STATISTICAL COMPUTATION ASSIGNMENT-2

Submitted by: Meghal Dani
Roll no.: MT17144

SC - ASSIGNMENT 2

MEGHAL DANI
MT17144

Q1 $N = 32$

$k = 4$

Let $u_1, u_2, u_3, \& u_4$ be mean of each group

$$u_1 = \frac{62+81+75+50+67+40+46+36}{8}$$

$= 57.125$

$$u_2 = \frac{72+49+53+68+39+69+40+15}{8}$$

$= 50.625$

$$u_3 = \frac{42+52+31+70+22+61+68+76}{8}$$

$= 52.75$

$$u_4 = \frac{80+57+87+54+28+29+52+45}{8}$$

$= 54$

$\overline{X}$ (overall mean) $= \frac{1716}{32} = 53.625.$

degree of freedom $= K-1 = 3$

$N-k = 32 - 4 = 28$

$$\boxed{f\ score = \frac{\Sigma\, u_j(\overline{x}_j - \overline{x})^2 / (k-1)}{\Sigma\Sigma\, (x - \overline{x}_j)^2 / (N-k)}}$$ → formula used

$$\sum_{j=1}^{4} n_j(\bar{x}_j - \bar{x})^2 = 8(57.125-53.625)^2 + 8(50.625-53.625)^2 + 8(52.75-53.627)^2$$
$$+ 8(54-53.625)^2$$
$$= 177.25$$

$$\sum_{j=1}^{4}\sum (x-\bar{x}_j)^2 = \left[(62-57.125)^2 + (81-57.125)^2 + (75-57.125)^2 + (50-57.125)^2 + (67-57.125)^2 + (40-57.125)^2 + (46-57.125)^2 + (36-57.125)^2\right]$$
$$+ \left[(42-50.625)^2 + (49-50.625)^2 + (53-50.625)^2 + (68-50.625)^2 + (39-50.625)^2 + (69-50.625)^2 + (40-50.625)^2 + (15-50.625)^2\right] +$$
$$+\left[(42-52.75)^2 + (52-52.75)^2 + (31-52.75)^2 + (70-52.75)^2 + (22-52.75)^2 + (61-52.75)^2 + (68-52.75)^2 + (76-52.75)^2\right]$$
$$+\left[(80-54)^2 + (57-54)^2 + (87-54)^2 + (54-54)^2 + (28-54)^2 + (29-54)^2 + (52-54)^2 + (45-54)^2\right]$$
$$= 10379.48438$$

$$F = \frac{177.25/3}{10379.48438/28} = 0.159$$

1) Computed answer : $F = 0.159(3,28)$, $p > 0.05$ (accept null hypothesis)

2) Null hypothesis :- There will be no difference some where in history scores between students of four groups with different academic major.

3) Alternate hypothesis:- There will be difference in history scores between 4 groups.

4) probability level chosen :- $p = 0.05$ as there is little risk involved if either Type I or a type II major is made.

5) degrees of freedom :- 3,28

6) Yes, there is significant difference found between 4 groups in terms of performance on history exam.

7) Students with different academic major performed differently in history exam

Q2

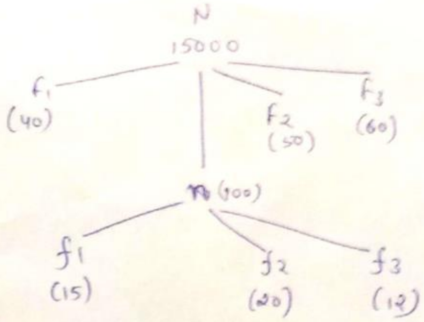
$N = 15000$

$n = 100$

$f_1 = 40$ $\quad f_1 = 15$

$f_2 = 50$ $\quad f_2 = 20$

$f_3 = 60$ $\quad f_3 = 12.$



Hypergeometric-Distribution

p-value of enrichment of f_1 $= \sum_{i=15}^{40} \dfrac{{}^{40}C_i \; {}^{14960}C_{100-i}}{{}^{15000}C_{100}}$ [summation of hyper-geometric distⁿ]

$= \boxed{2.681 \times 10^{-23}}$ (using R-code & online calculat

p-value of enrichment of f_2 $= \sum_{i=20}^{50} \dfrac{{}^{50}C_i \; {}^{14950}C_{100-i}}{{}^{15000}C_{100}}$

