NAME: ANSARI SAAD AHMED

FCS2122007

SMTH 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.** / Miss. **Ansari Saad Ahmed** Roll No. **FCS2122007** Has successfully completed the necessary course of experiments in the subject of **Statistical Methods and Hypothesis Testing** 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
Mrs. Soni Yadav
(SUBJECT NAME)

Examination Date:
Examiner’s Signature & Date:

Head of the Department
Prof. Manoj Singh

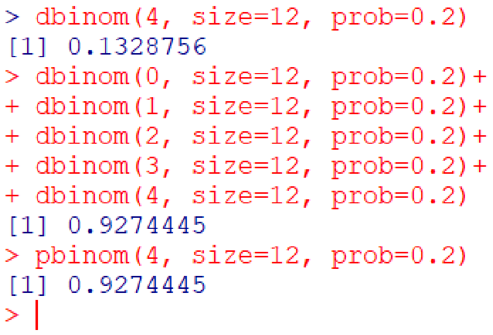
College Seal
And
Date

Practical 1

Aim: Binomial Distribution The binomial distribution is a discrete probability distribution. It describes the outcome of 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 successful outcomes in an experiment of independent trials is as follows.

Problem:

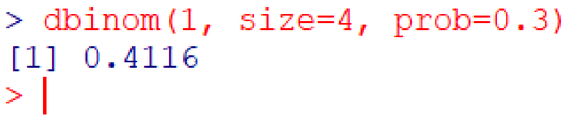
Suppose there are twelve multiple-choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or fewer correct answers if a student attempts to answer every question randomly.



```
> 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
> 
```

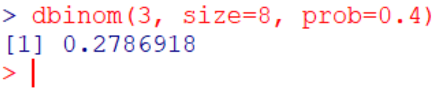
Exercise

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.



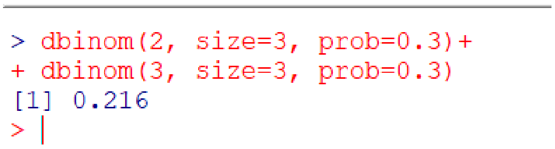
```
> dbinom(1, size=4, prob=0.3)
[1] 0.4116
> 
```

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?



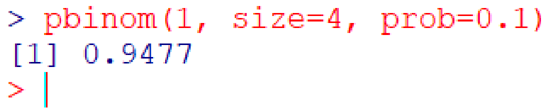
```
> dbinom(3, size=8, prob=0.4)
[1] 0.2786918
> 
```

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.



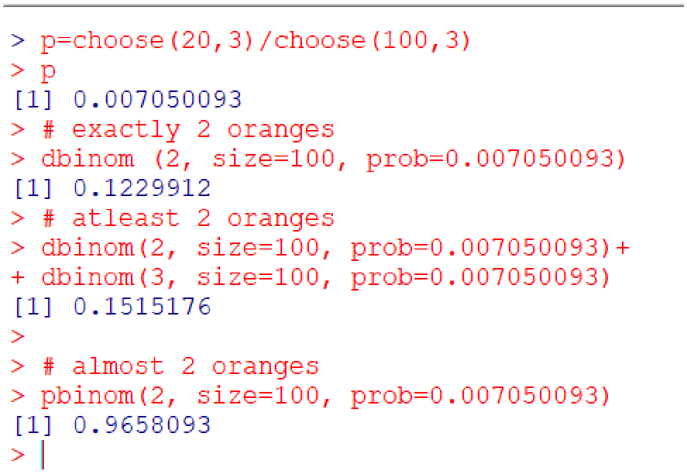
```
> dbinom(2, size=3, prob=0.3)+
+ dbinom(3, size=3, prob=0.3)
[1] 0.216
> 
```

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?



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

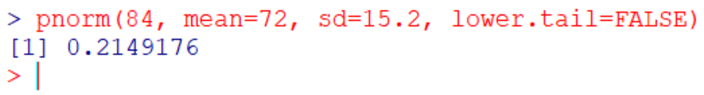
Problem 5: A basket contains 20 good oranges and 80 bad oranges. 3 oranges are drawn at random from this basket. Find the probability that out of 3 a) exactly 2 b) at least 2 c) at most 2 are good orange.



```
> p=choose(20,3)/choose(100,3)
> p
[1] 0.007050093
> # exactly 2 oranges
> dbinom (2, size=100, prob=0.007050093)
[1] 0.1229912
> # atleast 2 oranges
> dbinom(2, size=100, prob=0.007050093)+
+ dbinom(3, size=100, prob=0.007050093)
[1] 0.1515176
>
> # almost 2 oranges
> pbinom(2, size=100, prob=0.007050093)
[1] 0.9658093
> 
```

Practical 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?



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

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

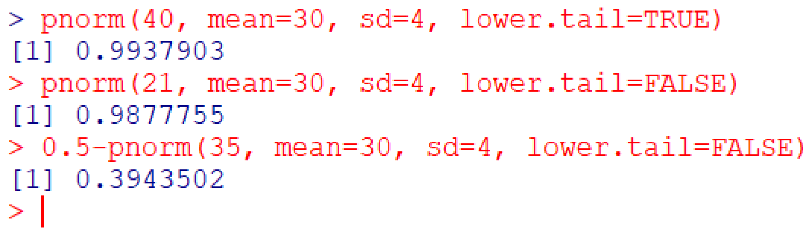
Exercise

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

a) P(x < 40)

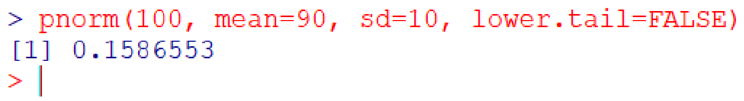
b) P(x > 21)

c) P(30 < X < 35)



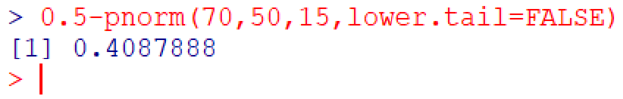
```
> pnorm(40, mean=30, sd=4, lower.tail=TRUE)
[1] 0.9937903
> pnorm(21, mean=30, sd=4, lower.tail=FALSE)
[1] 0.9877755
> 0.5-pnorm(35, mean=30, sd=4, lower.tail=FALSE)
[1] 0.3943502
> |
```

Problem 2: 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?



