STATISTICAL COMPUTATION ASSIGNMENT 1
Submitted by: Meghal Dani

Q1 Here the standard deviation of population is unknown & assuming data follows normal distribution,
- sample size > 40
- standard dev. of sample is known.

We use t-statistics in such a case, and calculate t-score :-

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ = sample mean
μ = population mean
s = standard deviation of sample
n = sample size.

degree of freedom = $n-1$

(a) Given :- $\mu = 300$
$\bar{x} = 190$
$s = 50$
$n = 20$

Solution :- degree of freedom = $n-1$
$= 19$

$$t = \frac{190 - 300}{50/\sqrt{20}}$$

$$= \frac{-110}{11.18}$$

$$= -9.83$$

p value from t-calculator is < 0.00001.

Thus the result is not significant & we reject the null hypothesis.

The probability of bulbs having average life of no more than 190 days < 0.00001

[p value calculated can be viewed from R program given below]

R code Q1a:

```r
m_population = 300 #mean of the population
s_sample = 50 #standard deviation of sample
n = 20   #sample size
m_sample = 190 #mean of the sample
t = (m_sample - m_population)/(s_sample/sqrt(n)) #t score
t
#p-value calculation
p_val = pt(-abs(t),df= n-1) #one sided t-test for finding lesser than or greater tha
p_val
```

Output:

```
Console   Terminal
~/
> m_population = 300 #mean of the population
> s_sample = 50 #standard deviation of sample
> n = 20   #sample size
> m_sample = 190 #mean of the sample
> t = (m_sample - m_population)/(s_sample/sqrt(n)) #t score
> t
[1] -9.838699
> #p-value calculation
> p_val = pt(-abs(t),df= n-1) #one sided t-test for finding lesser than or greater than
> p_val
[1] 3.417344e-09
```

(b) Given: $\bar{x} = 400$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{400 - 300}{50/\sqrt{20}}$$

$$= \frac{100}{11.18}$$

$$= 8.944$$

degree of freedom = 19

p value < 0.0001 for not more than 400

p value ≈ 0.999 for more than 400

R code Q1b:

```r
m_population = 300 #mean of the population
s_sample = 50 #standard deviation of sample
n = 20    #sample size
m_sample = 400 #mean of the sample
t = (m_sample - m_population)/(s_sample/sqrt(n)) #t score
t
#p-value calculation
p_val = pt(-abs(t),df= n-1) #one sided t-test for finding lesser than or greater tha
p_val
```

Output:

```
> m_population = 300 #mean of the population
> s_sample = 50 #standard deviation of sample
> n = 20    #sample size
> m_sample = 400 #mean of the sample
> t = (m_sample - m_population)/(s_sample/sqrt(n)) #t score
> t
[1] 8.944272
> #p-value calculation
> p_val = pt(-abs(t),df= n-1) #one sided t-test for finding lesser than or greater than
> p_val
[1] 1.537894e-08
> 
```

Q2 ~~me~~ Given:

mean of population = 200 days

standard deviation of population = 6

sample size = 20

standard deviation of sample = 60.

$$X^2 = [(n-1) * s^2] / \sigma^2$$

where σ = S.D. of population.

s = ~~no~~ S.D. of sample

n = sample size.

$$X^2 = \frac{[19 * 60 \times 60]}{6 \times 6}$$

$$X^2 = 1900$$

From X^2 table,

p-value < 0.001

Thus null hypothesis is rejected and we can say claim made by Philips was false.

Q3 To prove :- for a binomial distribution,

$$\text{mode} = \begin{cases} \lfloor (n+1)p \rfloor & ; \text{ if } (n+1)p \text{ is } 0 \text{ / non-integer} \\ (n+1)p \ \& \ (n+1)p - 1 & ; \text{ if } (n+1)p \in \{1, \dots n\} \\ n & ; \text{ if } (n+1)p = n+1 \end{cases}$$

Proof : for mode we find ratio $b(n,p;k+1) / b(n,p;k)$

$$a_k = P(X=k) = b(n,p;k)$$

$$= {}^nC_k \, p^k \, q^{n-k}$$

$$a_{k+1} = P(X=k+1) = b(n,p;k+1)$$

$$= {}^nC_{k+1} \, p^{k+1} \, q^{n-k-1}$$

$$\frac{a_{k+1}}{a_k} = \frac{{}^nC_{k+1} p^{k+1} q^{n-k-1}}{{}^nC_k \, p^k \, q^{n-k}}$$

$$= \frac{\cancel{n!}\, k!\,(n-k)!}{(k+1)!\,(n-k-1)!} \cdot \frac{p^{k+1} \, q^{n-k-1}}{\cancel{n!} \; p^k \, q^{n-k}}$$

