SIES

College of Arts,
Science &
Commerce

RISE WITH EDUCATION

NAAC REACCREDITED - ‘A’ GRADE

ISO 9001 : 2008

S.I.E.S College of Arts, Science and Commerce
Sion(W), Mumbai – 400 022.

CERTIFICATE

This is to certify that Mr. / Miss. **NADAR RAMTILAK SAIT SANKARALINGAM** Roll No. **FCS2122070** Has successfully completed the necessary course of experiments in the subject of **STATISTICAL METHODS AND TESTING OF HYPOTHESIS** during the academic year **2021 – 2022** complying with the requirements of **University of Mumbai**, for the course of **F.Y.BSc. Computer Science [Semester-2]**

Prof. In-Charge
Soni Yadav
(STATISTICAL METHODS AND TESTING OF HYPOTHESIS)

Examination Date:
Examiner’s Signature & Date:

Head of the Department
Prof. Manoj Singh

College Seal
And
Date

INDEX

Date :- 02-03-2022

Practical No. 1

Aim :- Binomial Distribution

The **binomial distribution** is a discrete probability distribution. If the probability of a successful trial is *p*, then the probability of having *x* successful outcomes in an experiment of *n* independent trials is as follows.

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{(n-x)} \quad where\ x = 0, 1, 2, ..., n$$

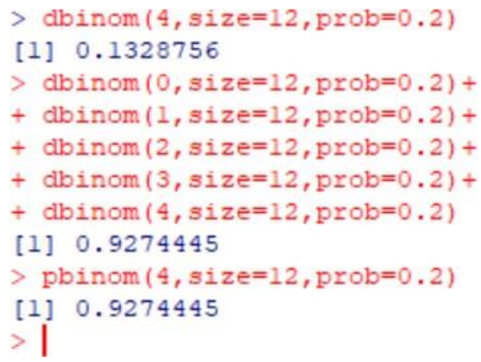
Problem :-

Suppose there are twelve multiple-choice questions in an English class quiz. Each questions have five possible answers, and only one of them is correct

I. Find the probability of having four or fewer correct answers if a student attempts to answer every question randomly.

II. Exactly four correct answers.

Code :-



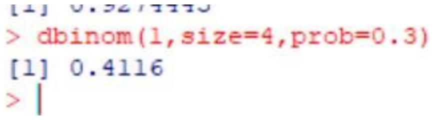
```
> dbinom(4,size=12,prob=0.2)
[1] 0.1328756
> dbinom(0,size=12,prob=0.2)+
+ dbinom(1,size=12,prob=0.2)+
+ dbinom(2,size=12,prob=0.2)+
+ dbinom(3,size=12,prob=0.2)+
+ dbinom(4,size=12,prob=0.2)
[1] 0.9274445
> pbinom(4,size=12,prob=0.2)
[1] 0.9274445
> |
```

Exercise :-

Problem 1 :-

In a store, out of all the people who came there, thirty percent bought a shirt. If four people came in the store together then find the probability of one of them buying a shirt.

Code :-

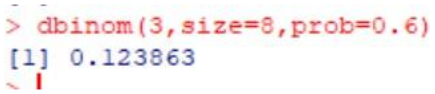


```
[1] 0.9274445
> dbinom(1,size=4,prob=0.3)
[1] 0.4116
> |
```

Problem 2 :-

In a hospital, sixty percent of patients are dying of a disease. If eight patients got admitted to the hospital for that disease on a certain day, what are the chances of three surviving?

Code :-

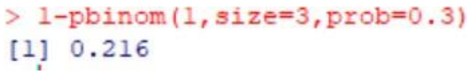


```
> dbinom(3,size=8,prob=0.6)
[1] 0.123863
```

Problem 3 :-

In a restaurant, seventy percent of people order Chinese food and thirty percent for Italian food. A group of three persons enters the restaurant. Find the probability of at least two of them ordering Italian food.

Code :-

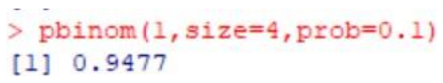


```
> 1-pbinom(1,size=3,prob=0.3)
[1] 0.216
```

Problem 4 :-

In an exam, only ten percent of students can qualify. If a group of 4 students has appeared, find the probability that at most one student will qualify?

Code :-



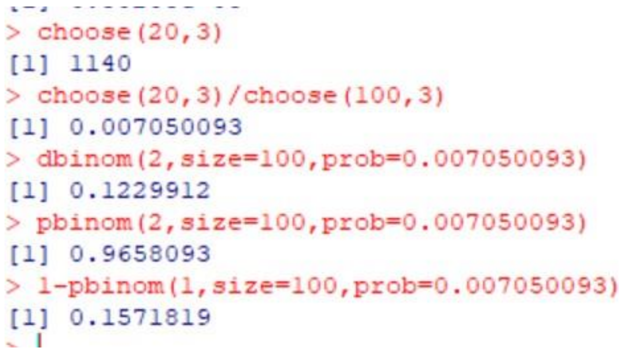
```
> pbinom(1,size=4,prob=0.1)
[1] 0.9477
```

Problem 5 :-

A basket contains 20 good oranges and 80 bad oranges. 3 oranges are drawn at random from this basket. Find the probability that out of 3

i) exactly 2
ii) at least 2
iii) at most 2 are good oranges.

Code :-



```
> choose(20,3)
[1] 1140
> choose(20,3)/choose(100,3)
[1] 0.007050093
> dbinom(2,size=100,prob=0.007050093)
[1] 0.1229912
> pbinom(2,size=100,prob=0.007050093)
[1] 0.9658093
> 1-pbinom(1,size=100,prob=0.007050093)
[1] 0.1571819
```

Date :- 17-03-2022

Practical No. 2

Aim :- Normal Distribution

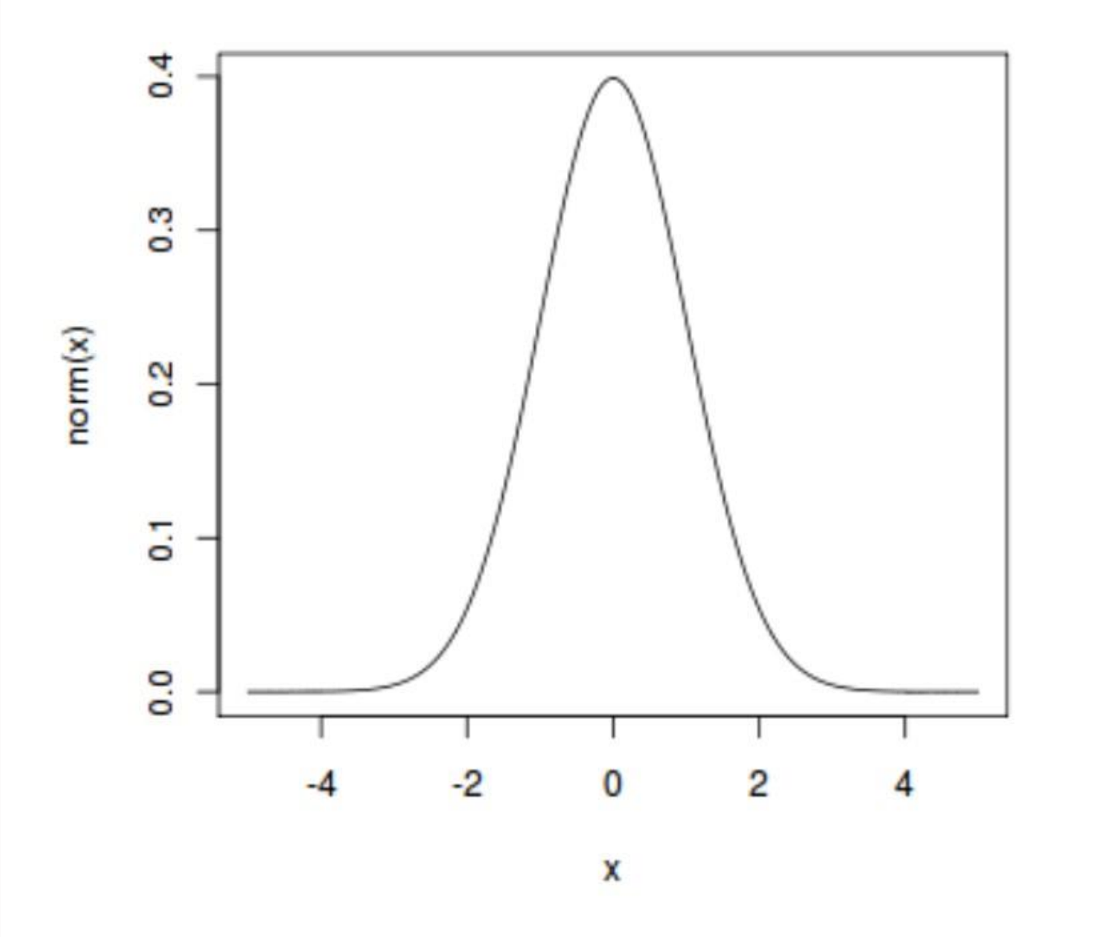
The normal distribution is defined by the following probability density function, where μ is the population mean and σ^2 is the variance.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable X follows the normal distribution, then we write:

$$X \sim N(\mu, \sigma^2)$$

In particular, the normal distribution with μ = 0 and σ = 1is called the standard normal distribution, and is denoted as N(0,1). It can be graphed as follows.



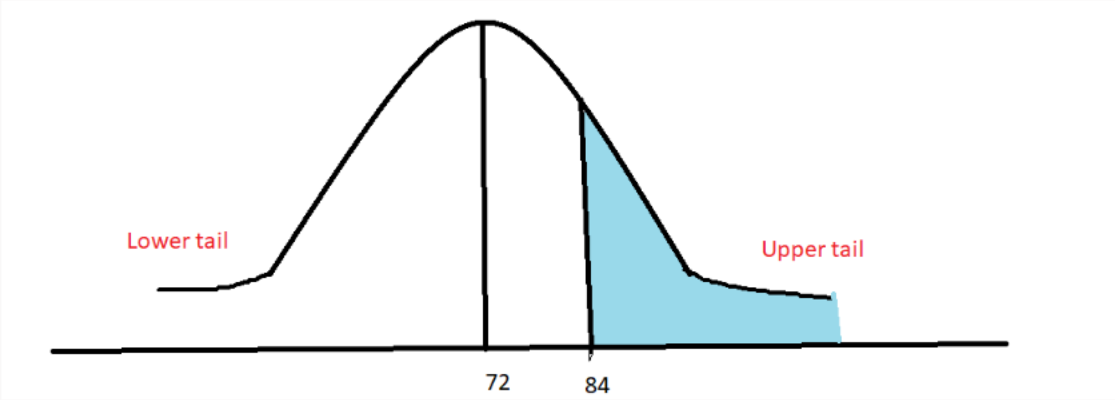
The normal distribution is important because of the **Central Limit Theorem**, which states that the population of all possible samples of size n from a population with mean μ and variance σ^2 approaches a normal distribution with meanμandσ2⁄nwhennapproaches infinity.

Problem :-

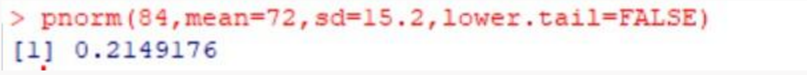
Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

Solution :-

We apply the function pnorm of the normal distribution with mean 72 and standard deviation 15.2. Since we are looking for the percentage of students scoring higher than 84, we are interested in the upper tail of the normal distribution.



