card. Speed is given in miles per hour averaged over the trip. The data are provided in the link

<https://dasl.datadescription.com/datafile/chicago_taxi/?_sfm_cases=4+59943&sf_paged=8>

Import the data in R (You may need to download and import)

*Solution: We saved the data in the local drive and imported in R using the R code below*

Answer (a)-(c) if your PUID is an even number otherwise answer (d)-(f)

1. Draw a pie chart to display the information regarding the method of payments.

> Q3=read.table("C:\\Users\\aryalgTaxi.txt", header=T)

> names(Q3)

[1] "date" "daytype" "seconds" "miles" "tips"

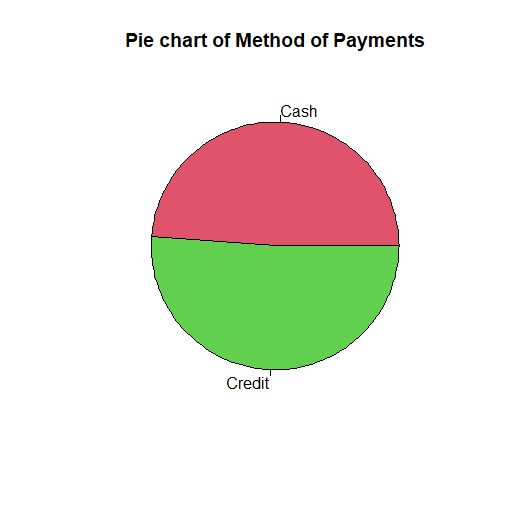
[6] "total" "payment\_type" "speed"

> dim(Q3)

[1] 13082 8

> attach(Q3)

> pie(table(payment\_type), col=c(2,3), main="Pie chart of Method of Payments")



1. Construct a 95% confidence interval for the tips amount based on the method of payments.

*Answer: Based on the R output below 95% confidence interval for the tips amount paid by cash is (0,0.006) and by credit card is (3.78, 3.95). You can also calculate the 95% CI for the difference in tips amount.*

> t.test(Cash$tips)$conf.int

[1] -0.0006482289 0.0057148956

attr(,"conf.level")

[1] 0.95

> Credit=subset(Q3, payment\_type=="Credit")

> t.test(Credit$tips)$conf.int

[1] 3.786408 3.951591

attr(,"conf.level")

[1] 0.95

1. Is there a significant difference in the tips amount by daytype (weekday vs, weekend)?

*Answer: Let and denote the average tips earnings during the weekday and weekend respectively. We are testing the following hypothesis*

*The R code below can be used to perfom the analysis*

> t.test(tips~daytype, data=Q3)

Welch Two Sample t-test

data: tips by daytype

t = 5.5765, df = 7642.2, p-value = 2.537e-08

alternative hypothesis: true difference in means between group weekday and group weekend is not equal to 0

95 percent confidence interval:

0.2084284 0.4343954

sample estimates:

mean in group weekday mean in group weekend

2.075906 1.754494

*Decision: Since p-value (0.0000000253) is smaller than 0.05 we reject the null hypothesis and conclude that the average tips amount during weekdays is different from the tips amount during weekends.*

*Remark: We may test for equality of variance and add that too. However, we will still get the same result.*

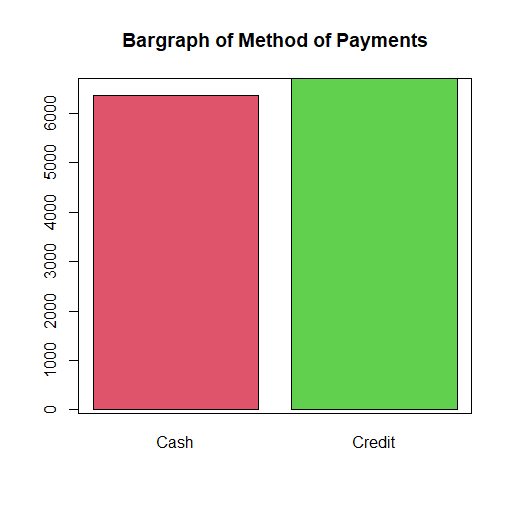
-----------------------------------------------------------------------------------------------------------

1. Draw a bar graph to display the payment information regarding the method of payments.

Answer: R code below produces a boxplot

> barplot(table(payment\_type), col=c(2,3), main="Bargraph of Method of Payments")

> box()



1. Construct a 90% confidence interval for the tips amount based on the method of payments.

*Answer: Based on the R output below 90% confidence interval for the tips amount paid by cash is (0,0.0052) and by credit card is (3.799, 3.938). You can also calculate the 90% CI for the difference in tips amount.*

> t.test(Cash$tips, conf.level=0.9)$conf.int

[1] -0.0001365989 0.0052032656

attr(,"conf.level")

[1] 0.9

> t.test(Credit$tips, conf.level=0.9)$conf.int

[1] 3.799690 3.938309

attr(,"conf.level")

[1] 0.9

1. Is there a significant difference in the tips amount by method of payments?

*Answer: Let and denote the average tips earnings by cash and credit payment respectively. We are testing the following hypothesis*

*The R code below can be used to perform the analysis*

*> t.test(tips~payment\_type, data=Q3)*

*Welch Two Sample t-test*

*data: tips by payment\_type*

*t = -91.703, df = 6725.9, p-value < 2.2e-16*

*alternative hypothesis: true difference in means between group Cash and group Credit is not equal to 0*

*mean in group Cash mean in group Credit*

*0.002533333 3.868999553*

*Decision: Since p-value (<0.00000001) is smaller than 0.05 we reject the null hypothesis and conclude that the average tips amount is different by payment type.*