$= \boxed{1.606 \times 10^{-31}}$

p-value of enrichment of f_3 $= \sum_{i=12}^{60} \dfrac{{}^{60}C_i \; {}^{14940}C_{100-i}}{{}^{15000}C_{100}}$

$= \boxed{4.198 \times 10^{-15}}$

Rcode:

f1:

```
f1 = sum(dhyper(15:40, 40, 14960, 100))
print(f1)                                   // o/p :- 2.681959e-23
f2 = sum(dhyper(20:50, 50, 14950, 100))
print(f2)                                   // o/p :- 1.606048e-31
f3 = sum(dhyper(12:60, 60, 14940, 100))
print(f3)                                   // o/p :- 4.198369e-15
```

Q3

	Math	Science	Total
fail	2	4	6
Pass	6	7	13
Total	8	11	19

$n_1 = 6$ ie Σ fail (Math) + Σ fail (Science) is the smallest marginal total.

Thus we look at following ordered pairs of (n_{11}, n_{12}) :-

(0,6) (1,5) (2,4) (3,3) (4,2) (5,1) (6,0)

~~$P_{n_{i,j}}$ = (Math fail)! (Math pass)! Sci~~

$$p_{n_{i,j}} = \frac{(\Sigma \text{fail})! \, (\Sigma \text{pass})! \, (\Sigma \text{Math})! \, (\Sigma \text{Science})!}{\text{Total}! \, n_{11}! \, n_{12}! \, n_{21}! \, n_{22}!}$$ formula used.

0	6
8	5

$$\frac{6! \, 13! \, 8! \, 11!}{19! \, 0! \, 6! \, 8! \, 5!} = 0.01702$$

1	5
7	6

$$\frac{6! \, 13! \, 8! \, 11!}{19! \, 1! \, 5! \, 7! \, 6!} = 0.13622$$

2	4
6	7

$$\frac{6! \, 13! \, 8! \, 11!}{19! \, 2! \, 4! \, 6! \, 7!} = 0.34055$$

3	3
5	8

$$\frac{6! \, 13! \, 8! \, 11!}{19! \, 3! \, 3! \, 5! \, 8!} = 0.34055$$

4	2
4	9

$$\frac{6! \, 13! \, 8! \, 11!}{19! \, 4! \, 2! \, 4! \, 9!} = 0.14189$$

5	1
3	10

$$\frac{6!\,13!\,8!\,11!}{19!\,5!\,1!\,3!\,10!} = 0.022703$$

6	0
2	11

$$\frac{6!\,13!\,8!\,11!}{19!\,6!\,0!\,2!\,11!} = 0.001031$$

n_{21}	$P_{n_{ij}}$
8	0.01702
7	0.13622
6	0.34055
5	0.34055
4	0.14189
3	0.022703
2	0.001031

Since $n_{12} = 6$, p-value is sum of probability ≤ 0.34055 $= 0.999964$
≈ 1

Thus we can't ~~accept~~ reject the null hypothesis, we accept null hypothesis and say fraction of pass in math is independent of science.

Rcode :-

```
Values ← c("Math", "Science")
Fail ← c(2,4)
Pass ← c(6,7)
data ← data.frame(Fail, Pass)
colnames(data) ← c("Fail", "Pass")
fisher.test(data)
```

O/P pvalue = 1

Q4

No. of sixes in set (5 throws)	Observed no. of sets (O_i)	Expected no. of sets (E_i)
0	4	$^5C_0 (1/6)^0 (5/6)^5 \times 40 = 16.075$
1	6	$^5C_1 (1/6)^1 (5/6)^4 \times 40 = 16.075$
2	7	$^5C_2 (1/6)^2 (5/6)^3 \times 40 = 6.43$
3	5	$^5C_3 (1/6)^3 (5/6)^2 \times 40 = 1.286$
4	10	$^5C_4 (1/6)^4 (5/6)^1 \times 40 = 0.128$
5	8	$^5C_5 (1/6)^5 (5/6)^0 \times 40 = 0.005$
	$\Sigma = 40$	

Method 1

Applying Chi-square Test :-

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \quad \text{formula used}$$

$$= \frac{(4-16.075)^2}{16.075} + \frac{(6-16.075)^2}{16.075} + \frac{(7-6.43)^2}{6.43} + \frac{(5-1.286)^2}{1.286} + \frac{(10-0.128)^2}{0.128} + \frac{(8-0.005)^2}{0.005}$$

$$= 13571.543$$

p-value < 0.00001
≈ 0.