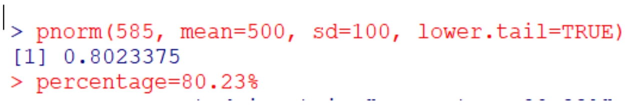
```
> pnorm(100, mean=90, sd=10, lower.tail=FALSE)
[1] 0.1586553
> |
```

Problem 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?



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

Problem 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 705. Tom takes the test and scores 5854. Will he be admitted to this university?

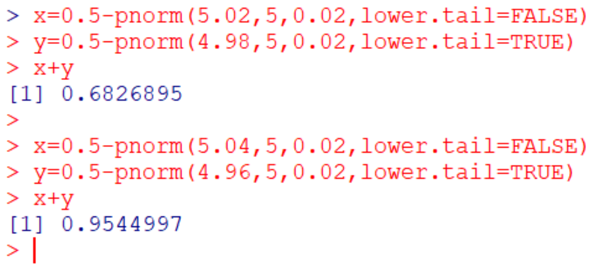


```
> pnorm(585, mean=500, sd=100, lower.tail=TRUE)
[1] 0.8023375
> percentage=80.23%
```

Problem 5: The length of similar components produced by a company are approximately 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 If between 4.96 and 5.04 cm?

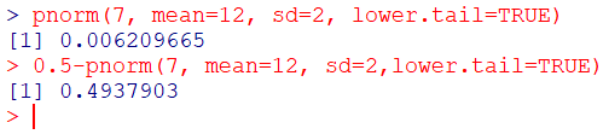


```
> x=0.5-pnorm(5.02,5,0.02,lower.tail=FALSE)
> y=0.5-pnorm(4.98,5,0.02,lower.tail=TRUE)
> x+y
[1] 0.6826895
> 
> x=0.5-pnorm(5.04,5,0.02,lower.tail=FALSE)
> y=0.5-pnorm(4.96,5,0.02,lower.tail=TRUE)
> x+y
[1] 0.9544997
> 
```

Problem 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

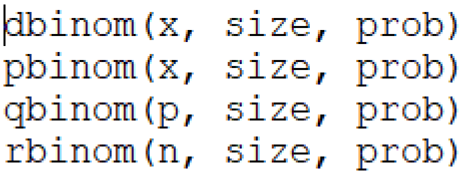
b) Between 7 and 12 months



```
> pnorm(7, mean=12, sd=2, lower.tail=TRUE)
[1] 0.006209665
> 0.5-pnorm(7, mean=12, sd=2,lower.tail=TRUE)
[1] 0.4937903
> 
```

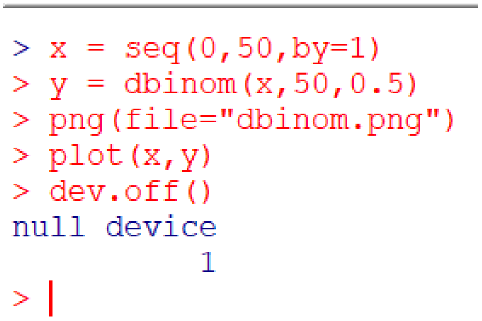
Practical 3

Aim: Property plotting of Binomial Distribution.

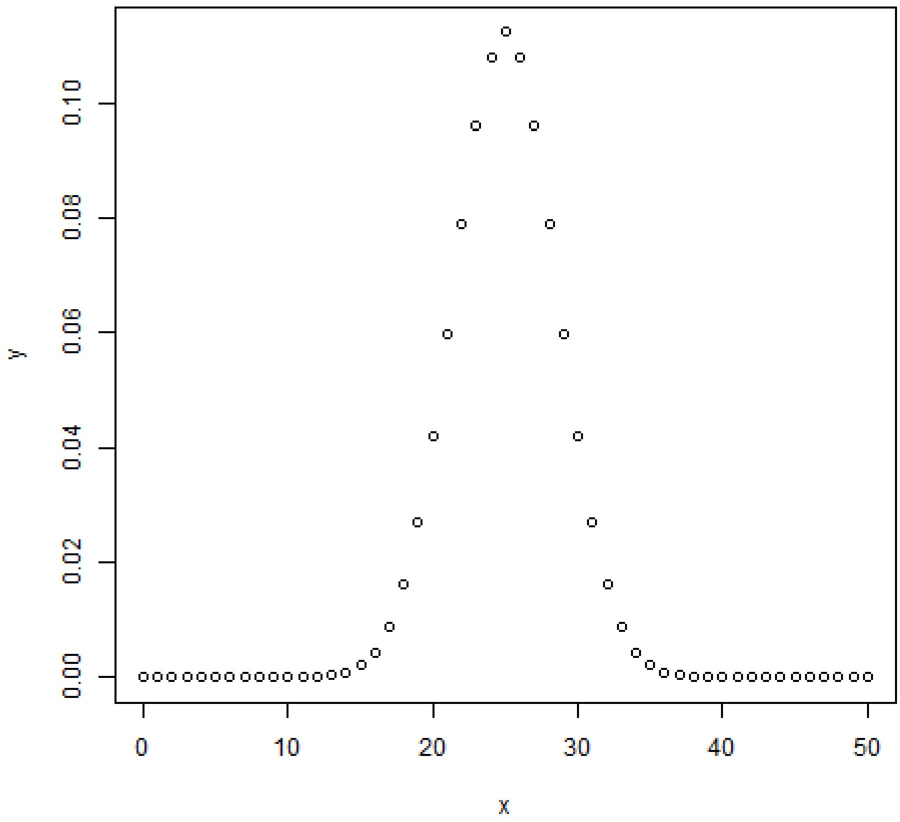


```
dbinom(x, size, prob)
pbinom(x, size, prob)
qbinom(p, size, prob)
rbinom(n, size, prob)
```

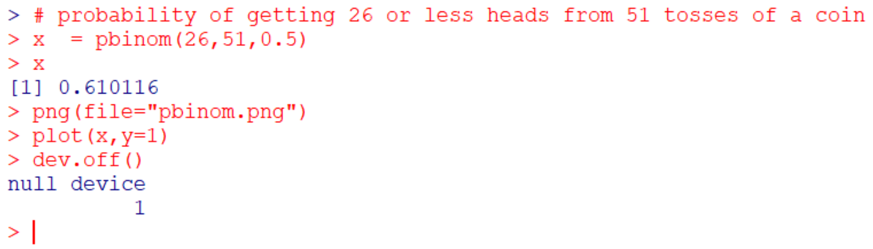
dbinom(): This function gives the probability density distribution at each point.