*Remark: We may test for equality of variance and add that too. However, we will still get the same result.*

**4)** A small-scale clinical trial is conducted to study the effect of drug in reduction of excess body weight. Age and gender (1 -male, 0-female) were recorded at the baseline. The percent excess body weight loss (EWL) was recorded after 3 months into the study. The dataset is provided below:

|  |  |  |
| --- | --- | --- |
| age | gender | EWL |
| 49 | 0 | 14.2 |
| 54 | 1 | 25.4 |
| 37 | 0 | 14.1 |
| 43 | 0 | 20 |
| 57 | 1 | 11.7 |
| 48 | 1 | 16.6 |
| 34 | 0 | 15.9 |
| 51 | 0 | 17.4 |
| 54 | 0 | 22.8 |
| 45 | 0 | 16.7 |
| 36 | 1 | 12.7 |
| 57 | 1 | 15 |
| 44 | 1 | 8.4 |
| 56 | 1 | 11.2 |
| 44 | 1 | 17.3 |
| 47 | 1 | 20.5 |
| 44 | 0 | 6.7 |
| 52 | 0 | 29.4 |
| 51 | 1 | 21.9 |
| 44 | 0 | 23.6 |
| 53 | 0 | 23.8 |
| 55 | 1 | 7.4 |
| 30 | 0 | 23.1 |
| 47 | 1 | 16.8 |
| 26 | 1 | 14.1 |
| 56 | 0 | 24.6 |
| 28 | 0 | 17.8 |
| 34 | 1 | 27.8 |
| 43 | 1 | 10.6 |
| 55 | 1 | 26.8 |
| 52 | 0 | 15.7 |
| 54 | 0 | 23.7 |

1. Write R code to draw a random number between 1 and 10.

> sample(1:10,1) # This is how you draw a random number

[1] 3

If it is an odd number answer (b) and (c) otherwise answer (c) and (d).

*We used R code below to import the data from C: drive*

> read.csv( "C:\\Users\\aryalg \\weight.csv")

> head(Q4)

age gender EWL

1 49 0 14.2

2 54 1 25.4

3 37 0 14.1

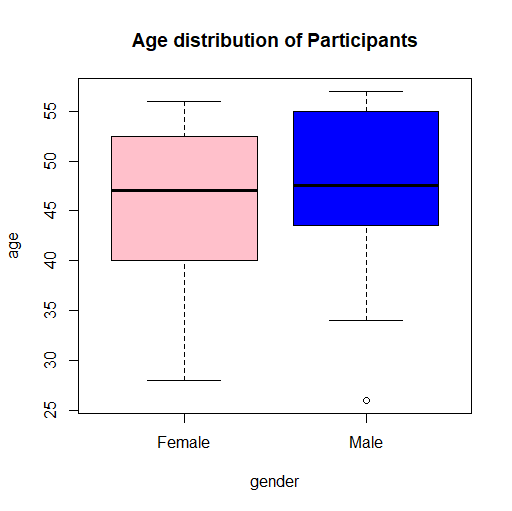
4 43 0 20.0

5 57 1 11.7

6 48 1 16.6

1. Display the age distribution based on gender.

> boxplot(age~gender, col=c("pink","blue"), main="Age distribution of Participants", names=c("Female", "Male"))



1. Is there is a significant difference in age between male and female?

*Let and denote the average age of male and female participants We are testing the following hypothesis*

*The R code below can be used to perform the analysis*

> t.test(age~gender)

Welch Two Sample t-test

data: age by gender

t = -0.5489, df = 29.992, p-value = 0.5871

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

-8.261235 4.761235

sample estimates:

mean in group 0 mean in group 1

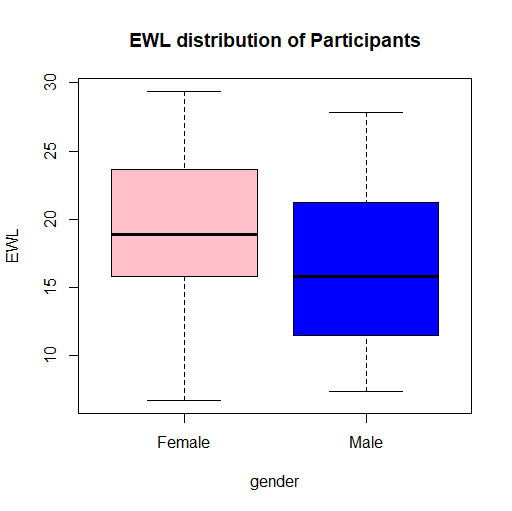
45.375 47.125

*Conclusion: Since p-value(0.5871) is higher than 0.05 we fail to reject the null hypothesis and conclude that we don’t have enough evidence to prove that there is a significant difference on age based on gender.*

*Remark: We may test for equality of variance and add that too. However, we will still get the same result.*

1. Display the EWL distribution by gender.

> boxplot(EWL~gender, col=c("pink","blue"), main="EWL distribution of Participants", names=c("Female", "Male"))



1. Is there a significant difference in EWL based on gender?

*Let and denote the average EWL of male and female participants We are testing the following hypothesis*

*The R code below can be used to perform the analysis*

> t.test(EWL~gender)

Welch Two Sample t-test

data: EWL by gender

t = 1.3347, df = 29.422, p-value = 0.1922

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

-1.504631 7.167131

sample estimates:

mean in group 0 mean in group 1

19.34375 16.51250

*Conclusion: Since p-value(0.1922) is higher than 0.05 we fail to reject the null hypothesis and conclude that we don’t have enough evidence to prove that there is a significant difference on EWL based on gender.*

*Remark: We may test for equality of variance and add that too. However, we will still get the same result.*