Mathematical Sciences

(Including Statistics)

: Minimum Sample Size Required for Asymptotic Convergence to Normality :

Soumya Mukherjee*, Subhadeep Chaudhuri, Swaraj Bose & Suraj Maiti

Department of Statistics

St. Xavier's College(Autonomous)

30, Park Street, Kolkata: 700016

Keywords: **Central Limit theorem, Slutsky's theorem, Convergence to normality, moment measures**

*Email id: soumyamukherjee1997jpbs@gmail.com

ABSTRACT

We know, from theory that the distribution of various ***statistics*** *converges to a* ***normal distribution****. However, the size of the sample required for this convergence depends on a number of factors like the* ***distribution of the parent population****,* ***value(s) of its parameter(s)*** *and also on the* ***statistic under consideration****. In this paper, we have investigated the* ***minimum sample size required for this asymptotic convergence*** *of some moment based statistics based on samples drawn from a number of distributions with varying parameters. For computation, R software has been used.*

: Minimum Sample Size Required for Asymptotic Convergence to Normality :

➢ Introduction:

We know, from the **Central Limit Theorem** that if $X_1, X_2 \ldots X_n$ be an i.i.d sample from some distribution with mean μ and variance σ^2, then $\boldsymbol{T_1} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ **,** where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ follows asymptotically a standard normal distribution. Further, **Slutsky's theorem** says if the **variance** σ^2 is replaced by the corresponding **sample variance** s^2, then also this **asymptotic normality** holds i.e. $\boldsymbol{T_2} = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ follows **asymptotically a standard normal distribution**.

However, the minimum sample size n required depends on the distribution from which the sample is drawn and also for a given distribution on the values of its parameters.

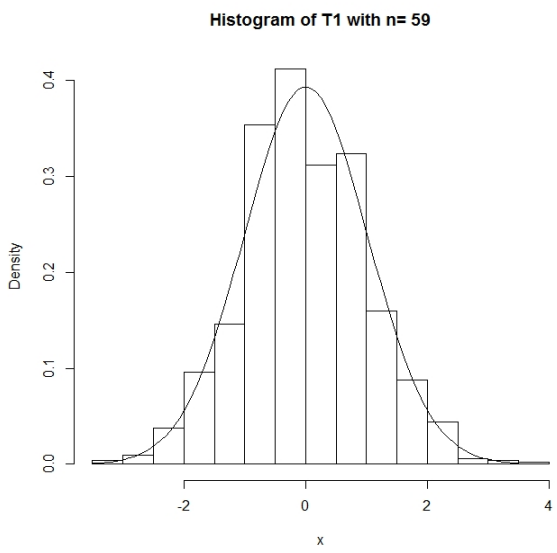
In this paper, we have considered **the convergence to normality** of some statistics based on moments. The parent distributions considered are **Binomial**, **Poisson**, **Exponential** and **Gamma** with varying parameter values. For the testing of normality, the **Shapiro - Wilks test** has been used. The sample size required in each case has been reported along with the corresponding **p-value** obtained. For computation, **R software** has been used.

➢ Results obtained for the statistic T_1:

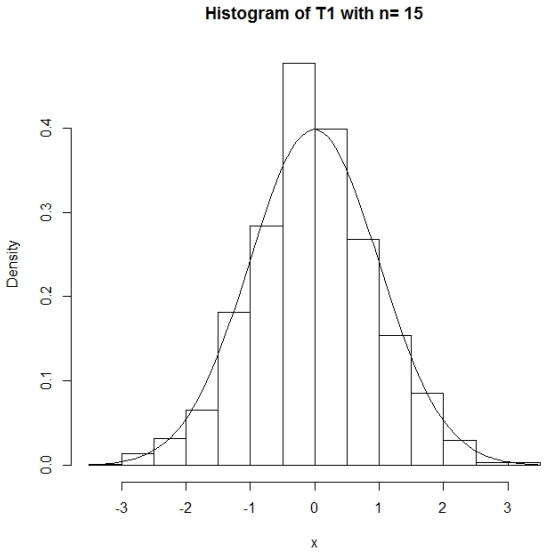
1. Let us consider the parent distribution as **Binomial** with **parameters n=10** and varying values of **p** i.e. $\mathbf{X_i \sim B\ (1\ , p)\quad i = 1, 2, \ldots, n}$

(i) p=0.1 n=59 p-value=0.07642046
(ii) p=0.2 n=33 p-value=0.1119601
(iii) p=0.3 n=20 p-value=0.05097259
(iv) p=0.4 n=17 p-value=0.0612432
(v) p=0.5 n=15 p-value=0.06842878
(vi) p=0.6 n=22 p-value=0.06896388
(vii) p=0.7 n=25p-value=0.2274666
(viii) p=0.8 n=38 p-value=0.1570726
(ix) p=0.9 n=66 p-value=0.07685392

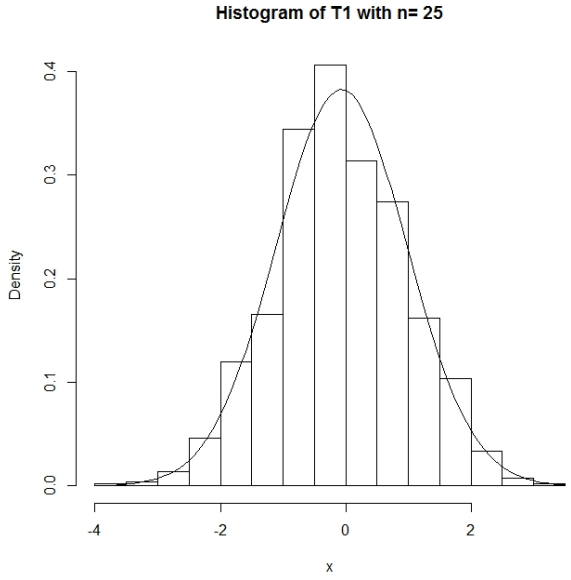
The histogram of the distribution of the statistic $\mathbf{T_1}$ along with the **normal density curve** is given below for **p=0.1, 0.5, 0.7 and 0.9**



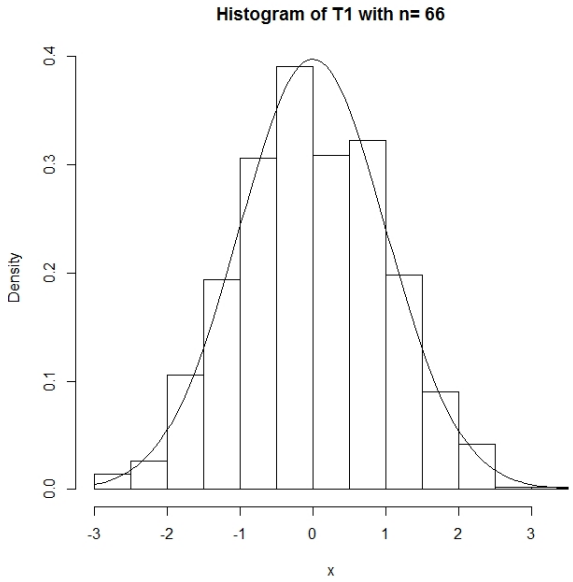
p=0.1



p=0.5



p=0.7



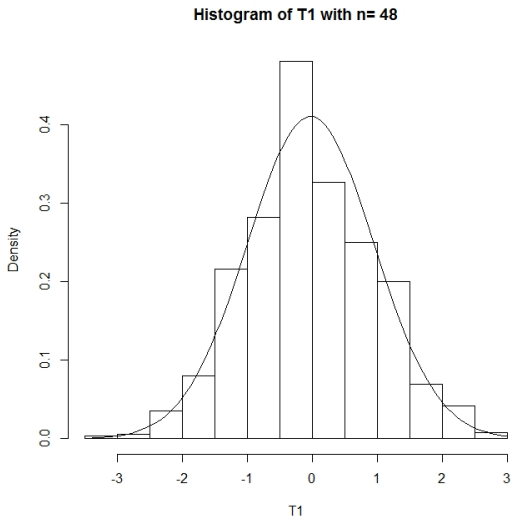
p=0.9

2. Let us consider the parent distribution as **Poisson** with varying values of the **parameter λ.** i.e $\mathbf{X_i \sim P(\lambda)}$, $\mathbf{i = 1, 2, \dots, n}$