$$= \frac{\cancel{k!}\,(n-k)\cancel{(n-k-1)!}}{(k+1)\cancel{k!}\;\cancel{(n-k-1)!}} \cdot \frac{p \cdot \cancel{p^k} \cdot \cancel{q^{n-k-1}}}{\cancel{p^k} \cdot q \cdot \cancel{q^{n-k-1}}}$$

$$= \frac{(n-k)}{(k+1)} \cdot \frac{p}{q}$$

$$= \frac{n-k}{k+1} \cdot \frac{p}{1-p}$$

~~if~~ comparing this ratio to 1 we get :-

$$(k+1)(1-p) = (n-k)p$$

$$k - \cancel{kp} + 1 - p = np - \cancel{kp}$$

$$k = np + p - 1$$

$$\underline{k = (n+1)p - 1}$$

from here we can follow :-

if $K > (n+1)p - 1$ $\Rightarrow$ $a_{K+1} > a_K$

if $K = (n+1)p - 1$ $\Rightarrow$ $a_{K+1} = a_K$

if $K < (n+1)p - 1$ $\Rightarrow$ $a_{K+1} < a_K$

(i) Now, when $(n+1)p$ is 0/ non-integer, we will have single mode for the binomial distribution

$\lfloor (n+1)p - 1 \rfloor + 1$

$= \lfloor (n+1)p \rfloor$

(ii) when $(n+1)p$ is integer, distribution will be bimodal & mode will exist for both a_{K+1} & a_K values.

ie ~~at~~ $(n+1)p$ & $(n+1)p - 1$

(iii) when. $(n+1)p = n+1$

$K = (n+1)p - 1$

$= n+1-1$

$= n$

where K is the mode of distribution.

Hence proved.

Q4

Given :-

$f(x) = \text{constant} \cdot x^{\alpha-1} (1-x)^{\beta-1}$

To find : constant term.

To prove : constant $= \frac{1}{\Gamma(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) * \Gamma(\beta)}$

Proof :-

$\text{PDF} = \int_0^1 f(x)dx = 1$

$\Rightarrow \int_0^1 \text{constant} \cdot x^{\alpha-1} (1-x)^{\beta-1} dx = 1$

$\Rightarrow \text{constant} \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = 1 \qquad \text{———} (1)$

We know that,

$$\beta(m,n) = \frac{\cancel{m}(m-1)!\,(n-1)!}{(m+n-1)!}$$

$$= \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

$$= \int_0^1 u^{m-1} (1-u)^{n-1} du$$

eq^n(1) is of the form $\beta(m,n)$ or $\beta(\alpha,\beta)$. Thus can be written as;

$\Rightarrow \int \text{constant} \cdot \beta(\alpha,\beta) = 1$

$\Rightarrow \text{constant} \cdot \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = 1$

$\Rightarrow \text{constant} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$

Hence Proved

Q5 To prove: μ (Beta Distribution) $= \dfrac{\alpha}{\alpha+\beta}$

Proof :-

$$E(x) = \int x \cdot f(x)\, dx$$

$$= \int x \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1} (1-x)^{\beta-1} dx$$

$$= \int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha} (1-x)^{\beta-1} dx$$

$$= \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!} \cdot \beta(\alpha+1, \beta)\, dx$$

$$= \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}$$

$$= \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,\cancel{(\beta-1)!}} \cdot \frac{\alpha!\,\cancel{(\beta-1)!}}{(\alpha+\beta)!}$$

$$= \frac{\cancel{(\alpha+\beta-1)!} \cdot \alpha\,\cancel{(\alpha-1)!}}{\cancel{(\alpha-1)!}\;(\alpha+\beta)\,\cancel{(\alpha+\beta-1)!}}$$

$$E(x) = \frac{\alpha}{\alpha+\beta}$$

Hence Proved.