>pnorm(84,mean=72,sd=15.2,lower.tail=FALSE)[1]0.21492



```
> pnorm(84,mean=72,sd=15.2,lower.tail=FALSE)
[1] 0.2149176
```

Answer :-

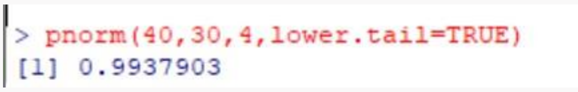
The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

Exercise :-

1.X is a normally distributed variable with mean μ =30 and standard deviation σ = 4. Find

a) P(x < 40)

Code :-



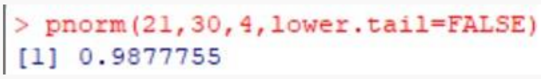
```
> pnorm(40,30,4,lower.tail=TRUE)
[1] 0.9937903
```

Answer :-

The probability of x less than 40 is 0.9937903.

b) P(x > 21)

Code :-



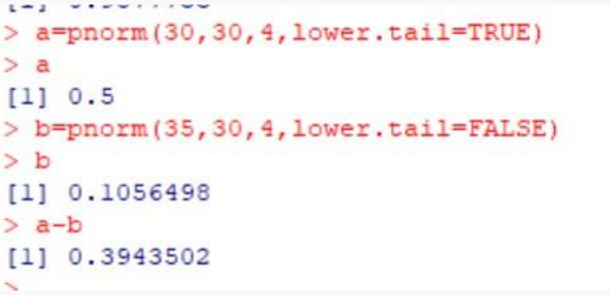
```
> pnorm(21,30,4,lower.tail=FALSE)
[1] 0.9877755
```

Answer :-

The probability of x greater than 21 is 0.9877755.

c) P(30 < x < 35)

Code :-



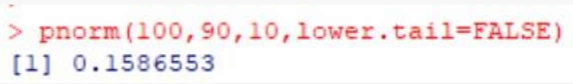
```
> a=pnorm(30,30,4,lower.tail=TRUE)
> a
[1] 0.5
> b=pnorm(35,30,4,lower.tail=FALSE)
> b
[1] 0.1056498
> a-b
[1] 0.3943502
```

Answer :-

The probability of x lying between 30 and 35 is 0.3943502.

2. A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

Code :-

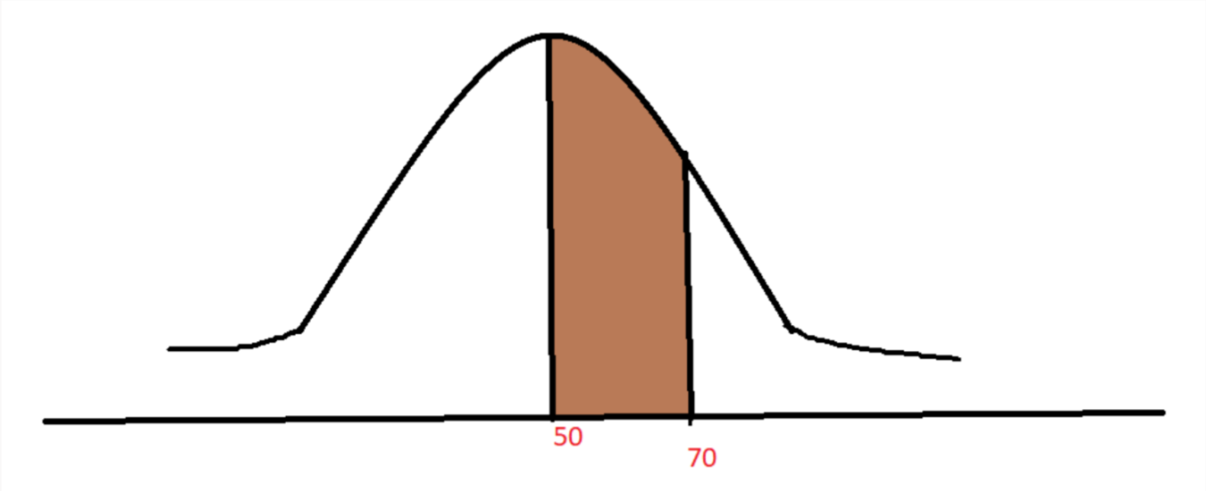


```
> pnorm(100,90,10,lower.tail=FALSE)
[1] 0.1586553
```

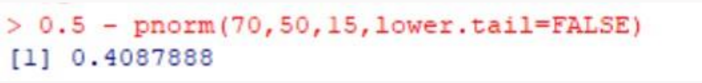
Answer :-

The probability of a car picked at random is travelling at more than 100 km/hr is 0.1586553.

3. For a certain type of computers, the length of time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.



Code :-



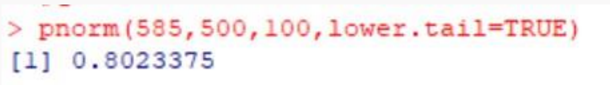
```
> 0.5 - pnorm(70,50,15,lower.tail=FALSE)
[1] 0.4087888
```

Answer :-

The probability that the length of time will be between 50 and 70 hours is 0.4087888

4. Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. Tom wants to be admitted to this university and he knows that he must score better than at least 70% in the test. Tom takes the test and scores 585. Will he be admitted to this university?

Code :-



```
> pnorm(585,500,100,lower.tail=TRUE)
[1] 0.8023375
```

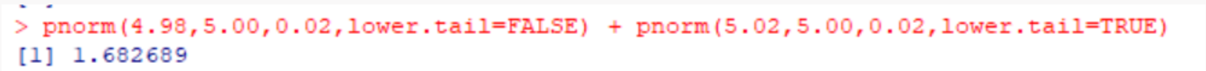
Answer :-

Yes, Tom will be admitted to the university as he scored 80.23%.

5. The length of similar components produced by a company are approximated by a normal distribution model with a mean of 5 cm and a standard deviation of 0.02 cm. If a component is chosen at random.

a) what is the probability that the length of this component is between 4.98 and 5.02 cm?

Code :-



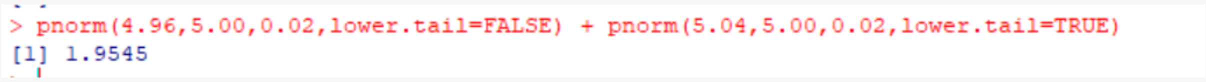
```
> pnorm(4.98,5.00,0.02,lower.tail=FALSE) + pnorm(5.02,5.00,0.02,lower.tail=TRUE)
[1] 1.682689
```

Answer :-

The probability that the length of the component chosen is between 4.98 and 5.02cm is 1.682689.

b) what is the probability that the length of this component is between 4.96 and 5.04 cm?

Code :-



```
> pnorm(4.96,5.00,0.02,lower.tail=FALSE) + pnorm(5.04,5.00,0.02,lower.tail=TRUE)
[1] 1.9545
```

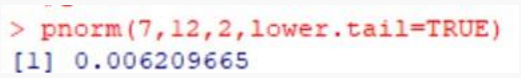
Answer :-

The probability that the length of the chosen component is between 4.96 and 5.04cm is 1.9545.

6.The length of life of an instrument produced by a machine has a normal distribution with a mean of 12 months and standard deviation of 2 months. Find the probability that an instrument produced by this machine will last

a) less than 7 months.

Code :-



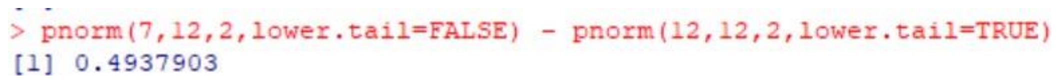
```
> pnorm(7,12,2,lower.tail=TRUE)
[1] 0.006209665
```

Answer :-

The probability that an instrument produced by the machine will last less than 7 months is 0.006209665.

b) between 7 and 12 months

Code :-



```
> pnorm(7,12,2,lower.tail=FALSE) - pnorm(12,12,2,lower.tail=TRUE)
[1] 0.4937903
```

Answer :-

The probability that an instrument produced by the machine will last between 7 and 12 months is 0.4937903.

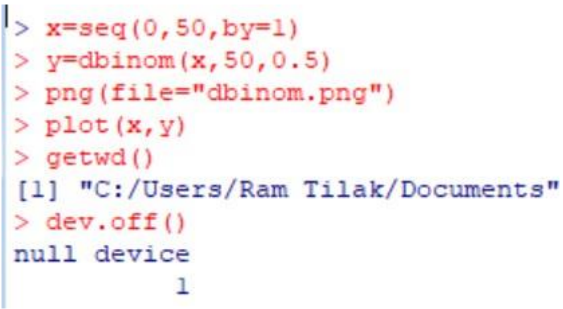
Date :- 23-03-2022

Practical No. 3

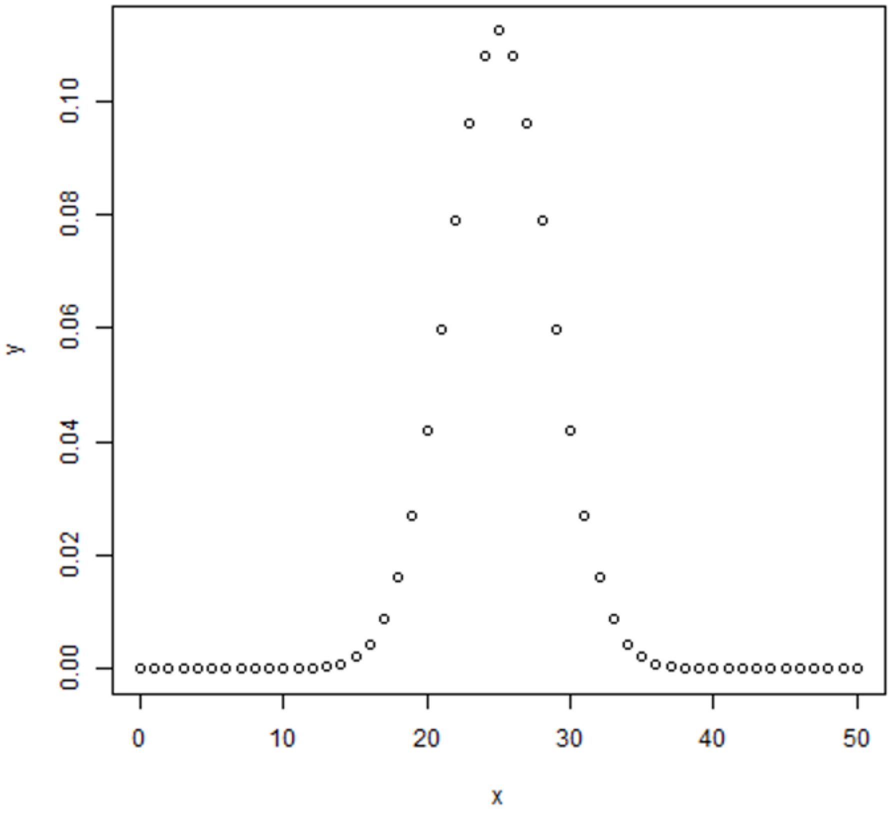
Aim :- Property plotting Binomial distributions

dbinom()

This function gives the probability density distribution at each point.

