Name:Mohnish Kalaimani
Roll No:FCS2122061
FYBsc Computer Science
Statistical Methods and Testing of Hypothesis Journal

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. **Mohnish Kalaimani** Roll No.**FCS2122061** Has successfully completed the necessary course of experiments in the subject of **Statistical Method and Testing of Hypothesis**during the academic year 2021 – 2022complying with the requirements of University of Mumbai, for the course of **F.Y.BSc. Computer Science [Semester-2]**

Prof. In-Charge
Mrs.Soni Yadav
(Statistical Method and Testing of Hypothesis)

Examination Date:
Examiner's Signature & Date:

Head of the Department:
Prof.Manoj Singh

College Seal
And

Practical No	Aim
1	Problem based on binomial distribution
2	Problem based on normal distribution
3	Property plotting of binomial distribution
4	Property plotting of normal distribution
5	Problem based on pdf,cdf,pmf, for discrete and continuous distribution
6	Z test, t test
7	Non-Parametric tests-I (Sign Test,Wilcoxon Test)
8	Non-Parametric tests-II (Kruskal Wallis Test,Mann Whitney U Test)
9	Chi Square Test of independence

Practical No 1

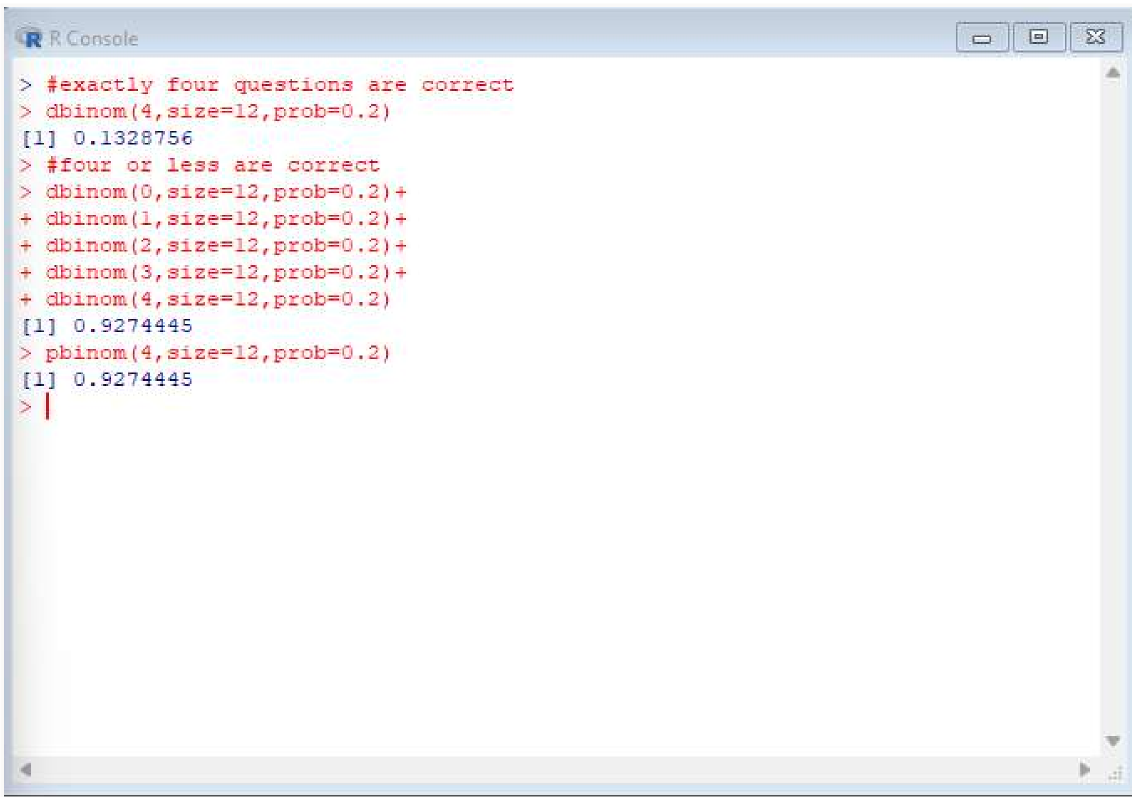
Binomial Distribution

The binomial distribution is a discrete probability distribution. It describes the outcome of n independent trials in an experiment. Each trial is assumed to have only two outcomes,either success or failure. 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.

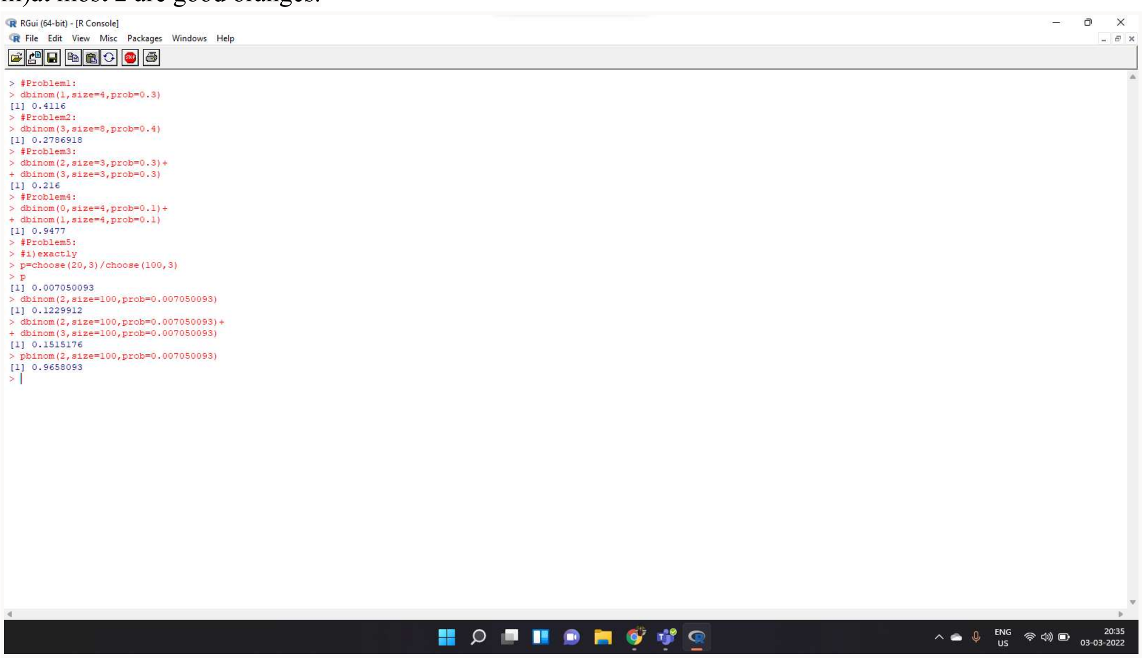
Problem

Suppose there are twelve multiple-choice questions in an English class quiz. Each question
has five possible answers, and only one of them is correct. Find the probability of having four or
fewer correct answers if a student attempts to answer every question randomly.

Solution:



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.
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?
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.
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?
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.



Practical No 2

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

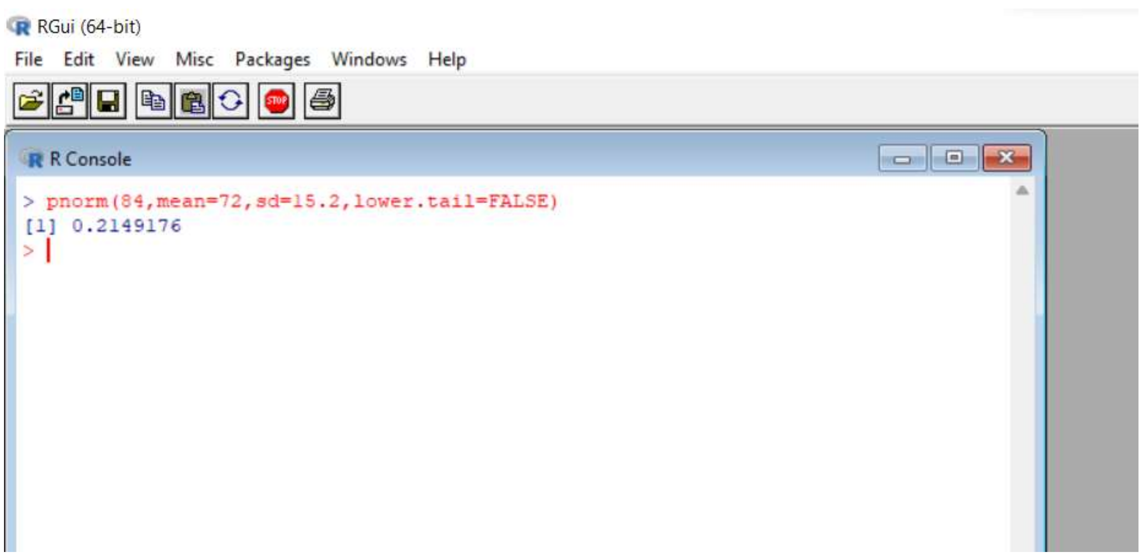
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

In particular, the normal distribution with $\mu = 0$ and $\sigma = 1$ is called the standard normal distribution, and is denoted asN(0,1). It can be graphed as follows.