Thus, we reject the null hypothesis & say dice is biased

Method-2

Applying G-test :-

$$G = 2 \sum_{i=1}^{6} O_i \log\left(\frac{O_i}{E_i}\right) \quad \text{formula used.}$$

$$= 2\left[4\log\left(4/16.075\right) + 6\log\left(6/16.075\right) + 7\log\left(7/6.43\right) + 5\log\left(5/1.286\right) + 10\log\left(\frac{10}{0.128}\right) + 8\log\left(8/0.005\right)\right]$$

$$= 85.55$$

p value < 0.00001 ≈ 0.

Thus we reject the null hypothesis.

Q5(a)

No. of Dice	Observed Count	Expected Count
1	20	28.33
2	20	28.33
3	20	28.33
4	40	28.33
5	40	28.33
6	30	28.33
	$\Sigma = 170$	

Expected count of each $E(p) = \frac{170}{6} = 28.33$.

$\left[\because p = \text{probability} \times \text{count} = \frac{1}{6} \times 170 \right]$

Applying chi-square Test :-

$$X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(20-28.33)^2}{28.33} \times 3 + \frac{(40-28.33)^2}{28.33} \times 2 + \frac{(30-28.33)^2}{28.33}$$

$$= 17.03$$

p-value = 0.004 < 0.05

Therefore, we reject the null hypothesis and conclude that dice is biased.

Q5(b) The dice is similar to mine if there is found correlation in test of independence.

$$DF \text{ (degree of freedom)} = (\text{no. of groups} - 1)(\text{sets} - 1)$$
$$= (2-1)(6-1)$$
$$= 5$$

Applying Chi-square test of inpendence:-

	Count(1)	count(2)	count(3)	count(4)	count(5)	count(6)	Total
Me	20	20	20	40	40	30	170
Friend	10	5	10	20	30	20	95
Total	30	25	30	60	70	50	265

$E_{Me,1} = \frac{30 \times 170}{265} = 19.24$ $\quad E_{f,1} = \frac{30 \times 95}{265} = 10.75$

$E_{Me,2} = \frac{25 \times 170}{265} = 16.03$ $\quad E_{f,2} = \frac{25 \times 95}{265} = 8.96$

$E_{Me,3} = \frac{30 \times 170}{265} = 19.24$ $\quad E_{f,3} = \frac{30 \times 95}{265} = 10.75$

$E_{Me,4} = \frac{60 \times 170}{265} = 38.49$ $\quad E_{f,4} = \frac{60 \times 95}{265} = 21.50$

$E_{Me,5} = \frac{70 \times 170}{265} = 44.90$ $\quad E_{f,5} = \frac{70 \times 95}{265} = 25.09$

$E_{Me,6} = \frac{50 \times 170}{265} = 32.07$ $\quad E_{f,6} = \frac{50 \times 95}{265} = 17.92$

$$X^2 = \frac{(20-19.24)^2}{19.24} + \frac{(20-16.03)^2}{16.03} + \frac{(20-19.24)^2}{19.24} + \frac{(40-38.49)^2}{38.49} + \frac{(40-44.90)^2}{44.90} + \frac{(30-32.07)^2}{32.07}$$
$$+ \frac{(10-10.75)^2}{10.75} + \frac{(5-8.96)^2}{8.96} + \frac{(10-10.75)^2}{10.75} + \frac{(20-21.5)^2}{21.5} + \frac{(30-25.09)^2}{25.09} + \frac{(20-17.92)^2}{17.92}$$
$$= 4.67.$$

p (df - 5) > 0.05

Thus we accept the null hypothesis & say there is no correlation between me & my friend's dice ie, they are not similar but independent

$\boxed{Q5(c)}$ To find correlation with my Dice & friend(B) Dice, we will again use test of independence but fischer's test of independence as the number is small

Contigency table is :-

	count 1	count 2	count 3	count 4	count 5	count 6	Total
Your Dice	2	2	3	4	4	3	18
Friend Dice	1	2	2	3	3	3	14
Total	3	4	5	7	7	6	32

pvalue = Σ (all ordered pairs possible)

= 1 [using R program below] (also verified from online calculator)

Probability = 9.4×10^{-3}

Thus we accept the null hypothesis & say there is no correlation between friend (B) & my dice. ie, they are independent.