Q6 (i) $I = \int_0^{\infty} e^{-t}\, dt$

$= \left[-e^{-t}\right]_0^{\infty}$

$= -e^{-\infty} + e^{-0}$

$= -0 + 1$

$= 1$

(ii) $I = \int_{-\infty}^{\infty} e^{-x^2} dx$

$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)^2$

$= \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)$

Using property of definite integral $\int_a^b f(x)\, dx = \int_a^b f(y)\, dy$:-

$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right)$

$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dx\, dy$

Since e^{-t^2} is even function :-

$\int_{-\infty}^{\infty} e^{-t^2} dt = 2\int_0^{\infty} e^{-t^2} dt$

$\therefore I^2 = 4\int_0^{\infty}\int_0^{\infty} e^{-(x^2+y^2)}\, dx\, dy$

Taking $y = xs$ & differentiating y w.r.t. s we get :-

$dy = x\,ds$

$$I^2 = 4\int_0^\infty \left[\int_0^\infty e^{-(x^2+y^2)}\,dy \right] dx$$

$$= 4\int_0^\infty \left[\int_0^\infty e^{-x^2(1+s^2)}\,x\,ds \right] dx.$$

$$= 4\int_0^\infty \left[\int_0^\infty e^{-x^2(1+s^2)}\,x\,dx \right] ds.$$

$$= 4\int_0^\infty \left[\frac{1}{2}\int_0^\infty e^{-x^2(1+s^2)}\,d(x^2) \right] ds$$

$$= 4\int_0^\infty \left[\frac{1}{-2(1+s^2)}\, e^{-x^2(1+s^2)} \right]_0^\infty ds.$$

$$= 4\int_0^\infty \left[-\frac{1}{2}\left(\frac{e^{-\infty}}{1+s^2} - \frac{e^{-0}}{1+s^2} \right) \right] ds$$

$$= 4\int_0^\infty -\frac{1}{2}\left[\frac{0-1}{1+s^2} \right] ds$$

$$= 4 \left(\frac{1}{2}. \int_0^\infty \frac{ds}{1+s^2} \right)$$

$$= 2\left[\tan^{-1} s \right]_0^\infty$$

$$= 2\left[\frac{\pi}{2} - 0 \right]$$

$$I^2 = \pi$$

$\boxed{I = \sqrt{\pi}.}$ $\longrightarrow$ Answer

Q7 Gamma function :

$$\Gamma(n) = \int_0^\infty t^{n-1} e^{-t}\, dt$$

It is a good approximation for factorial as given below :-

$$\Gamma(n+1) = \int_0^\infty t^n e^{-t}\, dt$$

using ILATE :-

$$\Gamma(n+1) = \left[-t^n e^{-t}\right]_0^\infty + \int_0^\infty n t^{n-1} e^{-t}\, dt$$

$$= \lim_{t\to\infty} \left(-t^n e^{-t}\right) - 0.e^{-0} + n\int_0^\infty t^{n-1} e^{-t}\, dt$$

as at $x \to \infty$
$e^{-\infty} \to 0$

$$\Gamma(n+1) = n\int_0^\infty t^{n-1} e^{-t}\, dt$$

$$\Gamma(n+1) = n\Gamma(n) \qquad \text{———— (A)}$$

we calculate $\Gamma(1)$:

$$\Gamma(1) = \int_0^\infty t^{1-1} e^{-t}\, dt$$

$$= \int_0^\infty t^0 . e^{-t}\, dt = \int_0^\infty e^{-t} dt = \left[-e^{-t}\right]_0^\infty$$

$$= -e^{-\infty} + e^{-0} = -0+1$$

$$= 1$$

Given $\Gamma(1) = 1$ & $\Gamma(n+1) = n\Gamma(n)$

$$\Gamma(n) = (n-1)\,\Gamma(n-2) = (n-1)(n-2)\Gamma(n-3) = ---$$

$$= (n-1)(n-2)(n-3) ---- 3.2.1$$

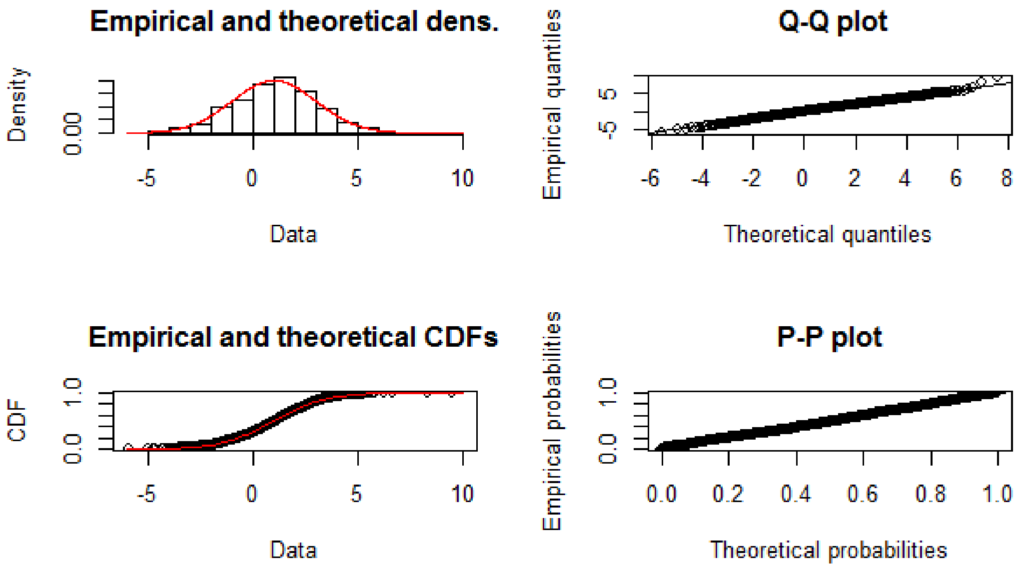
$$= (n-1)!$$

Hence Proved, that Gamma function can be used for approximation of factorial.