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

Problem

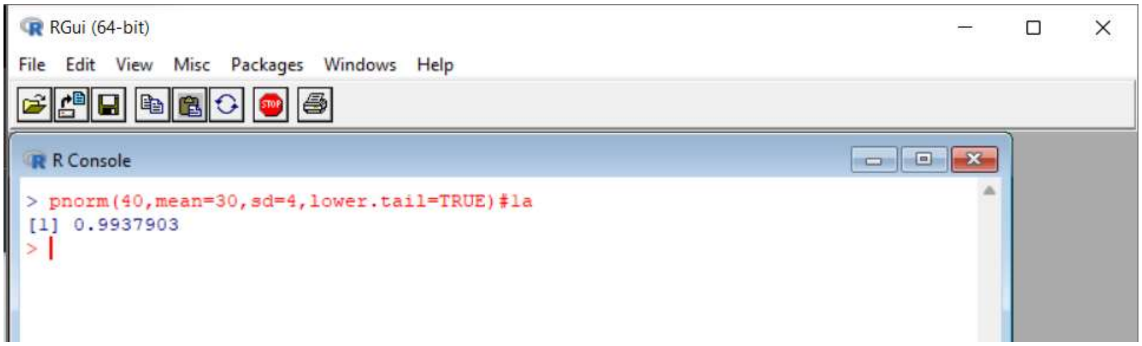
Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72,and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?



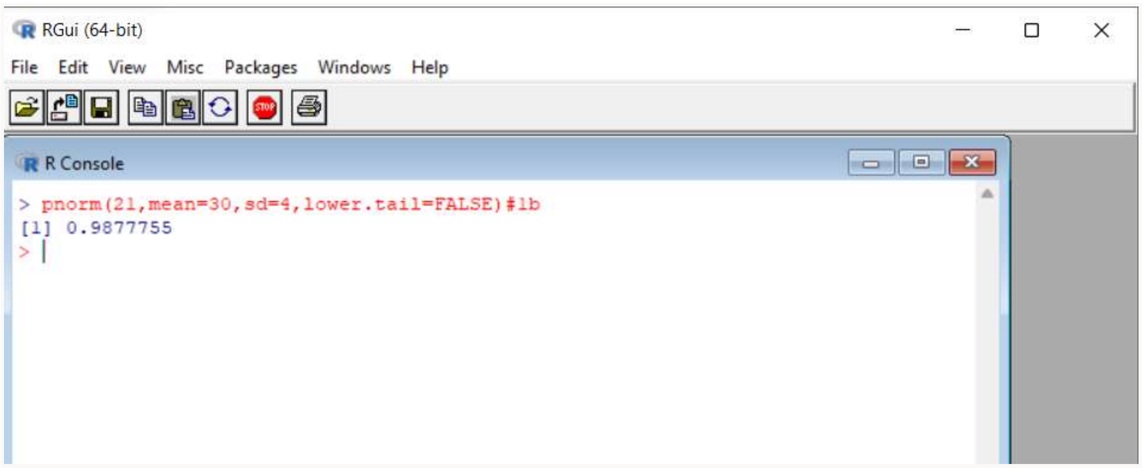
Answer:The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

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

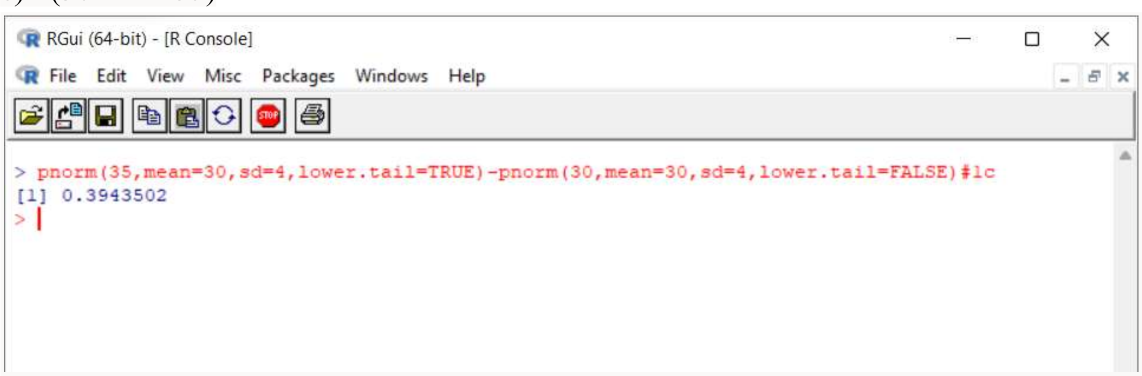
a) P(x < 40)



b) P(x > 21):

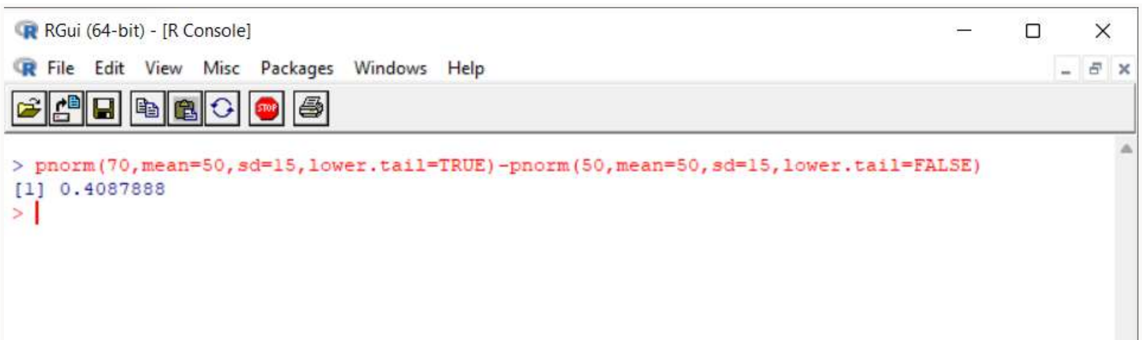


c) P(30 < x < 35)

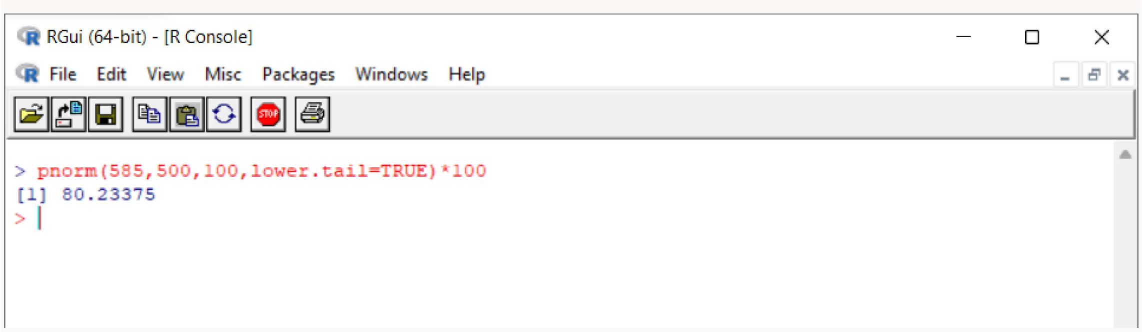


2.A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

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.



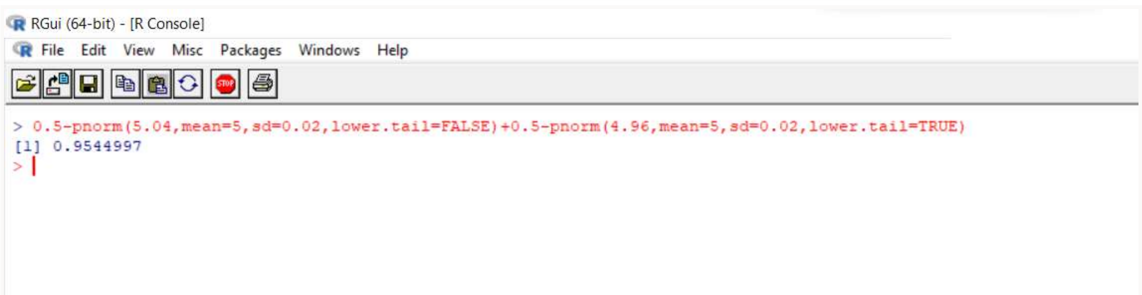
4.Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. Tom wants to be admitted to this university and he knows that he must score better than at least 70% of the students who took the test.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



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?

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



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.

b) between 7 and 12 months.

