MATHEMATICAL STATISTICS
LAB FILE

MSMA 109

2022-2023

Submitted To:

Dr. Satyabrata Adhikari
Dr. Goonjan Jain
Ms. Shruti
Ms. Neha
Department of Applied Mathematics

Submitted By:

RITIKA GUPTA
2K22/MSCMAT/54

INDEX

S. No.	TITLE	DATE	REMARKS
1	Descriptive Statistics	23/08/2022	
2	Plotting Ogive curve	30/08/2022	
3	Bayes Theorem	06/09/2022	
4	Joint, Marginal, Conditional Probability and CDF	06/09/2022	
5	Binomial Distribution	13/09/2022	
6	Poisson Distribution	20/09/2022	
7	Normal Distribution	11/10/2022	
8	Correlation Analysis	18/10/2022	
9	Regression Analysis	18/10/2022	
10	Hypothesis Testing: t-test	01/11/2022	

Practical-1–Descriptive-Statistics.R

2022-09-13

```r
#NAME: RITIKA GUPTA
#ROLL NO.:
#COURSE: MSC MATHEMATICS




#Question 1. Create a set of first 5 prime numbers
P=c(2,3,5,7,11)
P
```

```
## [1]  2  3  5  7 11
```

```r
#Question 2. create a set of 10 random numbers from 50 to 100
R=sample(50:100,10)
R
```

```
##  [1] 58 68 60 55 56 86 78 69 77 64
```

```r
#Question 3. calculate mean, median, mode of data
V=sample(1:100,15)
V
```

```
##  [1] 33 90 45 59 39 18 51 56 73 92 71 41 32  3 17
```

```r
mean(V)
```

```
## [1] 48
```