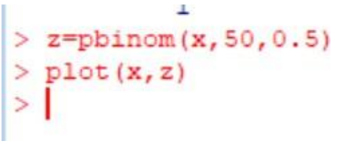


```
> x=seq(0,50,by=1)
> y=dbinom(x,50,0.5)
> png(file="dbinom.png")
> plot(x,y)
> getwd()
[1] "C:/Users/Ram Tilak/Documents"
> dev.off()
null device
          1
```

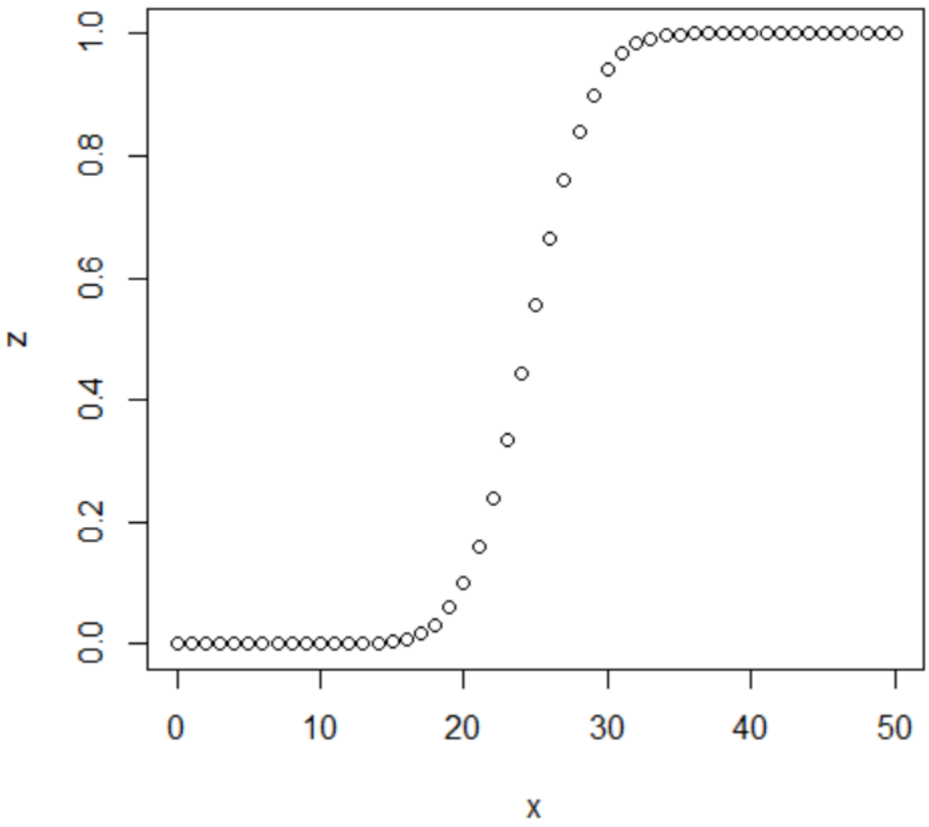


pbinom()

This function gives the cumulative probability of an event. It is a single value representing the probability.

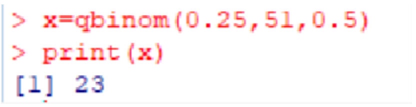


```
> z=pbinom(x,50,0.5)
> plot(x,z)
> 
```



qbinom()

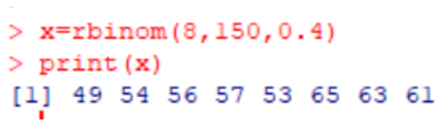
This function takes the probability value and gives a number whose cumulative value matches the probability value.



```
> x=qbinom(0.25,51,0.5)
> print(x)
[1] 23
```

rbinom()

This function generates required number of random values of given probability from a given sample.



```
> x=rbinom(8,150,0.4)
> print(x)
[1] 49 54 56 57 53 65 63 61
```

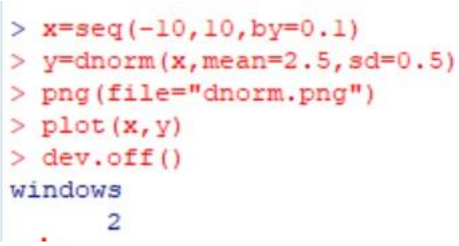
Date :- 23-03-2022

Practical No. 4

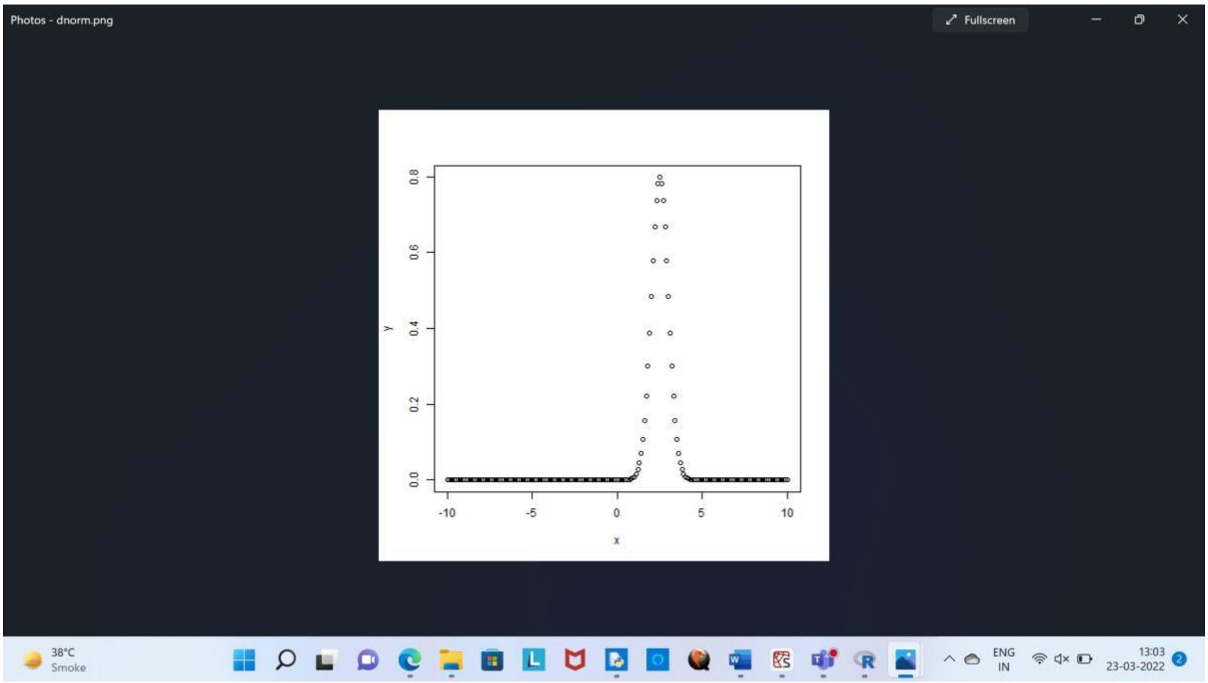
Aim :- Property plotting of normal distribution

dnorm()

This function gives height of the probability distribution at each point for a given mean and standard deviation.

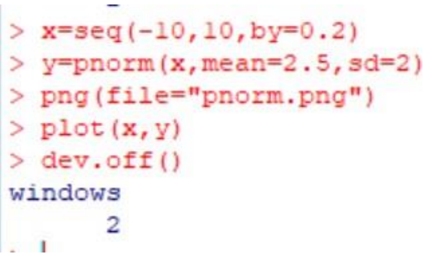


```
> x=seq(-10,10,by=0.1)
> y=dnorm(x,mean=2.5,sd=0.5)
> png(file="dnorm.png")
> plot(x,y)
> dev.off()
windows
      2
```

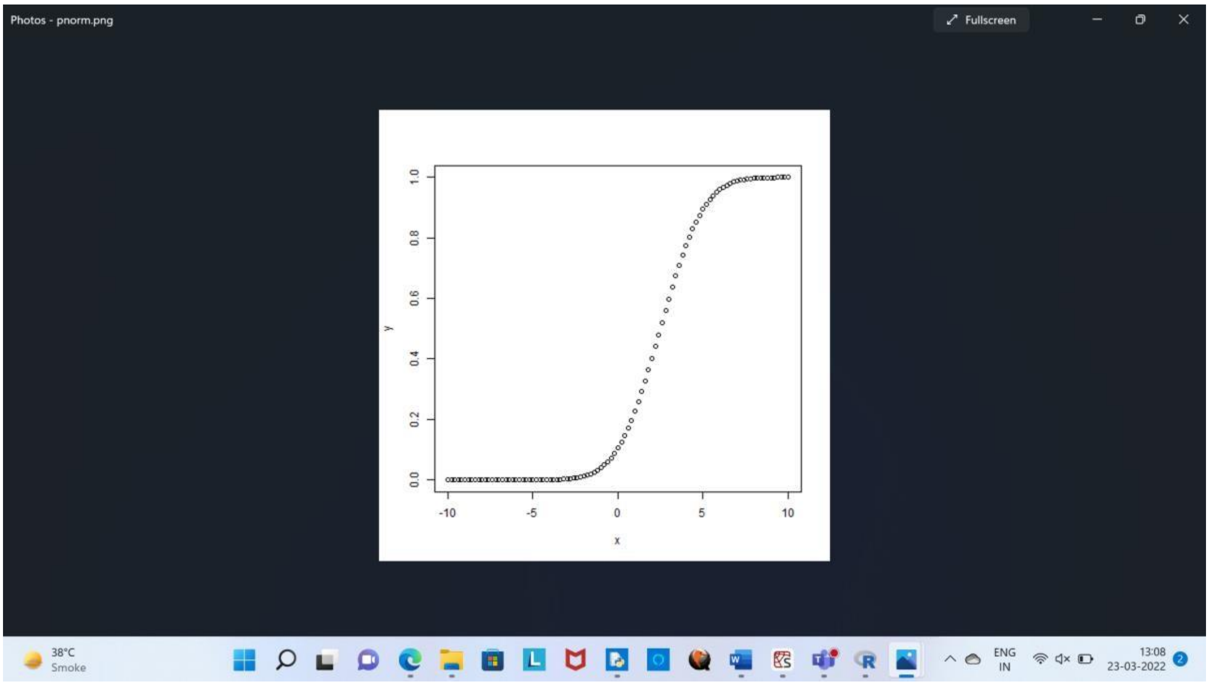


pnorm()

This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".

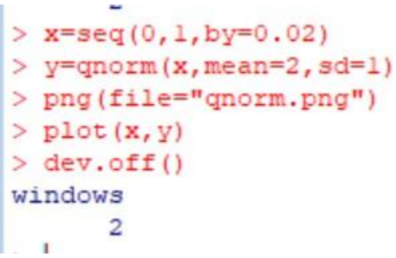


```
> x=seq(-10,10,by=0.2)
> y=pnorm(x,mean=2.5,sd=2)
> png(file="pnorm.png")
> plot(x,y)
> dev.off()
windows
      2
```

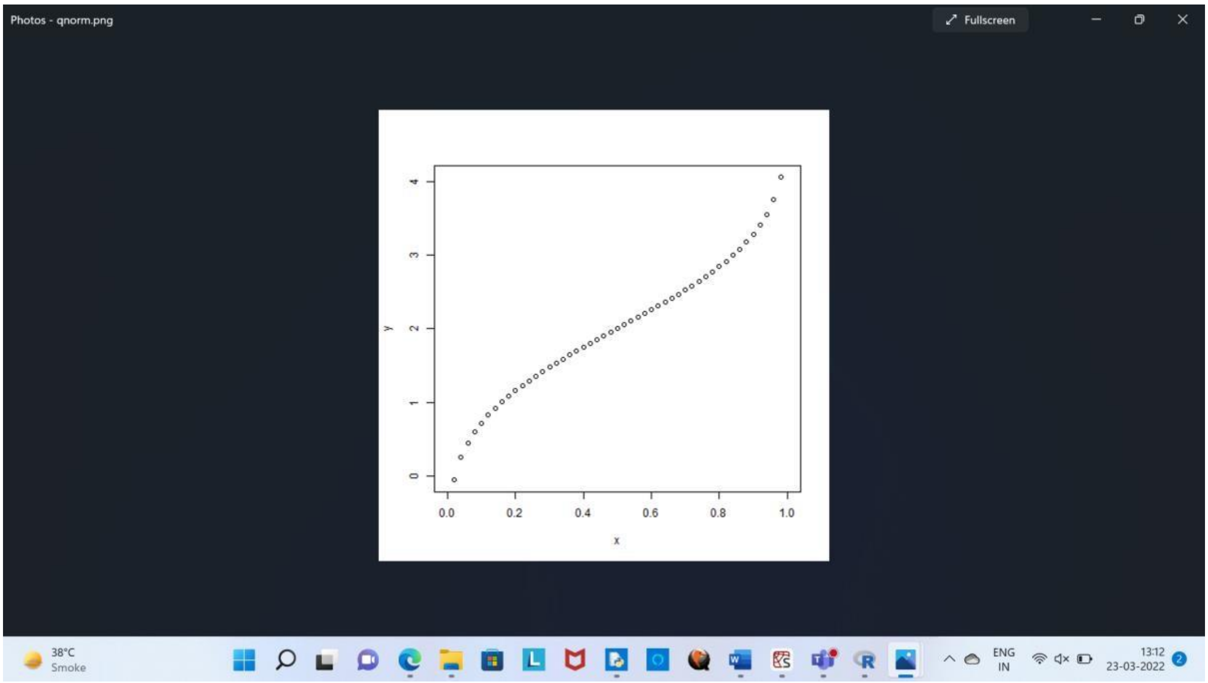


qnorm()

This function takes the probability value and gives a number whose cumulative value matches the probability value.