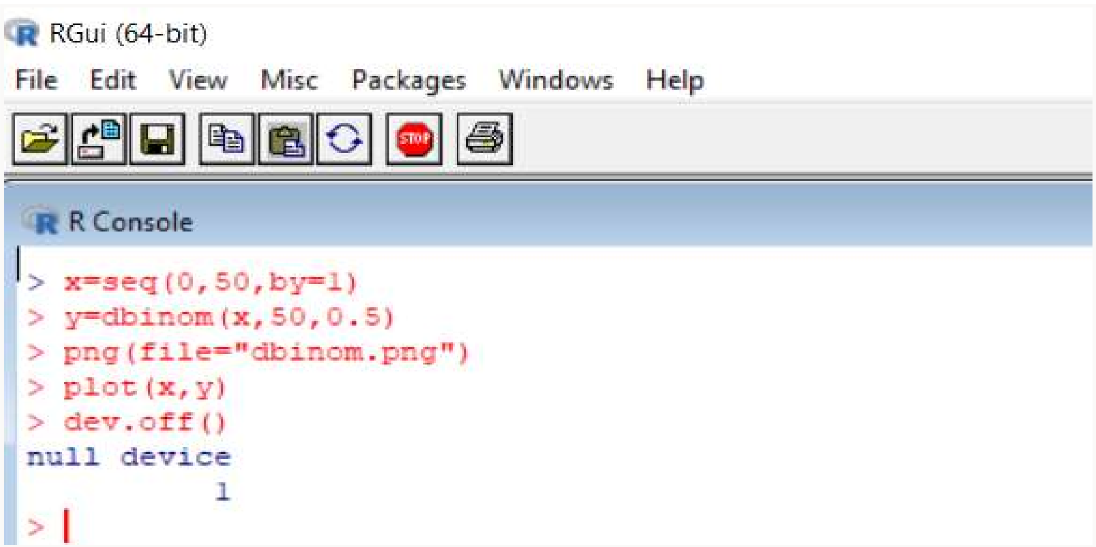
Practical No 3

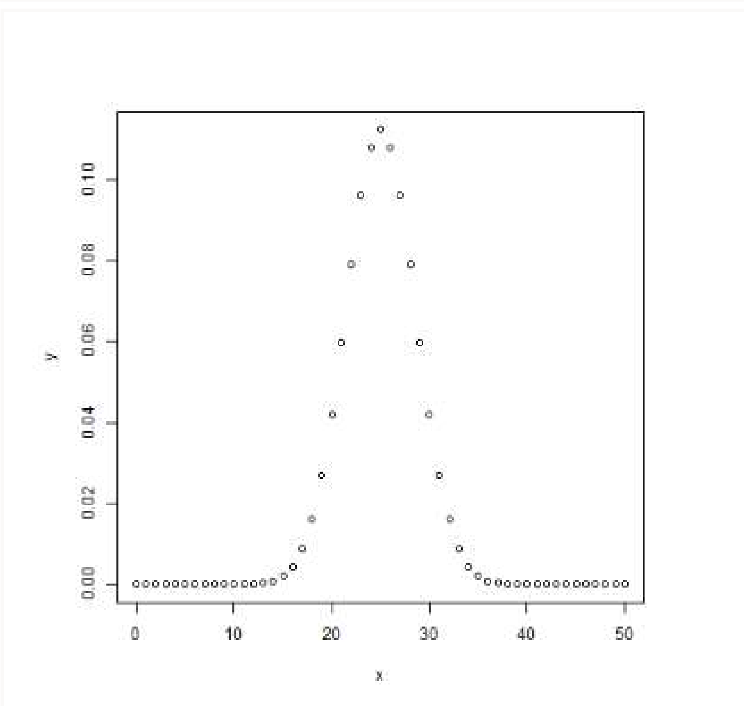
Aim:Property plotting of Binomial Distribution

dbinom(x, size, prob)
pbinom(x, size, prob)
qbinom(p, size, prob)
rbinom(n, size, prob)

<u>dbinom()-</u>

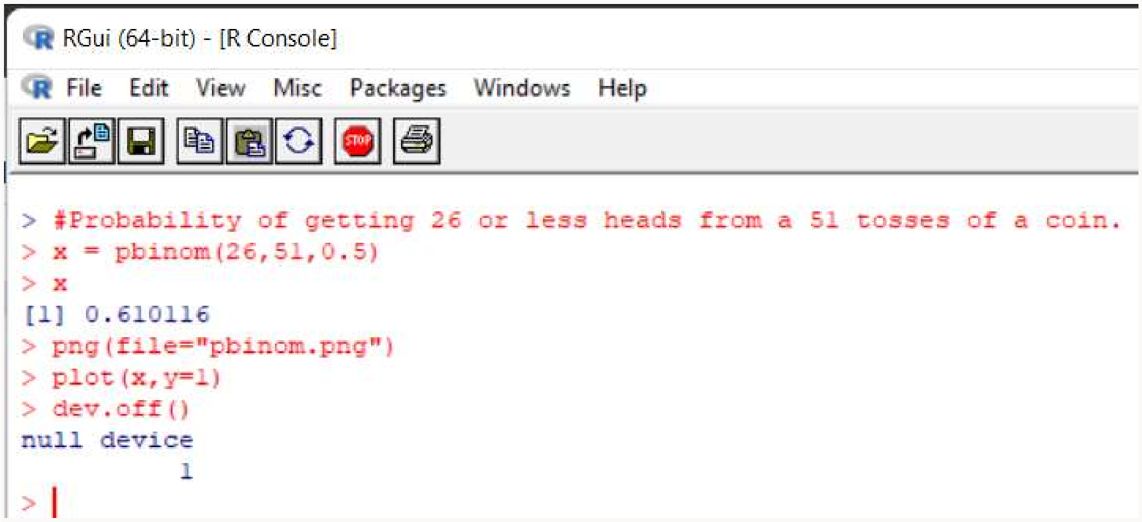
This function gives the probability density distribution at each point.



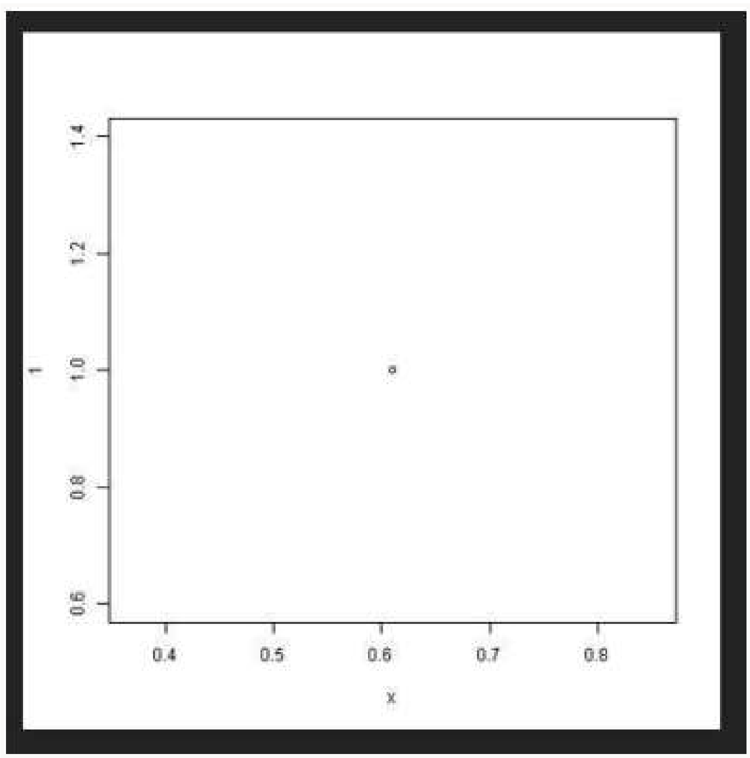


pbinom()

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

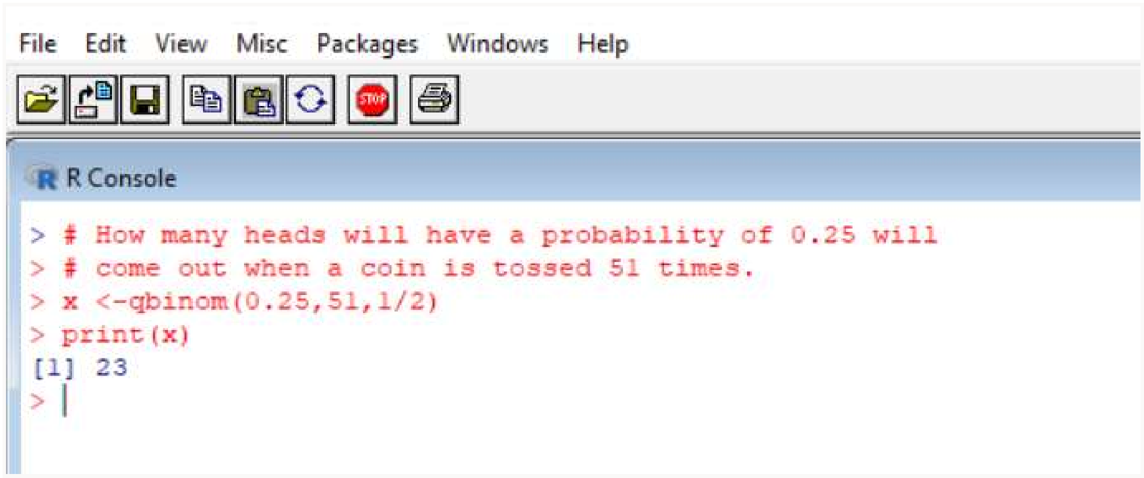


```
RGui (64-bit) - [R Console]
File  Edit  View  Misc  Packages  Windows  Help

> #Probability of getting 26 or less heads from a 51 tosses of a coin.
> x = pbinom(26,51,0.5)
> x
[1] 0.610116
> png(file="pbinom.png")
> plot(x,y=1)
> dev.off()
null device
          1
>
```