Q8) R Code : using the same code we can guess distribution of all the data file(d1,d5,d7,d8,d10)

```r
#input data : data taken into consideration are d1, d5, d7, d8, d10
data = read.table("d8.txt",header = FALSE,sep = "\n")

#function to calculate mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

#calculation of parameters of data
mean = mean(as.matrix(data))            #mean of data
median = median(as.matrix(data))        #median of data
mod_data = getmode(as.matrix(data))     #mode of data
sd= sd(as.matrix(data))                 #standard deviation
variance = var(as.matrix(data))         #variance
maxi = max(as.matrix(data))             #maximum value in data
min = min(as.matrix(data))              #minimum value in data
range = range(as.matrix(data))          #range of data

#Distibution check: normal / poisson / binomial
library(fitdistrplus)
temp <- c(as.matrix(data))

FITN <- fitdist(temp, "norm")     #normal
FITP <- fitdist(temp, "pois")     #poisson
fitBinom=fitdist(temp, dist="binom", fix.arg=list(size=1000), start=list(prob=0.5)
fitUnif = fitdist(temp, dist="unif") #uniform


summary(FITN)
summary(FITP)
summary(fitBinom)
summary(fitUnif)
```

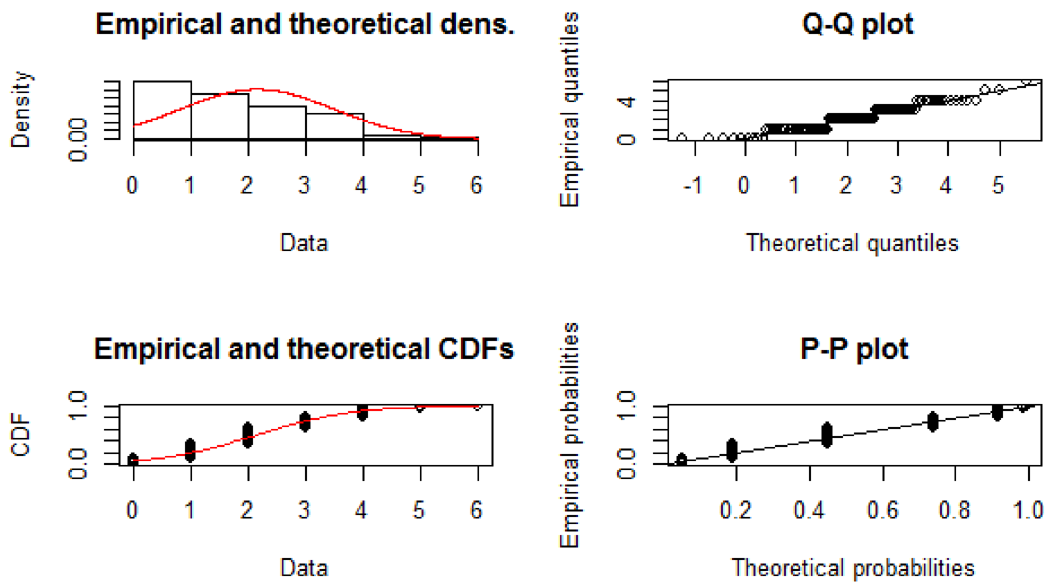
Output for d1.txt :plot of normal distribution and parameter values respectively