$\boxed{\text{Rcode}}$:-

```
Me ← c(2,2,3,4,4,3)
friend ← c(1,2,2,3,3,3)
data ← data.frame(Me, friend)
fisher.test(data)            // O/P :- p-value = 1
```

Q6) **R Code For data 1 :**

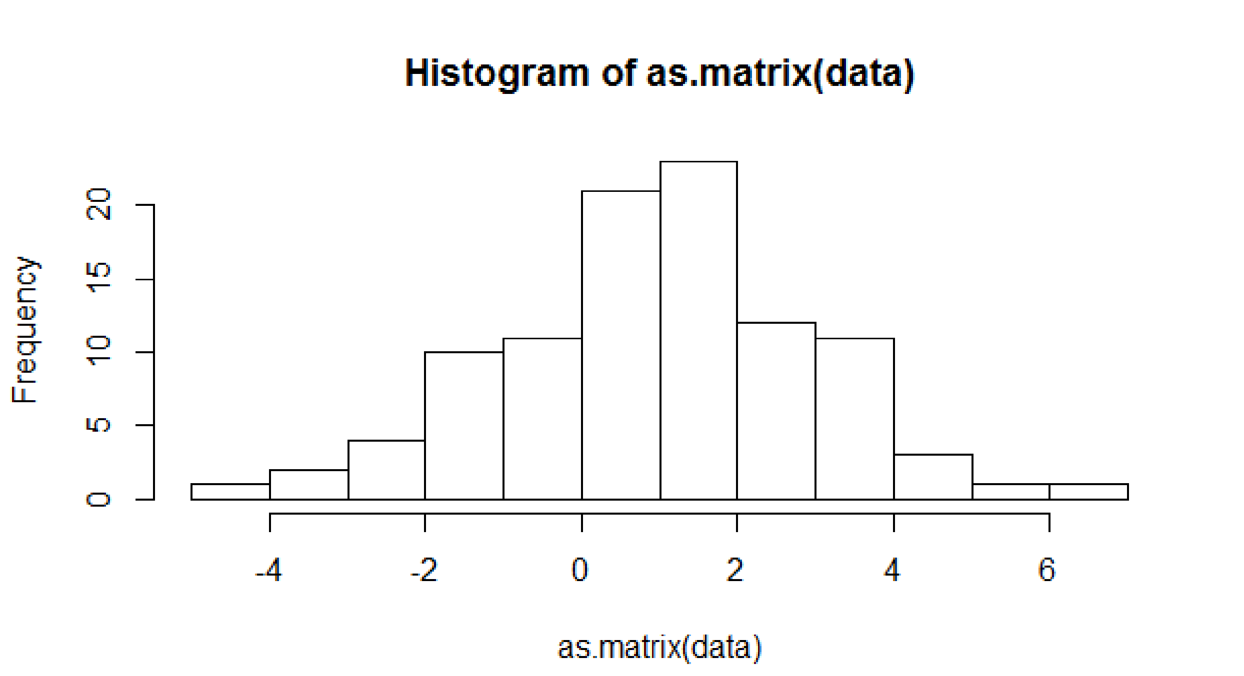
```
data = read.table("a2_d1.txt",header = FALSE,sep = "\n")
h <- hist(as.matrix(data))
h$breaks
h$equidist
m <- mean(as.matrix(data))
s_dev <- sd(as.matrix(data))

h$counts #observed_counts
#test for normal distribution
#find expected counts :
ex <- c()
for(i in 1:length(h$breaks)-1){
  ex_1 = pnorm(h$breaks[i+1],m,s_dev)
  ex_2 = pnorm(h$breaks[i],m,s_dev)
  exp = ex_1 - ex_2
  ex <- append(ex,exp)
}
exp_num <- length(as.matrix(data)) * ex
X <- sum((h$counts - exp_num)*(h$counts - exp_num) / (exp_num))
dof <- length(h$breaks) - 3    ## degree of freedom
p <- pchisq(X, df = dof)

#using library
library(fitdistrplus)
temp <- c(as.matrix(data))
FITN <- fitdist(temp,"norm")
plot(FITN)
```

Output:

P-value : 0.1233339 (accept the null hypothesis and say the distribution is normal/ gaussian)

The plots generated are as follows which shows Q-Q plot following expected values and is a straight line proving the distribution is gaussian.