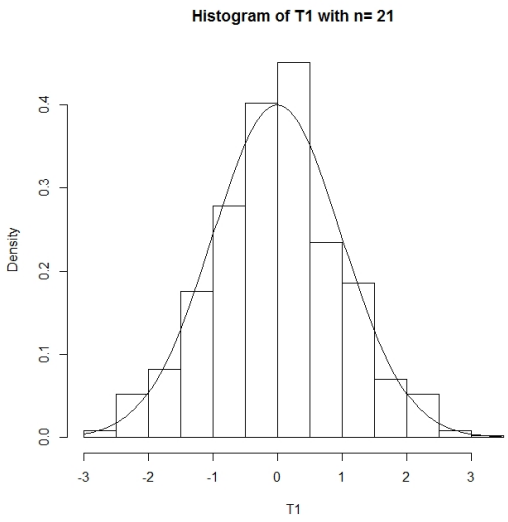
(i) λ = 0.5 n=86 p-value = 0.06156697
(ii) λ = 1 n=48 p-value = 0.06647147
(iii) λ = 1.5 n=38 p-value = 0.06867479
(iv) λ = 2 n=26 p-value = 0.1105512

(v) λ = 2.5 n=21 p-value = 0.08764987
(vi) λ = 3 n=19 p-value = 0.05012738
(vii) λ = 3.5 n=16 p-value = 0.05938056
(viii) λ = 4 n=13 p-value = 0.05614385
(ix) λ = 4.5 n=12 p-value = 0.1145419
(x) λ = 5 n=10 p-value = 0.06975946

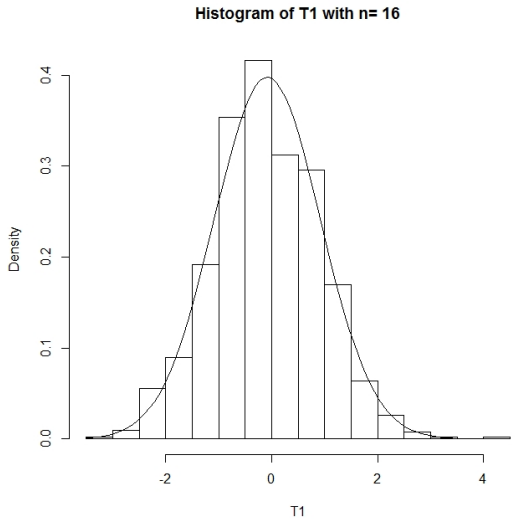
The histogram of the distribution of the statistic $\mathbf{T_1}$ along with the **normal density curve** is given below for **λ=1, 2.5, 3.5 and 5**



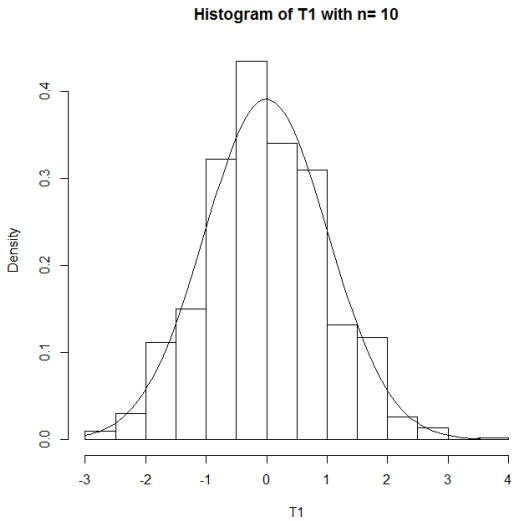
λ = 1



λ = 2.5



λ = 3.5



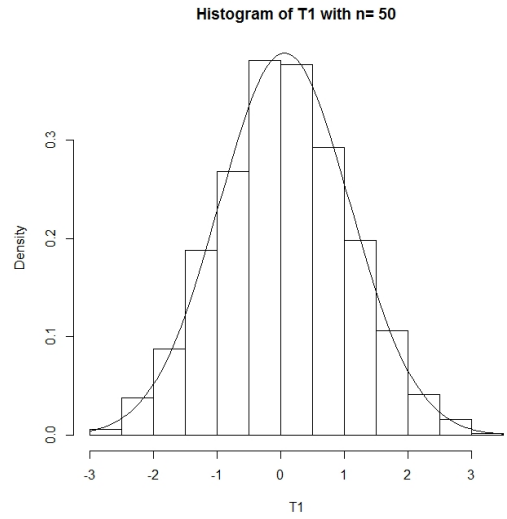
λ = 5

2. Let us consider the parent distribution as **Gamma** with varying values of the parameters
α and θ i.e. $\mathbf{X_i \sim G(\alpha, \theta)} \quad \mathbf{i = 1, 2, \dots, n}$

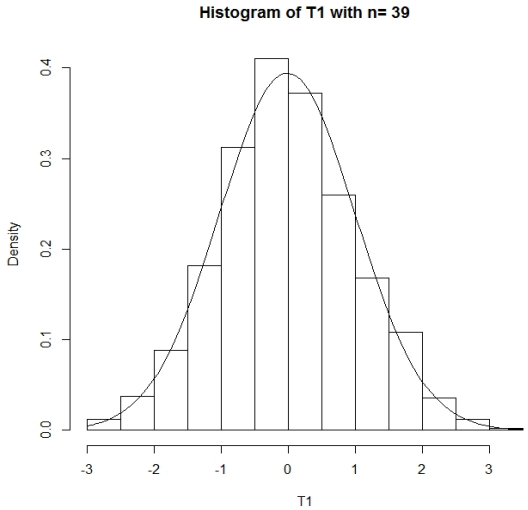
(i) α=1, θ=1 n=50 p-value=0.07065069

(ii) α=1, θ=2 n=39 p-value=0.1041021
(iii) α=1, θ=3 n=61 p-value=0.05817784
(iv) α=2, θ=1 n=29 p-value=0.060071
(v) α=3, θ=1 n=23 p-value=0.09012639