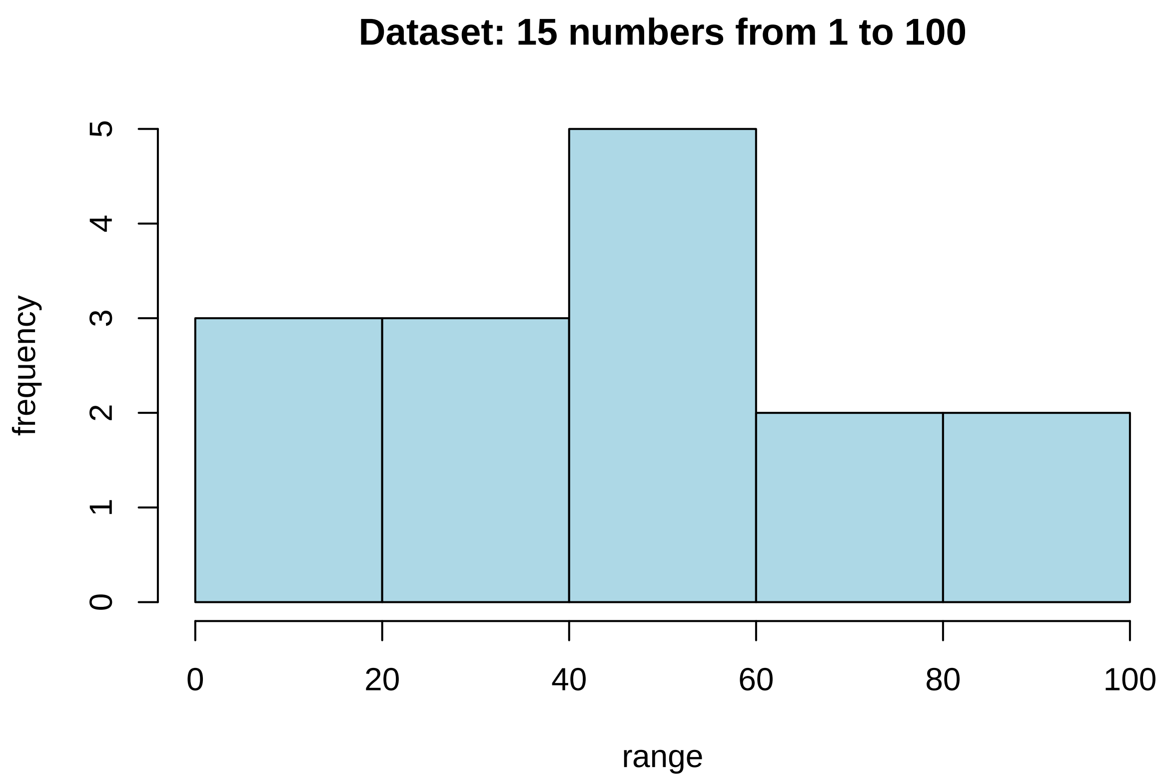
```r
V1=sort(V) #sorting data in ascending order
V1
```

```
##  [1]  3 17 18 32 33 39 41 45 51 56 59 71 73 90 92
```

```r
median(V1)
```

```
## [1] 45
```

```r
hist(V1,col = "lightblue",xlab = "range",ylab = "frequency",
     main = "Dataset: 15 numbers from 1 to 100")
```




```r
#Question 4. Statistical inferences of data in excel
library(readxl)
data= read_excel("marks_data.xlsx")
View(data)

#mean, median of data in excel
mean1=mean(data$'Marks 1')
median1=median(sort(data$'Marks 1'))
mean2=mean(data$'Marks 2')
median2=median(sort(data$'Marks 2'))

print(paste("Mean of first row = ",mean1))
```

```
## [1] "Mean of first row =  181.645161290323"
```

```r
print(paste("Median of first row = ",median1))
```

```
## [1] "Median of first row =  125"
```

```r
print(paste("Mean of second row = ",mean2))
```

```
## [1] "Mean of second row =  78.8064516129032"
```

```r
print(paste("Median of second row = ",median2))
```

```
## [1] "Median of second row =  77"
```

```r
#standard deviation and variation
sd1=sd(data$`Marks 1`)
```

```
sd2=sd(data$`Marks 2`)
var1=var(data$`Marks 1`)
var2=var(data$`Marks 2`)

print(paste("Standard deviation of 1st row = ",sd1))
```

```
## [1] "Standard deviation of 1st row =  159.93155794216"
```

```
print(paste("Variance of 1st row = ",var1))
```

```
## [1] "Variance of 1st row =  25578.1032258065"
```

```
print(paste("Standard deviation of 2nd row = ",sd2))
```

```
## [1] "Standard deviation of 2nd row =  13.2600134611262"
```

```
print(paste("Variance of 2nd row = ",var2))
```

```
## [1] "Variance of 2nd row =  175.827956989247"
```

```
#kurtosis and skewness
library(moments)
k1=kurtosis(data$`Marks 1`)
k2=kurtosis(data$`Marks 2`)
s1=skewness(data$`Marks 1`)
s2=skewness(data$`Marks 2`)

print(paste("Kurtosis of 1st row = ",k1))
```

```
## [1] "Kurtosis of 1st row =  2.77608997027731"
```

```
print(paste("Skewness of 1st row = ",s1))
```

```
## [1] "Skewness of 1st row =  0.958734494896294"
```

```
print(paste("Kurtosis of 2nd row = ",k2))
```

```
## [1] "Kurtosis of 2nd row =  1.84218052212265"
```

```
print(paste("Skewness of 2nd row = ",s2))
```

```
## [1] "Skewness of 2nd row =  0.274812460487586"
```

```
#sample() takes a sample of the specified size from the elements of x using either
#with or without replacement.
#Argument replace=TRUE when size is greater than length of dataset.

#c() command is a function which combines its arguments to form a vector.

#sort() arranges data given in the argument in an ascending order

#mean() function calculates the arithmetic mean of the data given in its argument.

#median() function calculates the sample median of the data given in its argument.

#hist() function computes a histogram of the given data values.
#xlab, ylab arguments produce axes labels, 'main' argument produces title, and
#col argument is used to give colour to fill the bars of the histogram.
```

```r
#read_excel() is used to import excel data in R progam

#sd() function calculates standard deviation of data in its argument

#var() function calculates variance of data in its argument

#kurtosis() function calculates kurtosis of data in its argument

#skewness() function calculates skewness of data in its argument




#**********practice codes**************

S=sample(1:10)
S
```

```
##  [1]  8 10  5  1  9  3  2  4  6  7
```

```r
x=1:100
x1=sample(x,10)
x2=sample(50:100,10)
x1
```

```
##  [1] 65 54 13  5 77 61 22 46 52 67
```

```r
x2
```

```
##  [1] 64 68 52 93 71 65 90 99 88 73
```

```r
length(x1)
```

```
## [1] 10
```

```r
length(x2)
```

```
## [1] 10
```

```r
y=c(2,4,6,8,10)
y
```

```
## [1]  2  4  6  8 10
```

```r
y1=c("H","T")
sample(1:2,10,replace = TRUE)
```

```
##  [1] 2 1 2 2 2 2 1 1 1 1
```

```r
#**********practice codes**************
```