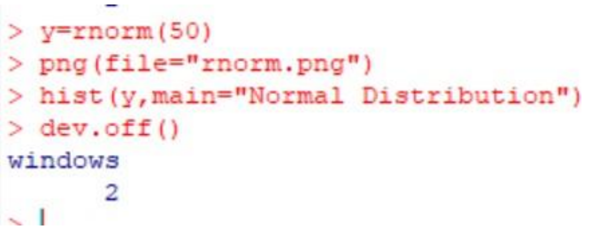


```
> x=seq(0,1,by=0.02)
> y=qnorm(x,mean=2,sd=1)
> png(file="qnorm.png")
> plot(x,y)
> dev.off()
windows
      2
```

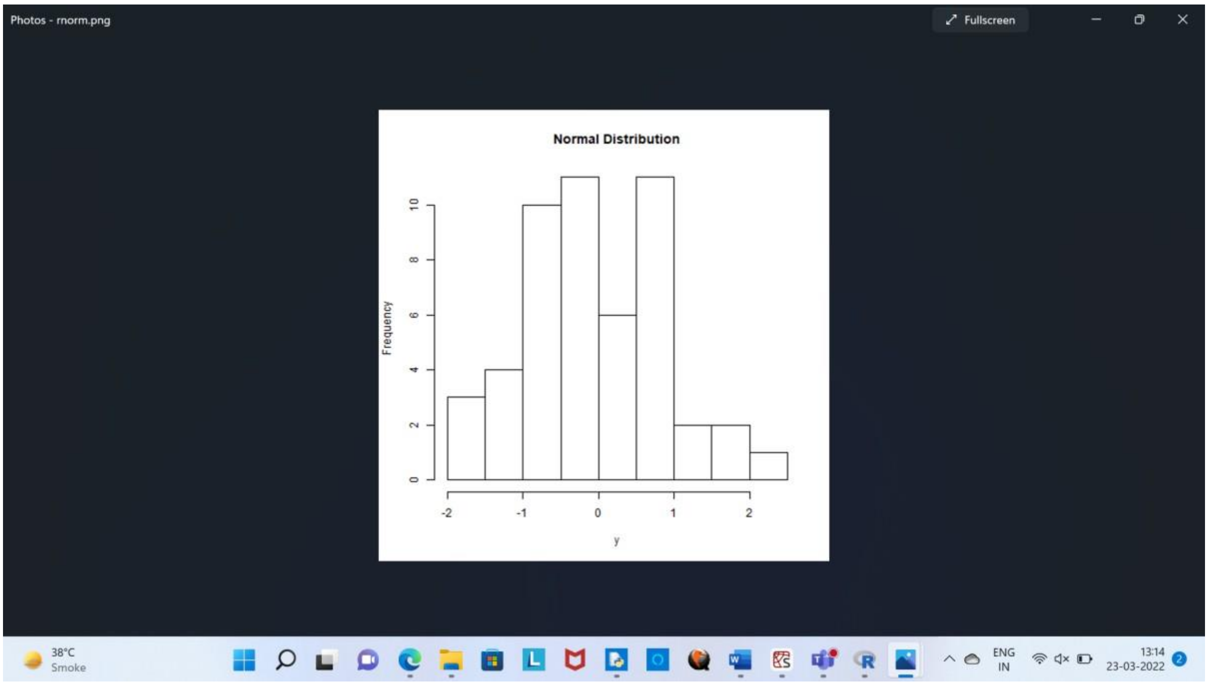


rnorm()

This function is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.



```
> y=rnorm(50)
> png(file="rnorm.png")
> hist(y,main="Normal Distribution")
> dev.off()
windows
      2
> |
```

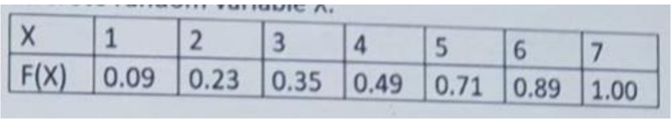


Date :- 01-04-2022

Practical No. 5

Aim :- Problems based on pdf, cdf, pmf, for discrete and continuous distribution.

Q.1. Following is the cumulative distribution function of a discrete random variable X.



X	1	2	3	4	5	6	7
F(X)	0.09	0.23	0.35	0.49	0.71	0.89	1.00

Find

a) p.m.f of X

b) Mean

c) Standard Deviation

d) P(2<=x<=6)

e) P(x=4/x>=2)

Sol:-

X	1	2	3	4	5	6	7
F(X)	0·09	0·23	0·35	0·49	0·71	0·89	1·00
P(X)	0·09	0·14	0·12	0·14	0·22	0·18	0·11
X·P(X)	0·09	0·28	0·36	0·56	1·1	1·08	0·77
X²·P(X)	0·09	0·56	1·08	2·24	5·5	6·48	5·39

$$\text{Mean } (E(x)) = \sum_{k=1}^{n} x_k \cdot P(x_k)$$

$$= 1(0.09) + 2(0.14) + 3(0.12) + 4(0.14) + 5(0.22)$$
$$+ 6(0.18) + 7(0.11)$$
$$= 0.09 + 0.28 + 0.36 + 0.56 + 1.1 + 1.08 + 0.77$$
$$= 4.24$$

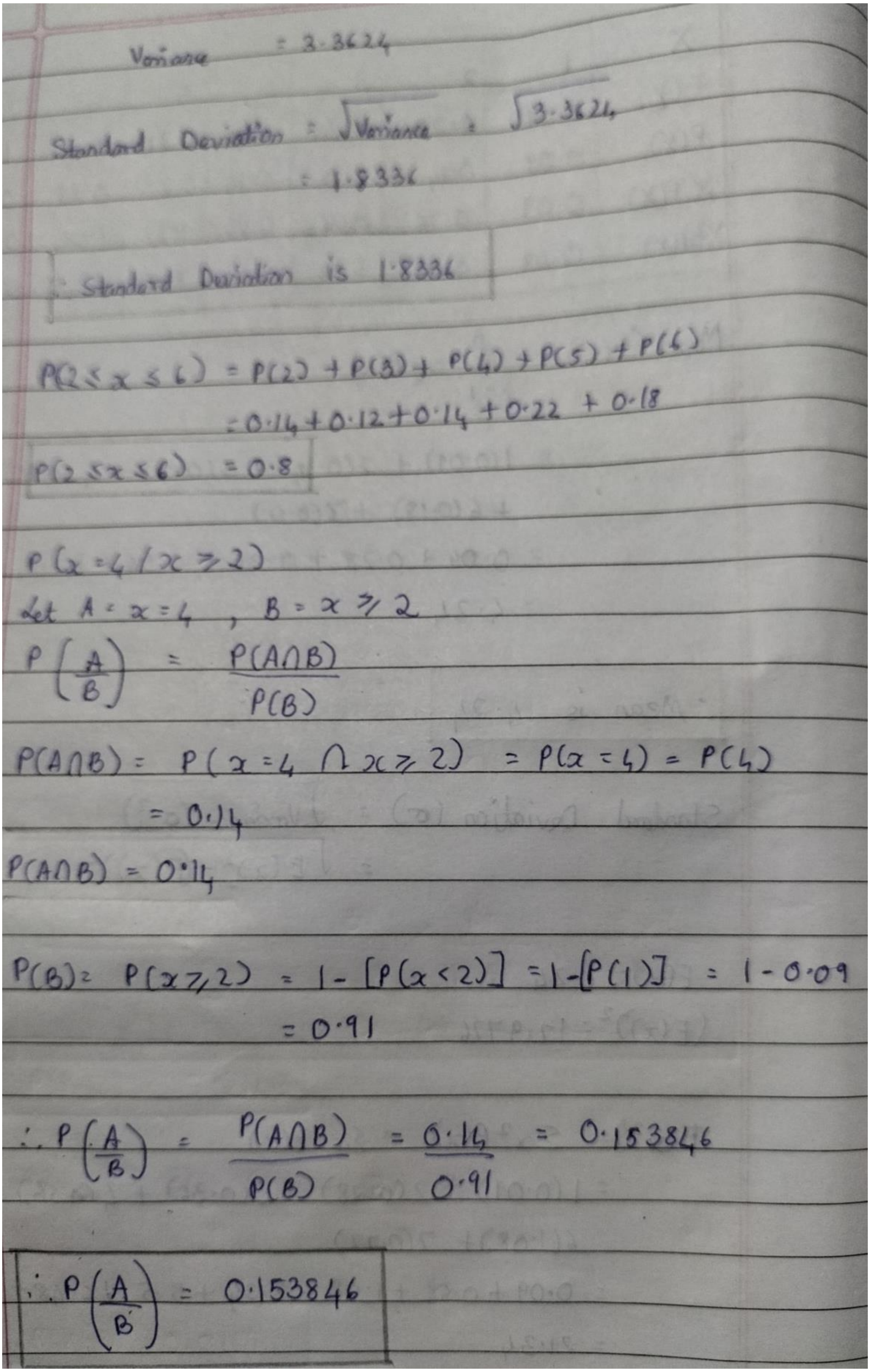
$$\boxed{\therefore \text{Mean is } 4.24}$$

$$\text{Standard Deviation } (\sigma) = \sqrt{\text{Variance } (\sigma^2)}$$
$$= \sqrt{E(x)^2 - (E(x))^2}$$

$$E(x) = 4.24$$
$$(E(x))^2 = 17.9776$$

$$E(x^2) = \sum x^2 P(x) = \sum x \cdot E(x)$$
$$= 1(0.09) + 2(0.28) + 3(0.36) + 4(0.56) + 5(1.1) +$$
$$6(1.08) + 7(0.77)$$
$$= 0.09 + 0.56 + 1.08 + 2.34 + 5.5 + 6.48 + 5.39$$
$$= 21.34$$

$$\text{Variance} = E(x^2) - (E(x))^2 = 21.34 - 17.9776$$



Variance $= 3.3624$

Standard Deviation $= \sqrt{\text{Variance}} = \sqrt{3.3624}$

$= 1.8336$

$\boxed{\therefore \text{Standard Deviation is } 1.8336}$

$P(2 \leq x \leq 6) = P(2) + P(3) + P(4) + P(5) + P(6)$

$= 0.14 + 0.12 + 0.14 + 0.22 + 0.18$

$\boxed{P(2 \leq x \leq 6) = 0.8}$

$P(x = 4 / x \geq 2)$

Let $A = x = 4$, $B = x \geq 2$

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

$P(A \cap B) = P(x = 4 \cap x \geq 2) = P(x = 4) = P(4)$

$= 0.14$

$P(A \cap B) = 0.14$

$P(B) = P(x \geq 2) = 1 - [P(x < 2)] = 1 - [P(1)] = 1 - 0.09$

$= 0.91$

$$\therefore P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{0.14}{0.91} = 0.153846$$