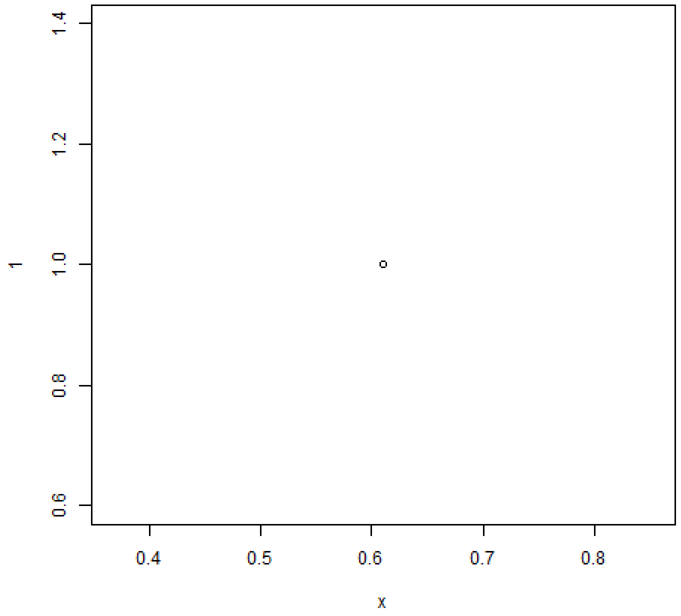
```
> x = seq(0,50,by=1)
> y = dbinom(x,50,0.5)
> png(file="dbinom.png")
> plot(x,y)
> dev.off()
null device
          1
> 
```



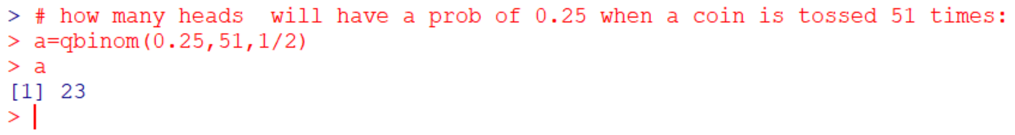
pbinom(): This function gives the cumulative probability of an event. It is a single value representing the probability.



```
> # probability of getting 26 or less heads from 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
> 
```

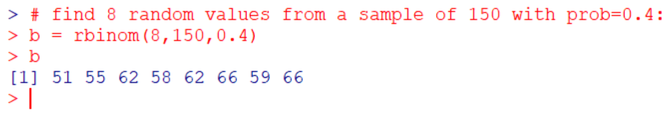


qbinom(): This function takes the probability value and gives a number whose cumulative value matches the probability value.



```
> # how many heads  will have a prob of 0.25 when a coin is tossed 51 times:
> a=qbinom(0.25,51,1/2)
> a
[1] 23
> 
```

rbinom(): This function generates required number of random values of given probability from a given sample.



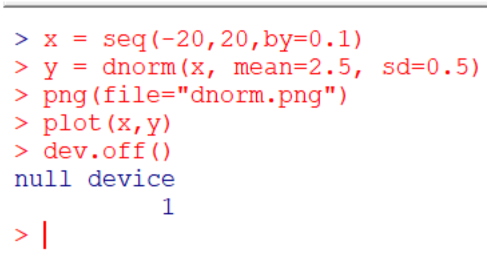
```
> # find 8 random values from a sample of 150 with prob=0.4:
> b = rbinom(8,150,0.4)
> b
[1] 51 55 62 58 62 66 59 66
> 
```

Practical 4

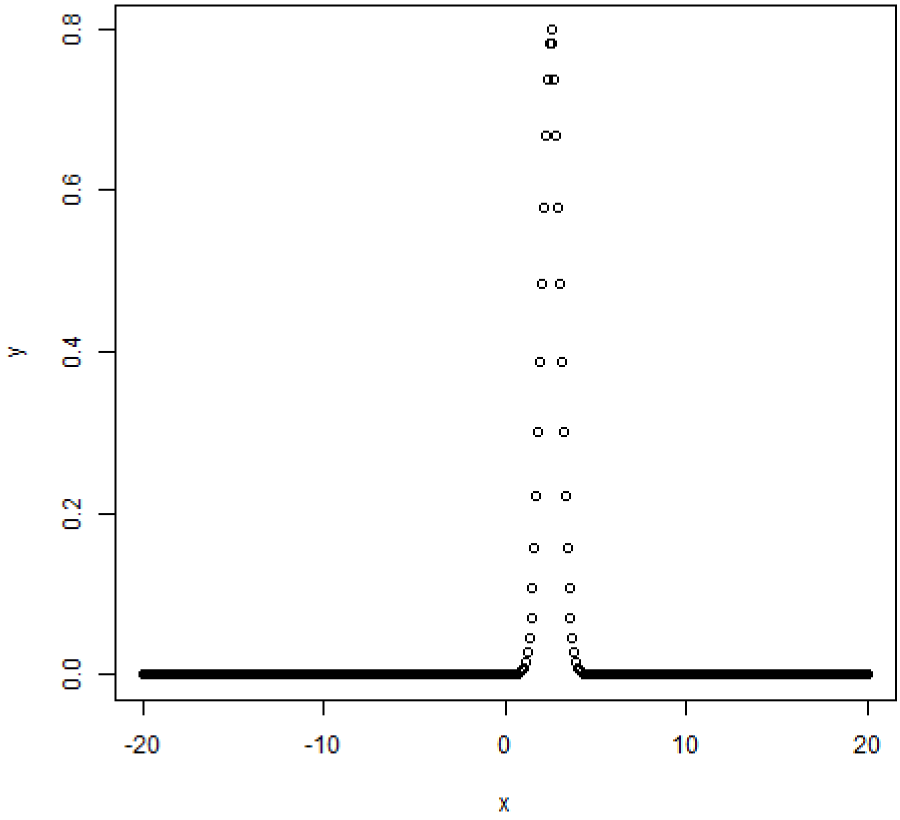
Aim: Property plotting of Normal Distribution.