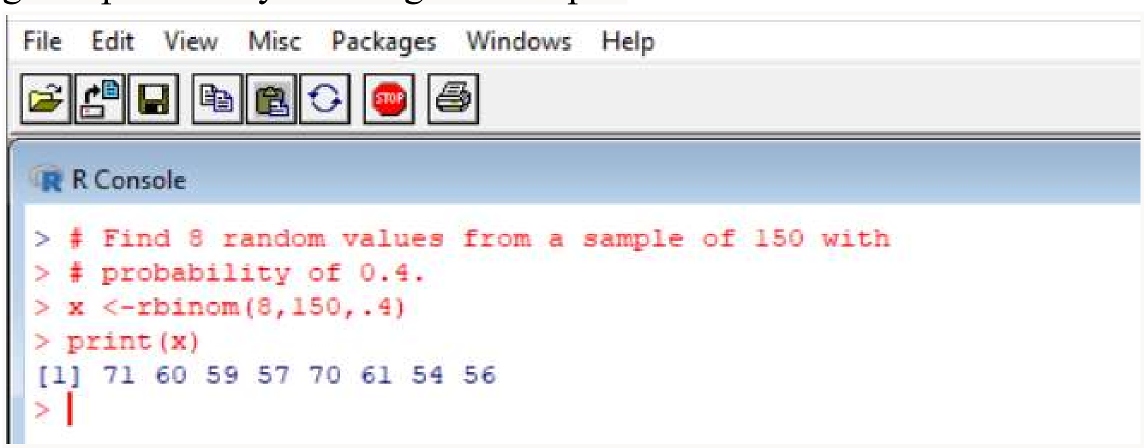
qbinom()

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



rbinom()

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

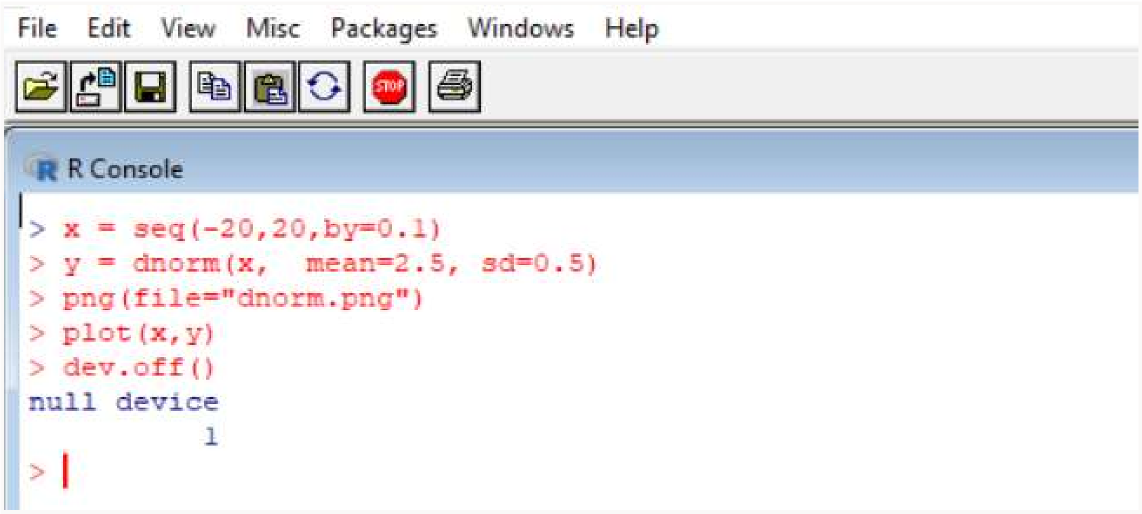


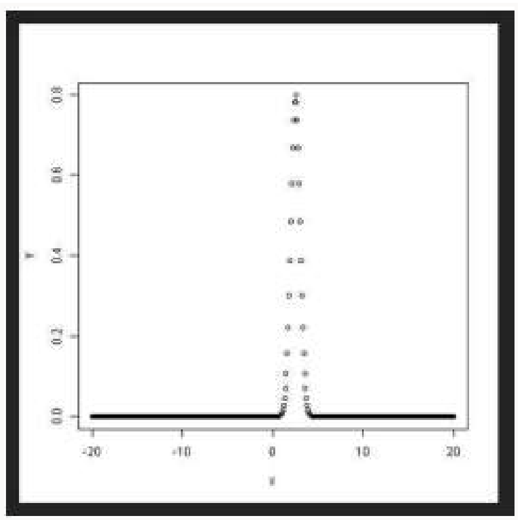
Practical No 4

Aim: Property plotting of Normal Distribution.

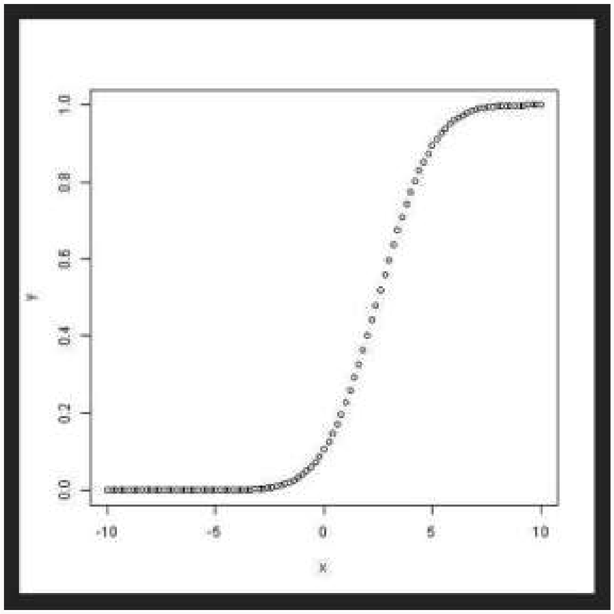
```
dnorm(x, mean, sd)
pnorm(x, mean, sd)
qnorm(p, mean, sd)
rnorm(n, mean, sd)
```

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

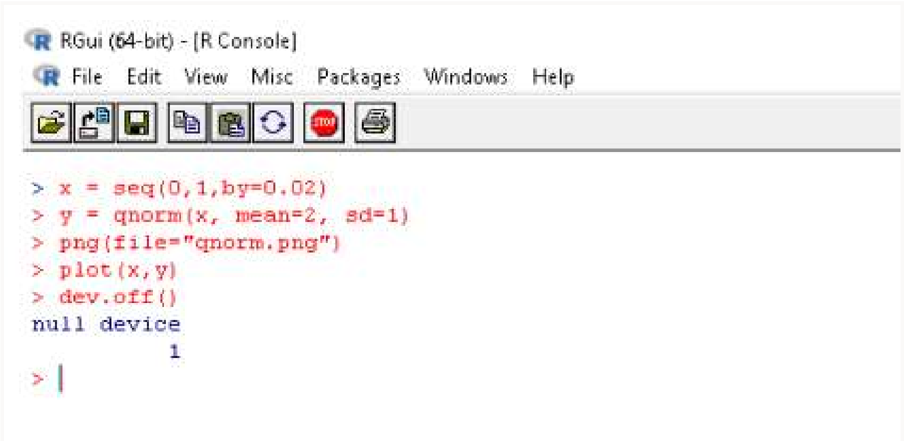


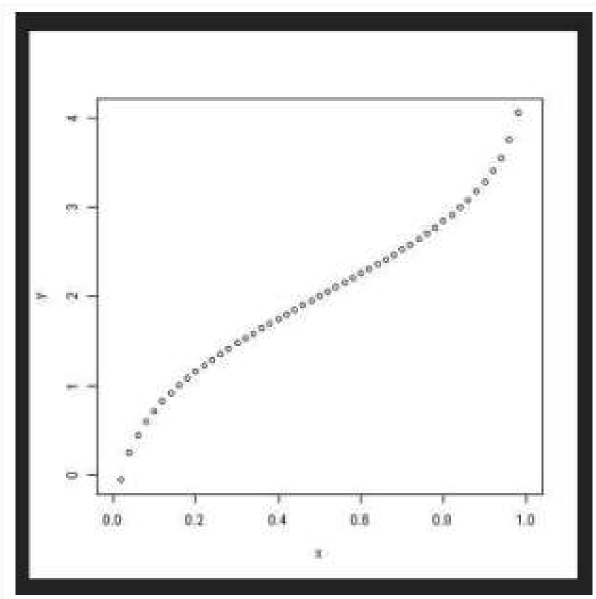


pnorm()- This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".