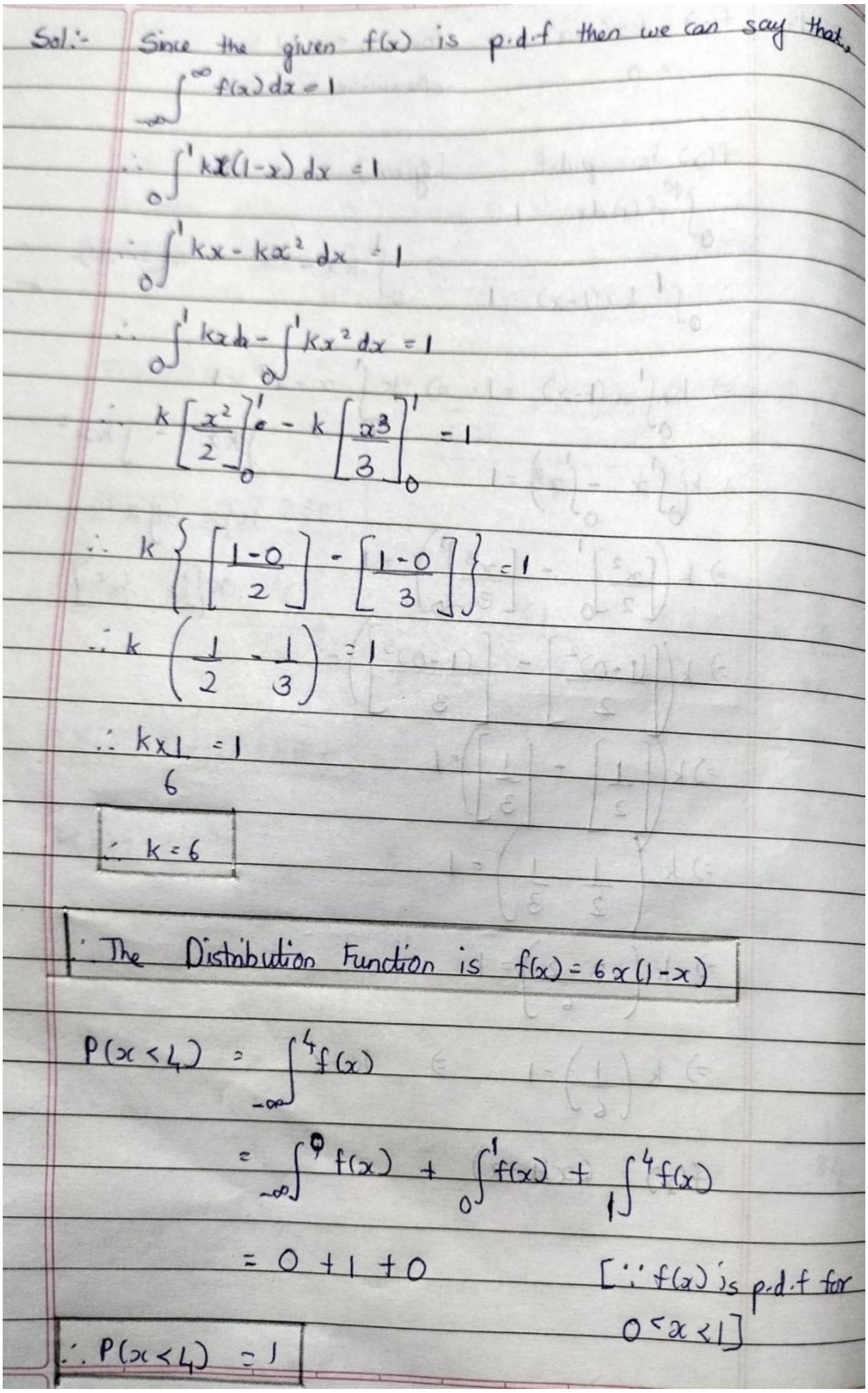
$$\boxed{\therefore P\left(\frac{A}{B}\right) = 0.153846}$$

Q.2. Let X be continuous random variable with p.d.f

f(x) = kx(1-x) , for 0<x<1

= 0 , otherwise

Find k, Distribution Function of X, P(x<4)



Sol:- Since the given f(x) is p.d.f then we can say that,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\therefore \int_{0}^{1} kx(1-x)\,dx = 1$$

$$\therefore \int_{0}^{1} kx - kx^2\,dx = 1$$

$$\therefore \int_{0}^{1} kx\,dx - \int_{0}^{1} kx^2\,dx = 1$$

$$\therefore k\left[\frac{x^2}{2}\right]_0^1 - k\left[\frac{x^3}{3}\right]_0^1 = 1$$

$$\therefore k\left\{\left[\frac{1-0}{2}\right] - \left[\frac{1-0}{3}\right]\right\} = 1$$

$$\therefore k\left(\frac{1}{2} - \frac{1}{3}\right) = 1$$

$$\therefore k \times \frac{1}{6} = 1$$

$$\boxed{\therefore k = 6}$$

$\boxed{\therefore \text{The Distribution Function is } f(x) = 6x(1-x)}$

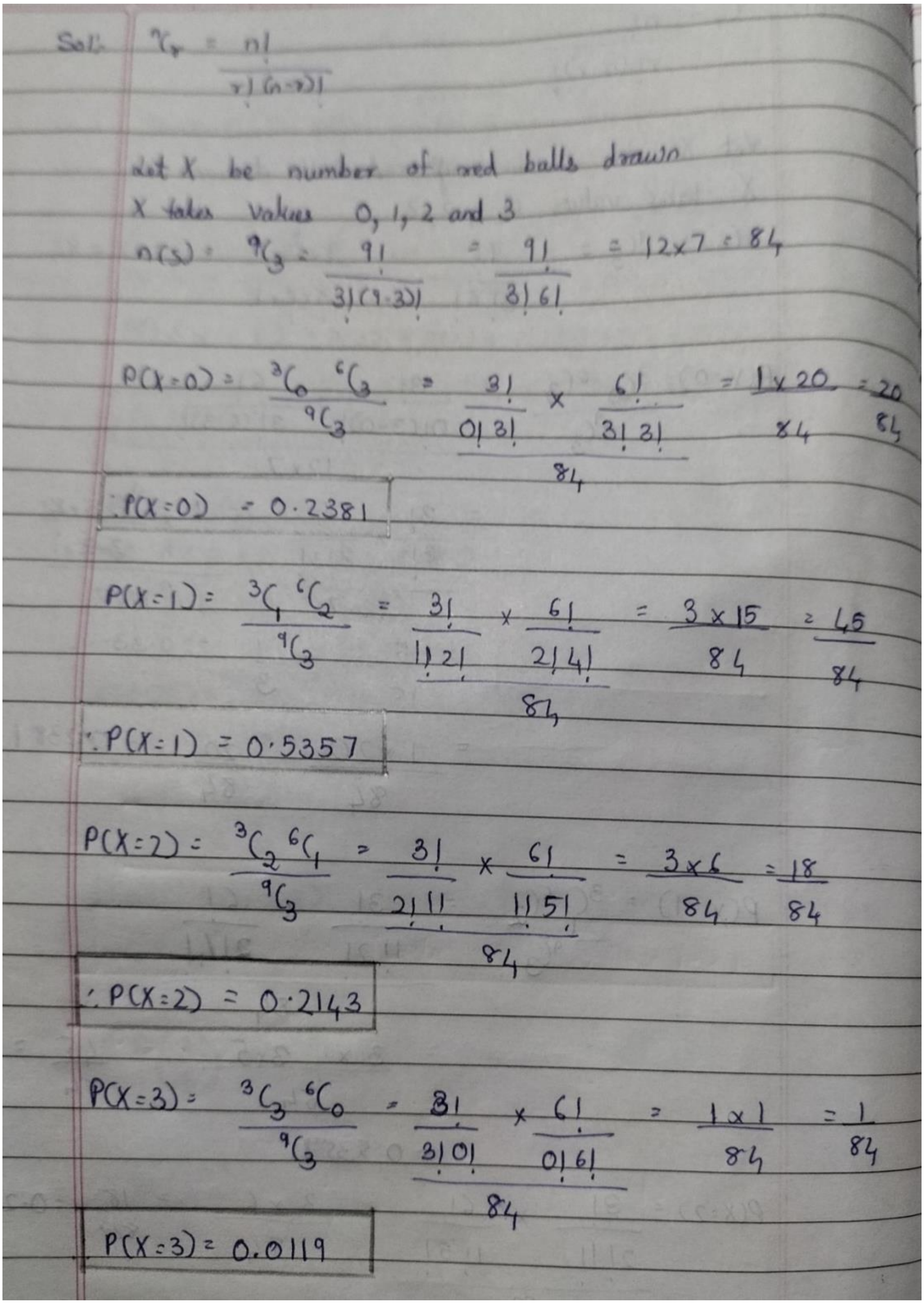
$$P(x<4) = \int_{-\infty}^{4} f(x)$$

$$= \int_{-\infty}^{0} f(x) + \int_{0}^{1} f(x) + \int_{1}^{4} f(x)$$

$$= 0 + 1 + 0 \qquad [\because f(x) \text{ is p.d.f for } 0<x<1]$$

$$\boxed{\therefore P(x<4) = 1}$$

Q.3. A bag contains 6 green and 3 red balls. Three balls are drawn at random without replacement. What is expected number of red balls that will be drawn?



Sol: ${}^nC_r = \frac{n!}{r!(n-r)!}$

Let X be number of red balls drawn

X takes values 0, 1, 2 and 3

$$n(S) = {}^9C_3 = \frac{9!}{3!(9-3)!} = \frac{9!}{3!6!} = 12\times7 = 84$$

$$P(X=0) = \frac{{}^3C_0\,{}^6C_3}{{}^9C_3} = \frac{\frac{3!}{0!3!}\times\frac{6!}{3!3!}}{84} = \frac{1\times20}{84} = \frac{20}{84}$$

$$\boxed{\therefore P(X=0) = 0.2381}$$

$$P(X=1) = \frac{{}^3C_1\,{}^6C_2}{{}^9C_3} = \frac{\frac{3!}{1!2!}\times\frac{6!}{2!4!}}{84} = \frac{3\times15}{84} = \frac{45}{84}$$

$$\boxed{\therefore P(X=1) = 0.5357}$$

$$P(X=2) = \frac{{}^3C_2\,{}^6C_1}{{}^9C_3} = \frac{\frac{3!}{2!1!}\times\frac{6!}{1!5!}}{84} = \frac{3\times6}{84} = \frac{18}{84}$$

$$\boxed{\therefore P(X=2) = 0.2143}$$

$$P(X=3) = \frac{{}^3C_3\,{}^6C_0}{{}^9C_3} = \frac{\frac{3!}{3!0!}\times\frac{6!}{0!6!}}{84} = \frac{1\times1}{84} = \frac{1}{84}$$

$$\boxed{P(X=3) = 0.0119}$$

Date :- 01-04-2022

Practical No. 6

Aim :- Z test , t test

Lower Tail Test of Population Mean with Known Variance

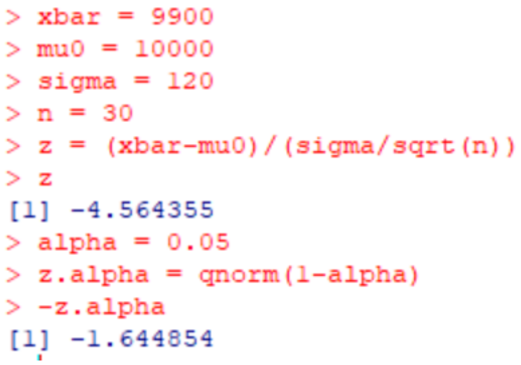
Problem :-

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At .05 significance level, can we reject the claim by the manufacturer?