Practical 2: Ogive Curve

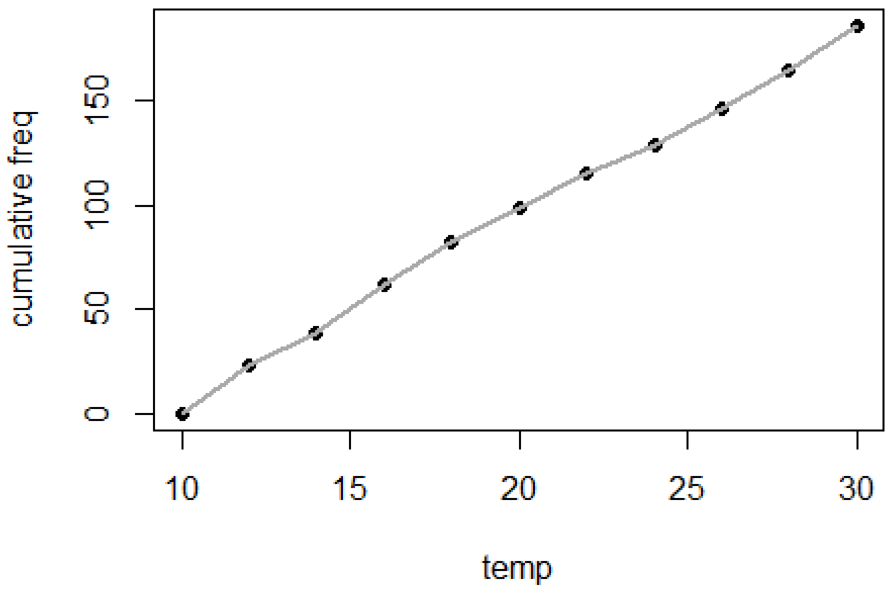
RITIKA GUPTA

2K22/MSCMAT/54

23-08-2022

```
#sample dataset
y=sample(10:30,200,replace = T)

#class interval

#seq(a,b,c) generates a sequence from a to b by common difference c.
a=seq(10,30,2)
#cut(x,y) divides the range of x into intervals corresponding to y.
y1=cut(y,a)


#frequency table
f=table(y1)
f

## y1
## (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28]
(28,30]
##      23      16      23      20      17      16      14      17      19
21

#cumulative frequency
cf=cumsum(f)
cf

## (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28]
(28,30]
##      23      39      62      82      99     115     129     146     165
186

s=c(0,cf)
s

##         (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26]
(26,28]
##       0      23      39      62      82      99     115     129     146
165
## (28,30]
##     186
```

```
plot(a,s,xlab="temp",ylab="cumulative freq",col="black",pch=19)
lines(a,s,col="darkgray",lwd=2)
```



Practical 3: Bayes Theorem

RITIKA GUPTA

2K22/MSCMAT/54

30-08-2022

```
#Q1
#There are two identical urns containing 4 white, 3 red balls and, 3 white, 7
red balls respectively.
#An un is choosen at random and a ball is drawn from it. Find the probability
that the ball is white.
#If the ball drawn is white, what is the probability that it is drawn from
the first urn?

library(LaplacesDemon)

## Warning: package 'LaplacesDemon' was built under R version 4.2.2

PrA=c(1/2,1/2)
PrBA=c(4/7,3/10)
R1=BayesTheorem(PrA,PrBA)

prob_white=(PrA[1]*PrBA[1])+(PrA[2]*PrBA[2])
prob_first_white=R1[1]
prob_second_white=R1[2]

print(paste("P(ball is white) = ",prob_white))

## [1] "P(ball is white) =  0.435714285714286"

print(paste("P(I urn | white ball) = ",prob_first_white))

## [1] "P(I urn | white ball) =  0.655737704918033"

print(paste("P(II urn | white ball) = ",prob_second_white))

## [1] "P(II urn | white ball) =  0.344262295081967"

#Q2
#What's the probability of going to hell conditional on consorting given that
a person consort.
#6 people consorted out of 9 who went to hell. 5 people consorted out of 7
who went to heaven.
#75% of people go to hell. 25% of people go to heaven.
#A is the event when person goes to heaven #B is the event when person goes
to hell
```

```r
PrX=c(3/4,1/4)
PrYX=c(2/3,5/7)
R2=BayesTheorem(PrX,PrYX)
prob_hell_consort=R2[1]
prob_heaven_consort=R2[2]

print(paste("P(hell | person consorts) = ",prob_hell_consort))

## [1] "P(hell | person consorts) =  0.736842105263158"

print(paste("P(heaven | person consorts) = ",prob_heaven_consort))

## [1] "P(heaven | person consorts) =  0.263157894736842"

#Q3
#Suppose doctors are asked to report the no. of cases of small pox & chicken pox and
#the symptoms are observed using the result of the survey.
#We find the probability that a patient has spots given that they have
#small pox or chicken pox is 20% and 80% respectively.
#Public health statistic inform us that the small pox in general population is 0.001 and chicken pox is 0.01.
#What's the probability of the patient having the small pox given that they have a spot on their face.
#What's the probability of the patient having chicken pox & a spot on their face.

PrS=c(0.001,0.01)
PrTS=c(0.2,0.8)
R3=BayesTheorem(PrS,PrTS)
Prob_small_spot=R3[1]
Prob_chicken_spot=R3[2]

print(paste("P(small pox | spots) = ",Prob_small_spot))

## [1] "P(small pox | spots) =  0.024390243902439"

print(paste("P(chicken pox | spots) = ",Prob_chicken_spot))

## [1] "P(chicken pox | spots) =  0.975609756097561"
```

Practical 4:

Joint, Marginal, Conditional Probabilities and Cumulative Distribution Function

RITIKA GUPTA

2K22/MSCMAT/54

06-09-2022

```r
#Joint, Marginal, Conditional Probability and CDF

#Create a joint probability table and use it to compute marginal and
#conditional probabilities

#P(X=0,Y=1) = 1/3
#P(X=1,Y=1) = 1/3
#P(X=-1,Y=1) = 1/3

#we have two random variables X and Y with respective sample spaces
#X={-1,0,1} and Y={-1,0,1}


#JOINT PROBABILITY

#The joint probability table for these random variables is given by,

P=matrix(c(0,0,0,0,0,1/3,1/3,0,1/3),ncol = 3)
rownames(P)=c("X=-1","X=0","X=1")
colnames(P)=c("Y=-1","Y=0","Y=1")
P

##      Y=-1       Y=0       Y=1
## X=-1    0 0.0000000 0.3333333
## X=0     0 0.0000000 0.0000000
## X=1     0 0.3333333 0.3333333

#To display P(X=-1,Y=1), we will display element stored in the
#1st row and 3rd column of p, i.e.,

print(paste("P(-1,1)= ",P[1,3]))

## [1] "P(-1,1)=  0.333333333333333"

#MARGINAL PROBABILITY
```

```r
#Marginal probability of X at x is given by
#P(X=x) = P(X=x,Y=-1) + P(X=x,Y=0) + P(X=x,Y=1)
Px=apply(P,1,sum)
Px

##      X=-1       X=0       X=1
## 0.3333333 0.0000000 0.6666667

#Marginal probability of Y at y is given by
#P(Y=y) = P(X=-1,Y=y) + P(X=0,Y=y) + P(X=1,Y=y)
Py=apply(P,2,sum)
Py

##      Y=-1       Y=0       Y=1
## 0.0000000 0.3333333 0.6666667

#CONDITIONAL PROBABILITY

#Compute the conditional probability P(X=-1|Y=1) and display it.
P_xy=P[1,3]/Py[3]
print(paste("P(X=-1|Y=1)= ",P_xy))

## [1] "P(X=-1|Y=1)=  0.5"

#CUMMULATIVE DISTRIBUTION FUNCTION

#Find the CDF of X
CDF_x=cumsum(Px)
CDF_x

##      X=-1       X=0       X=1
## 0.3333333 0.3333333 1.0000000

#Find the CDF of Y
CDF_y=cumsum(Py)
CDF_y

##      Y=-1       Y=0       Y=1
## 0.0000000 0.3333333 1.0000000
```

Practical 5: Binomial Distribution

RITIKA GUPTA

2K22/MSCMAT/54

13-09-2022

```r
#Binomial distribution

#Q1. The probability that a person can achieve a target is 3/4. The count
#of tries is 5, what is the probability that he will attain the target at
#least thrice?

#X is the discrete random Variable denoting successful tries
#X can have values: 0,1,2,3,4,5
#To find P(X>=3),

x=3:5
px=sum(dbinom(x,5,3/4))
print(paste("P(X>=3) = ",px))

## [1] "P(X>=3) =  0.896484375"

#Q2. Find the probability of getting at least 5 heads on tossing an
#unbiased coin 6 times.

#Y is a discrete random variable denoting number of successes, i.e. heads
#Y can have values: 0,1,2,3,4,5,6
#To find P(Y>=5),

y=5:6
py=sum(dbinom(y,6,1/2))
print(paste("P(Y>=5) = ",py))

## [1] "P(Y>=5) =  0.109375"
```

Practical 6: Poisson Distribution

RITIKA GUPTA

2K22/MSCMAT/54

20-09-2022

```
#Poisson Distribution

#Q1. A car hire firm has two cars which it hires out day by day. The number
#of demands of a car on each day is distributed as a poission distribution
#with average number of cars demanded per day as 1.5. Calculate the
proportion of
#days on which neither car is used and the proportion of days on which some
#demand is refused.

#Let X be a discrete random variable denoting number of cars demanded on a
day
#Possible values of X: 0,1,2,...
#Given, average no. of cars demanded per day, u = 1.5
#To find P(X=0) and P(X>2)

u=1.5
x=0:2
p1=dpois(x,u)[1]
print(paste("P(neither car is used) = P(X=0) = ",p1))

## [1] "P(neither car is used) = P(X=0) =  0.22313016014843"

p2=1-dpois(x,u)[1]-dpois(x,u)[2]-dpois(x,u)[3]
print(paste("P(some demand is refused) = P(X>2) = ",p2))

## [1] "P(some demand is refused) = P(X>2) =  0.191153169461942"

#Q2. If the probability of a bad reaction from medicine is 0.002, determine
#the chance that out of 1000 persons more than 3 will suffer a bad reaction
#from medicine.

#Let X be a random variable denoting number of people suffering a bad
reaction
#X can have values: 0,1,2,3,...
#To find P(X>3),

n=1000
p=0.002
```

```
U=n*p
X=0:3
prob=1-(dpois(X,U)[1]+dpois(X,U)[2]+dpois(X,U)[3]+dpois(X,U)[4])
print(paste("P(X>3) = ",prob))

## [1] "P(X>3) =  0.142876539501453"
```

