PRACTICALS ON STATISTICAL METHODS AND TESTING OF HYPOTHESIS USING R-SOFTWARE

INDEX

PRACTICAL 1

PRACTICAL BASED ON BINOMIAL DISTRIBUTION

1) If X~ Bino (10, 0.6).

Find a) P(X=0) b) P(X=2) c) P (X≤3) d) P(X>5)

2) Plot probability mass function (pmf) and distribution function for the following random variables X ~ Bino (8, 0.65)

3) A set of similar fair coins are tossed 640 times with the following result – no. of

Heads:	0	1	2	3	4	5	6
Frequency:	7	64	140	210	132	75	12

Fit the binomial distribution to the data.

4) Plot the pmf of X ~ Bino (30, 0.05) and comment on graph.

Q.1 SOLUTION

///OUTPUT

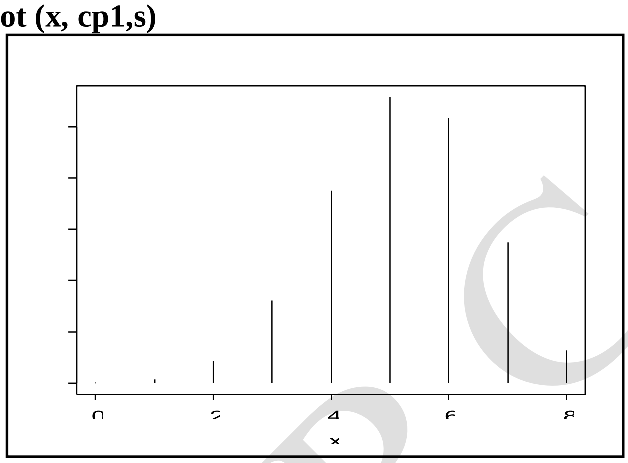
```
> #a]P(X=0)

> a1=dbinom (0,10,0.6)

> a1

[1] 0.0001048576

>  # b] P(X=2)

> b1=dbinom (2,10,0.6)

> b1

[1] 0.01061683

> #c] P(X<=3)

> c1=pbinom (3,10,0.6)

> c1

[1] 0.05476188

> #d] P(X>5)

> d1=1-pbinom (5,10,0.6)

> d1

[1] 0.6331033
```

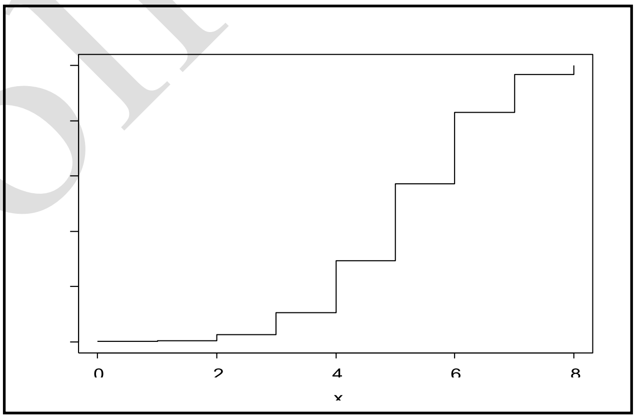
Q.2 SOLUTION

```
> n=8; p=0.65
> x=0: n
> bp=dbinom(x, n, p)
> d=data.frame("x-values"=x,"probabilities"=bp)
> d
```

```
         x.values     probabilities
  1            0     0.0002251875
  2            1      0.0033456434
  3            2      0.0217466823
  4            3      0.0807733916
  5            4      0.1875096590
  6            5      0.2785857791
  7            6      0.2586867948
  8            7      0.1372623809
  9            8      0.0318644813
```

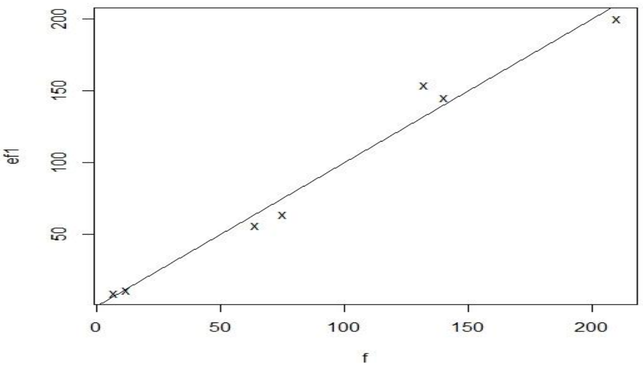
```
>  plot (x, bp,"h")
>cp=pbinom (x, n, p)
>  cp1=round (cp,4)
>  d1=data.frame(x, cp1)
>  plot (x, cp1,s)
```





Q.3 Solution

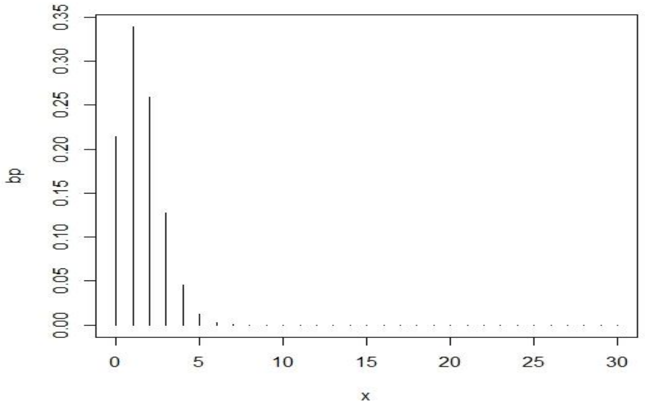
```
x=0:6

> f=c(7,64,140,210,132,75,12)

> m=sum(x*f)/sum(f)

> n=max(x)

> p=m/n

> q=1-p

> px=dbinom(x,n,p)

> px1=round(px,4)

> ef=sum(f)*px1

> ef1=round(ef,0)

> d=data.frame(x,f,"expected frequency"=ef1)
> d
  x   f     expected.frequency
1 0   7                  9
2 1  64                 56
3 2 140                145
4 3 210                200
5 4 132                154
6 5  75                 64
7 6  12                 11
> plot(f,ef1, pch="x")
> abline(0,1)
```

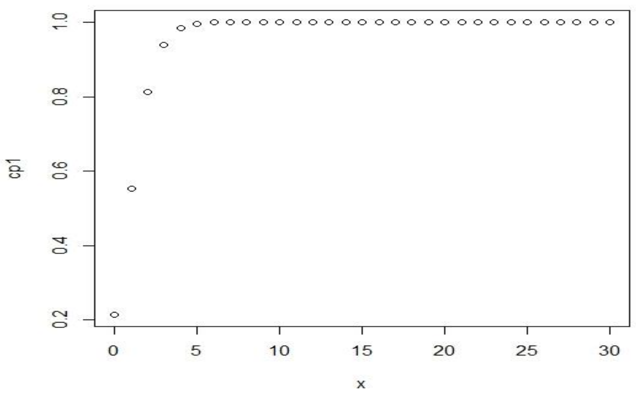


Q.4 Solution

```
n=30
> p=0.05
> x=0:n
> bp=dbinom(x,n,p)
> d=data.frame("x-values"=x,"probalities"=bp)
> d
   x.values  probalities
1         0 2.146388e-01
2         1 3.389033e-01
3         2 2.586367e-01
4         3 1.270496e-01
5         4 4.513605e-02
6         5 1.235302e-02
7         6 2.708997e-03
8         7 4.888415e-04
```

```
9          8 7.396944e-05
10          9 9.516536e-06
11         10 1.051828e-06
12         11 1.006534e-07
13         12 8.387780e-09
14         13 6.112552e-10
15         14 3.906518e-11
16         15 2.193133e-12
17         16 1.082138e-13
18         17 4.690382e-15
19         18 1.782894e-16
20         19 5.926516e-18
21         20 1.715570e-19
22         21 4.299675e-21
23         22 9.257674e-23
24         23 1.694769e-24
25         24 2.601619e-26
26         25 3.286255e-28
27         26 3.326169e-30
28         27 2.593504e-32
29         28 1.462502e-34
30         29 5.308539e-37
plot(x, bp,"h")
> cp1=pbinom(x,n,p)
> plots(x,cp1)
```





PRACTICAL 2

PRACTICAL BASED ON NORMAL DISTRIBUTION

1) Let X~ N (50,40). Find P (X≤60), P(X≥100) , P(10≤X≤20) and P(X≤k)=0.293.