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

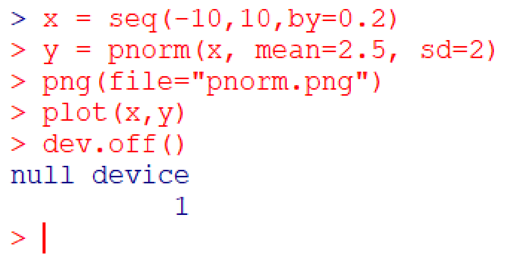
dnorm(): This function gives height of the probability distribution at each point for a given mean and standard deviation.



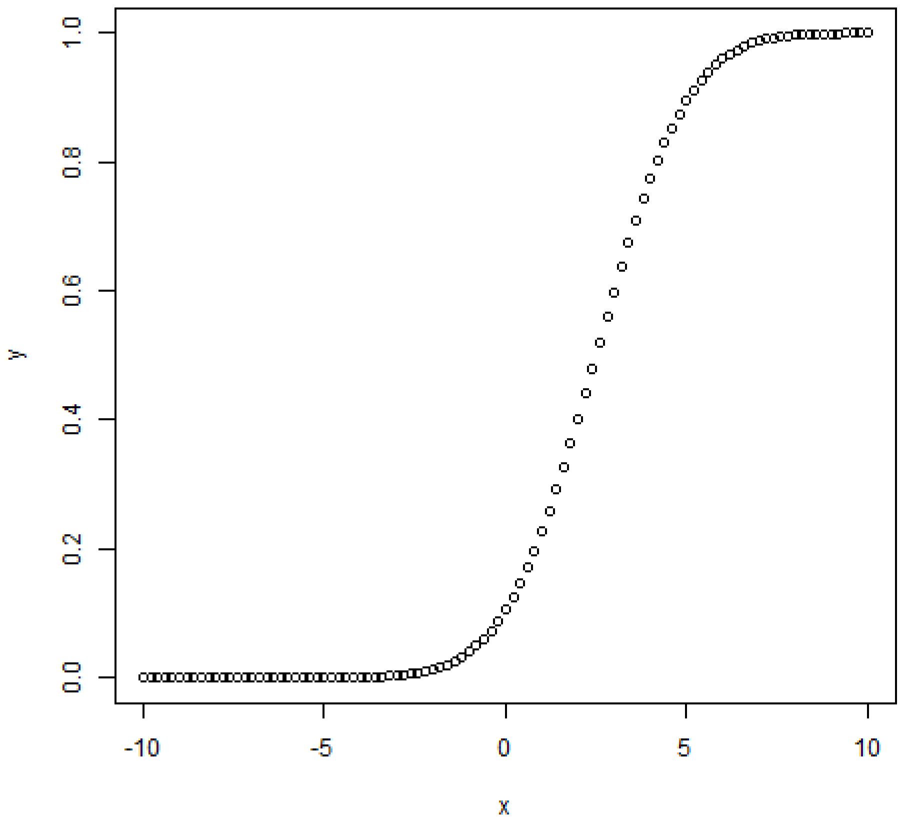
```
> x = seq(-20,20,by=0.1)
> y = dnorm(x, mean=2.5, sd=0.5)
> png(file="dnorm.png")
> plot(x,y)
> dev.off()
null device
          1
> |
```



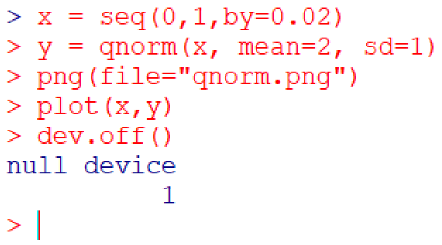
pnorm(): This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".



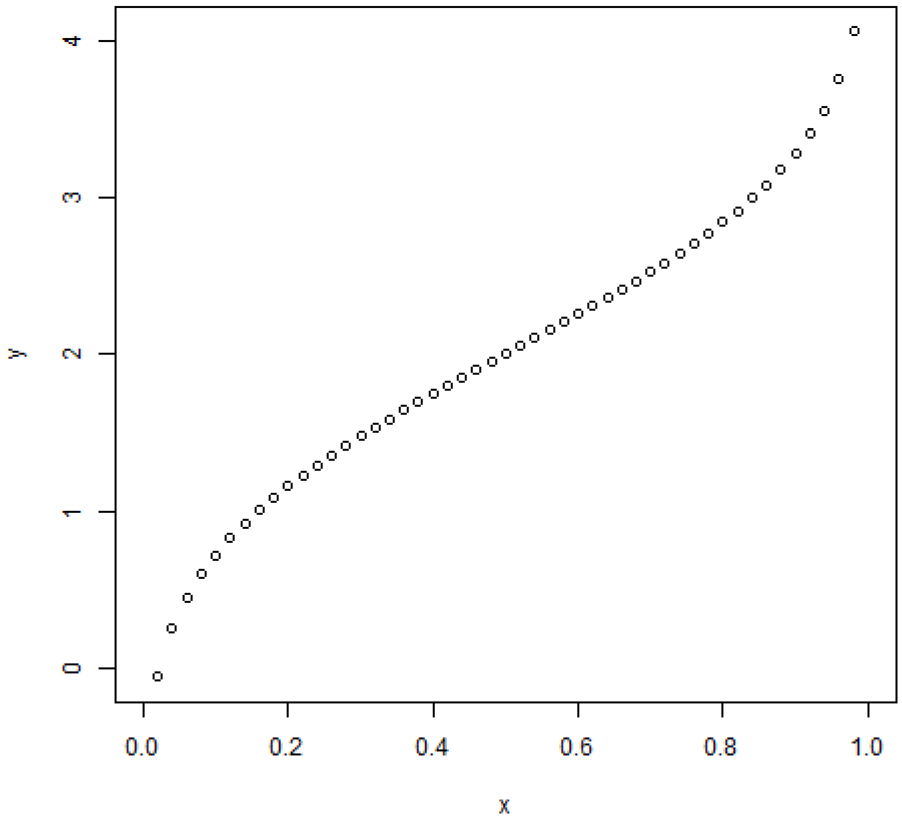
```
> x = seq(-10,10,by=0.2)
> y = pnorm(x, mean=2.5, sd=2)
> png(file="pnorm.png")
> plot(x,y)
> dev.off()
null device
          1
> 
```



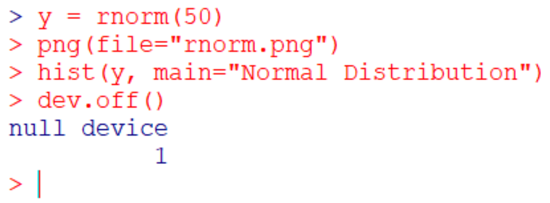
qnorm(): This function takes the probability value and gives a number whose cumulative valuematches the probability value.