Practical 7: Normal Distribution

RITIKA GUPTA

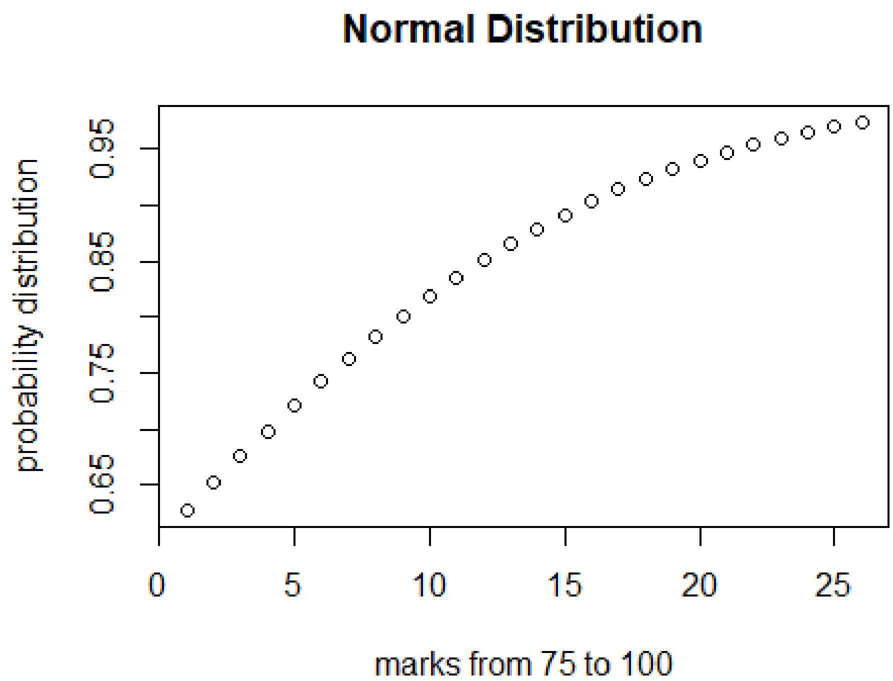
2K22/MSCMAT/54

11-10-2022

```
#Normal Distribution

#Q1. Suppose x is normally distributed with mean 17.96 and variance 375.67.
#X represents the widget weight. What is the probability that a randomly
#chosen widget weight weighs less than 19?

#To find F(19) = P(X<=19),
m=17.96
v=375.67
sd=sqrt(v)
pnorm(19,m,sd)

## [1] 0.521396

#Q2. Assume that the test scores of the exam fit the normal distribution.
#Furthuremore, the mean test score is 70 and the standard deviation is 15.4.
#(a) What is the percentage of students scoring 84 marks or less than in
exam?
#(b) What is the percentage of students scoring 80 marks or more than in
exam?
#(c) What is the percentage of students scoring more than 75 and
#less than equal to 100 in exam?

M=70
SD=15.4

#(a) To find 100*P(X<=84)
pnorm(84,M,SD)*100

## [1] 81.83489

#(b) To find 100*P(X>=80)
(1-pnorm(80,M,SD,lower.tail = FALSE))*100

## [1] 74.19441

#(c) To find 100*P(75<X<=100)
sum(dnorm(75:100,M,SD,log = FALSE))*100
```

```
## [1] 36.12412
```

```
F=pnorm(75:100,M,SD)
plot(F,xlab = "marks from 75 to 100",
     ylab = "probability distribution",
     main = "Normal Distribution")
```




```
#FUNCTIONS USED
#dnorm() gives the density function
#pnorm() gives the distribution function
```

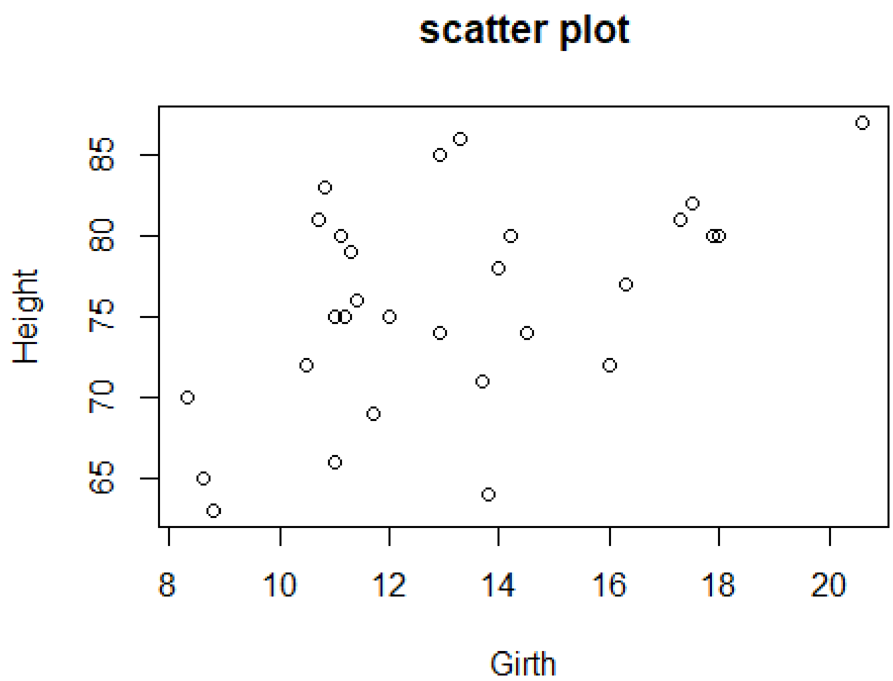
Practical 8: Correlation Analysis

RITIKA GUPTA

2K22/MSCMAT/54

18-10-2022

```
#Correlation Analysis

attach(trees)  #calling the pre-existing dataset trees
summary(trees)

##      Girth           Height       Volume
##  Min.   : 8.30   Min.   :63   Min.   :10.20
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##  Median :12.90   Median :76   Median :24.20
##  Mean   :13.25   Mean   :76   Mean   :30.17
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##  Max.   :20.60   Max.   :87   Max.   :77.00

#correlation coefficient of girth and height
cor(Girth,Height)

## [1] 0.5192801

#correlation coefficient of volume and height
cor(Volume,Height)

## [1] 0.5982497

#correlation coefficient of girth and volume
cor(Girth,Volume)

## [1] 0.9671194

#scatter plot of height vs girth
plot(Girth,Height,main="scatter plot")
```