2) Fit a normal distribution to the following data of height (in cms) of 200 Indianadult males

Height in cms	144-150	150-156	156-162	162-168	168-174	174-180	180-186
No of Adults	3	12	23	52	61	39	10

3) Find a) P(X ≤ 0.8) b) P (X > 0.5)
If i. **X**~**Norm**al(**2**, **1**. **5**) ii. **X**~**Norm**al(**0**, **1**)

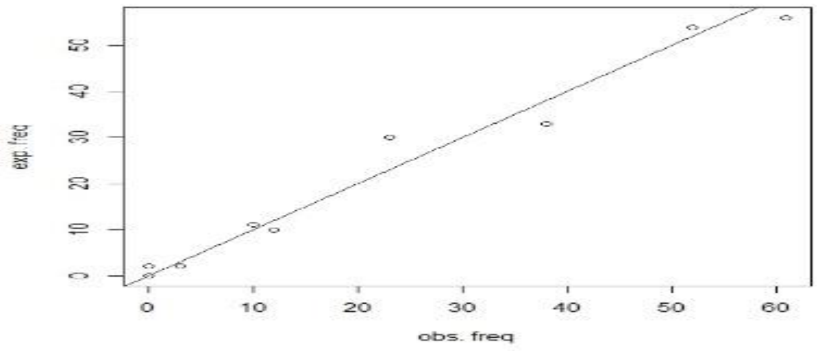
Plot pdf and distribution function.

Q.1 Solution

```
> mu=50
> sd=sqrt(40)
# P (X≤60)
> p1=pnorm(60,mu,sd)
> p1
[1] 0.9430769
# P(X≥100)
> p2=1-pnorm(100,mu,sd)
> p2
[1] 1.332268e-15
# P(10≤X≤20)
> p3=pnorm(20,mu,sd)-pnorm(10,mu,sd)
> p3
[1] 1.050591e-06
> l1=seq(144,180,6)
# P(X≤k)=0.293
> p4=qnorm(0.293,mu,sd)
> p4
[1] 46.55538
```

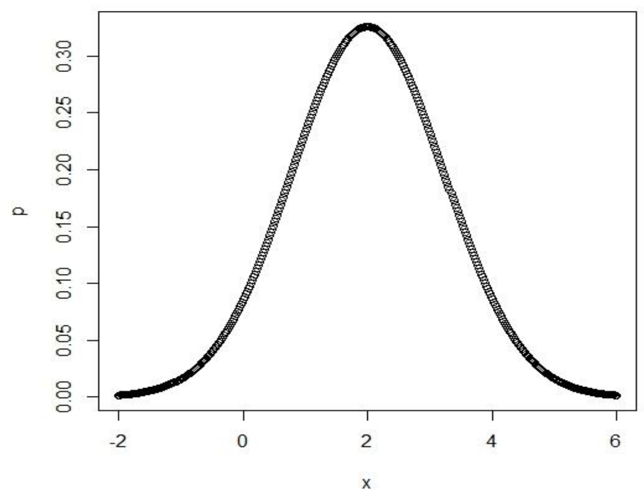
Q.2 Solution

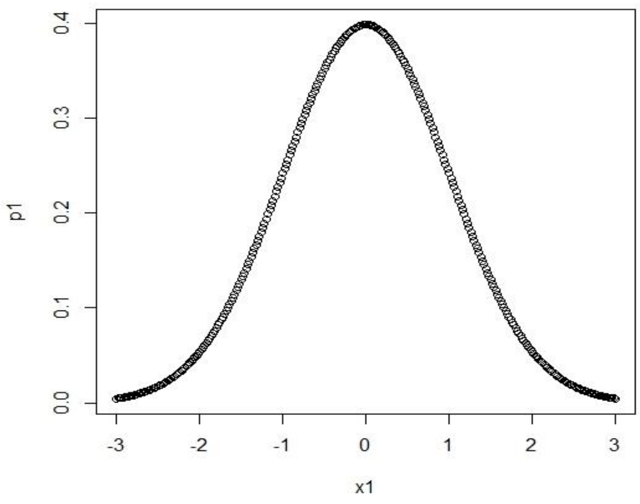
```
> u1=seq(150,186,6)
> f=c(3,12,23,52,61,38,10)
> x=(l1+u1)/2
> n=sum(f)
> k=length(f)
> m=sum(f*x)/n
> v=sum(f*(x-m)^2)/n
> sd=sqrt(v)
> l1=c(-9999,l1,186)
> cp=pnorm(l1,m,sd)
> p=diff(cp)
> p=c(p,1-cp[k+2])
> u1=c(144,u1,9999)
> f=c(0,f,0)
> ef=round(n*p,0)
> d=data.frame("Lower limit"=l1,"Upper limit"=u1, "Obs. freq"=f,
"prob"=p,"cum prob"=cp, "expfreq"=ef)
> d
  Lower.limit Upper.limit Obs..freq         prob     cum.prob expfreq
1       -9999         144         0 0.0009399578 0.0000000000       0
2         144         150         3 0.0086375755 0.0009399578       2
3         150         156        12 0.0478866245 0.0095775333      10
4         156         162        23 0.1514062428 0.0574641578      30
5         162         168        52 0.2734738985 0.2088704007      54
6         168         174        61 0.2824486590 0.4823442992      56
7         174         180        38 0.1668159923 0.7647929581      33
8         180         186        10 0.0562916410 0.9316089505      11
9         186        9999         0 0.0120994085 0.9879005915       2
> plot(f,ef,xlab="obs. freq",ylab="exp. freq","p")
> abline(0,1)
```



#Q.3 Solution

```
> a=pnorm(0.8,2,sqrt(1.5),lower.tail=1)
> a
[1] 0.1635934
> b=pnorm(0.5,2,sqrt(1.5),lower.tail=0)
> b
[1] 0.8896643
> x=seq(-2,6,by=0.02)
> p=dnorm(x,2,sqrt(1.5))
> plot(x,p)
> a1=pnorm(0.8,0,sqrt(1),lower.tail=1)
> a1
[1] 0.7881446
> b1=pnorm(0.5,0,sqrt(1),lower.tail=0)
> b1
[1] 0.3085375
> x1=seq(-3,3,by=0.02)
> p1=dnorm(x1,0,sqrt(1))
> plot(x1,p1)
```


p
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-2
0
2
4
6
x


p1
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x1

PRACTICAL 3

PRACTICAL BASED ON DISCRETE DISTRIBUTION

1) The Probability distribution of discrete random variable is given below:

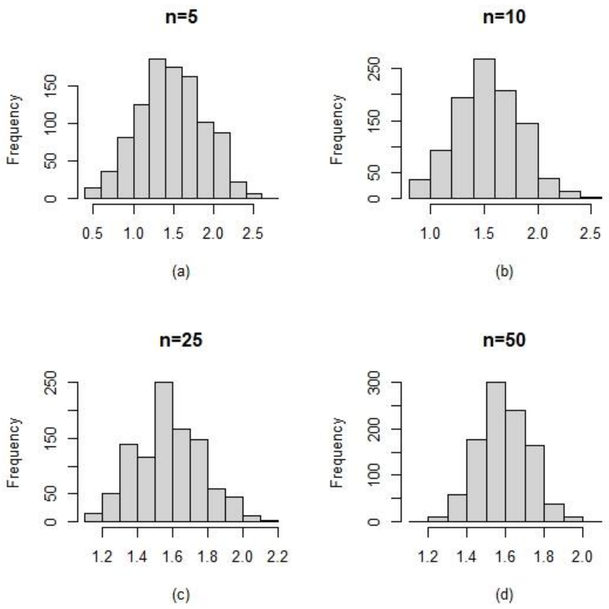
X	0	1	2	3
P(X=x)	0.1	0.4	0.3	0.2

For the sample of size 5, 10, 25, 50 find the sample mean, sample variance, Q1, Q2, Q3.

2) Draw a sample of size 5, 10, 25,50 from normal distribution with mean 10 and standard deviation 4. Describe the sample also plot the graph for it.

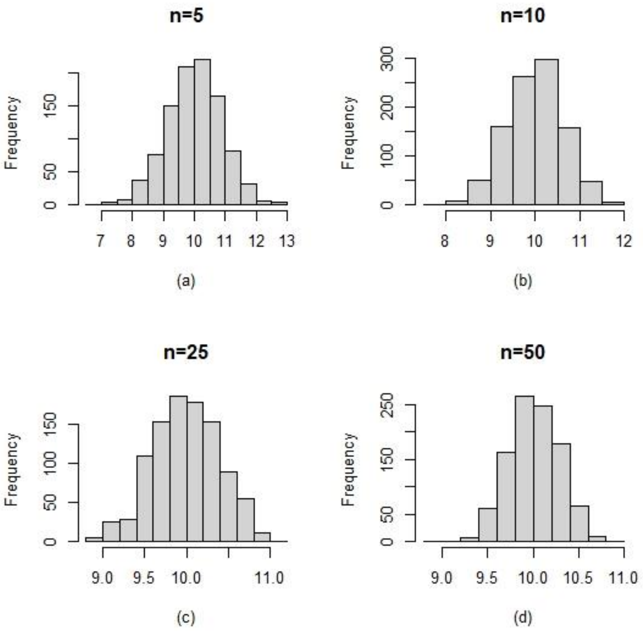
Q.1 Solution BINOMIAL DISTRIBUTION

```
set.seed(1)#for producing the same sequence of random variable every time
> n=50#sample size
> rep=1000#repetitions
> xv=c(0,1,2,3)#X values
> prob=c(0.1,0.4,0.3,0.2)#Probability Values #random sample from Discrete Distribution
> x1=sample(xv,n*rep,replace = TRUE,prob=prob);
> x=matrix(x1,rep,n)#arrangement of random numbers in matrix
> s.mean5=rowMeans(x[,1:5])#sample mean n=5
> s.mean10=rowMeans(x[,1:10])#sample mean n=10
> s.mean25=rowMeans(x[,1:25])#sample mean n=25
> s.mean50=rowMeans(x[,1:50])#sample mean n=50
> s.mean=data.frame(s.mean5,s.mean10,s.mean25,s.mean50) #bind all means
> apply(s.mean,2,mean)
 s.mean5         s.mean10       s.mean25      s.mean50
 1.57200         1.58730        1.58748        1.59980
> apply(s.mean,2,var)#Calculation of mean and variance
   s.mean5         s.mean10       s.mean25       s.mean50
0.17194795     0.08977849    0.03582307       0.017344771
> par(mfrow=c(2,2))
> hist(s.mean5,xlab = "(a)",main="n=5")
> hist(s.mean10,xlab = "(b)",main="n=10")
> hist(s.mean25,xlab = "(c)",main="n=25")
>  hist(s.mean50,xlab = "(d)",main="n=50")
>
```



Q.2 SOLUTION NORMAL DISTRIBUTION

```
> set.seed(25) #for producing the same sequence of random variable every time
> n=50#sample size
> rep=1000#repetitions
> x1=rnorm(rep*n,10,2)
> x=matrix(x1,rep,n)
> s.mean5=rowMeans(x[,1:5])#sample mean n=5
> s.mean10=rowMeans(x[,1:10])#sample mean n=10
> s.mean25=rowMeans(x[,1:25])#sample mean n=25
> s.mean50=rowMeans(x[,1:50])#sample mean n=50
> s.mean=data.frame(s.mean5,s.mean10,s.mean25,s.mean50) #bind all means
> apply(s.mean,2,mean)
  s.mean5  s.mean10  s.mean25  s.mean50
10.004742  9.994193  9.989217 10.002109
> apply(s.mean,2,var)#Calculation of mean and variance
   s.mean5    s.mean10    s.mean25    s.mean50
0.81704713 0.40690323 0.15585247 0.07559211
> par(mfrow=c(2,2))
> hist(s.mean5,xlab = "(a)",main="n=5")
> hist(s.mean10,xlab = "(b)",main="n=10")
> hist(s.mean25,xlab = "(c)",main="n=25")
> hist(s.mean50,xlab = "(d)",main="n=50")
```


n=5
Frequency
0
50
150
7
8
9
10
11
12
13
n=10
Frequency
0
100
200
300
8
9
10
11
12
n=25
Frequency
0
50
100
150
9.0
9.5
10.0
11.0
n=50
Frequency
0
50
150
250
9.0
9.5
10.0
10.5
11.0

(a)

(b)

(c)

(d)

PRACTICAL 4

PRACTICAL BASED ON CONFIDENCE INTERVAL

Q.1 A survey of 40 retired women revealed the mean age at which their income was maximum to be 45 years with a standard deviation of 6.3 years. Find 95% confidence interval for the mean age of maximum earnings of women who survived till they retire.

```
> n=40
> xbar=45
> s=6.3
> alp=0.05
> z=qnorm(alp/2,0,1,lower.tail=0)
> l=xbar-z*s/sqrt(n)
> u=xbar+z*s/sqrt(n)
> #95% CI for mean
> paste("(",l,",",u,")")
[1] "( 43.0476456482406 , 46.9523543517594 )"
```

Q.2 In a study of television viewing habits, order to obtain an interval estimate of the average number of hours per week that teenagers spend watching television programs, a random sample of 100 teenaged children is taken. Sample investigation revealed a mean of 9.2 hours with standard deviation of 3.2 hours. Obtain the desired interval estimate with confidence coefficient 0.99.

```
> n=100
> xbar=9.2
> s=3.2
 alp=0.01
 z=qnorm(alp/2,0,1,lower.tail=0)
> l=xbar-z*s/sqrt(n)
> u=xbar+z*s/sqrt(n)
> #99% CI for mean
> paste("(",l,",",u,")")
[1] "( 8.37573462286435 , 10.0242653771356 )"
```