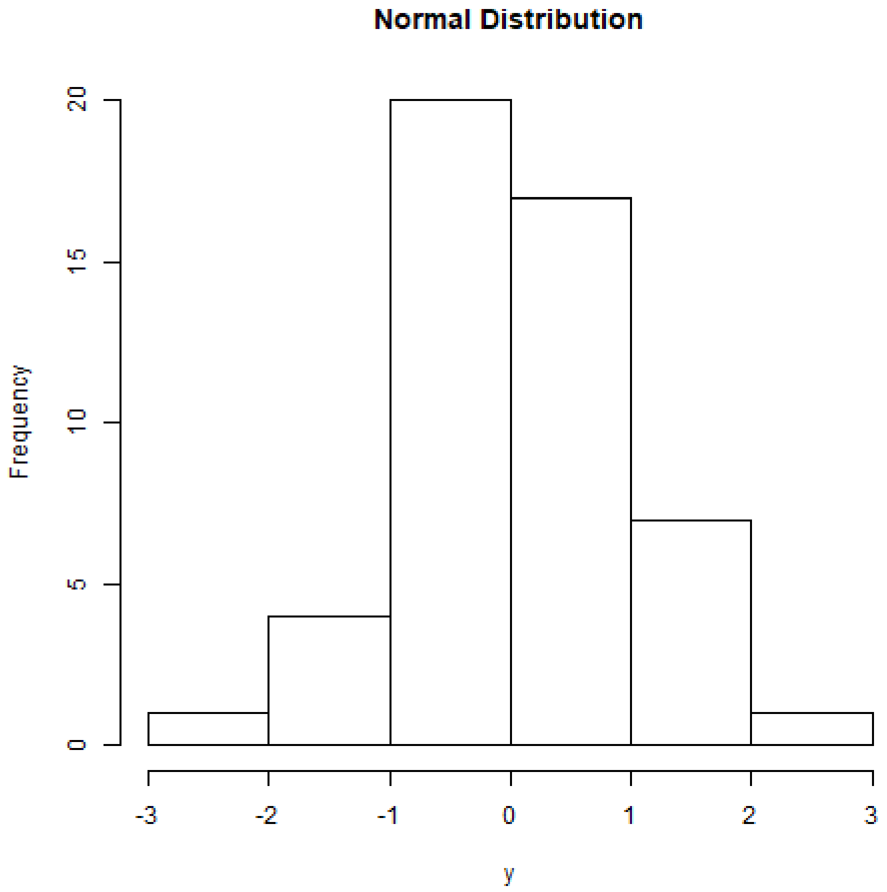
```
> x = seq(0,1,by=0.02)
> y = qnorm(x, mean=2, sd=1)
> png(file="qnorm.png")
> plot(x,y)
> dev.off()
null device
          1
> |
```



rnorm(): This function is used to generate random numbers whose distribution is normal. It takesthe sample size as input and generates that many random numbers. We draw a histogramto show the distribution of the generated numbers.

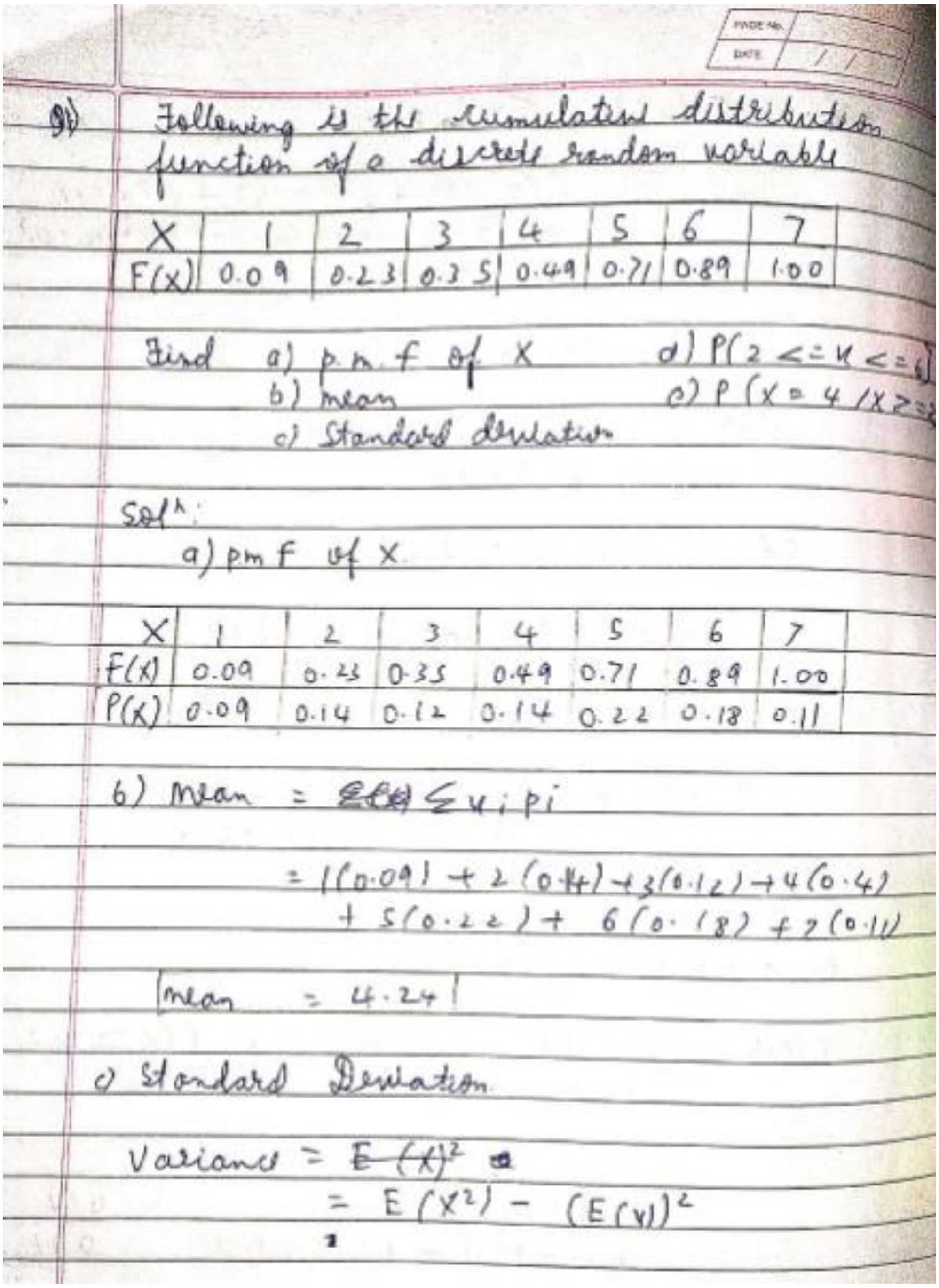


```
> y = rnorm(50)
> png(file="rnorm.png")
> hist(y, main="Normal Distribution")
> dev.off()
null device
          1
> |
```