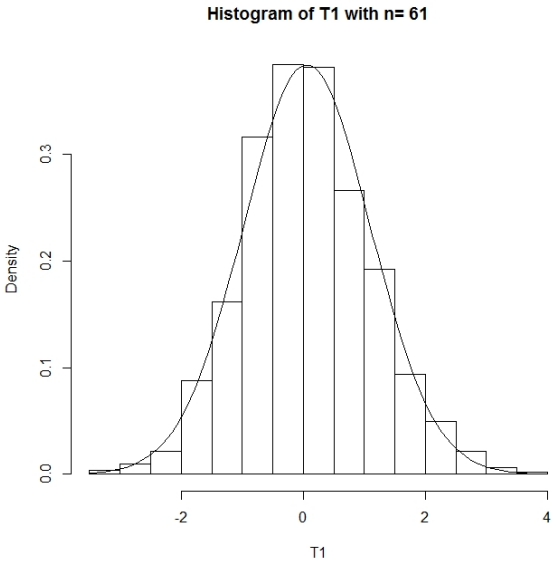
The histogram of the distribution of the statistic $\mathbf{T_1}$ along with the normal density curve is given below for the cases **(i),(ii),(iii) and (iv)**



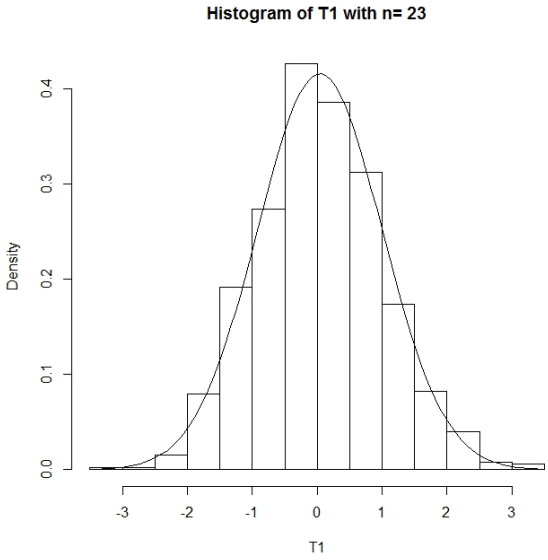
α=1, θ=1



α=1, θ=2



α=1, θ=1



α=3, θ=1

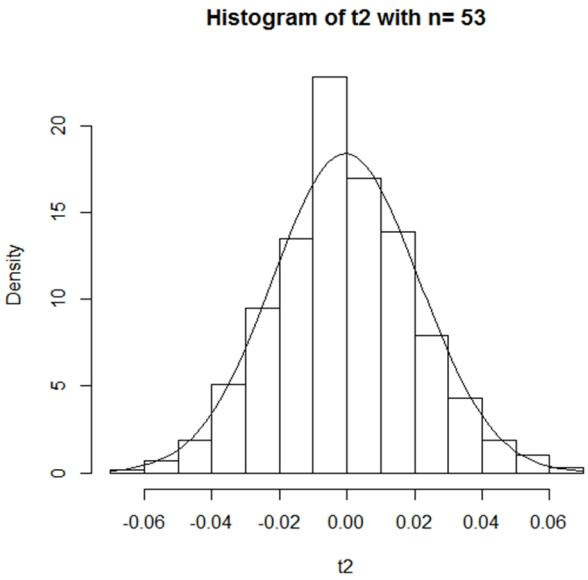
- **<u>Results obtained for the statistic T_2 :</u>**

1. Let us consider the parent distribution as **Binomial** with parameters **n=10** and varying values of **p**
   i.e. $X_i \sim Bin(10, p)$ $i = 1,2, \dots, n$

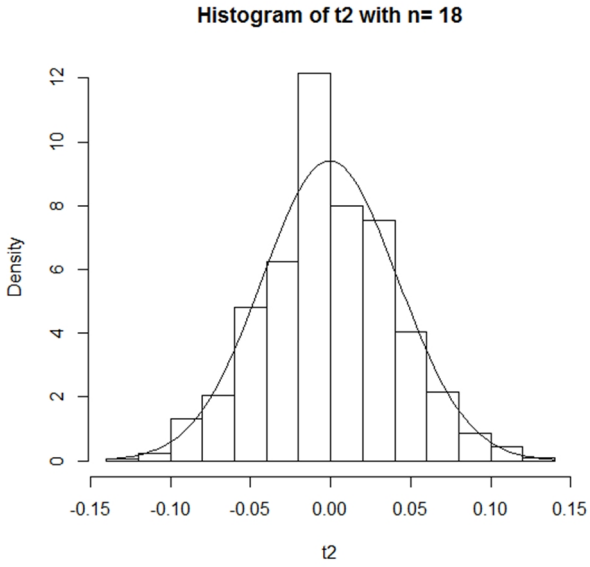
(i) p=0.1 n=53 p-value=0.06241203

(ii) p=0.2 n=24 p-value=0.05372757
(iii)p=0.3 n=24 p-value=0.07570789
(iv)p=0.4 n=20 p-value=0.05077569
(v) p=0.5 n=18 p-value=0.06900874
(vi)p=0.6 n=18 p-value=0.1041021
(vii) p=0.7 n=26 p-value=0.1807386
(viii) p=0.8 n=26 p-value=0.09113384
(ix) p=0.9 n=51 p-value=0.05349547

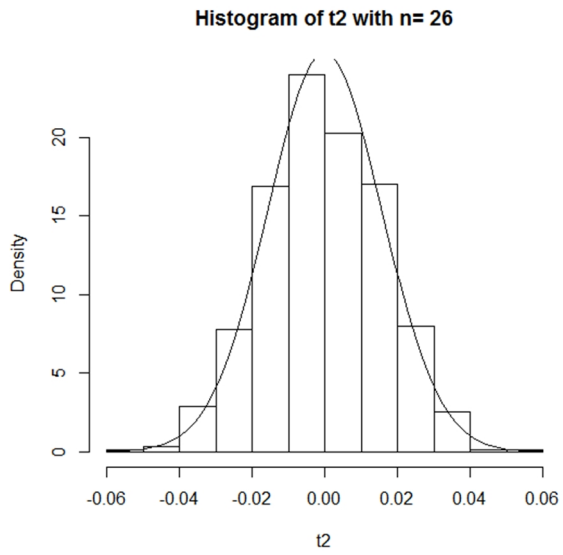
The histogram of the distribution of the statistic $\mathbf{T_2}$ along with the **normal density curve** is given below for **p=0.1, 0.5, 0.7 and 0.9**