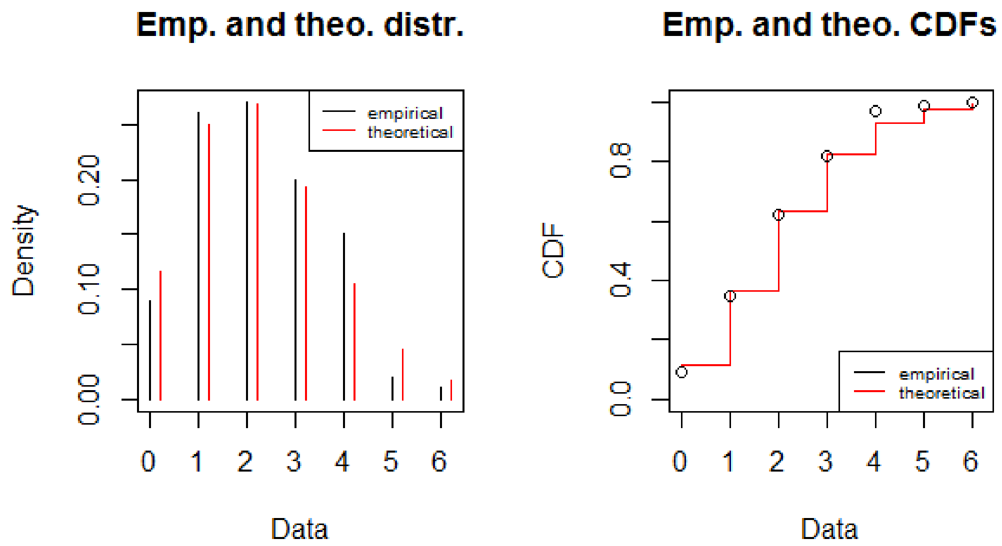
maxi	9.52761080185824
mean	1.01589150828133
median	1.07419927770765
min	-5.83887864396858
mod_data	2.08430584925726
range	num [1:2] -5.84 9.53
sd	1.99779388886411

We can clearly see that qqplot is a straight line when we tried to fit normal data to it. Hence d1 follows Normal Distribution.

<u>Output of d5.txt</u>: plot of normal distribution, parameter values and poisson distribution respectively.

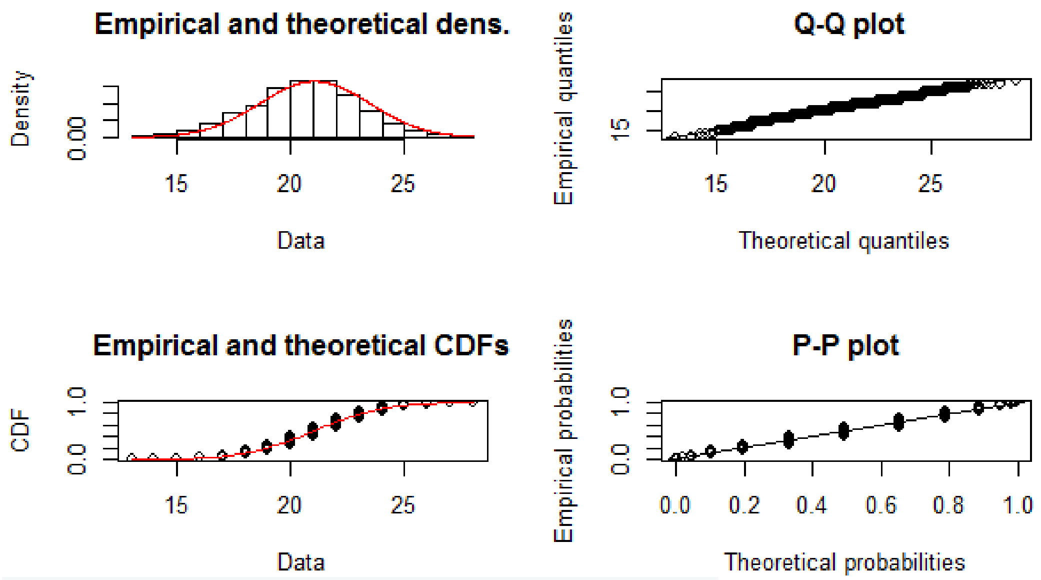


maxi	6L
mean	2.16
median	2
min	0L
mod_data	2L
range	int [1:2] 0 6
sd	1.32360084312424