scatter plot
Height
65
70
75
80
85
8
10
12
14
16
18
20
Girth

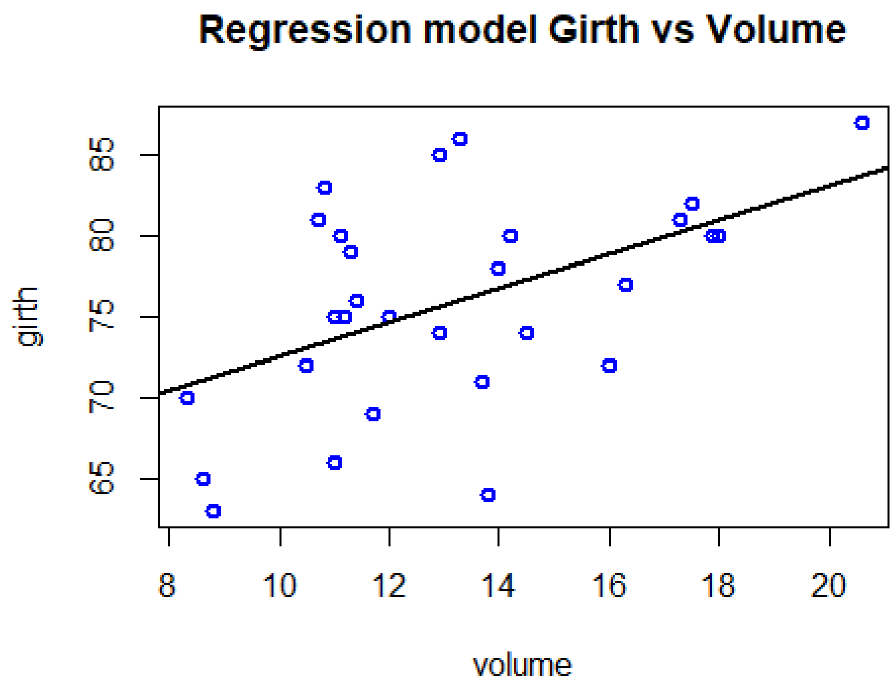
Practical 9: Regression Analysis

RITIKA GUPTA

2K22/MSCMAT/54

18-10-2022

```
# Regression Analysis

attach(trees)
summary(trees)

##      Girth           Height       Volume
##  Min.   : 8.30   Min.   :63   Min.   :10.20
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##  Median :12.90   Median :76   Median :24.20
##  Mean   :13.25   Mean   :76   Mean   :30.17
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##  Max.   :20.60   Max.   :87   Max.   :77.00

model=lm(Girth ~ Volume)
model

##
## Call:
## lm(formula = Girth ~ Volume)
##
## Coefficients:
## (Intercept)       Volume
##      7.6779       0.1846

coefficients(model)

## (Intercept)      Volume
##   7.6778570   0.1846321

plot(Girth,Height,main = "Regression model Girth vs Volume",col='blue',
     lwd=2,xlab="volume",ylab="girth")

#regression line between girth and height
abline(lm(Height~Girth),lwd=2)
```