Practical 5

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



PAGE No.

DATE / /

Q1) Following is the cumulative distribution function of a discrete random variable

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

Solⁿ:

a) pmf of 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
P(x)	0.09	0.14	0.12	0.14	0.22	0.18	0.11

b) mean = ~~E(x)~~ $\sum x_i p_i$

$= 1(0.09) + 2(0.14) + 3(0.12) + 4(0.4) + 5(0.22) + 6(0.18) + 7(0.11)$

$\boxed{\text{mean} = 4.24}$

c) Standard Deviation

Variance = ~~$E(X)^2$~~

$= E(X^2) - (E(X))^2$

Date / /

where $E(x^2) = \Sigma x^2 \cdot P(x)$

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

$= 21.34$

$Var = 21.34 - (4.24)^2$
$= 3.3624$

$S.d = \sqrt{var} = 1.8336$

$\therefore$ Standard deviation = 1.8336.

d) $P(2 \leq x \leq 6)$

$= P(2) + P(3) + P(4) + P(5) + P(6)$
$= 0.14 + 0.12 + 0.14 + 0.22 + 0.18$
$= 0.8$

e) $P(X = 4 / X \geq 2)$

$P(A/B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(X=4) \cap P(X \geq 2)}{P(X \geq 2}$

$= \dfrac{P(4)}{P(2) + P(3) + P(4) + P(5)} = \dfrac{0.14}{0.91}$

$\therefore P(X = 4 / X \geq 2) = 0.1538$

Q2) Let X be continuous random variable with pdf $F(x) = kx(1-x)$, for $0 \leq x < 1$
$= 0$ otherwise

Find k, Distribution function of x, $(0 < x < 4)$

Sol^n: Since given $f(x)$ is pdf, we can say that

$\int_0^1 f(x)\, dx = 1$

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

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

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

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

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

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

$= k\left[\frac{1}{2} - \frac{1}{3}\right] = k\left[\frac{1}{6}\right]$

$\boxed{k = 6}$

d) $P(X < 4)$

$= \int_0^4 6(x - x^2)$

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

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

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

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

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

$= 2(-40)$
$= -80$

Q2) A bag contains 6 green and 3 red balls Three balls are drawn at random without replacement What is the expected no. of red balls that will be drawn?

Solⁿ Let x denote no. of red balls

$\therefore x \to 0, 1, 2, 3.$

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

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

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

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

x	0	1	2	3
$F(X=x)$	$\frac{20}{84}$	$\frac{45}{84}$	$\frac{18}{84}$	$\frac{1}{84}$

$E x = \Sigma x \cdot P(x)$

$= 0\left(\frac{20}{84}\right) + 1\left(\frac{45}{84}\right) + 2\left(\frac{18}{84}\right) + 3\left(\frac{1}{84}\right)$

$= 84/84 = 1$

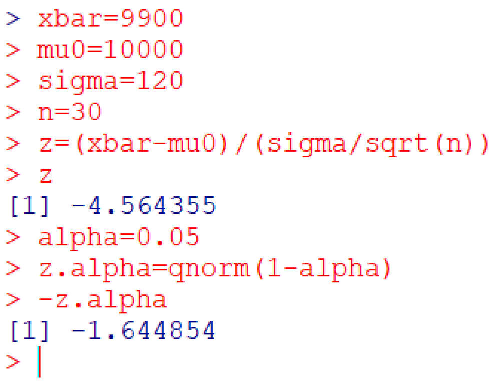
$\therefore$ Expected no. of red balls drawn is 1

Practical 6

Aim: Z test, T test

Lower Tail Test of Population Mean with Known Variance.

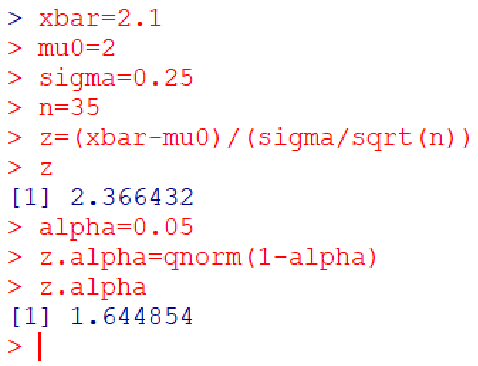
Problem: Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At .05 significance level, can we reject the claim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



```
> 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
> 
```

Upper Tail Test of Population Mean with Known Variance

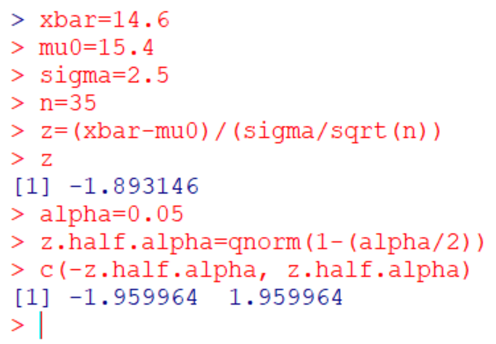
Problem: Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At .05 significance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



```
> 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
> 
```

Two-Tailed Test of Population Mean with Known Variance.