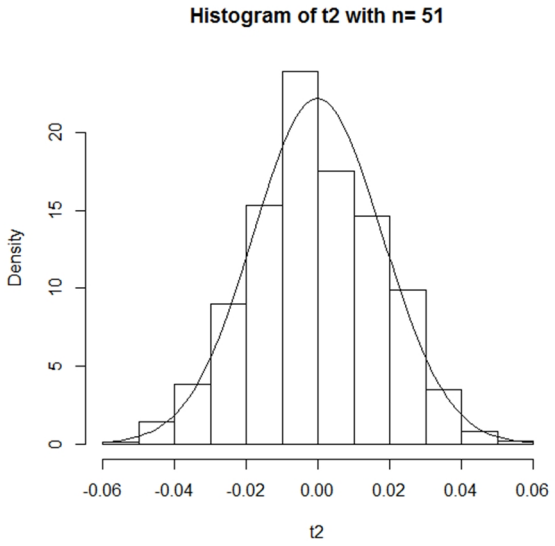
p=0.1



p=0.5



p=0.7

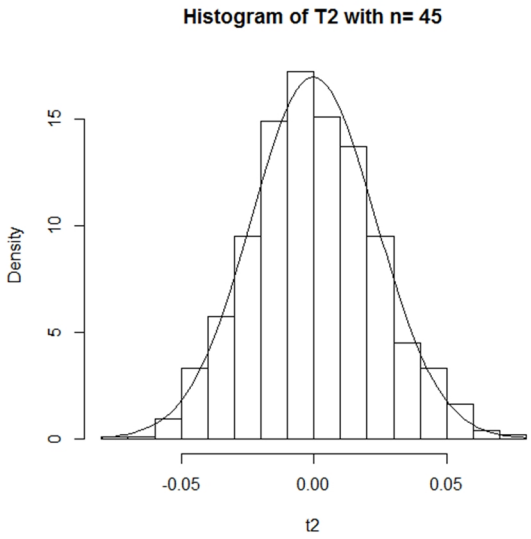


p=0.9

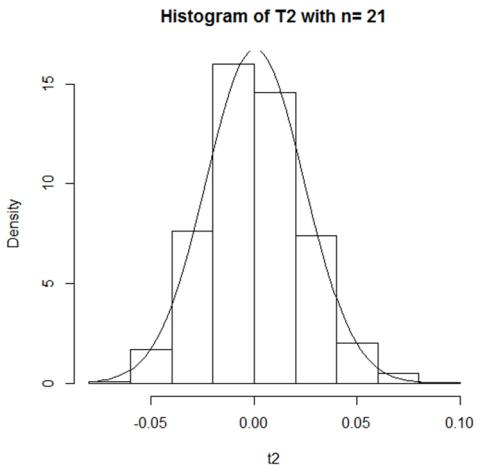
2. Let us consider the parent distribution as **Gamma** with varying values of the parameters
α and θ i.e. $\mathbf{X_i \sim G(\alpha, \theta)\ i = 1, 2, \dots, n}$

(i) α=1, θ=1 n=45 p-value=0.1810247
(ii) α=1, θ=2 n=21 p-value=0.060071
(iii) α=1, θ=3 n=22 p-value=0.1592041
(iv) α=2, θ=1 n=58 p-value=0.9078991
(v) α=3, θ=1 n=58 p-value=0.2048856

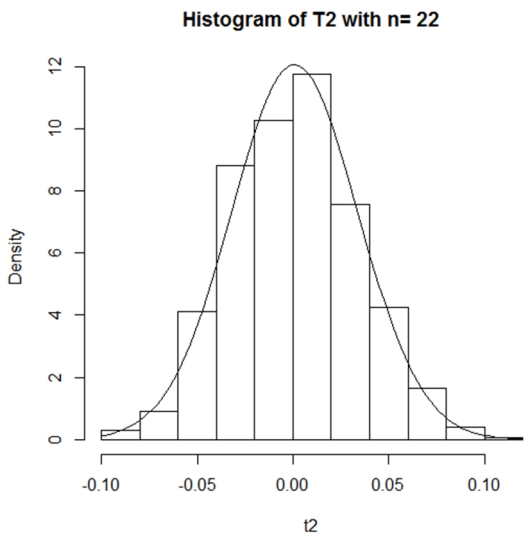
The histogram of the distribution of the statistic $\mathbf{T_2}$ along with the normal density curve is given below for the cases **(i),(ii),(iii),(iv) and (v)**



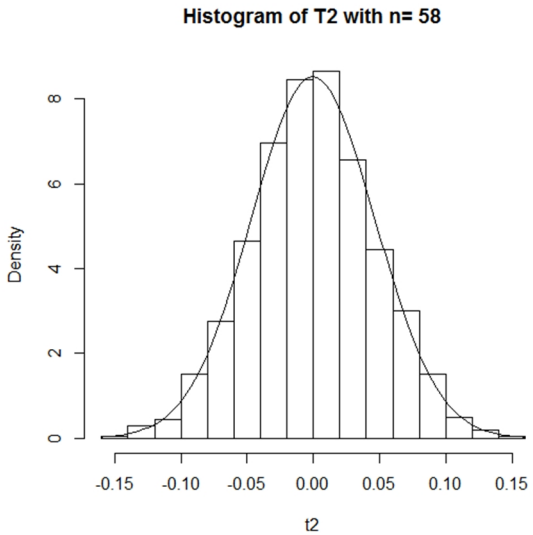
α=1, θ=1



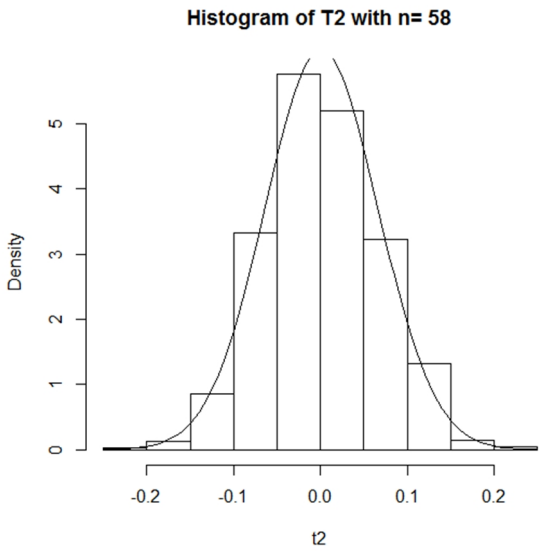
α=1, θ=2



α=1, θ=3



α=2, θ=1



$\alpha=3, \theta=1$

3. Let us consider the parent distribution as **Exponential** with varying values of the parameter λ

i.e. $\mathbf{X_i \sim E}\ (\boldsymbol{\lambda})\ i = 1,2, \ldots, n$

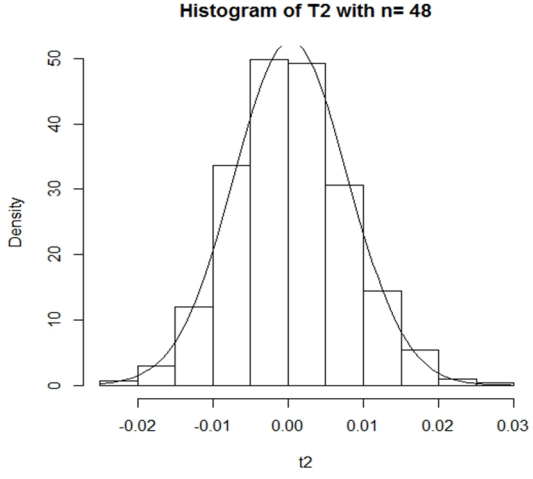
(i) $\lambda = 0.5$ n=48 p-value = 0.07201049