Regression model Girth vs Volume
girth
85
80
75
70
65
8
10
12
14
16
18
20
volume

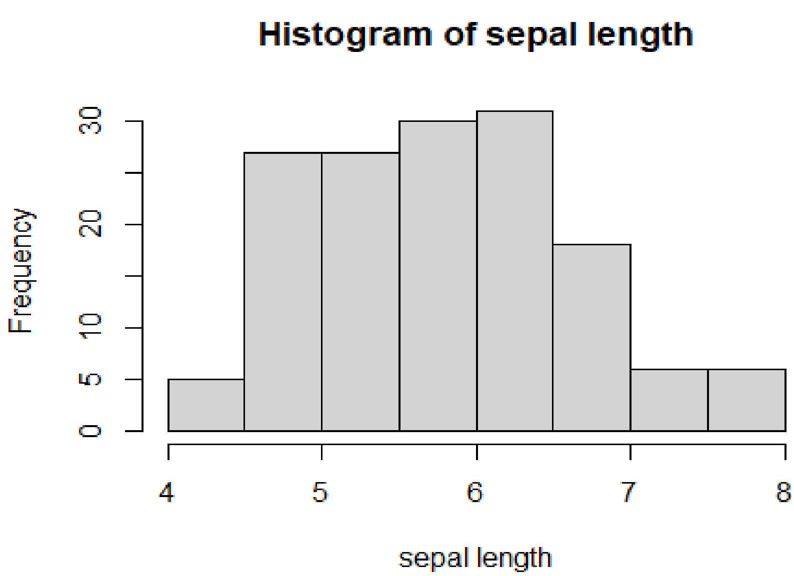
Practical 10: Hypothesis Testing

RITIKA GUPTA

2K22/MSCMAT/54

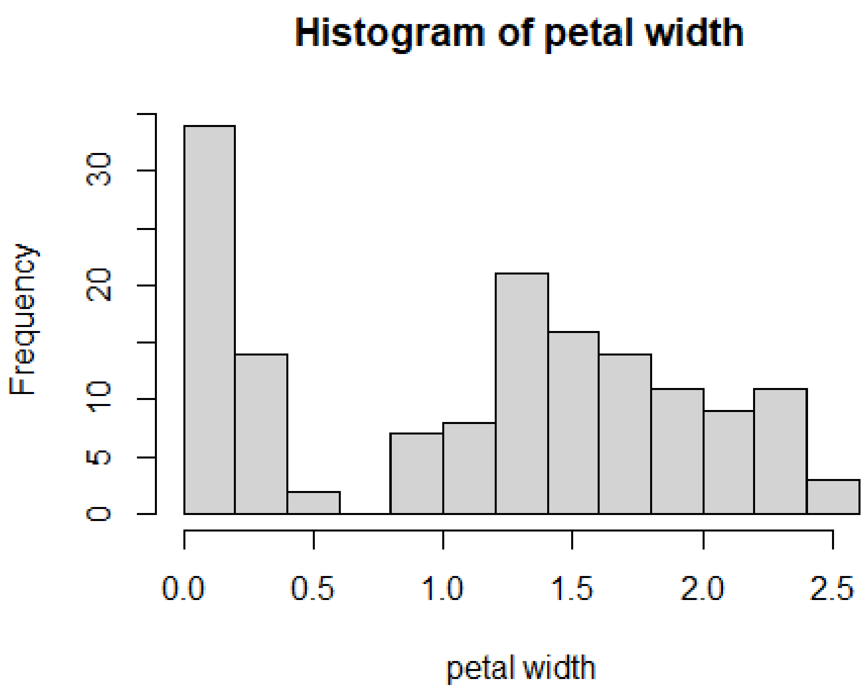
01-11-2022

```
#Q1. To test mean of data
#sample t-test and confidence intervals are parametric method appropriate
#for examining a single numeric variable

attach(iris) #attach the iris dataset
summary(iris)

##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  

#(i) Sepal Length of flower is chosen here for mean to be tested
sl=Sepal.Length

#draw the histogram to predict the mean sepal length
hist(sl,xlab="sepal length",main = "Histogram of sepal length")
```



```r
# we will test the null hypothesis that mean is 6 ,i.e.,
# Ho:mu=6
# one-sided confidence interval for mu
t.test(sl, mu=6, alt="greater", conf=0.95)

##
##  One Sample t-test
##
## data:  sl
## t = -2.3172, df = 149, p-value = 0.9891
## alternative hypothesis: true mean is greater than 6
## 95 percent confidence interval:
##  5.731427      Inf
## sample estimates:
## mean of x
##  5.843333

#A smaller p-value means that there is stronger evidence in favor of
#the alternative hypothesis.

#There is 95% chance that the mean will lie between 5.731427 to infinity

#p=0.9891 > 0.05, i.e., null hyp is not rejected.
#Also, we can see here that exact value of mean is 5.84333 which is less than 6.
#Thus greater p-value signifies lower evidence to support alternate hypothesis.

#testing for alternate hypothesis being that mean is less than 6
t.test(sl, mu=6, alt="less", conf=0.95)

##
##  One Sample t-test
##
## data:  sl
## t = -2.3172, df = 149, p-value = 0.01093
## alternative hypothesis: true mean is less than 6
## 95 percent confidence interval:
##      -Inf 5.95524
## sample estimates:
## mean of x
##  5.843333

#p=0.01093 < 0.05
#There is 95% chance that the mean will lie between -infinity to 5.95524
#i.e., null hypothesis is rejected.

#Here, we know that exact value of mean is less than 6 which supports
#alternate hypothesis

#two-tailed is by default
#confidence interval can be changed
#store it in an object
TEST_1=t.test(sl, mu=6,alt="greater", conf=0.95)
TEST_2=t.test(sl, mu=6,alt="less", conf=0.95)
```

```r
#(ii) Petal Width of flower is chosen here for mean to be tested
pw=Petal.Width

#draw the histogram to predict the mean petal width
hist(pw,xlab="petal width",main = "Histogram of petal width")
```




```r
# we will test the null hypothesis that mean is 1 ,i.e.,
# Ho:mu=1
# one-sided confidence interval for mu
t.test(pw, mu=1, alt="less", conf=0.90)

##
##  One Sample t-test
##
## data:  pw
## t = 3.2028, df = 149, p-value = 0.9992
## alternative hypothesis: true mean is less than 1
## 90 percent confidence interval:
##      -Inf 1.279448
## sample estimates:
## mean of x
##  1.199333

#Here, p=0.9992 > 0.05
#There is 90% chance that the mean will lie between -infinity to 1.279448
#Thus, null hypothesis is rejected.

#testing for alternate hypothesis being that mean is greater than 1
t.test(pw, mu=1, alt="greater", conf=0.90)

##
##  One Sample t-test
```