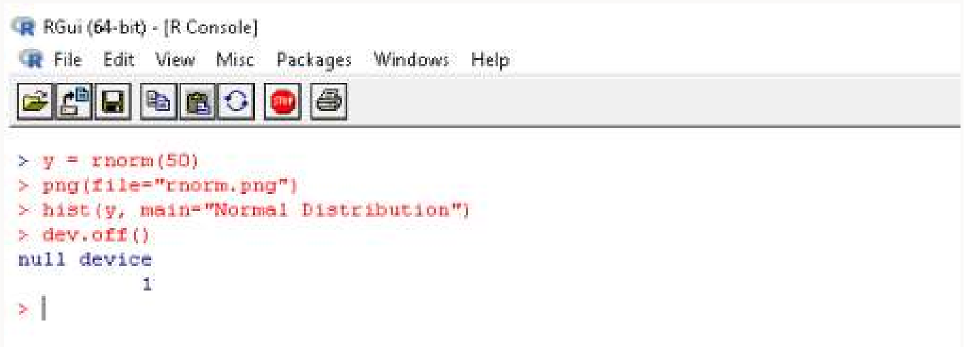


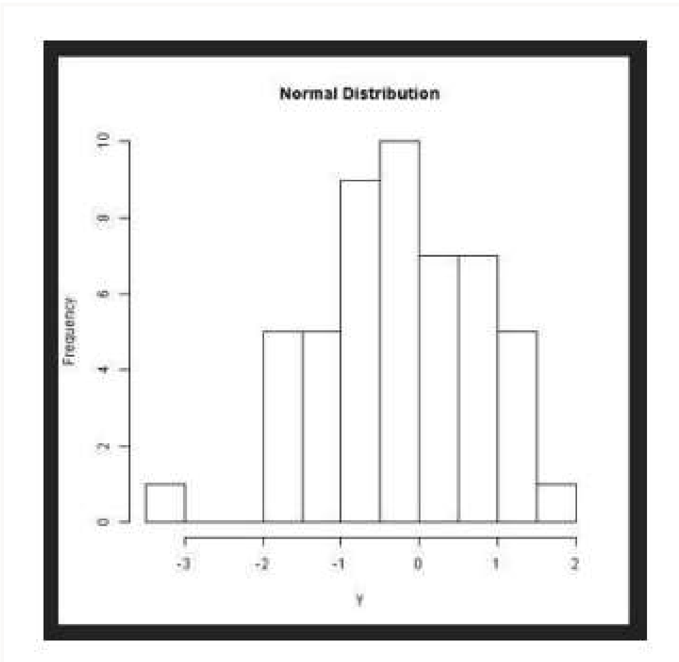
qnorm()- This function takes the probability value and gives a number whose cumulative value matches the probability value.





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.





Practical No 5

Q1)

Practical No. 5

Q1) 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) $p(2 \leq x \leq 6)$
e) $p(x=4/x \geq 2)$

Solⁿ: a)

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

b) $\therefore$ Mean $= \sum x_i p_i$

$= 1(0.09) + 2(0.14) + 3(0.12) + 4(0.14) + 5(0.22) + 6(0.18)$
$+ 7(0.11)$

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

c) $\therefore$ Variance $= E(x^2) - (E(x))^2$

where $E(x^2) = \sum x^2 p(x)$

$= 1^2(0.09) + 2^2(0.14) + 3^2(0.12) + 4^2(0.14) + 5^2(0.22) + 6^2(0.18) + 7^2(0.11)$

$\boxed{\therefore V(x) = 21.94}$

$\therefore V(x) = 21.34 - (4.24)^2 = 3.3624$

Standard Deviation $= \sqrt{V(x)} = \sqrt{3.3624} = 1.8337$

$\boxed{\therefore \sigma^2 = 1.8337}$

d) $p(2 \leq x \leq 6) = p(2) + p(3) + p(4) + p(5) + p(6)$
$= 0.14 + 0.12 + 0.14 + 0.22 + 0.18$

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

e) $p(x=4/x \geq 2)$

Comparing with $p(A/B)$

$A = x = 4$

$B = x \geq 2$

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

$p(B) = (x \geq 2) = p(2) + p(3) + p(4) + p(5) + p(6) + p(7)$
$= 0.14 + 0.12 + 0.14 + 0.22 + 0.18 + 0.11$
$= 0.91$

$\therefore p(x=4/x \geq 2) = \frac{p(A \cap B)}{p(B)} = \frac{x=4}{x \geq 2} = \frac{0.14}{0.91} = 0.1538$

$\boxed{p(x=4/x \geq 2) = 0.1538}$

Q2)

Q② 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, $p(x<4)$

Solⁿ: $f(x)$ is a pdf

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

$= k\int_0^1 x(1-x)\,dx$

$= k\int_0^1 x - x^2\,dx$

$= k\left[\int_0^1 x\,dx - \int_0^1 x^2\,dx\right]$

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

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

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

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

$= k\left[\frac{1}{6}\right]$

$= \frac{k}{6}$

RHS = 1

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

$\boxed{\therefore k = 6}$

2. $f(x) = 6x(1-x)$

$= 6x - 6x^2$

3. $p(x<4)$

$\int_0^4 6(x-x^2)\,dx$

$= 6\int_0^4 x - x^2\,dx$

$= 6\left[\int_0^4 x\,dx - \int_0^4 x^2\,dx\right]$

$= 6\left[\left[\frac{x^2}{2}\right]_0^4 - \left[\frac{x^3}{3}\right]_0^4\right]$

$= 6\left[\frac{4\times4-0}{2} - \left[\frac{4\times4\times4-0}{3}\right]\right]$

$= 6\left[8 - \frac{64}{3}\right] = 6\left[\frac{24-64}{3}\right] = 6\left[\frac{-40}{3}\right] = -80$

$\boxed{\therefore p(x<4) = -80}$

Q3)

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

Sol^n:

Total number (N) possibilities of drawing 3 balls out of 9 balls (6 green and 3 red)

$C(9,3) = {}^9C_3 = \frac{9!}{6! \times 3!} = \frac{9 \times 8 \times 7 \times 6!}{3 \times 2 \times 1 \times 6!} = 84$

X → The ball is a red ball

$p(X=0) = \frac{{}^3C_0 \, {}^6C_3}{{}^9C_3} = \frac{\frac{3!}{0! \times 3!} \times \frac{6!}{3! \times 3!}}{84} = \frac{\frac{6 \times 5 \times 4 \times 3!}{3 \times 2 \times 1 \times 3!}}{84} = 0.2381$

$p(X=1) = \frac{{}^3C_1 \, {}^6C_2}{{}^9C_3} = \frac{\frac{3!}{1! \times 2!} \times \frac{6!}{2! \times 4!}}{84} = \frac{\frac{3 \times 2!}{1 \times 2!} \times \frac{6 \times 5 \times 4!}{2 \times 1 \times 4!}}{84}$

$= \frac{\frac{3 \times 6 \times 5}{2 \times 1 \times 1}}{84} = 0.5357$

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

$= 0.2143$

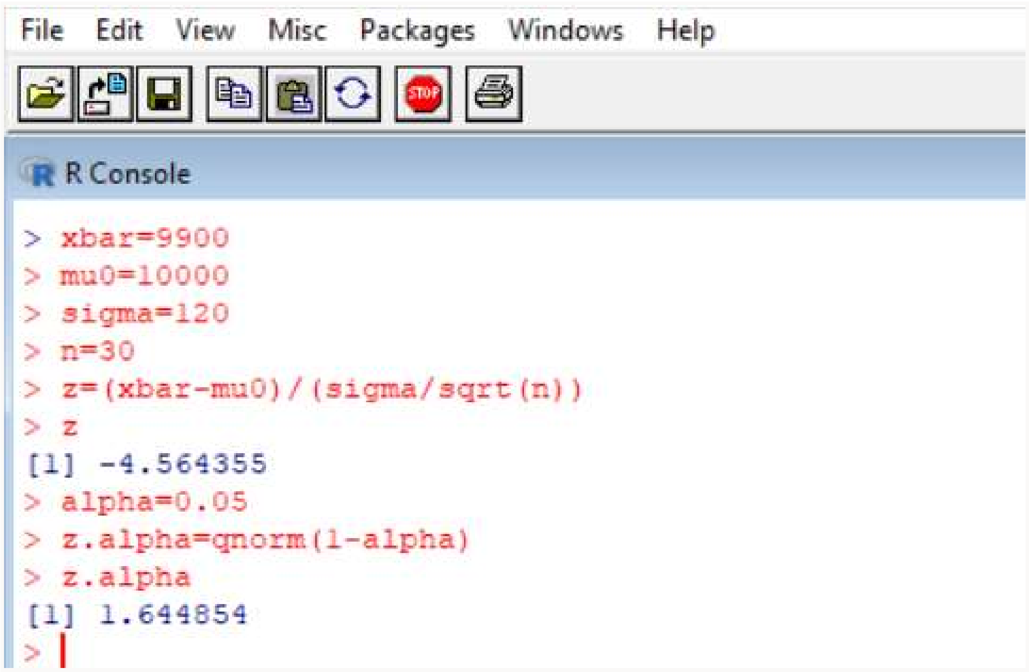
$p(X=3) = \frac{{}^3C_3 \, {}^6C_0}{{}^9C_3} = \frac{\frac{3!}{3! \times 0!} \times \frac{6!}{0! \times 6!}}{84} = \frac{1}{84} = 0.0119$

Practical No 6

Lower Tail Test of Population Mean with KnownVariance

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:



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

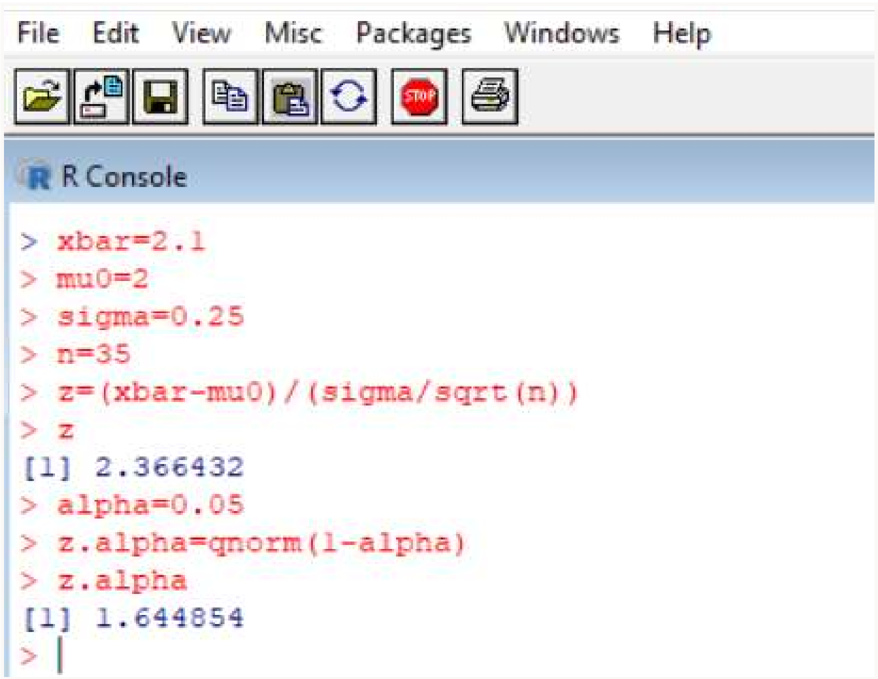
If Z > Zalpha True Reject H0-4.5644 > 1.6499 False Accept H0

Conclusion:There is evidence that 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:

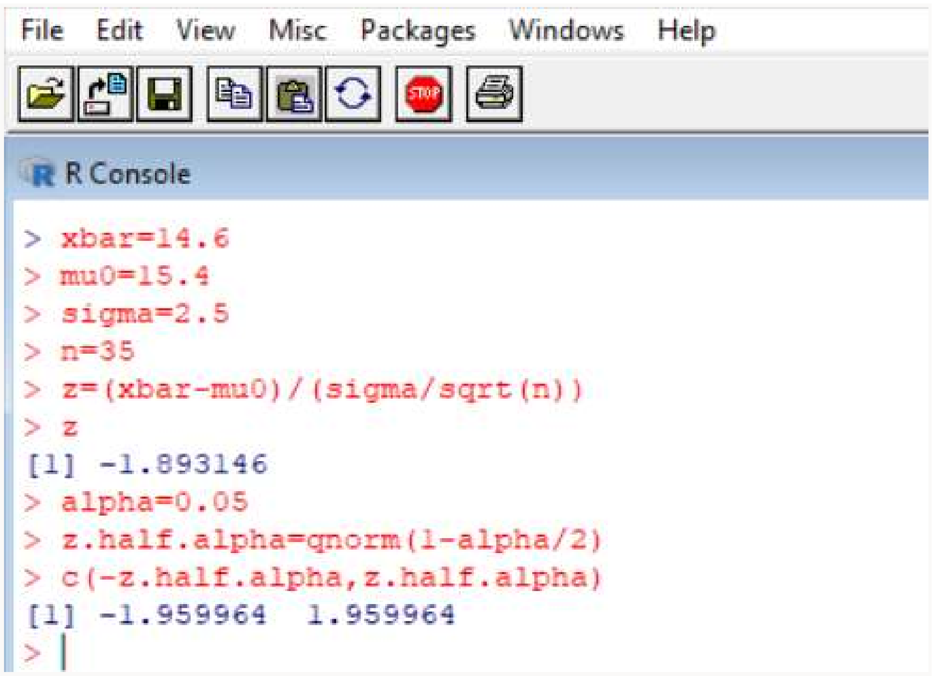


One sided, alpha=0.05, Z=2.3664, Z.alpha=1.6449

If Z > -Zalpha True Reject H0 2.3664 > -1.6499 False Accept H0

Conclusion:There is evidence that mean=2

Two-Tailed Test of Population Mean with KnownVariance
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 .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?
Solution:

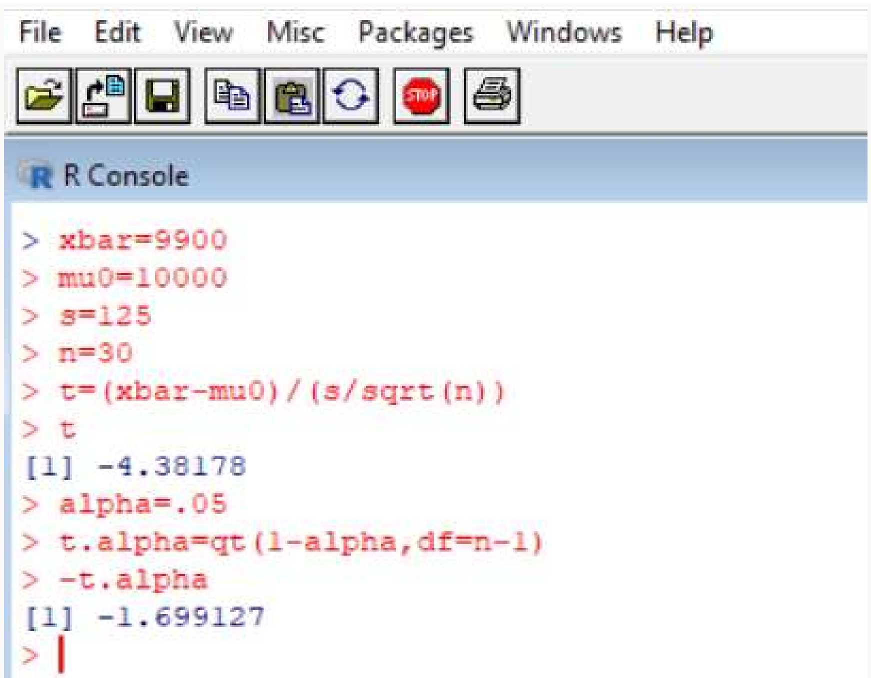


Two sided, alpha=0.05, Z=-1.8931, Z.alpha=1.9599
If |Z| > Zalpha/2 True Reject H0 1.8931 > 1.9599 False Accept H0
Conclusion:There is evidence that mean=15.4

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 anormally 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 .05 significance level, can we reject the claim by the manufacturer?

Solution:



One sided, alpha=0.05, t= -4.381 , t.alpha=1.6991

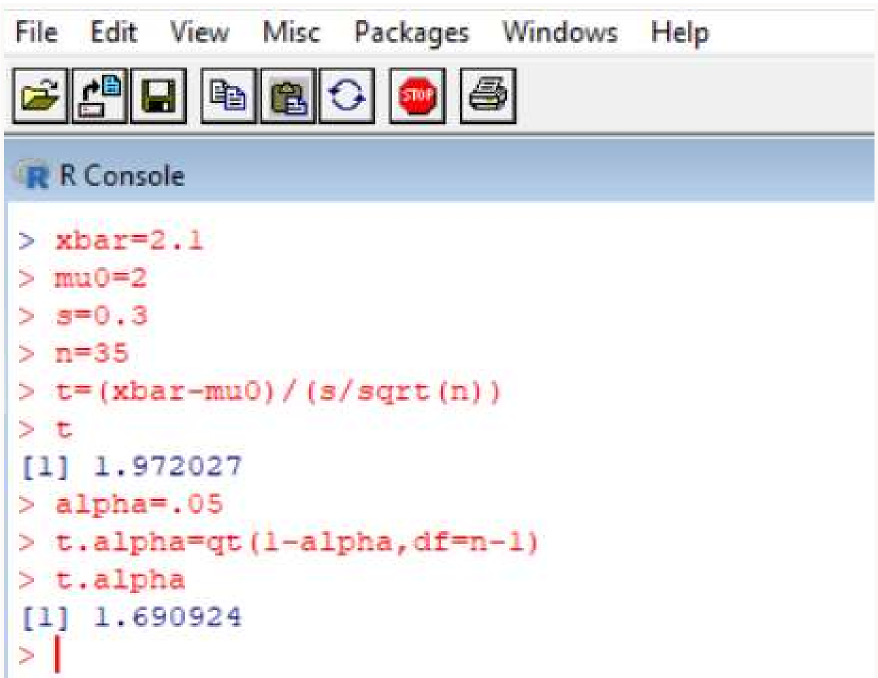
If t > talpha True Reject H0 -4.381 > 1.6991 False Accept H0

Conclusion:There is evidence that mean =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 deviationis 0.3 gram. At .05 significance level, can we reject the claim on food label?