Solution :-

Null Hypothesis :- H_0 : mean=10000

Alternative Hypothesis :- H_1 : mean >10000



```
> xbar = 9900
> mu0 = 10000
> sigma = 120
> n = 30
> z = (xbar-mu0)/(sigma/sqrt(n))
> z
[1] -4.564355
> alpha = 0.05
> z.alpha = qnorm(1-alpha)
> -z.alpha
[1] -1.644854
```

One sided, alpha=0.05, Z= -4.5644 , Z.alpha=1.6449

If Z > Z_{alpha} True Reject H_0

-4.5644 > 1.6449

As the condition fails we accept Null Hypothesis

Conclusion :-

There is sufficient evidence that the mean = 10000

Upper Tail Test of Population Mean with Known Variance

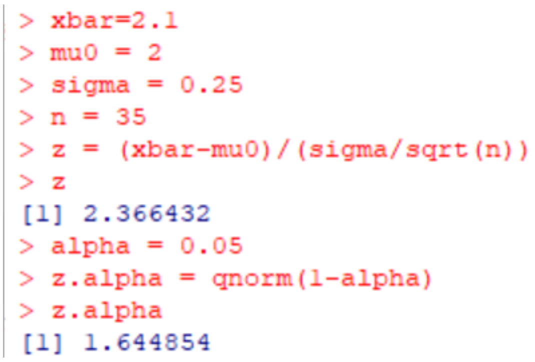
Problem :-

Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At .05 significance level, can we reject the claim on food label?

Solution :-

Null Hypothesis :- H_0 : Mean = 2.0 grams

Alternative Hypothesis :- H_1 : Mean > 2.1 grams



```
> xbar=2.1
> mu0 = 2
> sigma = 0.25
> n = 35
> z = (xbar-mu0)/(sigma/sqrt(n))
> z
[1] 2.366432
> alpha = 0.05
> z.alpha = qnorm(1-alpha)
> z.alpha
[1] 1.644854
```

One sided, alpha=0.05, Z= 2.366432 , Z.alpha=1.6449

If $Z > Z_{alpha}$ True Reject H_0

2.366432 > 1.6449

As the condition does not fail we reject Null Hypothesis

Conclusion :-

There is sufficient evidence that the mean > 2.1 grams

Two-Tailed Test of Population Mean with Known Variance

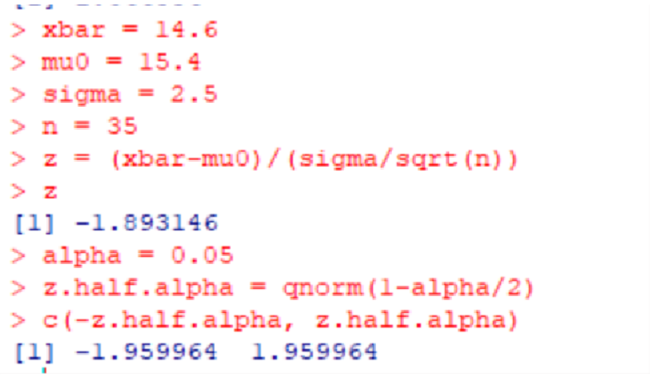
Problem :-

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At 0.05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution :-

Null Hypothesis :- $H_0 : \mu = \mu_0$

Alternative Hypothesis :- $H_1 : \mu != \mu_0$



```
> xbar = 14.6
> mu0 = 15.4
> sigma = 2.5
> n = 35
> z = (xbar-mu0)/(sigma/sqrt(n))
> z
[1] -1.893146
> alpha = 0.05
> z.half.alpha = qnorm(1-alpha/2)
> c(-z.half.alpha, z.half.alpha)
[1] -1.959964  1.959964
```

If $|Z| > Z_{alpha/2}$ is true then reject Null Hypothesis

$|1.893146| > 1.959964$

As the condition fails we accept Null Hypothesis

Conclusion :-

There is sufficient evidence that the mean penguin weight does not differ from last year.

Lower Tail Test of Population Mean with Unknown Variance

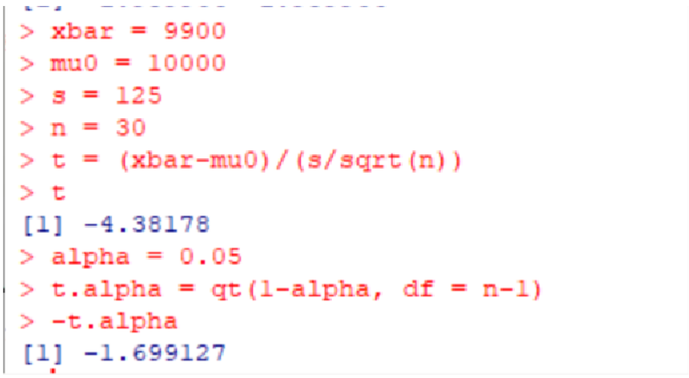
Problem :-

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a normally distributed sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the sample standard deviation is 125 hours. At 0.05 significance level, can we reject the claim by the manufacturer?

Solution :-

Null Hypothesis :- H_0 : Mean = 10000

Alternative Hypothesis :- H_1 : Mean < 10000



```
> xbar = 9900
> mu0 = 10000
> s = 125
> n = 30
> t = (xbar-mu0)/(s/sqrt(n))
> t
[1] -4.38178
> alpha = 0.05
> t.alpha = qt(1-alpha, df = n-1)
> -t.alpha
[1] -1.699127
```

If $t < t_{alpha}$ is true then reject Null Hypothesis

-4.38178 < -1.699127

As condition remains true we reject Null Hypothesis

Conclusion :-

There is sufficient evidence that mean is less than 10000.

Upper Tail Test of Population Mean with Unknown Variance

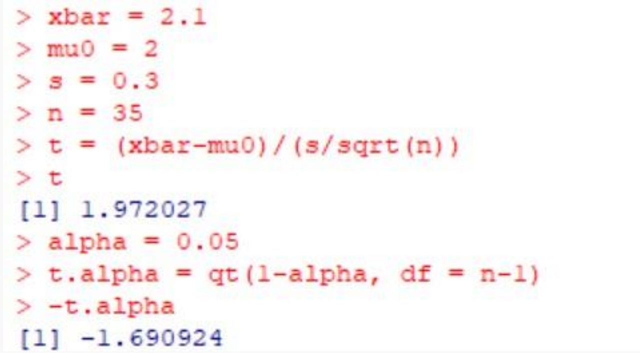
Problem :-

Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the sample standard deviation is 0.3 gram. At .05 significance level, can we reject the claim on food label?

Solution :-

Null Hypothesis :- H_0 : Mean = 2.0 grams

Alternative Hypothesis :- H_1 : Mean > 2.0 grams



```
> xbar = 2.1
> mu0 = 2
> s = 0.3
> n = 35
> t = (xbar-mu0)/(s/sqrt(n))
> t
[1] 1.972027
> alpha = 0.05
> t.alpha = qt(1-alpha, df = n-1)
> -t.alpha
[1] -1.690924
```

If t > t_{alpha} is true then reject Null Hypothesis

1.972027 > -1.690924

As condition remains true we reject Null Hypothesis

Conclusion :-

There is sufficient evidence that mean is more than 2.0 grams.

Two-Tailed Test of Population Mean with Unknown Variance

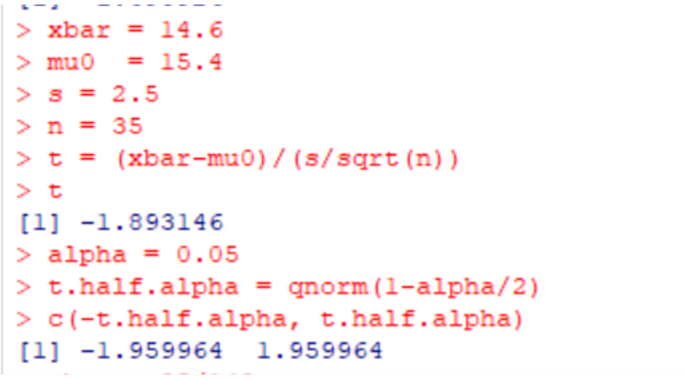
Problem :-

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the sample standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution :-

Null Hypothesis :- $H_0 : \mu = \mu_0$

Alternative Hypothesis :- $H_1 : \mu \, != \mu_0$



```
> xbar = 14.6
> mu0  = 15.4
> s = 2.5
> n = 35
> t = (xbar-mu0)/(s/sqrt(n))
> t
[1] -1.893146
> alpha = 0.05
> t.half.alpha = qnorm(1-alpha/2)
> c(-t.half.alpha, t.half.alpha)
[1] -1.959964  1.959964
```

If $|t| > t_{alpha/2}$ is true then reject Null Hypothesis

$|1.893146| > 1.959964$

As the condition fails we accept Null Hypothesis

Conclusion :-

There is sufficient evidence that the mean penguin weight does not differ from last year.

Lower Tail Test of Population Proportion

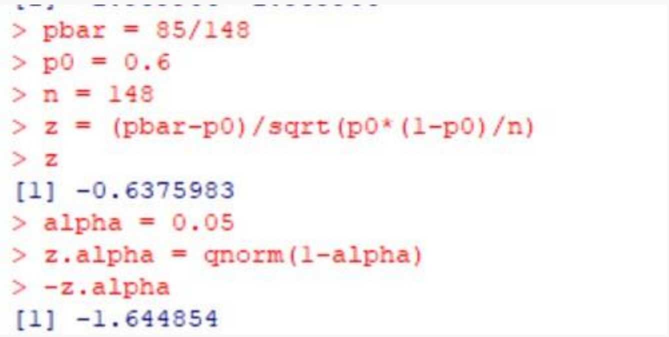
Problem :-

Suppose 60% of citizens voted in last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?

Solution :-

Null Hypothesis :- H0 : H > H0 i.e. %of citizens voted last election > %of citizens voted this election.

Alternative Hypothesis :- H1 : H = H0 i.e. %of citizens voted last election = %of citizens voted this election.



```
> pbar = 85/148
> p0 = 0.6
> n = 148
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z
[1] -0.6375983
> alpha = 0.05
> z.alpha = qnorm(1-alpha)
> -z.alpha
[1] -1.644854
```

If $Z > Z_{alpha}$ then reject Null Hypothesis

-0.6375983 > -1.644854

As the condition remains true we reject Null Hypothesis

Conclusion :-

There is sufficient evidence that we can reject the null hypothesis that the proportion of voters in the population is above 60% this year.

Upper Tail Test of Population Proportion

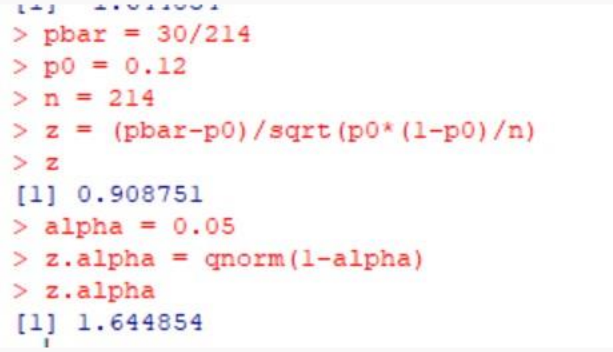
Problem :-

Suppose that 12% of apples harvested in an orchard last year was rotten. 30 out of 214 apples in a harvest sample this year turns out to be rotten. At 0.05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?