Q.3 For the following data find the 99% confidence interval

20,16,26,27,23,22,18,24,25,19,18,28,25,27,22

```
> x=c(20,16,26,27,23,22,18,24,25,19,18,28,25,27,22)
> xbar=mean(x)
> n=length(x)
> s=sd(x)
> alp=.1
> t=qt(alp/2,n-1,lower.tail=0)
> l=xbar-t*s/sqrt(n)
> u=xbar+t*s/sqrt(n)
> #90% CI for mean
> paste("(",l,",",u,")")
[1] "( 20.9506711391254 , 24.3826621942079 )"
```

Q.4 For the following data with two samples of different size. Calculate the 95% Confidence interval for difference of means

74,77,74,73,79,76,82,72,75,78,77,78,76,76

70,75,74,70,69,72,76,72,72,77,77,72,75,78,72,74,75

OUTPUT

```
> x=c(74,77,74,73,79,76,82,72,75,78,77,78,76,76)
> y=c(70,75,74,70,69,72,76,72,72,77,77,72,75,78,72,74,75)
> n1=length(x)
> n2=length(y)
> xbar=mean(x)
> ybar=mean(y)
> s1=sd(x) #SD of X
> s2=sd(y)
> s=sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2))
> t=qt(.05/2,n1+n2-2,lower.tail=0)
> #i)
> l=xbar-ybar-t*s*sqrt(1/n1+1/n2)
> u=xbar-ybar+t*s*sqrt(1/n1+1/n2)
> #95% CI for difference of means
> paste("(",l,",",u,")")
[1] "( 0.733919803307959 , 4.63582809585169 )"
```

Q.5 For a given sample of 100, 35 are working as professor. Construct a 95% confidence interval for the probability that almost most of the education people from the samples working as a professor.

```
> n=100
> p=0.35
> q=1-p
> s=sqrt((p*q)/n)
> alp=0.05
> z=qnorm(alp/2,0,1,lower.tail=0)
> l=p-z*(s/sqrt(n))
> u=p+z*(s/sqrt(n))
> paste("(",l,",",u,")")
[1] "( 0.340651567608909 , 0.359348432391091 )"
```

PRACTICAL 5

PRACTICAL BASED ON t- test, F- test

1) **(One Sample t-test):** A sample of 13 students from a government school has thefollowing scores in a test.

89 88 78 76 78 78 86 83 82 76 72 77 92.

Do this data support that?

i) The mean mark of the school students is 80? Test at 5% level.

ii) The mean mark of the school students is more than 75? Test at 1% level.

iii) The mean mark of the school students is less than 85? Test at 10% level.

2) **(Two Sample t-test):** The yield of two varieties of mango (in tons) on twoindependent sample of 10 and 12 plants are given below.

Variety-A: 22 24 26 23 26 30 32 34

Variety-B: 28 25 26 30 32 30 33 28 30 35

i) Test whether the yield of Variety-A is not equal to Variety-B at 2% level of significance.

ii) Test whether the difference between yields of Variety-A is less than Variety-B by 2 tonesat 5% level of significance.

iii) Test whether the difference between yield of Variety-A is more than Variety-B by 0.5 tones at 10% level of significance.

iv) Test whether the yield of Variety-A is not equal to Variety-B at 5% level of significance assume unequal variances of both samples.

3) **(Paired t-test):** A new variety of health drink in the market for weight of infants. A sample of 10 babies was selected and was given the above diet for a month and the weights were observed before (X) and after (Y) the diet given.

X : 6.6 6.85 6.75 7.2 6.75 6.65 6.7 7.3 6.9 6.6

Y : 6.9 7.3 7 7.6 6.85 7.3 6.7 7.45 7.3 6.5

i) Examine whether there is significant difference between before and after the healthy drink diet at 5% level of significance.

ii) Examine whether the weight gain after the healthy drink diet is more than 0.2 kg at 1% level of significance.

iii) Examine whether the weight loss after the healthy drink diet is less than 0.5 kg at 10% level of significance.

4) **(F- test):** The yield of two varieties of mango (in tons) on two independent sample of 10and 12 plants are given below.

Variety-A: 22 24 26 23 26 30 32 34

Variety-B: 28 25 26 30 32 30 33 28 30 35

i) Test whether the variance of variety-A is not equal to Variety-B at 5% level ofsignificance.

ii) Test whether the variance of variety-A is greater than Variety-B at 10% level of significance.

iii) Test whether the variance of variety-A is less than Variety-B at 1% level of significance.

SOLUTION

Q.1 One Sample test

i) Here we test, $H_0 : \mu = 80$ against $H_1 : \mu \neq 80$.

```
       x=c(89,88,78,76,78,78,86,83,82,76,72,77,92)    #data

     t.test(x,mu=80) #by default alternative is two sided and level is 5%
```

Output

```
One Sample t-test

data:  x

t = 0.68885, df = 12, p-value = 0.504

alternative hypothesis: true mean is not equal to 80

95 percent confidence interval:77.50427 84.80342

sample estimates:mean of x 81.15385
```

R Output gives the test statistic t, degrees of freedom and P-value.

Here P-value is 0.504>0.05, hence we do not have enough evidence to reject H_0 (i.e. Accept H_0). Output also gives additional information about the confidence interval with sample estimate of μ. Here 95% confidence interval is (77.50427, 84.80342) which also support the decision taken from P-value as 80 is included in the confidence interval.

ii) Here we test, $H_0 : \mu \leq 75$ against $H_1 : \mu > 75$.

```
t.test(x,mu=75,alternative = "greater",cof.level=0.99)
```

Output

```
One Sample t-test

data:  x

t = 3.6739, df = 12, p-value = 0.001592

alternative hypothesis: true mean is greater than 75

95 percent confidence interval:78.16846   Inf

sample estimates:mean of x  81.15385
```

Here P-value is 0.001592<0.01, hence we reject H_0 (i.e. Accept H_1). Output also gives one sided confidence interval with sample estimate of μ which support the decision taken from P-value.

iii) Here we test, $H_0 : \mu \geq 85$ against $H_1 : \mu < 85$.

```
x=c(89,88,78,76,78,78,86,83,82,76,72,77,92)

t.test(x,mu=85,alternative = "less",cof.level=0.9)
```

Output

```
One Sample t-test

data:  x

t = -2.2962, df = 12, p-value = 0.02024

alternative hypothesis: true mean is less than 85

95 percent confidence interval: -Inf 84.13923

sample estimates:mean of x   81.15385
```

Q.2 SOLUTION: TWO SAMPLE t TEST

i) Here we test, $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$

```
x=c(22,24,26,23,26,30,32,34)              #first sample data
y=c(28,25,26,30,32,30,33,28,30,35)        #second sample data
t.test(x,y,var.equal = TRUE, conf.level = 0.98)#by default c=0 and alternative
#hypothesis is two sided
```

OUPUT

```
Two Sample t-testdata:  x and y
t = -1.4607, df = 16, p-value = 0.1634
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval: -7.129169  1.979169
sample estimates: mean of x mean of y 27.125    29.700
```

Here P-value is 0.1634>0.02, hence we do not have enough evidence to reject H_0 (i.e. Accept H_0). Output also give confidence interval of difference of means with sample estimates of μ_1 and μ_2 which support the decision taken from P-value.

ii) Here we test, $H_0: \mu_1 - \mu_2 \geq 2$ against $H_1: \mu_1 - \mu_2 < 2$

```
t.test(x,y,var.equal = TRUE, mu=2,alternative = "less", conf.level = 0.95)
```

Output:

```
      Two Sample t-test
data:  x and y
t = -2.5953, df = 16, p-value = 0.009763
Alternative hypothesis: true difference in means is less than 2
95 percent confidence interval: -Inf 0.5026423
Sample estimates: mean of x mean of y: 27.125   29.700
```