Solution:



One sided, alpha=0.05, t= 2.366 , t.alpha=-1.6909

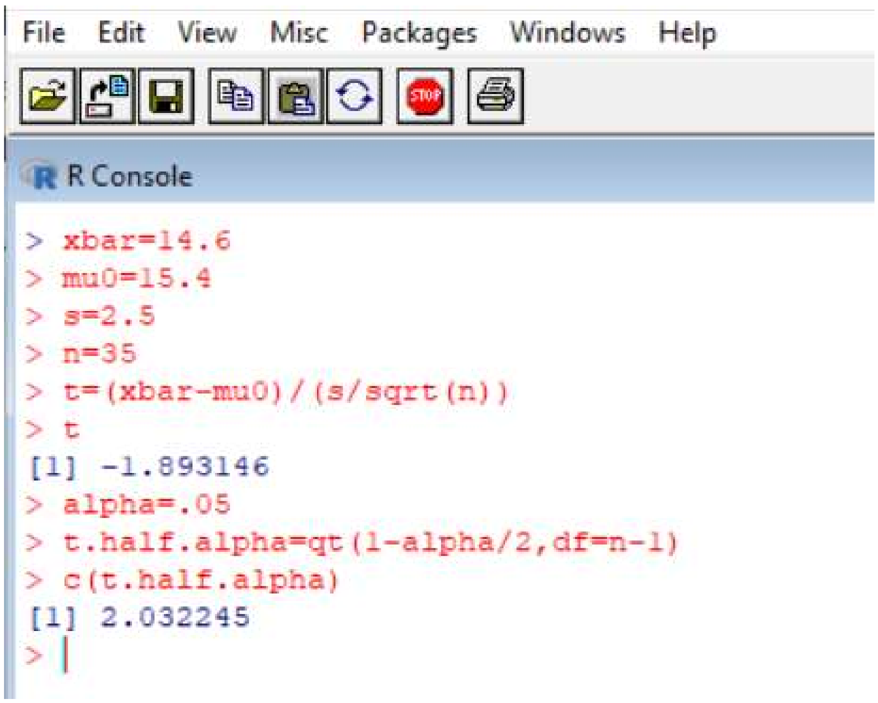
If t > talpha,n-1 True Reject H0 1.972 < -1.6909 False Accept H0

Conclusion:There is evidence that mean =2

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:



Two sided, alpha=0.05, t= -1.893 , t.alpha/2=2.0322

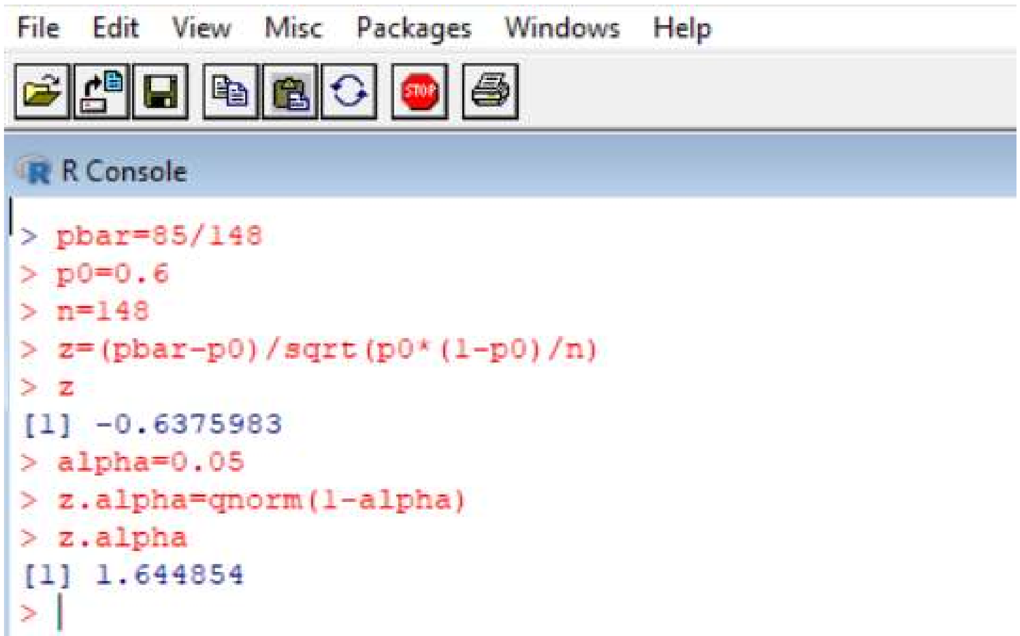
If |t|> t(alpha/2),n-1 True Reject H0 1.893 > 2.0322 False Accept H0

Conclusion:There is evidence that mean =15.4kg

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:

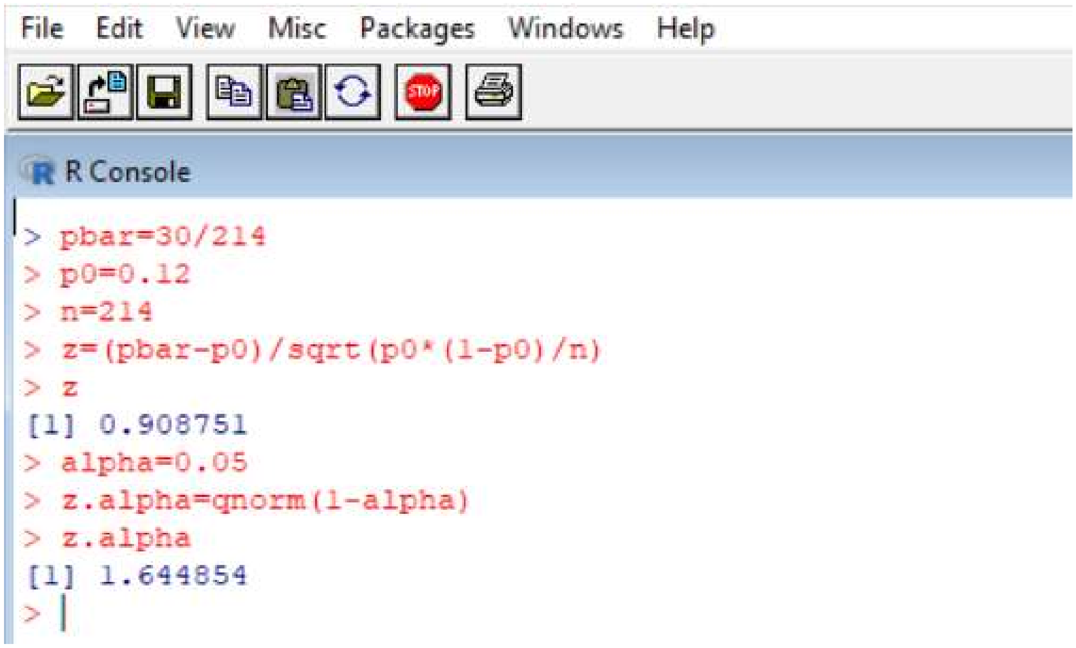


One sided, alpha=0.05, z= -0.6375 , z.alpha=1.6448

If Z> Zalpha True Reject H0 -0.6375 > 1.6448 False Accept H0

Conclusion:There is evidence that π=0.6

Upper Tail Test of Population Proportion ProblemSuppose 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 .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?
Solution:

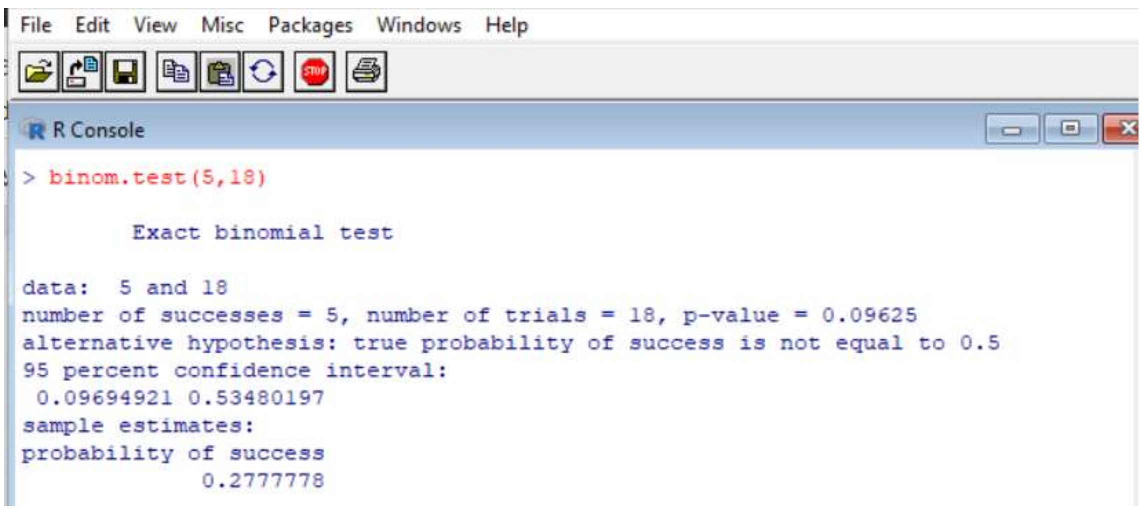


One sided, alpha=0.05, z= 0.9087 , z.alpha=1.6448
If Z> Zalpha True Reject H0 0.9087 > 1.6448 False Accept H0
Conclusion:There is evidence that π=0.12

PRACTICAL NO.7

Sign Test ExampleA: A soft drink company has invented a new drink, and would like to find out if it will be as popular as the existing favoritedrink. 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?



Conclusion: 0.09625 <= 0.05 False accept Ho

At .05 significance level, we do not reject the notion that the 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 theWilcoxon 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 namedimmer, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in thedata frame columnsY1andY2.>library(MASS)#loadtheMASSpackage>head(immer) LocVarY1Y21UFM81.080.72UFS105.482.3.....