(ii) $\lambda = 1$ n=41 p-value = 0.249688

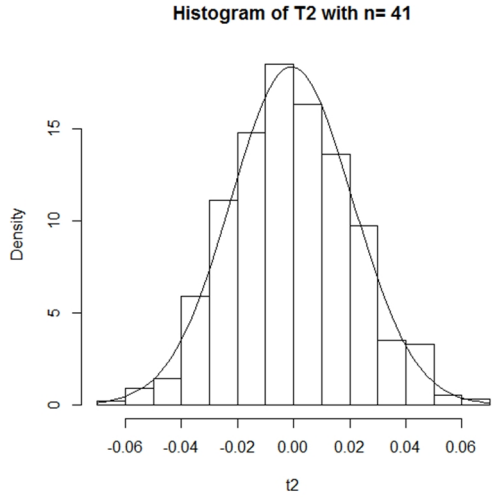
(iii) $\lambda = 1.5$ n=63 p-value = 0.111947

(iv) $\lambda = 2$ n=51 p-value = 0.2126678

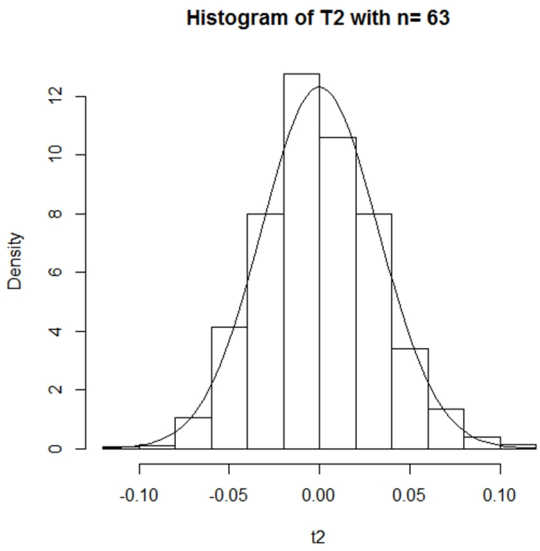
The histogram of the distribution of the statistic $\mathbf{T_2}$ along with the normal density curve is given below for the cases (i),(ii),(iii) and (iv)



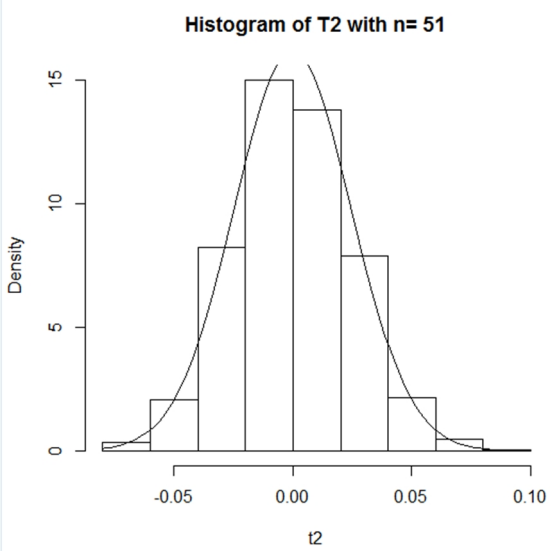
$\lambda = 0.5$



$\lambda = 1$



λ = 1.5



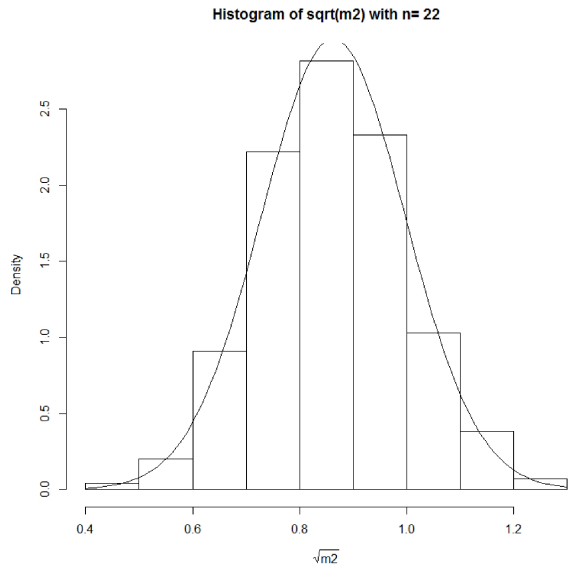
λ = 2

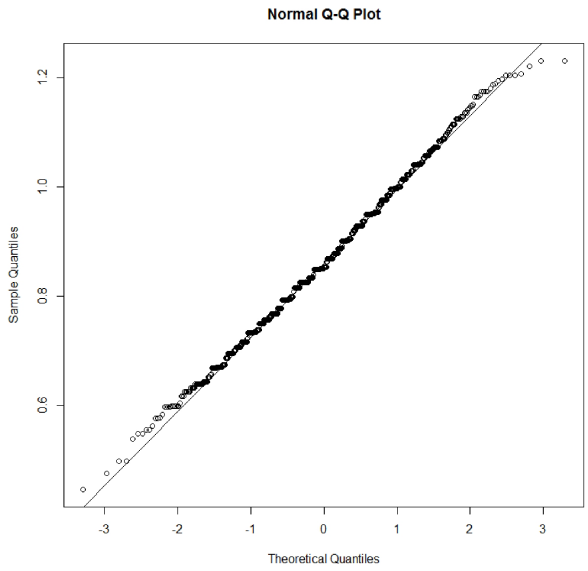
- **<u>Results obtained for the statistic $T_3 = \sqrt{m_2}$:</u>**

1. Let us consider the parent distribution as **Binomial with parameters k=5** and varying values of p i.e. $\mathbf{X_i \sim B\ (5, p)\ i = 1, 2, \dots, n}$

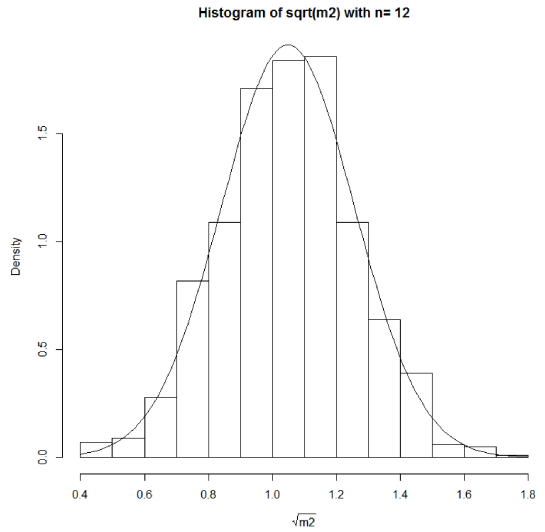
(i) p=0.2 n=22 p-value= 0.06675213
(ii) p=0.5 n=12 p-value= 0.09012639
(iii) p=0.8 n=30 p-value= 0.06250352

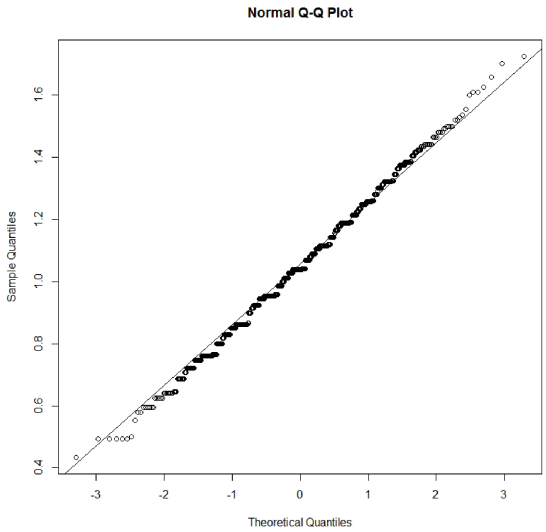
The histogram of the distribution of the statistic $\mathbf{T_3}$ along with the normal density curve is given below for each value of p. The quantile –quantile plot is also given.