Solution :-

Null Hypothesis :- H0 : H < H0 i.e. %of apples were rotten last year > %of apples were rotten last year.

Alternative Hypothesis :- H1 : H = H0 i.e. %of apples were rotten last year = %of apples were rotten last year.



```
> pbar = 30/214
> p0 = 0.12
> n = 214
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z
[1] 0.908751
> alpha = 0.05
> z.alpha = qnorm(1-alpha)
> z.alpha
[1] 1.644854
```

If $Z < Z_{alpha}$ then reject Null Hypothesis

0.908751 < -1.644854

As the condition remains true we reject Null Hypothesis

Conclusion :-

There is sufficient evidence that we can reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year.

Date :- 01-04-2022

Practical No. 7

Aim :- Non-parametric Test - 1

Sign Test

Example :-

A soft drink company has invented a new drink, and would like to find out if it will be as popular as the existing favourite drink. For this purpose, its research department arranges 18 participants for taste testing. Each participant tries both drinks in random order before giving his or her opinion.

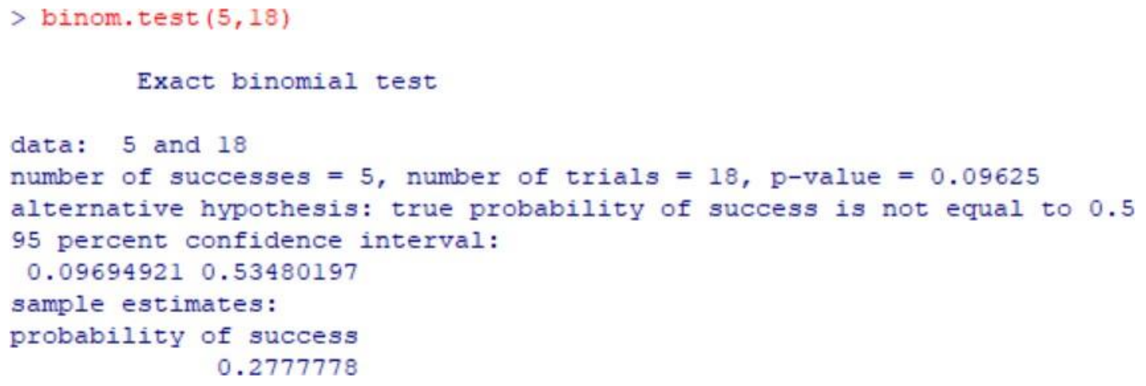
Problem :-

It turns out that 5 of the participants like the new drink better, and the rest prefer the old one. At .05 significance level, can we reject the notion that the two drinks are equally popular?

Solution :-

Null Hypothesis :- H_0 : Median difference is zero (Two drinks are equally popular)

Alternative Hypothesis :- H_1 : Median Difference is not Zero (Two drinks are not equally popular)



```
> binom.test(5,18)

        Exact binomial test

data:  5 and 18
number of successes = 5, number of trials = 18, p-value = 0.09625
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.09694921 0.53480197
sample estimates:
probability of success
             0.2777778
```

If p-value >= alpha, accept Alternative hypothesis (H_1)

0.09625 >= 0.05

As the condition fails, we accept H_0 i.e. Null Hypothesis.

Conclusion :-

There is sufficient evidence that two drinks are equally popular.

Wilcoxon Signed-Rank Test

Two data samples are matched if they come from repeated observations of the same subject. Using the Wilcoxon Signed-Rank Test, we can decide whether the corresponding data population distributions are identical without assuming them to follow the normal distribution.

Example :-

In the built-in data set named immer, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the data frame columns Y_1 and Y_2.

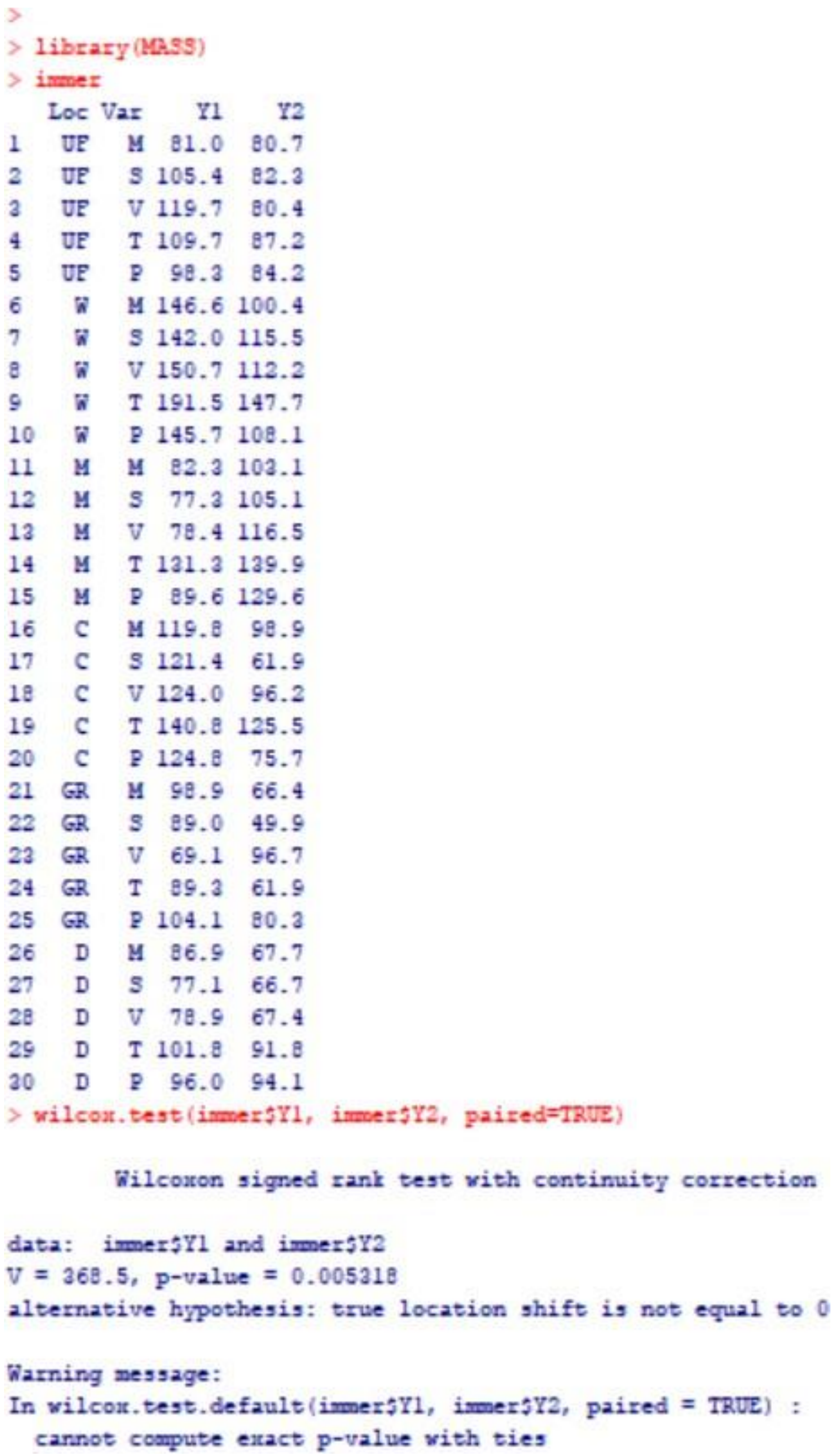
Problem :-

Without assuming the data to have normal distribution, test at 0.05 significance level if the barley yields of 1931 and 1932 in data set immer have identical data distributions.

Solution :-

Null Hypothesis :- H_0 : Median difference is equal (Y_1 and Y_2 are identical)

Alternative Hypothesis :- H_1 : Median difference is not equal (Y_1 and Y_2 are not identical)



```
>
> library(MASS)
> immer
   Loc Var    Y1    Y2
1   UF   M  81.0  80.7
2   UF   S 105.4  82.3
3   UF   V 119.7  80.4
4   UF   T 109.7  87.2
5   UF   P  98.3  84.2
6    W   M 146.6 100.4
7    W   S 142.0 115.5
8    W   V 150.7 112.2
9    W   T 191.5 147.7
10   W   P 145.7 108.1
11   M   M  82.3 103.1
12   M   S  77.3 105.1
13   M   V  78.4 116.5
14   M   T 131.3 139.9
15   M   P  89.6 129.6
16   C   M 119.8  98.9
17   C   S 121.4  61.9
18   C   V 124.0  96.2
19   C   T 140.8 125.5
20   C   P 124.8  75.7
21  GR   M  98.9  66.4
22  GR   S  89.0  49.9
23  GR   V  69.1  96.7
24  GR   T  89.3  61.9
25  GR   P 104.1  80.3
26   D   M  86.9  67.7
27   D   S  77.1  66.7
28   D   V  78.9  67.4
29   D   T 101.8  91.8
30   D   P  96.0  94.1
> wilcox.test(immer$Y1, immer$Y2, paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  immer$Y1 and immer$Y2
V = 368.5, p-value = 0.005318
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(immer$Y1, immer$Y2, paired = TRUE) :
  cannot compute exact p-value with ties
```

If p-value <= alpha, accept Alternative hypothesis (H_1)

0.005318 <= 0.05

As the condition is satisfied, we reject H_0 i.e. Null Hypothesis.

There is a median difference

Conclusion :-

There is sufficient evidence that Y_1 and Y_2 are not identical i.e Rejecting H_0

Date :- 01-04-2022

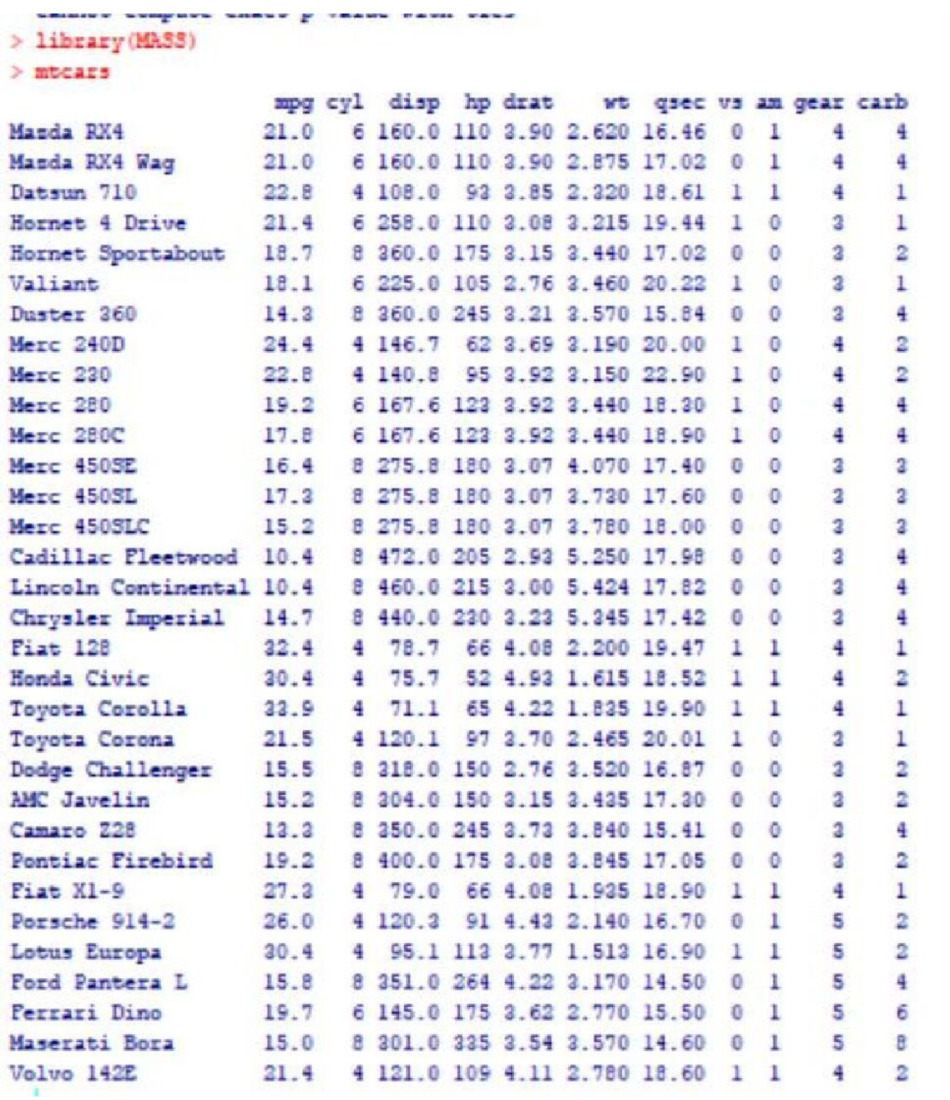
Practical No. 8

Aim :- Non-parametric Test - 2

Mann-Whitney-Wilcoxon Test

Two data samples are independent if they come from distinct populations and the samples do not affect each other. Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