Here P-value is 0.009763<0.05, hence we reject H_0 (i.e. Accept H_1). Output also gives one sided confidence interval of difference of means with sample estimates of μ_1 and μ_2 which support the decision taken from P-value.

iii) Here we test, $H_0: \mu_1 - \mu_2 \leq 0.5$ against $H_1: \mu_1 - \mu_2 > 0.5$

```
t.test(x,y,var.equal = TRUE, mu=0.5,alternative = "greater", conf.level = 0.9)
```

OUTPUT

```
Two Sample t-testdata:  x and y
t = -1.7444, df = 16, p-value = 0.9499
Alternative hypothesis: true difference in means is greater than 0.5
90 percent confidence interval: -4.931434 I nf
Sample estimates: mean of x mean of y 27.125    29.700
```

Here P-value is 0.9499>0.1, hence we do not have enough evidence to reject H_0 (i.e. Accept H_0). Output also give confidence interval of difference of means with sample estimates of μ_1 and μ_2 which support the decision taken from P-value.

iv) Here we test, $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$ where assumption of equality ofvariance of two sample does not hold.

```
t.test(x,y) #by default c=0, alternative hypothesis is two sided and los=5%

            #by default variances are not equal
```

```
OUTPUT

data:  x and y

t = -1.4037, df = 12.172, p-value = 0.1854

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -6.565645  1.415645

sample estimates: mean of x mean of y 27.125
```

Here P-value is 0.1854>0.05, hence we do not have evidence to reject H_0 (i.e. Accept H_0). Output also give confidence interval of difference of means with sample estimates of μ_1 and μ_2 which support the decision taken from P-value.

Q.3 SOLUTION: PAIRED T TEST

i) Here we test, $H_0: \mu_d = \mu_X - \mu_Y = 0$ against $H_1: \mu_d \neq 0$

```
 x=c(6.6,6.85,6.75,7.2,6.75,6.65,6.7,7.3,6.9,6.6) #Before Treatment Data

y=c(6.9,7.3,7,7.6,6.85,7.3,6.7,7.45,7.3,6.5)    #After Treatment Data

t.test(x,y,paired = TRUE) #by default c=0, alternative is two sided and los=5%
```

Output:

```
Paired t-test data:  x
and y

t = -3.6211, df = 9, p-value = 0.005563

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -0.42242786 -0.09757214

sample estimates:

mean of the differences -0.26
```

Here P-value is 0.005563<0.05, hence we reject H_0 (i.e. Accept H_1). Output also gives confidence interval and sample estimate of μ_d which also support the decision taken from P-Value.

ii) Here we test, $H_0: \mu_d = \mu_X - \mu_Y \leq 0.2$ against $H_1: \mu_d > 0.2$

```
x=c(6.6,6.85,6.75,7.2,6.75,6.65,6.7,7.3,6.9,6.6) #Before Treatment Data

y=c(6.9,7.3,7,7.6,6.85,7.3,6.7,7.45,7.3,6.5)     #After Treatment Data

t.test(x,y,paired = TRUE,mu=0.2,conf.level = 0.99,alternative = "greater")
```

```
OUTPUT Paired t-test

data:  x and y

t = -6.4065, df = 9, p-value = 0.9999

alternative hypothesis: true difference in means is greater than 0.2

99 percent confidence interval: -0.4625854      Inf

sample estimates:mean of the differences -0.26
```

Here P-value is 0.9999>0.01, hence we do not have evidence to reject H_0 (i.e. Accept H_0). Output also gives confidence interval and sample estimate of μ_d which also support the decision taken from P-value.

iii) Here we test, $H_0: \mu_d = \mu_X - \mu_Y \geq 0.5$ against $H_1: \mu_d < 0.5$

```
t.test(x,y,paired = TRUE,mu=0.5,conf.level = 0.9,alternative = "less")
```

```
OUTPUT Paired t-test

data:  x and y

t = -10.585, df = 9, p-value = 1.113e-06

alternative hypothesis: true difference in means is less than 0.5

90 percent confidence interval: -Inf -0.1606955

sample estimates: mean of the differences -0.26
```

Here P-value is <0.1, hence we reject H_0 (i.e. Accept H_1). Output also gives confidence intervaland sample estimate of μ_d which also support the decision taken from P-value.

Q.4 Solution: F TEST

i) Here we test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 \neq \sigma_2^2$

```
x=c(22,24,26,23,26,30,32,34)    #first sample data

y=c(28,25,26,30,32,30,33,28,30,35)    #second sample data

var.test(x,y) #by default alternative is two sided and los=5%
```

```
Output: F test to compare two variances

data:  x and y

F = 2.0141, num df = 7, denom df = 9, p-value = 0.3238 alternative hypothesis: true
ratio of variances is not equal to 1

95 percent confidence interval:0.4798759 9.7142569

sample estimates: ratio of variances 2.014062
```

Here P-value is 0.3238>0.05, Hence we do not have enough evidence to reject H_0.(i.e. Accept H_0). Output also gives 95% confidence interval for ratio of variance with their sample estimates which also support the decision taken from P-value.

ii) Here we test $H_0 : \sigma_1^2 \leq \sigma_2^2$ against $H_1: \sigma_1^2 > \sigma_2^2$

```
var.test(x,y,alternative = "greater",conf.level = 0.9)
```

```
F test to compare two variances

data:  x and y

F = 2.0141, num df = 7, denom df = 9, p-value = 0.1619 alternative hypothesis: true
ratio of variances is greater than 1

90 percent confidence interval:0.8039161  Inf

sample estimates: ratio of variances 2.014062
```

Here P-value is 0.1639>0.10, Hence we do not have enough evidence to reject H_0.(i.e. Accept H_0).

iii) Here we test $H_0 : \sigma_1^2 \geq \sigma_2^2$ against $H_1: \sigma_1^2 < \sigma_2^2$

```
var.test(x,y,alternative = "less",conf.level = 0.99)
```

```
F test to compare two variances
```

```
data:  x and y

F = 2.0141, num df = 7, denom df = 9, p-value = 0.8381 alternative
hypothesis: true ratio of variances is less than 1

99 percent confidence interval:0.00000 13.53198

sample estimates: ratio of variances 2.014062
```

Here P-value is 0.8381>0.01, Hence we do not have enough evidence to reject H_0.(i.e. Accept H_0).

PRACTICAL 6

PRACTICAL BASED ON ANALYSIS OF VARIANCE

1) **(ONE WAY):** The grade point average (GPA-4 point scale) of students participating in college sports program are compared .The data are as under.

Football	Tennis	Hockey
3.2	3.8	2.6
2.6	3.1	1.9
2.4	2.6	1.7
2.4	3.9	2.5
1.8	3.2	1.9

Do different sports have significant effect on GPA?

2) Suppose the National Transportation safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full size cars. If collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below. Test whether the mean pressure applied to the drivers head during a crash test is equal for each types of car at 5% level.

Compact Cars	Midsize Cars	Full size cars
643	469	484
655	427	456
702	525	402

3) **(TWO WAY):** Four varieties of wheat are planted at 3 different locations and their yields(units per plot)are recorded as below.:

Variety↓ Location→	Location 1	Location 2	Location 3
Variety1	14.3	7.6	19.2
Variety2	13.4	3.9	12.6
Variety3	18.4	13.4	15.1

Carry out analysis to check whether different locations or different varieties have significant effect on yield of wheat?