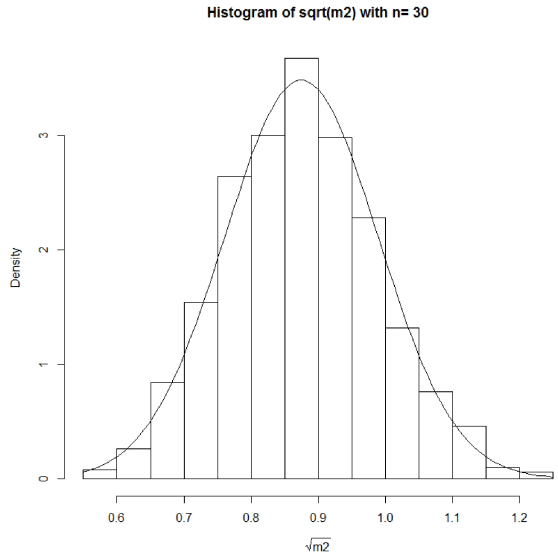


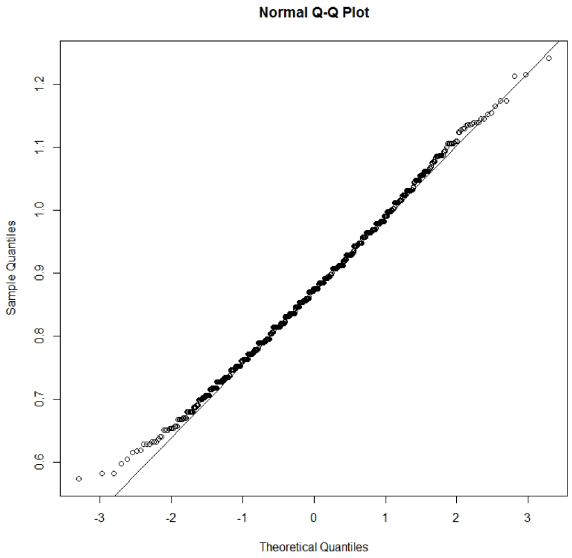
p=0.2





p=0.5



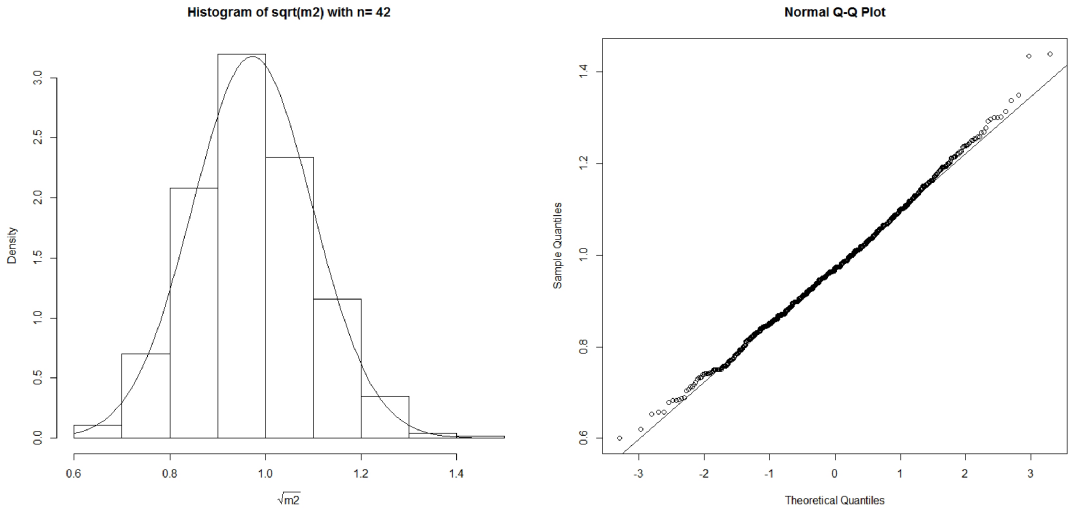


p=0.8

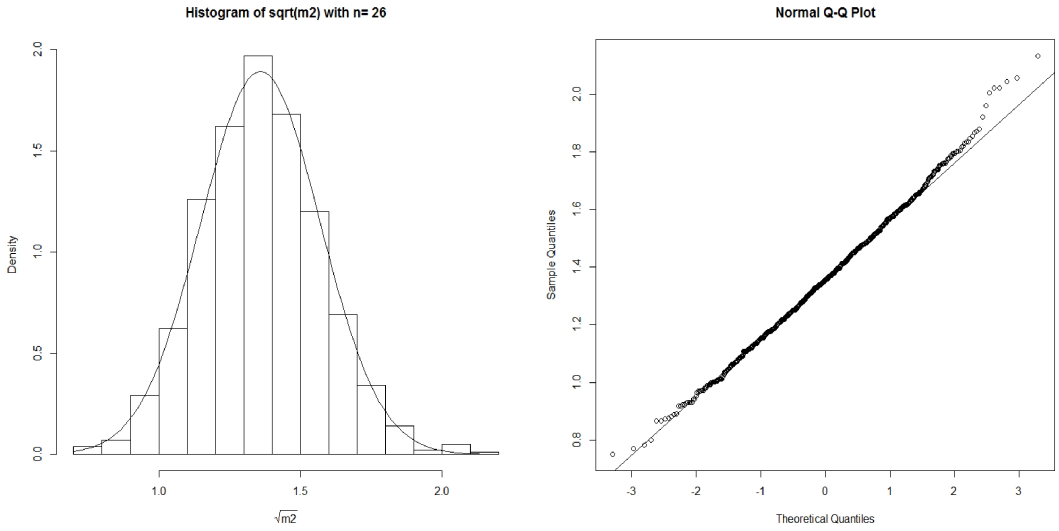
2. Let us consider the parent distribution as **Poisson** with varying values of the parameter λ.
   i.e. $\mathbf{X_i \sim P(\lambda)\ i = 1, 2, \dots, n}$

(i) λ = 1 n=42 p-value = 0.05952698
(ii) λ = 2 n=26 p-value = 0.0853228
(iii) λ = 5 n=22 p-value = 0.08496058
(iv) λ = 10 n=17 p-value = 0.07265746

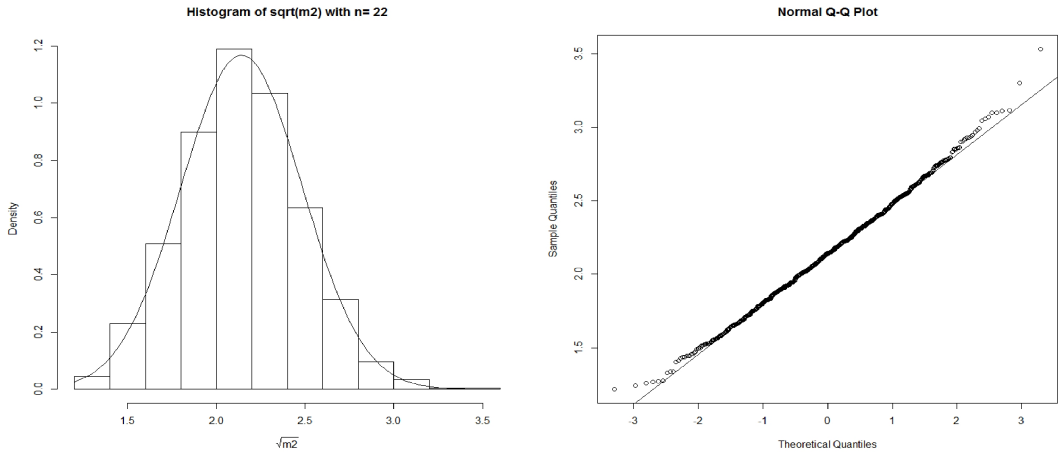
The histogram of the distribution of the statistic $\mathbf{T_3}$ along with the normal density curve is given below for each value of λ. The quantile –quantile plot is also given.