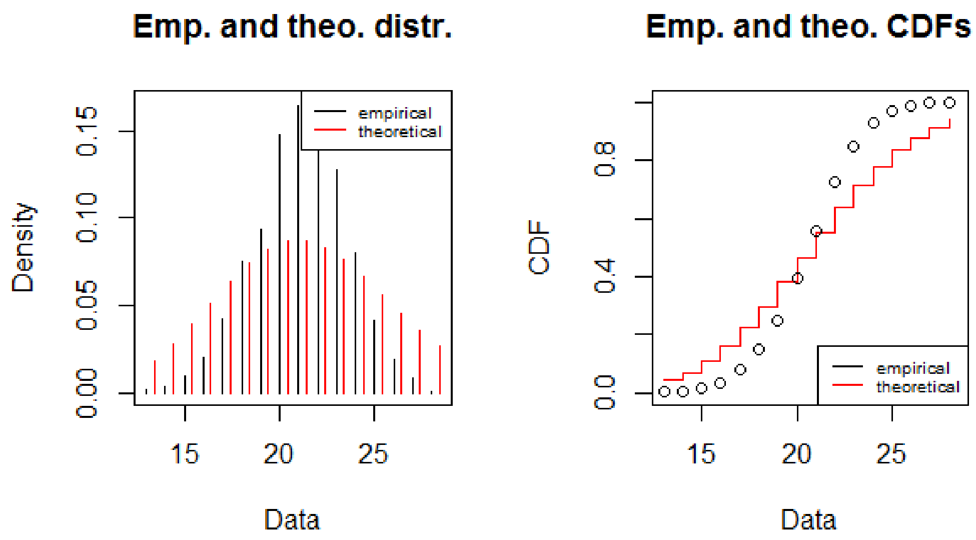


We can clearly see via plots that d5.txt do not follow normal distribution but follow Poisson Distribution.

Output of d7.txt: plot of normal distribution, parameter values and poisson distribution respectively.

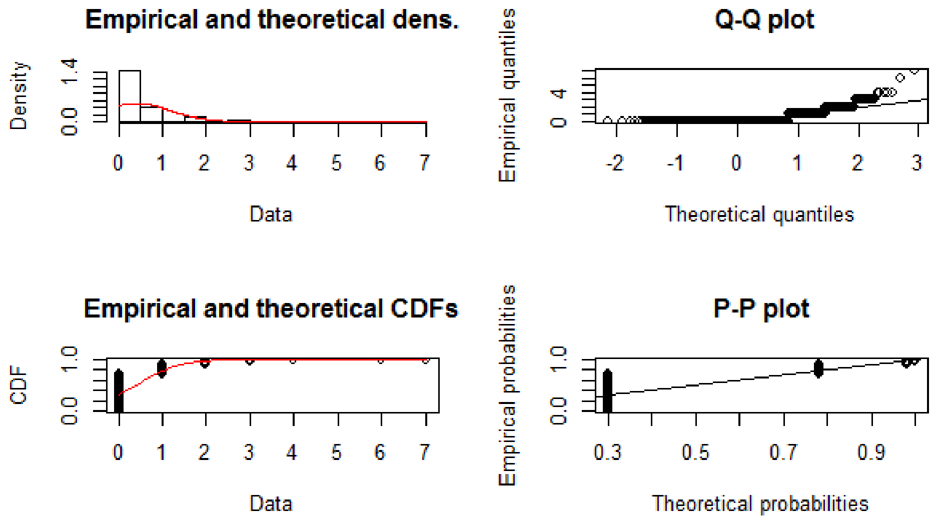


maxi	28L
mean	21.047
median	21
min	13L
mod_data	22L
range	int [1:2] 13 28
sd	2.42583504708702

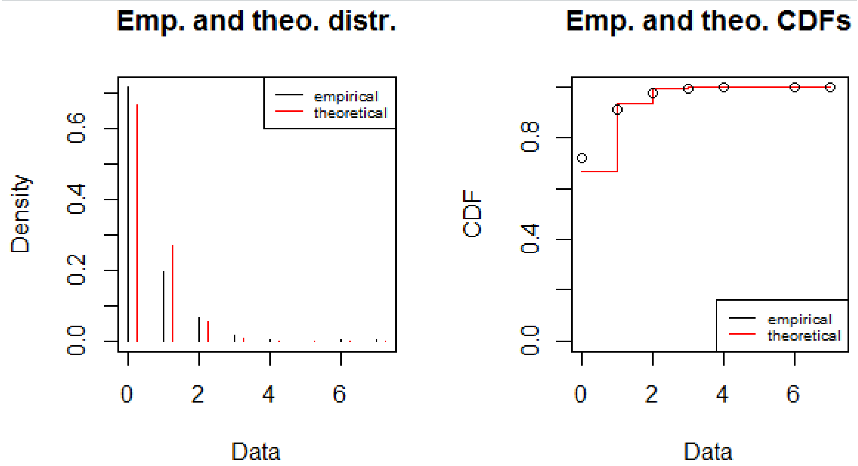


The above plots show that data follows normal distribution and not Poisson distribution as the data points deviate significantly from theoretical curve in qqplot in case of Poisson distribution.