Example :-



```
> library(MASS)
> mtcars
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

In the data frame column mpg of the data set mtcars, there are gas mileage data of various 1974 U.S. automobiles.

Meanwhile, another data column in mtcars, named am, indicates the transmission type of the automobile model (0 = automatic, 1 = manual). In other words, it is the differentiating factor of the transmission type.

In particular, the gas mileage data for manual and automatic transmissions are independent.

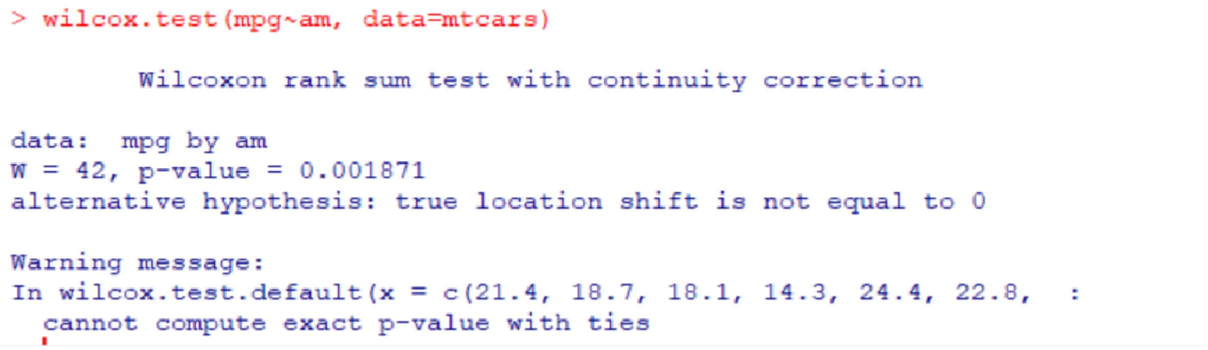
Problem :-

Without assuming the data to have normal distribution, decide at 0.05 significance level if the gas mileage data of manual and automatic transmissions in mtcars have identical data distribution.

Solution :-

Null Hypothesis :- H_0 : Median difference is zero (Two datas are identical)

Alternative Hypothesis :- H_1 : Median Difference is not Zero (Two datas are not identical)



```
> wilcox.test(mpg~am, data=mtcars)

        Wilcoxon rank sum test with continuity correction

data:  mpg by am
W = 42, p-value = 0.001871
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(21.4, 18.7, 18.1, 14.3, 24.4, 22.8,  :
  cannot compute exact p-value with ties
```

If p-value <= alpha, accept Alternative hypothesis (H_1)

0.001871 <= 0.05

As the condition is satisfied, we reject H_0 i.e. Null Hypothesis.

There is a median difference

Conclusion :-

There is sufficient evidence that the gas mileage data of manual and automatic transmissions in mtcar are non-identical populations.

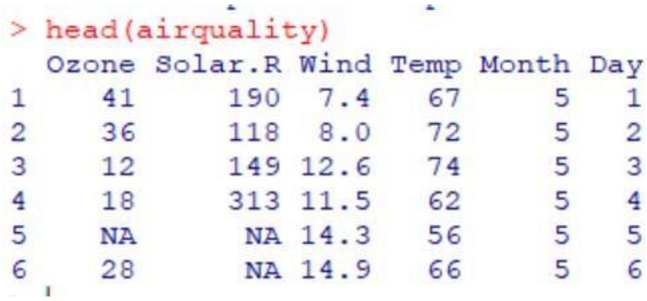
Kruskal-Wallis Test

A collection of data samples are independent if they come from unrelated populations and the samples do not affect each other. Using the Kruskal-WallisTest, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

Example :-

In the built-in data set named air quality, the daily air quality measurements in New York, May to September 1973, are recorded. The ozone density are presented in the data frame column Ozone.

Solution :-



```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

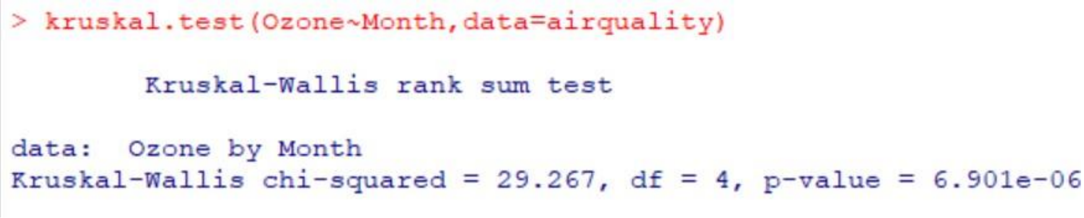
Problem :-

Without assuming the data to have normal distribution, test at 0.05 significance level if the monthly ozone density in New York has identical data distributions from May to September 1973.

Solution :-

Null Hypothesis :- H_0 : Median difference is zero (Two datas are identical)

Alternative Hypothesis :- H_1 : Median difference is zero (Two datas are not identical)



```
> kruskal.test(Ozone~Month,data=airquality)

        Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.267, df = 4, p-value = 6.901e-06
```

If p-value <= alpha, accept Alternative hypothesis (H_1)

$6.901e^{-6}$ <= 0.05

As the condition is satisfied, we reject H_0 i.e. Null Hypothesis.

There is a median difference

Conclusion :-

There is sufficient evidence that the monthly ozone density in New York from May to September 1973 are non-identical populations.

Date :- 01-04-2022

Practical No. 9

Aim :- Chi-squared Test of Independence

Chi-squared Test of Independence

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another. Assume f_{ij}, is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume e_{ij} to be the corresponding expected count if x and y are independent.

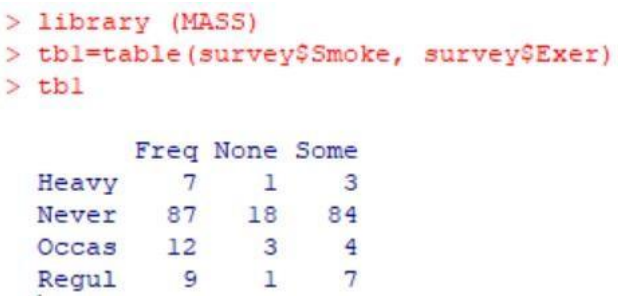
The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

$$\chi^2 = \sum_{i,\ j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

Example :-

In the built-in data set survey, the Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

We can tally the students smoking habit against the exercise level with the table function in R. The result is called the contingency table of the two variables.



```
> library (MASS)
> tbl=table(survey$Smoke, survey$Exer)
> tbl

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
```

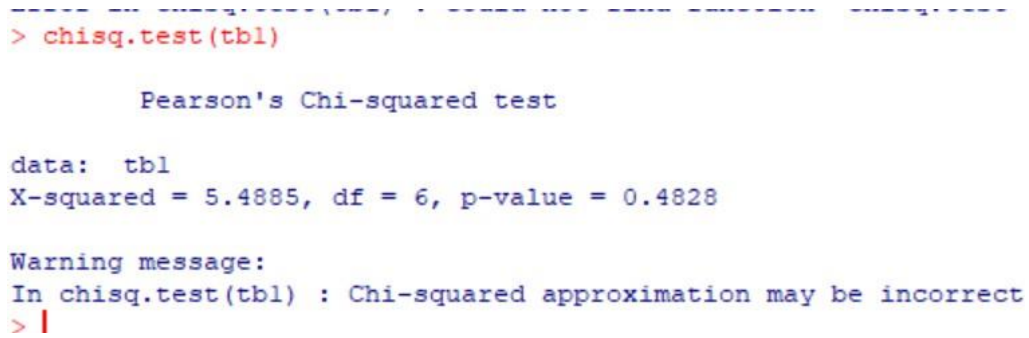
Problem :-

Test the hypothesis whether the students smoking habit is independent of their exercise level at 0.05 significance level.

Solution :-

Null Hypothesis :- H_0 : There is no association between student's smoking habit and their exercise level.

Alternative Hypothesis :- H_1 : There is association between student's smoking habit and their exercise level.



```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
> 
```

If p-value <= alpha then reject Null hypothesis

0.4828 <= 0.05

As this condition fails we accept Null hypothesis

Conclusion :-

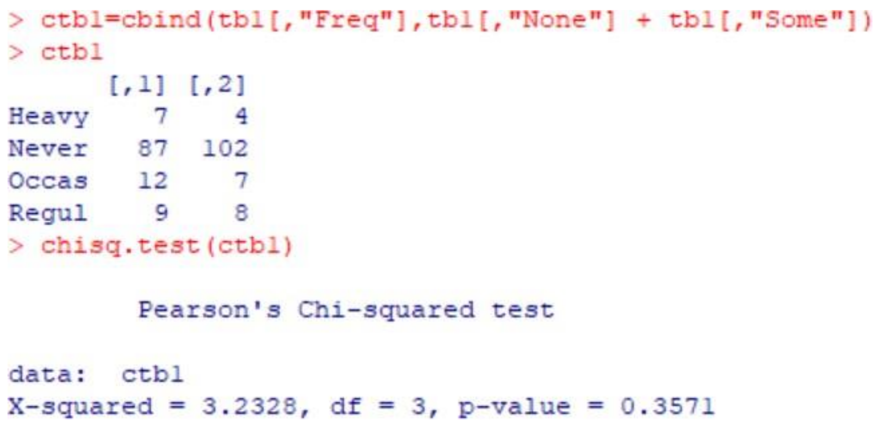
There is sufficient evidence that there is no association between smoking habit of students and their exercise level

Enhanced Solution

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl, and save it in anew table named ctbl. Then we apply the chisq.test function against ctbl instead.

Null Hypothesis :- H_0 : There is no association between student's smoking habit and their exercise level.

Alternative Hypothesis :- H_1 : There is association between student's smoking habit and their exercise level.



```
> ctbl=cbind(tbl[,"Freq"],tbl[,"None"] + tbl[,"Some"])
> ctbl
      [,1] [,2]
Heavy    7    4
Never   87  102
Occas   12    7
Regul    9    8
> chisq.test(ctbl)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571
```

If p-value <= alpha then reject Null hypothesis

0.3571 <= 0.05

As this condition fails we accept Null hypothesis

Conclusion :-

There is sufficient evidence that there is no association between smoking habit of students and their exercise level