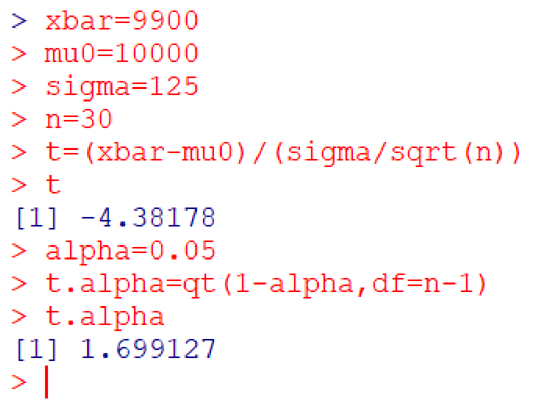
Problem: Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the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



```
> 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
> 
```

Lower Tail Test of Population Mean with Unknown Variance.

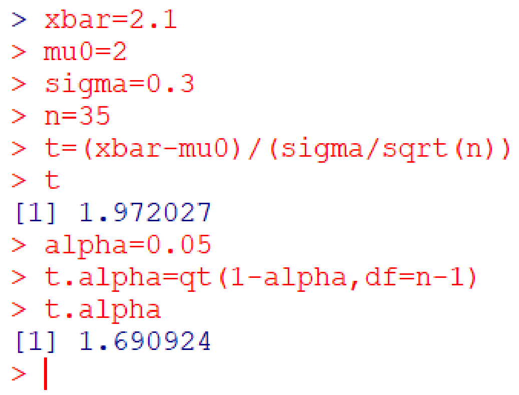
Problem: Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a normally distributed sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the sample standard deviation is 125 hours. At .05 significance level, can we reject the claim by the manufacturer?



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

Upper Tail Test of Population Mean with Unknown Variance.

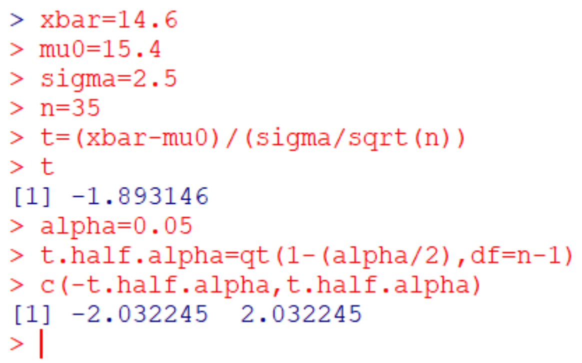
Problem: Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the sample standard deviation is 0.3 gram. At .05 significance level,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



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

Two-Tailed Test of Population Mean with Unknown Variance.

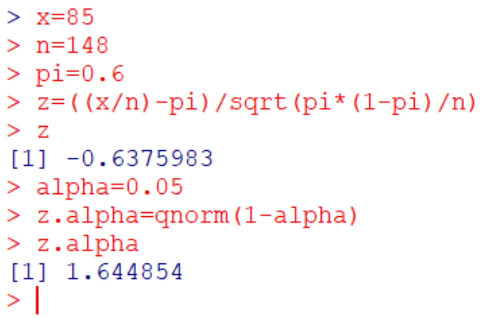
Problem: Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the sample standard deviation is 2.5 kg. At .05 significance level, can we reject the null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



```
> xbar=14.6
> mu0=15.4
> sigma=2.5
> n=35
> t=(xbar-mu0)/(sigma/sqrt(n))
> t
[1] -1.893146
> alpha=0.05
> t.half.alpha=qt(1-(alpha/2),df=n-1)
> c(-t.half.alpha,t.half.alpha)
[1] -2.032245  2.032245
> 
```

Lower Tail Test of Population Proportion

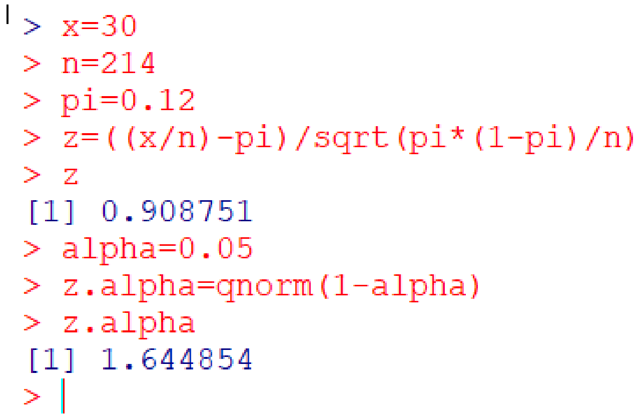
Problem: Suppose 60% of citizens voted in last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?



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

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 .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?



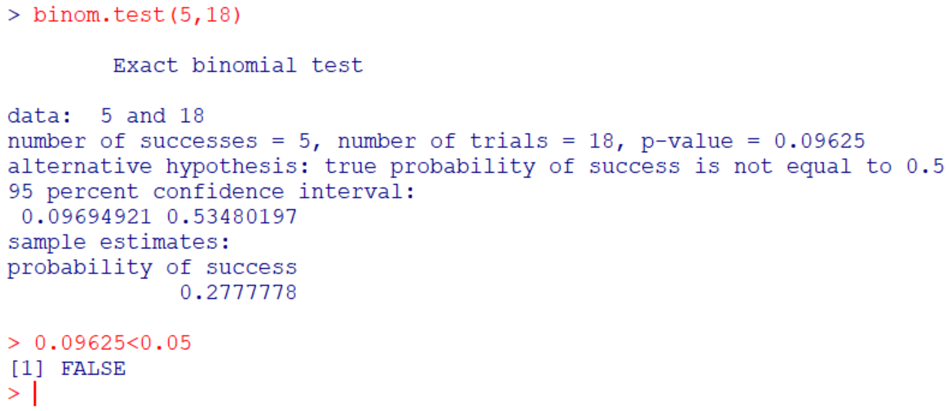
```
> x=30
> n=214
> pi=0.12
> z=((x/n)-pi)/sqrt(pi*(1-pi)/n)
> z
[1] 0.908751
> alpha=0.05
> z.alpha=qnorm(1-alpha)
> z.alpha
[1] 1.644854
> 
```

Practical 7

Aim: Non-parametric tests - I

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 well, and the rest prefer the old one. At .05 significance level, can we reject the notion that the two drinks are equally popular?