__Output of d8.txt:__normal distribution,parameter values, Poisson distribution respectively.



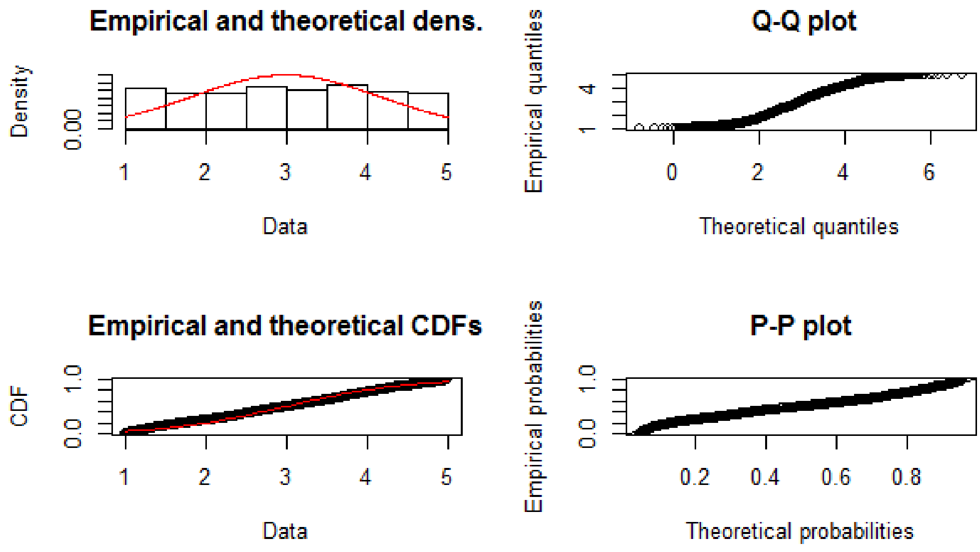
maxi	7L
mean	0.404
median	0
min	0L
mod_data	0L
range	int [1:2] 0 7
sd	0.772904509872585

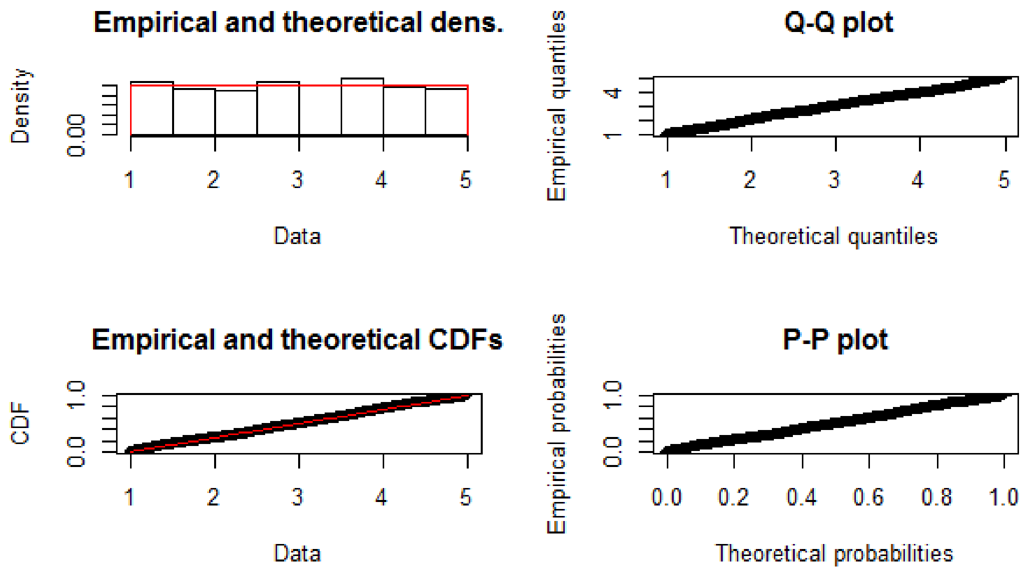


The data follows Poisson distribution and deviates from normal distribution plot.

__Output of d10.txt__: parameter values,plot of normal distribution and uniform distribution respectively

maxi	4.99785471893847
mean	2.99254500367492
median	3.03120387811213
min	1.00243655126542
mod_data	4.84567935112864
range	num [1:2] 1 5
sd	1.14560489898867





The data follows Uniform Distribution.

Q9 Geometric Distribution :-

(a) Derivation of mean :

$f(x) = pq^x = p(1-p)^x$ (given)

$E(x) = \sum_{x=0}^{\infty} x f(x)$

$= \sum_{x=0}^{\infty} x \cdot p q^x$

$= p \sum_{x=0}^{\infty} x \cdot q^x$

$= pq \sum_{x=0}^{\infty} x q^{x-1}$

$= pq \sum_{x=0}^{\infty} \frac{d}{dq}(q^x) = pq \frac{d}{dq}\left(\sum q^x\right)$

$= pq \frac{d}{dq}\left(\frac{1}{1-q}\right)$ [using summation of geometric series)

$= pq\left(\frac{1}{(1-q)^2}\right)$

$= \frac{pq}{p^2} = \frac{q}{p}$

$E(X) = \frac{q}{p}$

Hence Proved.