4) Four brands of flashlight batteries are to be compared by testing each brand in five flashlights. Twenty flashlights are randomly selected and divided randomly into four groups of five flashlights each. Then each group of flashlights uses a different brand of battery. The lifetimes of the batteries to the nearest hour are given as follows:

Brand A	Brand B	Brand C	Brand D
42	28	24	20
30	36	36	32
39	31	28	38
28	32	28	28
29	27	33	25

Preliminary data analysis indicate that the independent samples can from normal populations with equal standard deviations. At the 5% significance level, does there appear to be a difference in mean lifetime among the four brands of batteries?

5) **(TWO WAY):** A reputed marketing agency in India has three different training programs for its salesmen. The three programs are method – A, B, C. to access the success of the programs, 4 salesmen from each of the programs were sent to the field. The performances in terms of sales are given in the following table:

Salesmen	**Methods**		
	A	**B**	**C**
1	**4**	**6**	**2**
2	**6**	**10**	**6**
3	**5**	**7**	**4**
4	**7**	**5**	**4**

Test whether there is significant difference among methods and among salesmen.

6) **(TWO WAY):** An engineer suspects that surface finish of a metal part is influenced by type ofpaint used and drying time. Drying times are selected by him are 20, 25, 30 minutes and he randomly choses paint I, II. Conducted experiment yielded following data analyses it. Is thereany interaction present between paint and drying time?

Paint↓	**Drying Times(minutes)**		
	20	**25**	**30**
I	**74,64,50**	**73,61,44**	**78,85,92**
II	**92,86,68**	**98,73,88**	**66,45,85**

Q.1 ONE WAY

Solution . Here we apply ANOVA on way as GPA are classified according to one factor =sports

H0: The different sports have no significant effect on GPA

```
#data should be read treatment wise #To read treatments

>GPA=c(3.2,2.6,2.4,2.4,1.8,3.8,3.1,2.6,3.9,3.3,2.6,1.9,1.7,2.5,1.9)

> Sport=c(rep("Football",5),rep("Tennis",5),rep("Hockey",5))

> d=data.frame(Sport,GPA) # anova oneway

> av1=aov(GPA~Sport,data=d)

> summary(av1)
```

OUTPUT

```
              Df     Sum Sq       Mean Sq      F value       Pr(>F)

Sport         2      3.929        1.9647       8.456         0.00511 **

Residuals      12    2.788         0.2323
```

Interpretation: As F calculated is highly significant (**)Treatments differ significantly sports person's GPA differ according sport.

Interpretation: No sport shows significant difference in GPA means

Q.2 SOLUTION

H0: different locations or variety have no significant effect on yield of wheat

```
> #data should be read variety wise
> yield=c(14.3,13.4,18.4,7.6,3.9,13.4,19.2,12.6,15.1)
> loc=c(rep("L1",3),rep("L2",3),rep("L3",3))
> variety=c("V1","V2","V3","V1","V2","V3","V1","V2","V3")
> result=aov(yield~loc+variety)
> summary(result)
```

OUTPUT

```
            Df   Sum Sq    Mean Sq   F value   Pr(>F)
loc         2    103.79     51.89    6.389     0.0568 .
variety     2     49.79    24.89     3.065     0.1559
Residuals   4    32.49     8.12
```

Interpretation: The Calculated F ratio are not significant, as p value is > .05 The yield doesnot change significantly as location changes. Even the differences in varieties do not have significant influence on yield. Varieties do not differ significantly.

Q.3 SOLUTION

H0:The mean lifetimes of brands of batteries are equal.

```
>brand=c(42,30,39,28,29,28,36,31,32,27,24,36,28,28,33,20,32,38,28,25)
> battery=c(rep("Ba",5),rep("Bb",5),rep("Bc",5),rep("Bd",5))
> d=data.frame(battery,brand)
> av2=aov(brand~battery,data=d)
> summary(av2)
```

OUTPUT

```
            Df   Sum Sq    Mean Sq   F value   Pr(>F)
battery    3     68.2       22.73    0.739     0.544
Residuals   16    492.0     30.75
```

Interpretation: At 5% level of significance, there is not enough evidence to conclude that the mean difference lifetimes of the brands of batteries differ.

Q.4 SOLUTION

H01:There is no significant difference among the three programs.

H02: There is no significant difference among the three salesmen.

```
> sales=c(4, 6, 5, 7, 6, 10, 7, 5, 2, 6, 4, 4)
> met=c(rep("M1",4),rep("M2",4),rep("M3",4))
> sm=c("S1","S2","S3","S4","S1","S2","S3","S4","S1","S2","S3","S4")
> d=data.frame(sales,met,sm)
> r1=aov(sales~sm+met, data=d)
> summary(r1)
```

OUTPUT

```
                Df      Sum Sq          Mean Sq         F value         Pr(>F)
sm               3      17               5.667          3.4              0.0943 .
met              2      18              9.000            5.4            0.0456 *
Residuals        6       10              1.667
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

The null hypothesis is not rejected, that is we conclude that there is significant difference in the mean sales among the three programs.

The null hypothesis is rejected, that is we conclude that there does not exist significant difference in the mean sales among the four salesmen.

Q.5 SOLUTION

H0: there is no interaction present between paint and drying time

```
> DT=c(74,64,50,92,86,68,73,61,44,98,73,88,78,85,92,66,45,85)
> paint=c(rep("I",3),rep("II",3))
> DRT1=c(paint)
> DRT2=c(paint)
> DRT3=c(paint)
> DRT=c("DRT1","DRT2","DRT3")
> d=data.frame(DT,paint,DRT)
> fit=aov(DT~paint*DRT,data=d)
> summary(fit)
```

OUTPUT

```
                 Df     Sum Sq          Mean Sq         F value         Pr(>F)
   paint          1       356          355.6            1.250            0.285
   DRT           2       421            210.4            0.740            0.498
   paint:DT    2          315          157.4            0.553           0.589
   Residuals    12      3413            284.4
```

Interpretation: Interaction between drying time and paint is not significant. We canperform test for equality of paint means or for drying time means. Using error or error +interaction S.S.

i) H_{0A}: $\alpha_1 = \alpha_2 = \ldots\ldots\ldots \alpha_p = 0$ against H_{1A}: paints differ significantly .ii)Since calculated F ratio $< F_{\alpha,p-1,n-1}$, so H_{0A} is not rejected. We conclude that means of paints do not differ significantlyat confidence level 5 %.

ii) H_{0B}: $\beta_1 = \beta_2 = \ldots\ldots\ldots \beta_q = 0$ against H_{1B}: Drying times differ significantly.

ii) Here calculated F ratio $< F_{\alpha,q-1,n-1}$,so H_{0B} is not rejected. We conclude that means of Drying times do not differ significantly at 5 %.

PRACTICAL 7

PRACTICAL BASED ON NON – PARAMETRIC TEST I

1) **(Sign test)** It is known from the past experience that the median length of Sunfish in a particular polluted lake was 3.9 inches. During the past two years the lake was cleaned up and the conjecture is made that now median length is greater than 3.9 inches. A random sample of 10 sunfish selected from this lake showed lengths as 5.2, 4.1, 5.4, 5.7, 3.0, 6.3, 6.6, 2.8, 1.9, 4.5inches. Will you reject the null hypothesis at 10 % level of significance (l.o.s.) on the basis ofSign Test?

2) **(Wilcoxon Sign test):** A random sample of 10 infants showed the following pulse rates per minute: 110,121,125,122,112,117,129,114,124,127.Assuming that the distribution of pulse rates is symmetric. Is there any evidence to suggest that the median pulse rate of infants is more than 120 beats per minute? Use Wilcoxon's signed rank test at 5% l.o.s.

3) **(Wilcoxon Sign test for Paired sample):** Test scores of a group of 15 high – school students before &after a training program are as given below :

Score before	63	75	78	84	58	58	70	76	74	88	74	94	99	79	93
Score after	84	86	75	94	50	95	97	98	72	100	101	98	105	84	90

Use appropriate statistical test at 1%l.o.s to check if the training has any effect on the test scores.

Q .1 SOLUTION