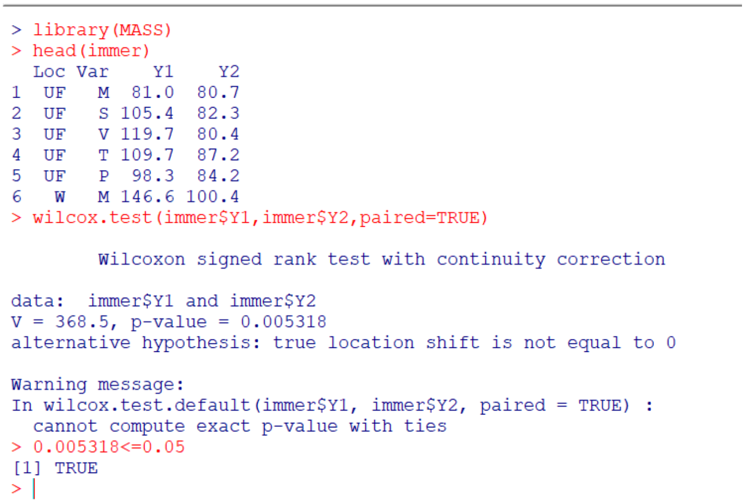
```
> binom.test(5,18)

        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

> 0.09625<0.05
[1] FALSE
> 
```

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 name dimmer, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the data frame.

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



```
> 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
> 0.005318<=0.05
[1] TRUE
> 
```

Practical 8

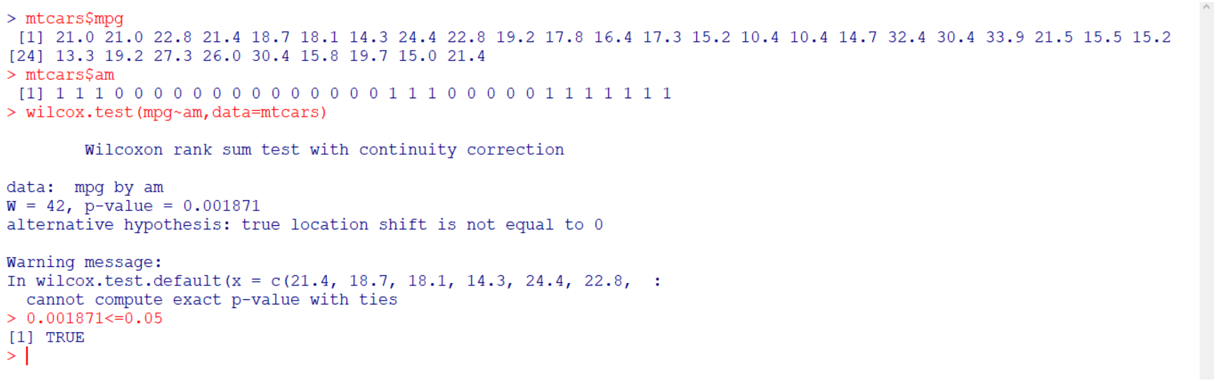
Aim: Non parametric tests – II

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

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.



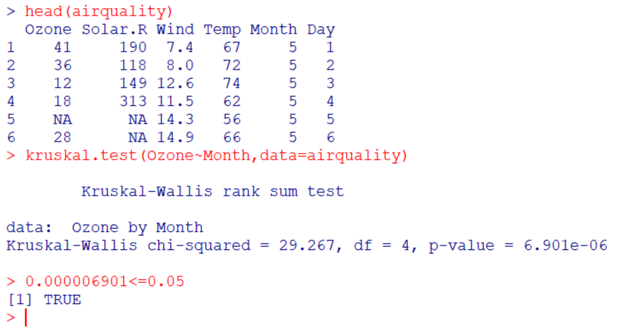
```
> mtcars$mpg
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2
[24] 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4
> mtcars$am
 [1] 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1
> wilcox.test(mpg~am,data=mtcars)

        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
[1] TRUE
> |
```

Kruskal-Wallis Test

A collection of data samples is independent if they come from unrelated populations and the samples do not affect each other. Using the Kruskal-Wallis Test, 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 is presented in the data frame column Ozone.

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.



```
> 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
[1] TRUE
> |
```

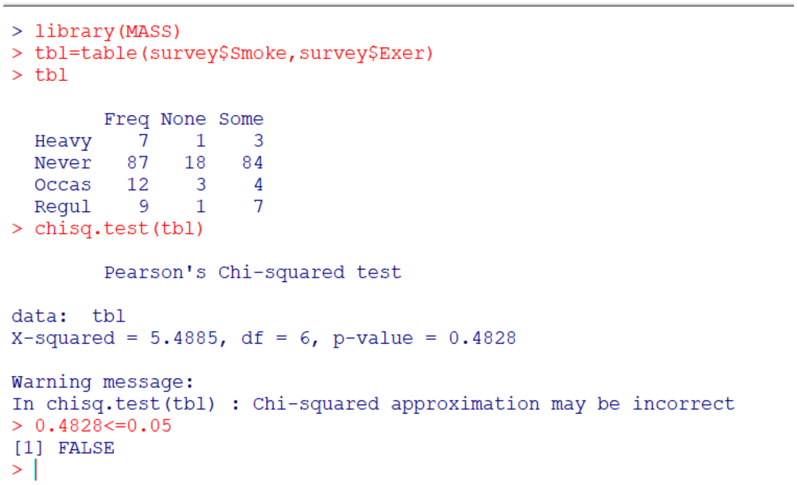
Practical 9

Aim: Chi Square Test of independence.

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

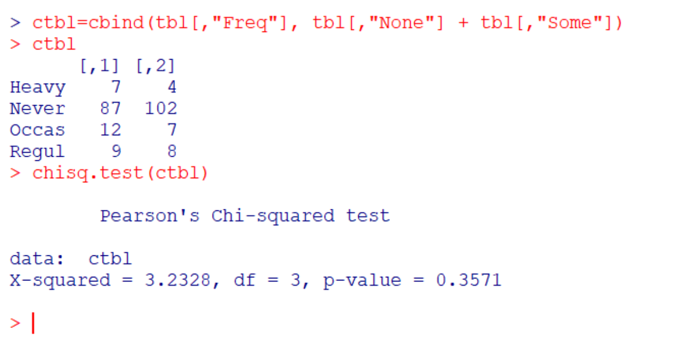
Example: In the built-in data set survey, the Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".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

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



```
> 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
> 0.4828<=0.05
[1] FALSE
> |
```

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 thechisq.testfunction against ctbl instead.



```
> 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

> |
```