Plots:



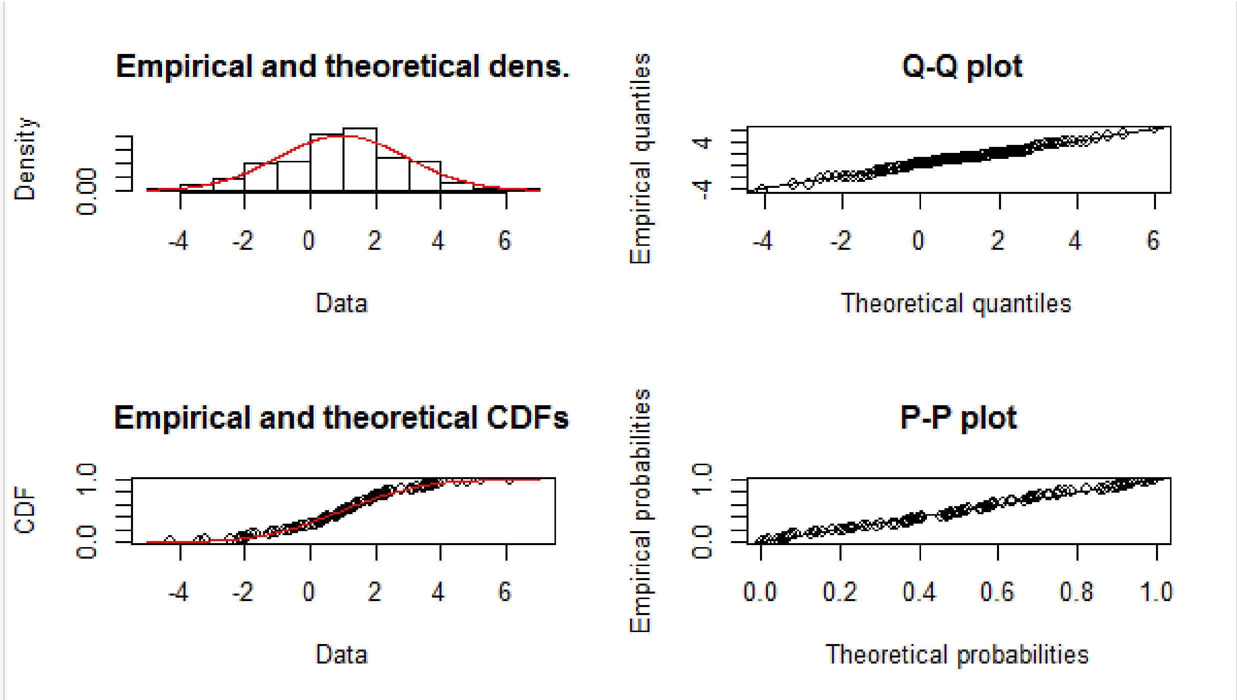


Figure a:Histogram of the data
Figure b:Fitdist output for normal distribution(Density plot, Q-Q plot, CDF plot and P-P plot)

R Code for Data 2:

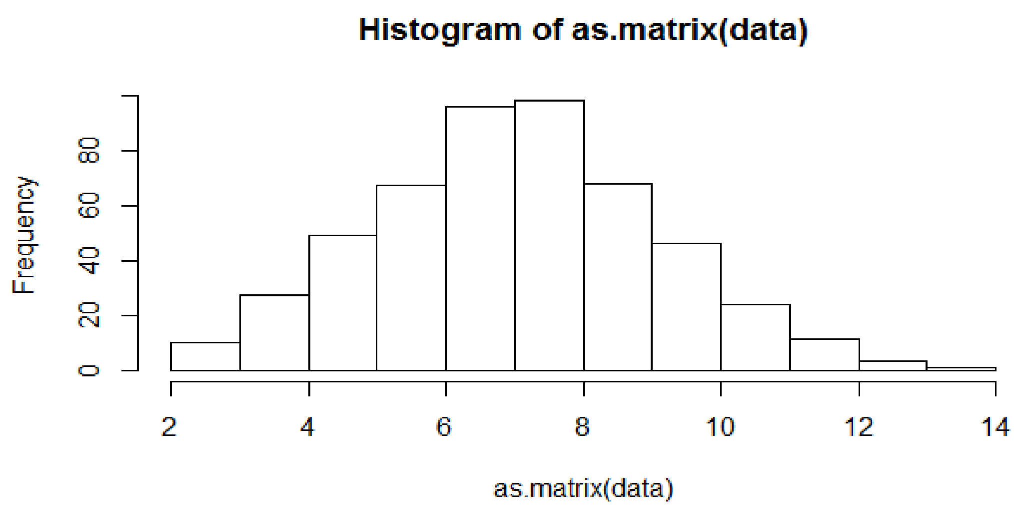
```
data = read.table("a2_d2.txt",header = FALSE,sep = "\n")
h <- hist(as.matrix(data))
h$breaks
h$equidist
m <- mean(as.matrix(data))
s_dev <- sd(as.matrix(data))

h$counts #observed_counts
#test for normal diistribution
#find expected counts :
ex <- c()
for(i in 1:length(h$breaks)-1){
  ex_1 = pnorm(h$breaks[i+1],m,s_dev)
  ex_2 = pnorm(h$breaks[i],m,s_dev)
  exp = ex_1 - ex_2
  ex <- append(ex,exp)
}
exp_num <- length(as.matrix(data)) * ex
X <- sum((h$counts - exp_num)*(h$counts - exp_num) / (exp_num))
dof <- length(h$breaks) - 3    ## degree of freedom
p <- pchisq(X, df = dof)

library(fitdistrplus)
temp <- c(as.matrix(data))
FITN <- fitdist(temp,"norm")
plot(FITN)


#test for poisson distribution
ex <- c()
for(i in 1:length(h$breaks)-1){
  ex_1 = ppois(h$breaks[i+1],m)
  ex_2 = ppois(h$breaks[i],m)
  exp = ex_1 - ex_2
  ex <- append(ex,exp)
}
exp_num <- length(as.matrix(data)) * ex
X <- sum((h$counts - exp_num)*(h$counts - exp_num) / (exp_num))
dof <- length(h$breaks) - 1    ## degree of freedom
p <- pchisq(X, df = dof)


#using library:
library(fitdistrplus)
temp <- c(as.matrix(data))
FITP <- fitdist(temp,"pois")
plot(FITP)
```

Output :
P-value for Normal Distribution : 0.9998739 (accept null hypothesis that the distribution is gaussian)

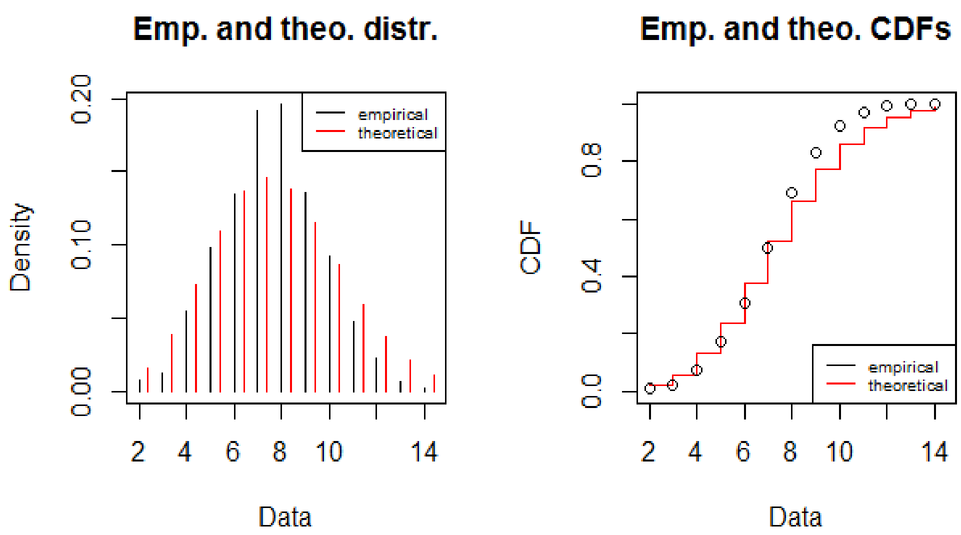
P-value for Poisson Distribution :0.9999729 (accept null hypothesis that the distribution is poisson)

Checking the above using fitdist library we get following plots.

Plots:
These plots deviate from normal and poisson distribution a little as seen in Q-Q plots respectively but, also follows them to some extent which proves the p-values calculated in above code using chi-square method.