```
>data=c(5.2, 4.1, 5.4, 5.7, 3.0, 6.3, 6.6, 2.8, 1.9, 4.5)
> SIGN.test(data, md=3.9, alternative="greater", conf.level=0.95)
```

OUTPUT

```
        One-sample Sign-Test
data:  data
s = 7, p-value = 0.1719
alternative hypothesis: true median is greater than 3.9
95 percent confidence interval:  2.978667      Inf
sample estimates: median of x    4.85
Achieved and Interpolated Confidence Intervals:

                          Conf.Level      L.E.pt          U.E.pt
Lower Achieved CI         0.9453          3.0000          Inf
Interpolated CI            0.9500         2.9787           Inf
Upper Achieved CI         0.9893         2.8000            Inf
```

Interpretation: Since p-value =0.1719 > 0.05 indicates one should not reject null hypothesis.

Q.2 SOLUTION

```
> x = c(110,121,125,122,112,117,129,114,124,127)
> wilcox.test(x, y=NULL, alternative='greater', mu=120, paired=FALSE, exact = NULL, correct =T,
conf.level=0.95)
```

OUTPUT

```
        Wilcoxon signed rank exact test
data:  x
```

```
V = 28, p-value = 0.5
alternative hypothesis: true location is greater than 120
```

Interpretation: Since p-value =0.1719 > 0.05 indicates one should not reject null hypothesis.

Q.3 SOLUTION

```
> x=c(63,75,78,84,58,58,70,76,74,88,74,94,99,79,93)
> y=c(84,86,75,94,50,95,97,98,72,100,101,98,105,84,90)
> wilcox.test(x,y,paired=TRUE, alternative='less',exact = NULL, correct=T, conf.level=0.99)
```

OUTPUT

```
        Wilcoxon signed rank test with continuity correction
data:  x and y
V = 13, p-value = 0.00412
alternative hypothesis: true location shift is less than 0
Warning message:
In wilcox.test.default(x, y, paired = TRUE, alternative = "less",  :
  cannot compute exact p-value with ties
```

Interpretation: Since p-value =0.0007523 < 0.05 indicates one should reject null hypothesis.

Interpretation: Since p-value =0.00412 < 0.01 indicates one should reject null hypothesis.

PRACTICAL 8

PRACTICAL BASED ON NON – PARAMETRIC TEST II

1) **(Kruskal Wallis Test):** Test if there exists a significance of difference between the scores of three groups when compared against each other for the following given data set. Use 5% l. o. s. Alsouse post-hoc test to find the exact significance.

Group 1	Group 2	Group 3
63	84	74
75	86	76
78	75	65
84	94	84
58	50	50
58	95	85
70	97	97
76	98	88
74	72	72
88	100	90
74	101	101
94	98	98
99	105	115
79	84	94
93	90	90

2) **(Chi- Square Goodness of fit test):** A shop owner claims that an equal number of customers come into his shop each weekday. To test this hypothesis, a researcher records the number of customers that come into the shop in a given week and finds the following:

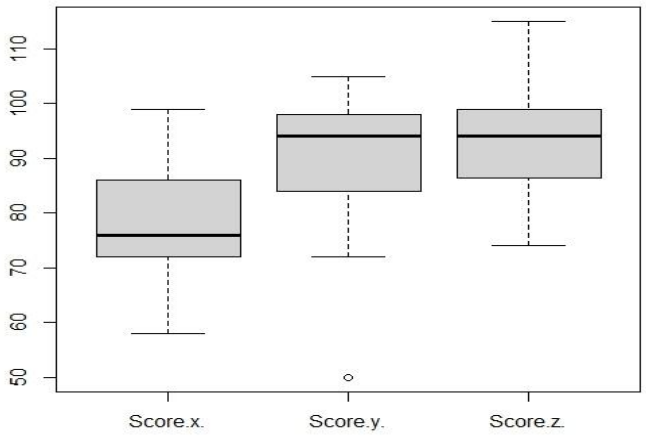
Days	Monday	Tuesday	Wednesday	Thursday	Friday
No. of customer	50	60	40	47	53

3) **(Test in r × c Contingency table)** Using the data given below decide whether we can conclude that standard of a salesman has efficient effect on HD performance if field selling at 5% level of significance

	Performance in field			Total
	Disappointing	Satisfactory	Excellent	
Poor dressed	21	15	6	42
Well dressed	24	35	26	85
Very well dressed	35	80	58	173
Total	80	130	90	300

Q.1 SOLUTION

```
> data<-read.csv("C:/Users/hp/OneDrive/Desktop/Statistics/kw.csv")
> data
   Score.x. Score.y. Score.z.
1       63      84      74
2       75      86      76
3       78      75     100
4       84      94      84
5       58      50     110
6       58      95      85
7       70      97      97
8       76      98      88
9       74      72      95
10      88     100      90
11      74     101     105
12      94      98      98
13      99     105     115
14      79      84      94
15      93      90      90
> boxplot(data)
```



```
> kwtest<-kruskal.test(data)
> kwtest
```

OUTPUT

```
        Kruskal-Wallis rank sum test

data:  data
Kruskal-Wallis chi-squared = 10.044, df = 2, p-value = 0.006593
```

Interpretation: Since p-value =0.006593 < 0.01 indicates one should reject null hypothesis and conclude that there exists significance of difference between the scores of three group at 1% l.o.s. To find exact significance of difference we used post-hoc test comparison

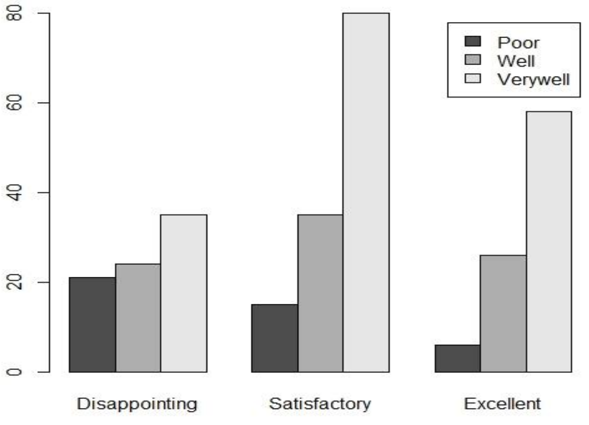
Q.2 SOLUTION

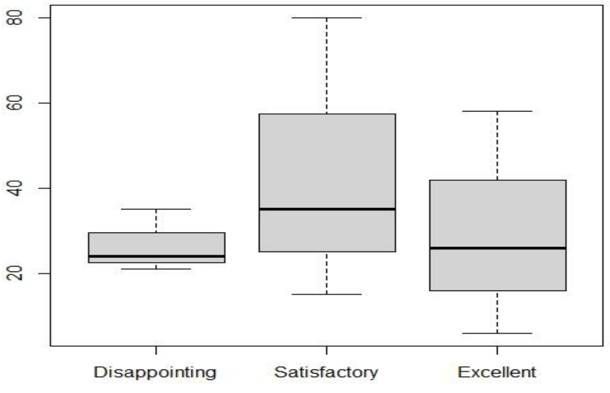
```
> obs<-c(50,60,40,47,53)
> exp<-c(0.2,0.2,0.2,0.2,0.2)
> chisq.test(x=obs, p=exp)
      Chi-squared test for given probabilities
data:  obs
X-squared = 4.36, df = 4, p-value = 0.3595
```

Q.3 SOLUTION

H0: The attributes are independent.