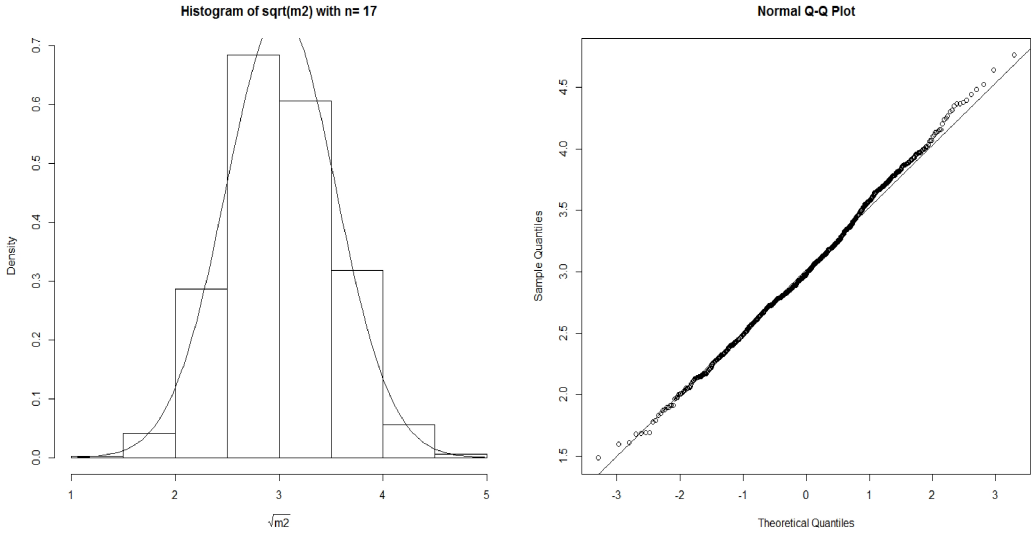
λ = 1



λ = 2



λ = 5



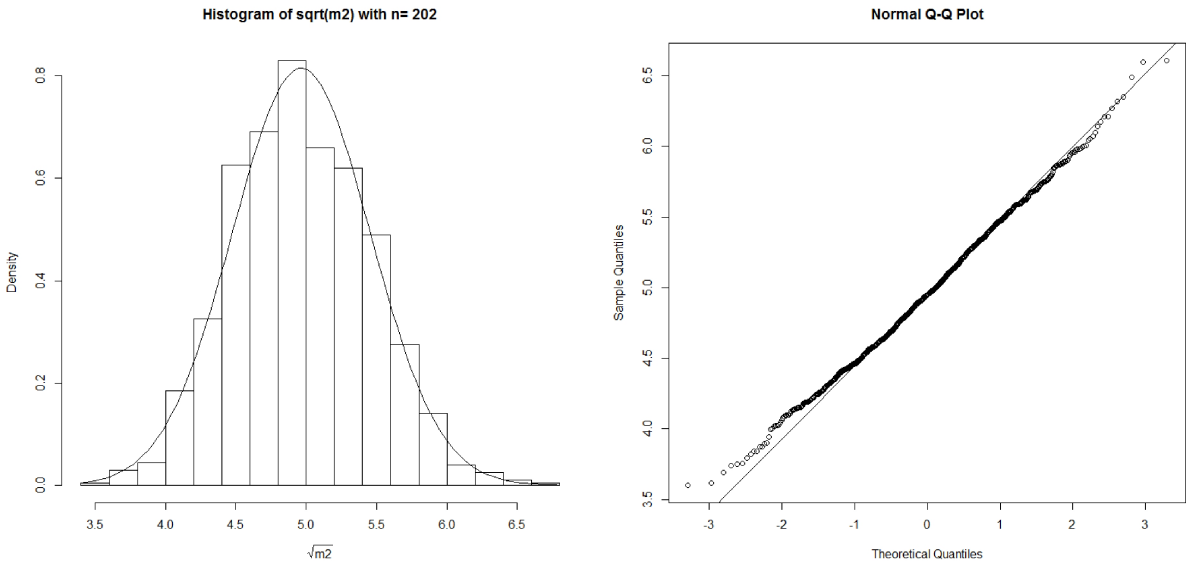
λ = 10

1. Let us consider the parent distribution as **Exponential** with varying values of the parameter λ

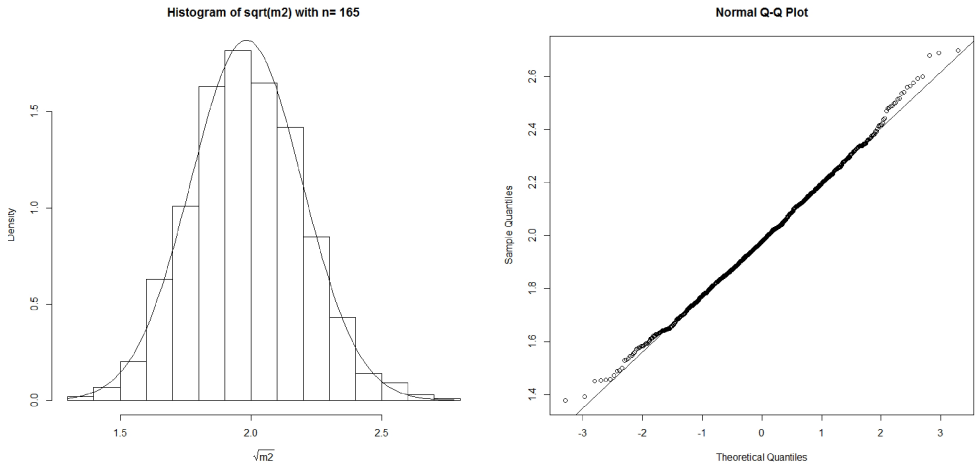
   i.e. $\mathbf{X_i \sim E}$ (λ) i = 1,2, … , n

(i) λ = 0.2 n= 202 p-value = 0.0936308
(ii) λ = 0.5 n=165 p-value = 0.08096232
(iii) λ = 1 n=180 p-value = 0.06368205
(iv) λ = 3 n=234 p-value = 0.05446721