.
Figure: Histogram of the data



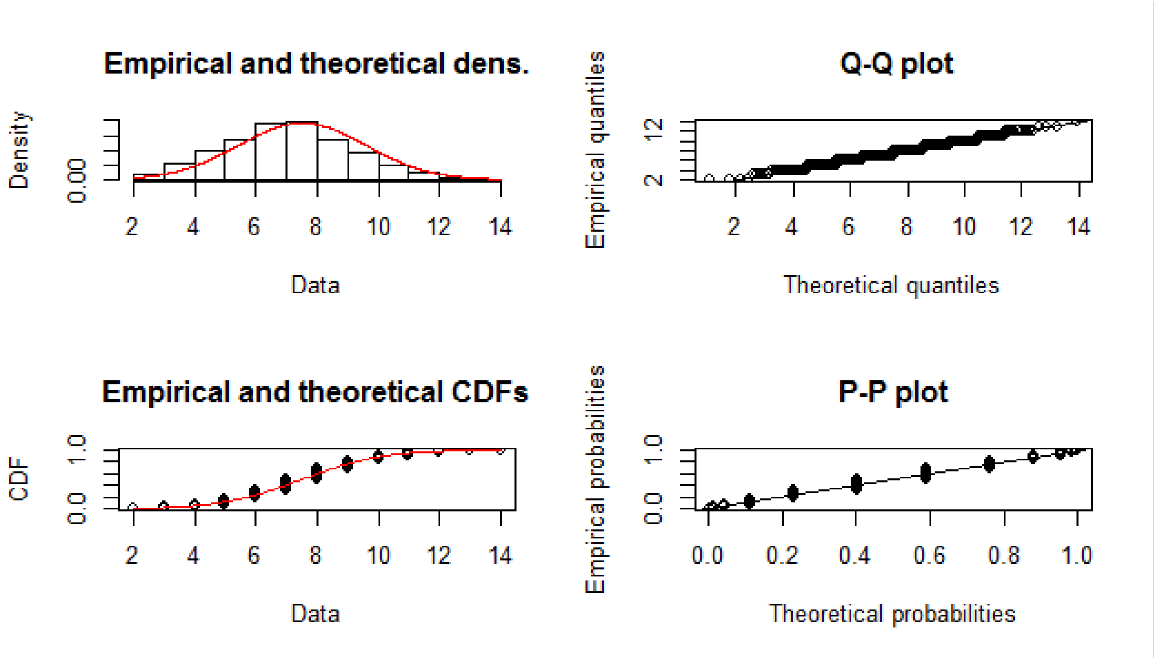


Figure a: Fitdist output for poisson distribution(Density plot and CDF plot)

Figure b: Fitdist output for normal distribution(Density plot, Q-Q plot, CDF plot and P-P plot)

Q7 CSE = 300 ECE = 200 CB = 30

SC: CSE : 30 ECE : 20 CB : 10

GT : CSE : 50 ECE : 20 CB : 5

AI : CSE : 100 ECE : 30 CB : 8

(i) likelihood of combination in all three subjects :-

SC :- $$\frac{{}^{300}C_{30} \times {}^{200}C_{20} \times {}^{30}C_{10}}{{}^{530}C_{60}} = 0.000078$$

GT :- $$\frac{{}^{300}C_{50} \times {}^{200}C_{20} \times {}^{30}C_{5}}{{}^{530}C_{75}} = 0.002$$

AI :- $$\frac{{}^{300}C_{100} \times {}^{200}C_{30} \times {}^{30}C_{8}}{{}^{530}C_{138}} = 0.00000025$$

(ii) P-value of enrichment of CB in each subject :-

SC :- $$\sum_{i=10}^{30} \frac{{}^{30}C_{i} \; {}^{500}C_{60-i}}{{}^{580}C_{60}} = 0.0008088$$

GT : $$\sum_{i=5}^{30} \frac{{}^{30}C_{i} \; {}^{500}C_{75-i}}{{}^{530}C_{75}} = 0.422$$

AI : $$\sum_{i=8}^{30} \frac{{}^{30}C_{i} \; {}^{500}C_{138-i}}{{}^{530}C_{138}} = 0.54$$

R-code :- (for p-value enrichment of CB using hypergeometric distribution)

SC: `sc = sum(dhyper(10:30, 30, 500, 60))` O/P :- 0.0008088517

GT : `gt = sum(dhyper(5:30, 30, 500, 75))` O/P :- 0.4229842

AI : `ai = sum(dhyper(8:30, 30, 500, 138))` O/P :- 0.5407916