Problem: Without assuming the data to have normal distribution, test at .05 significance level if the barley yields of 1931 and 1932 in data setimmerhave identical data distributions.

Solution:

File Edit View Misc Packages Windows Help

R Console

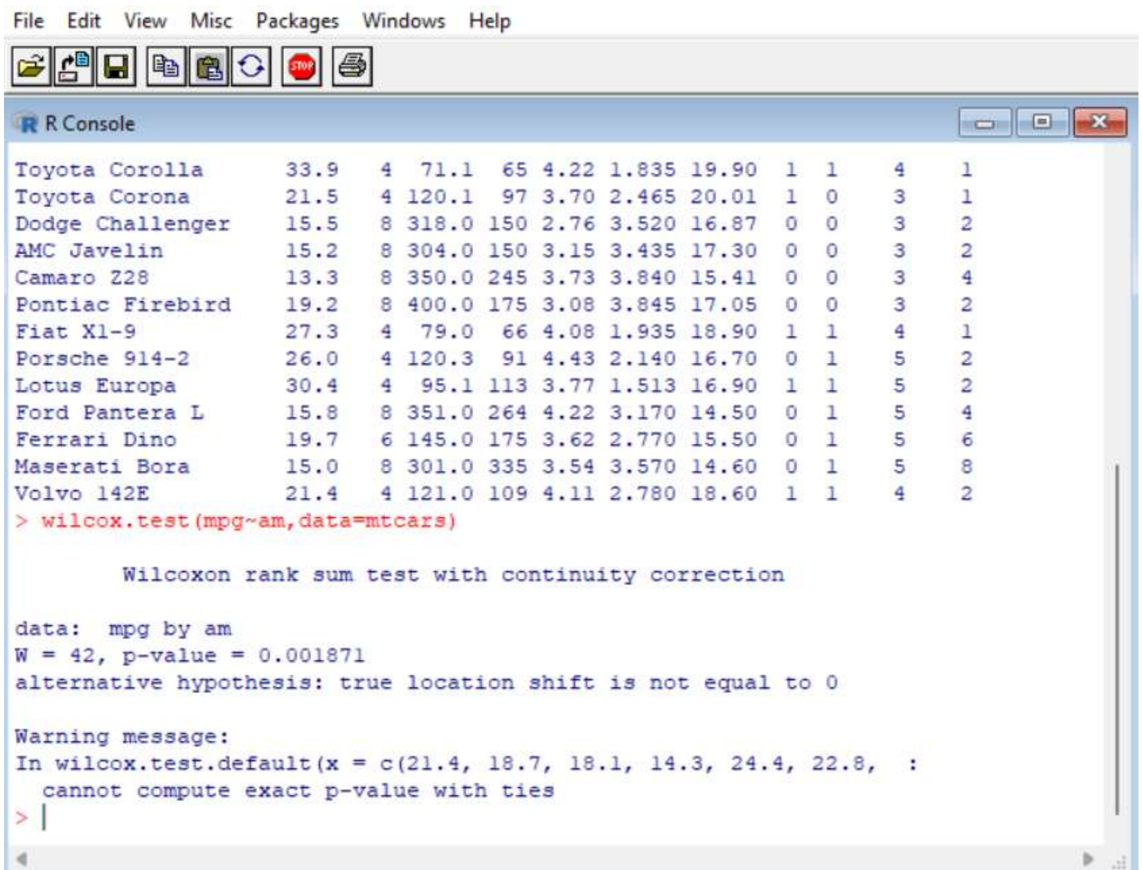
```
> library(MASS)
> head(immer)
  Loc Var    Y1    Y2
1  UF   M  81.0  80.7
2  UF   S 105.4  82.3
3  UF   V 119.7  80.4
4  UF   T 109.7  87.2
5  UF   P  98.3  84.2
6   W   M 146.6 100.4
> wilcox.test(immer$Y1,immer$Y2,paired=TRUE)

        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
> 
```

Conclusion:0.005318 <= 0.05 true reject Ho At .05 significance level, we conclude that the barley yields of 1931 and 1932 from the data set immer are nonidentical populations.

Practical No 8

Non parametric TestII:Mann-Whitney-Wilcoxon Test Two data samples are independent if they come from distinct populations and the samples do not affect each other. Using theMann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identicalwithoutassuming them to follow thenormal distribution.ExampleIn thedata frame columnmpgof the data setmtcars, there are gas mileage data of various 1974 U.S. automobiles.>mtcars$mpg[1]21.021.022.821.418.7...Meanwhile, another data column inmtcars, namedam, indicates the transmission type of the automobile model (0 = automatic, 1 = manual). In other words, it is the differentiating factor of the transmission type.>mtcars$am[1]11100000... In particular, the gas mileage data for manual and automatic transmissions are independent.

Problem: Without assuming the data to have normal distribution, decide at .05 significance level if the gas mileage data of manual and automatic transmissions in mt cars have identical data distribution.
Solution:

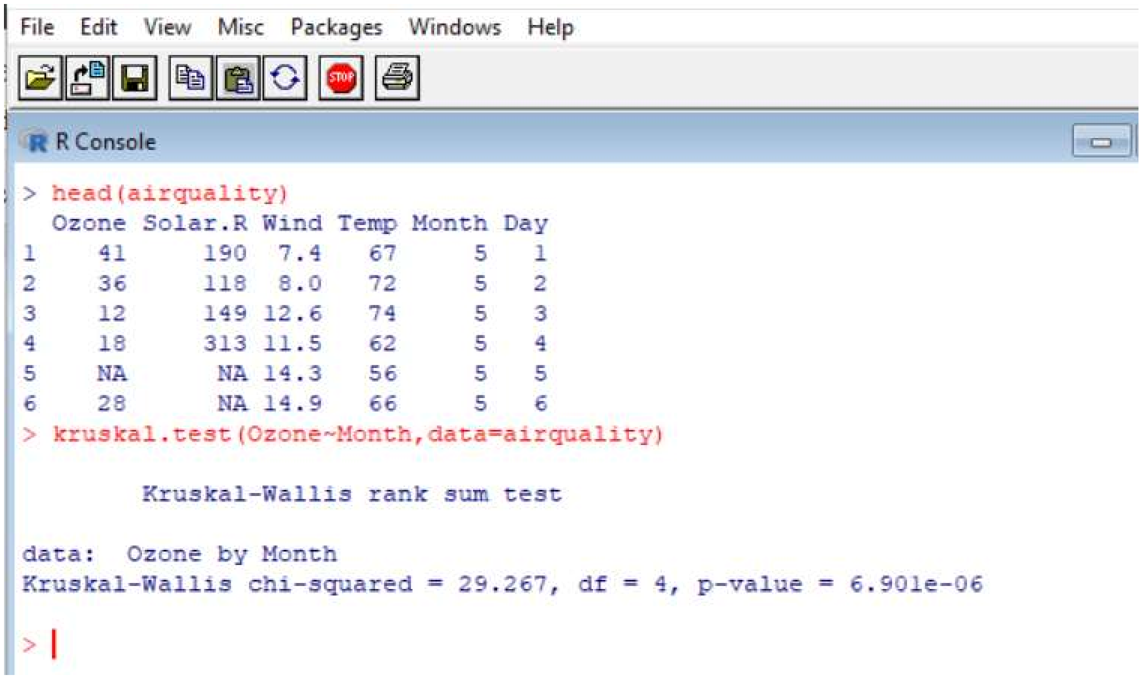



```
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28          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
> 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
> |
```

0.001871 <= 0.05 true reject Ho
Answer:At .05 significance level, we conclude 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 theKruskal-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 namedairquality, the daily air quality measurements in New York, May to September 1973, are recorded. The ozone density are presented in the data frame columnOzone.>head(air quality)OzoneSolar.RWindTempM onthDay1411907.46751 2361188.07252....

Problem: Without assuming the data to have normal distribution, test at .05 significance level if the monthly ozone density in New York has identical data distributions from May to September 1973.



File Edit View Misc Packages Windows Help

R Console

```
> 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
> 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

> 
```

0.000006901 < 0.05

Answer:At .05 significance level, we conclude that the monthly ozone density in New York from May to September 1973 arenonidenticalpopulations.

Practical No 9

Chi-squared Test of Independence Two random variablesxandyare calledindependentif the probability distribution of one variable is not affected by the presence of another.Assumefijis the observed frequency count of events belonging to bothi-th category ofxandj-th category ofy. Also assumeeijto be the corresponding expected count ifxandyare independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the followingChi-squaredtest statistics is less than a given significance levelα.ExampleIn the built-in data setsurvey, theSmokecolumn records the students smoking habit, while theExercolumn records their exercise level. The allowed values inSmokeare "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As forExer, they are "Freq" (frequently), "Some" and "None".We can tally the students smoking habit against the exercise level with thetablefunction in R. The result is called thecontingency tableof the two variables.>library(MASS)#loadtheMASSpackage>tbl=table(survey$Smoke,survey$Exer)>tbl#thecontingencytableFr eqNoneSomeHeavy713Never871884Occas1234Regul917

Problem:Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

File Edit View Misc Packages Windows Help

R Console

```
> 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
> 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
>
```

Answer:As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.