(b) Variance Derivation :

$Var(X) = E(X^2) - [E(X)]^2$

$= E[X(X-1)] + E(X) - [E(X)]^2$ ——— (A)

$E(X) = \frac{q}{p}$

$[E(X)]^2 = \frac{q^2}{p^2}$

where $q = 1-p$

$$E[X(X-1)] = \sum_{K=1}^{\infty} x(x-1)\, p q^{x-1}$$

$$= p \sum_{K=1}^{\infty} (x-1)x \cdot q^{x-1} = p \sum_{K=1}^{\infty} \frac{d}{dq}\left[(x-1) q^{x}\right]$$

$$= p \frac{d}{dq}\left(\sum_{K=1}^{\infty} (x-1) q^{x} \right) = p \frac{d}{dq}\left(q^2 \sum_{K=2}^{\infty} (x-1) q^{x-2} \right)$$

$$= p \frac{d}{dq}\left(q^2 \frac{d}{dq}\left(\sum_{K=2}^{\infty} q^{K-1} \right)\right) = p \frac{d}{dq}\left(q^2 \frac{d}{dq}\left(\sum_{K=1}^{\infty} q^{K} \right)\right)$$

$$= p \frac{d}{dq}\left(q^2 \frac{d}{dq}\left(\frac{1}{1-q} - 1 \right)\right)$$

$$= p \frac{d}{dq}\left[\frac{q^2}{(1-q)^2} \right] = p \left[\frac{(1-q)^2 \cdot 2q + 2(1-q)(q^2)}{(1-q)^4} \right]$$

$$= p \left[\frac{-2q^2 + 2q}{(1-q)^4} \right] = p \left[\frac{-2q(q-1)}{(1-q)^4} \right] = p \left[\frac{+2q(1-q)}{(1-q)^4} \right]$$

$$= p \left[\frac{+2q}{(1-q)^3} \right] = p \left[\frac{2(1-p)}{(1-1+p)^3} \right] = \frac{2q}{p^2}$$

$$\mathrm{Var}(X) = \frac{2q}{p^2} + \frac{q}{p} - \frac{q^2}{p^2} \qquad \left[\text{from } eq^n (A)\right]$$

$$= \frac{2q + qp - q^2}{p^2} = \frac{2(1-p) + (1-p)p - (1-p)^2}{p^2}$$

$$= \frac{2 - 2p - p + p^2 - 1 - p^2 + 2p}{p^2}$$

$$= \frac{1-p}{p^2}$$

Hence proved

Q80 $\ln n! = \ln 1 + \ln 2 + --- \ln n$

The RHS of above equation minus $\frac{1}{2}(\ln 1 + \ln n) = \frac{1}{2}\ln n$ is approximation by trapezoidal rule of integral.

Thus,

$$\ln n! - \frac{1}{2}\ln n \approx \int_1^n \ln x\,dx$$

$$I = \int_1^n \ln x\,dx$$

$$= [x\ln x]_1^n - \int_1^n x.\frac{d}{dx}(\ln x)\,dx \qquad [\text{using ILATE}]$$

$$= [x\ln x]_1^n - \int_1^n x.\frac{1}{x}\,dx$$

$$= [x\ln x]_1^n - \int_1^n dx.$$

$$= [x\ln x - x]_1^n$$

$$= n\ln n - n - 1\ln 1 + 1$$

$$= n\ln n - n + 1$$

Thus,

$$\ln n! \approx n\ln n - n + 1$$

Hence Proved.

R code and outputs for log(200!) and log(20!) follows :

```r
#using stirling's approximation
n <- as.integer(readline(prompt="Enter an integer: "))
ans = n*log10(n) - n +1
ans

#without approximation
f=factorial(200)
ans1=log10(f)
```

Answer for log(20!):

```
Console   Terminal
~/
> #using stirling's approximation
> n <- as.integer(readline(prompt="Enter an integer: "))
Enter an integer: 20
> ans = n*log10(n) - n +1
> ans
[1] 7.0206
>
```

Answer for log(200!) :

```
Console   Terminal
~/
> #using stirling's approximation
> n <- as.integer(readline(prompt="Enter an integer: "))
Enter an integer: 200
> ans = n*log10(n) - n +1
> ans
[1] 261.206
>
> #without approximation
> f=factorial(200)
Warning message:
In factorial(200) : value out of range in 'gammafn'
> ans1=log10(f)
>
```