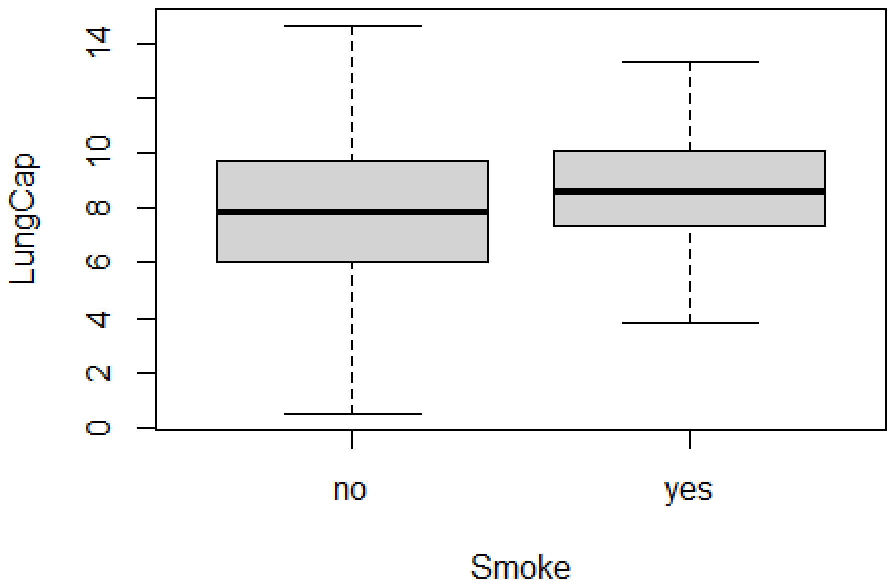
```
##
## data:  pw
## t = 3.2028, df = 149, p-value = 0.0008321
## alternative hypothesis: true mean is greater than 1
## 90 percent confidence interval:
##  1.119219      Inf
## sample estimates:
## mean of x
##  1.199333
```

```
#p=0.00083211 < 0.05, i.e., null hypothesis is not rejected.
#There is 95% chance that the mean will lie between 1.119219 and infinity

#Also, we can see here that exact value of mean is 1.199 > 1
#Thus lower p-value provides stronger evidence in favor of
#the alternative hypothesis.



#Q2. To examine the difference in means of two populations
#we will explore the relationship between mean lung capacity of smokers and
#non-smokers

data1=read.table(file.choose(), header=T) #reads the text file in system
attach(data1)
summary(data1)
```

```
##     LungCap            Age            Height          Smoke          
##  Min.   : 0.507   Min.   : 3.00   Min.   :45.30   Length:725        
##  1st Qu.: 6.150   1st Qu.: 9.00   1st Qu.:59.90   Class :character  
##  Median : 8.000   Median :13.00   Median :65.40   Mode  :character  
##  Mean   : 7.863   Mean   :12.33   Mean   :64.84                     
##  3rd Qu.: 9.800   3rd Qu.:15.00   3rd Qu.:70.30                     
##  Max.   :14.675   Max.   :19.00   Max.   :81.80                     
##     Gender           Caesarean        
##  Length:725         Length:725        
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
```

	LungCap	Age	Height	Smoke	Gender	Caesarean
1	6.475	6	62.1	no	male	no
2	10.125	18	74.7	yes	female	no
3	9.550	16	69.7	no	female	yes
4	11.125	14	71.0	no	male	no
5	4.800	5	56.9	no	male	no
6	6.225	11	58.7	no	female	no
7	4.950	8	63.3	no	male	yes
8	7.325	11	70.4	no	male	no
9	8.875	15	70.5	no	male	no

```
#Take two variables for the test, 1st is the LungCap which is numeric variable,
#2nd is Smoke which is categorical variable, "yes" means the individual is a
#smoker and "no" means he/she is non-smoker

#examine the plot before conducting the test
boxplot(LungCap ~ Smoke,data1)
```




```
#test the hypothesis that mean lung capacity of smokers = non smokers
#Ho= mean lungcap of smokers = mean lungcap of non smokers
#assume non equal variances

t.test(LungCap~Smoke,data1, mu=0, alt="two.sided", conf=0.95,
       var.equal=FALSE,paired=FALSE)

##
##  Welch Two Sample t-test
##
## data:  LungCap by Smoke
## t = -3.6498, df = 117.72, p-value = 0.0003927
## alternative hypothesis: true difference in means between group no and group y
es is not equal to 0
## 95 percent confidence interval:
##  -1.3501778 -0.4003548
## sample estimates:
##  mean in group no mean in group yes
##          7.770188          8.645455

#Since p=0.0003927 < 0.05, null hypothesis is rejected,
#i.e., average lung capacity of smokers and non-smokers is not equal

#check variance of two groups
var(data1$LungCap[data1$Smoke=="yes"])

## [1] 3.545292

var(data1$LungCap[data1$Smoke=="no"])

## [1] 7.431694
```