```
> mytable=matrix(c(21,15,6,24,35,26,35,80,58), byrow=TRUE, ncol=3)
> colnames(mytable)=c("Disappointing", "Satisfactory","Excellent")
> rownames(mytable)=c("Poor","Well","Verywell")
> mytable
                Disappointing          Satisfactory    Excellent
Poor             21                     15                6
Well              24                     35               26
Verywell         35                     80               58
> chisq.test(mytable,correct=FALSE)
      Pearson's Chi-squared test
data:  mytable
X-squared = 16.516, df = 4, p-value = 0.002399
>barplot(mytable, beside=T, legend=T)
>boxplot(mytable)
```





```
> chisq.test(mytable)
      Pearson's Chi-squared test
data:  mytable
X-squared = 16.516, df = 4, p-value = 0.002399
```

Interpretation:

Since p-value is less than l.o.s hence we reject Ho at 5% level of significance.

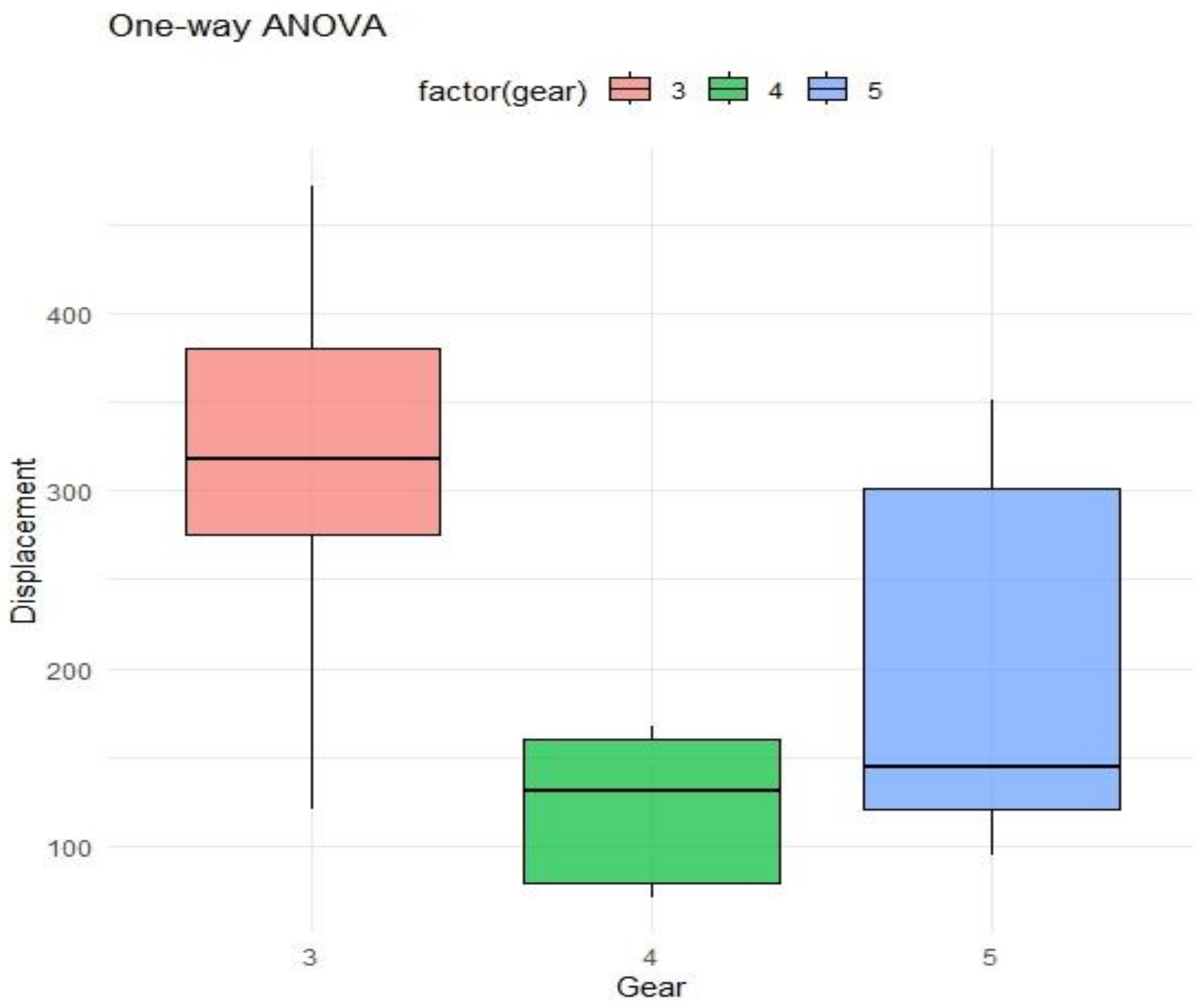
PRACTICAL 8

PRACTICAL BASED ON POST HOC TEST ON ONE WAY ANALYSIS

The mtcars(motor trend car road test) dataset is used which consist of 32 car brands and 11 attributes. The dataset comes preinstalled in dplyr package in R.

Ho: There is no difference between average displacement for different gear.

```
> library(dplyr)
Error in library(dplyr) : there is no package called ‘dplyr’
> head(mtcars)
                    Mpg   cyl    disp   hp     drat    wt     qsec   vs     am    gear   carb
Mazda RX4            21.0   6     160    110    3.90   2.620  16.46  0      1      4      4
Mazda RX4 Wag       21.0    6      160   110    3.90   2.875  17.02  0      1      4      4
Datsun 710          22.8    4     108     93    3.85   2.320  18.61  1      1      4      1
Hornet 4 Drive       21.4   6      258   110    3.08   3.215  19.44  1      0      3      1
Hornet Sportabout   18.7    8      360   175    3.15   3.440  17.02  0      0      3      2
Valiant              18.1   6      225   105    2.76   3.460  20.22  1      0      3      1
> mtcars_aov<-aov(mtcars$disp~factor(mtcars$gear))
> summary(mtcars_aov)
               Df Sum Sq Mean Sq F value   Pr(>F)
factor(mtcars$gear)  2 280221  140110   20.73 2.56e-06 ***
Residuals          29 195964    6757
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mtcars_aov2<-aov(mtcars$disp~factor(mtcars$gear)*factor(mtcars$am))
> summary(mtcars_aov2)
               Df Sum Sq Mean Sq F value   Pr(>F)
factor(mtcars$gear)  2 280221  140110  20.695 3.03e-06 ***
factor(mtcars$am)    1   6399    6399   0.945    0.339
Residuals          28 189565    6770
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
 > plot1<-ggplot(mtcars, aes(x=factor(gear), y=disp,fill=factor(gear)))+
+ geom_boxplot(color="black",alpha=0.7)+
+ labs(title="One-way ANOVA",x="Gear",y="Displacement")+
+ theme_minimal()+
+ theme(legend.position="top")
> plot1
```


One-way ANOVA
factor(gear)
3
4
5
400
300
200
100
Displacement
3
4
5
Gear