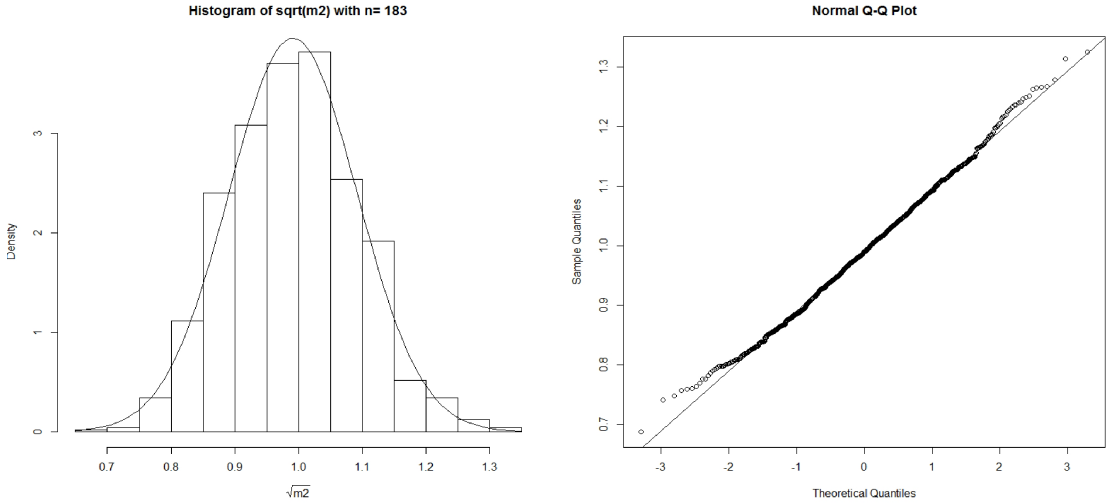
The histogram of the distribution of the statistic $\mathbf{T_3}$along with the normal density curve is given below for each λ. The quantile –quantile plot is also given.



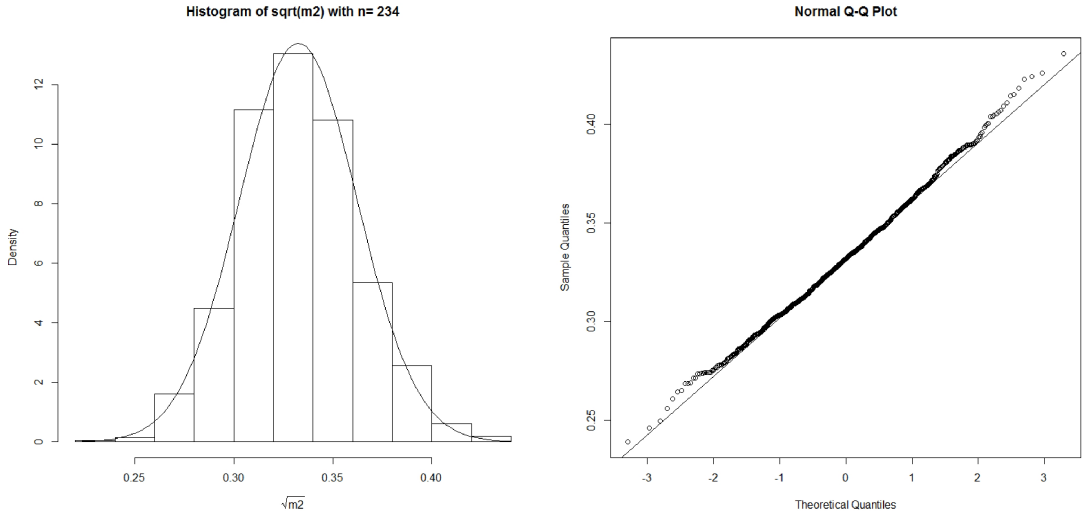
λ = 0.2



$\lambda = 0.5$



$\lambda = 1$

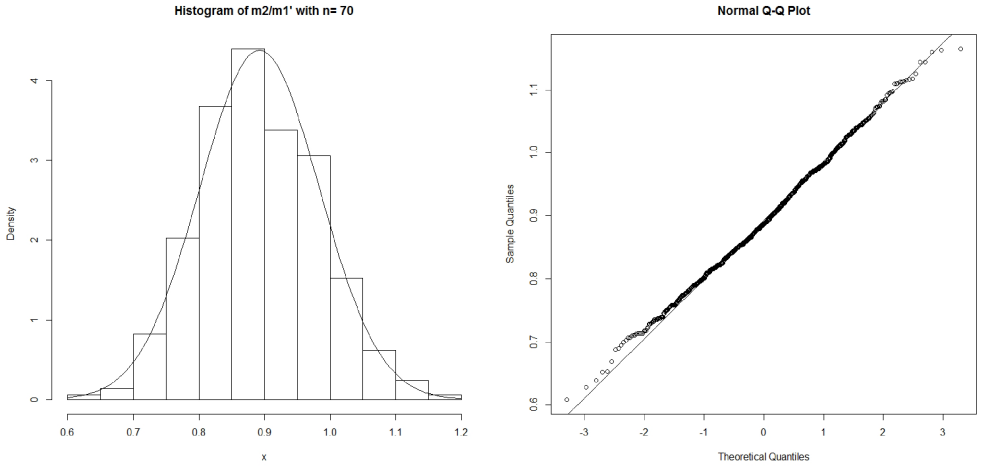


$\lambda = 3$

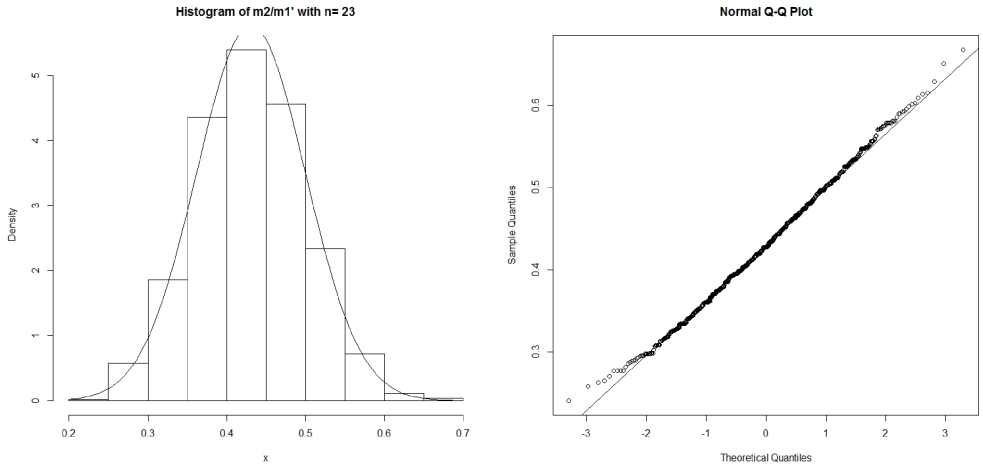
- **Results obtained for the statistic $T_4 = \sqrt{\overline{m_2}}/m_1'$:**

1. Let us consider the parent distribution as **Binomial** with parameters **k=5** and varying values of p i.e. $\mathbf{X_i \sim B}\ (\mathbf{5}, \mathbf{p})$ i = 1,2, … , n
   (i) p=0.2 n=70 p-value= 0.07065069
   (ii) p=0.5 n=23 p-value= 0.1143442
   (iii) p=0.8 n=37 p-value= 0.1830881

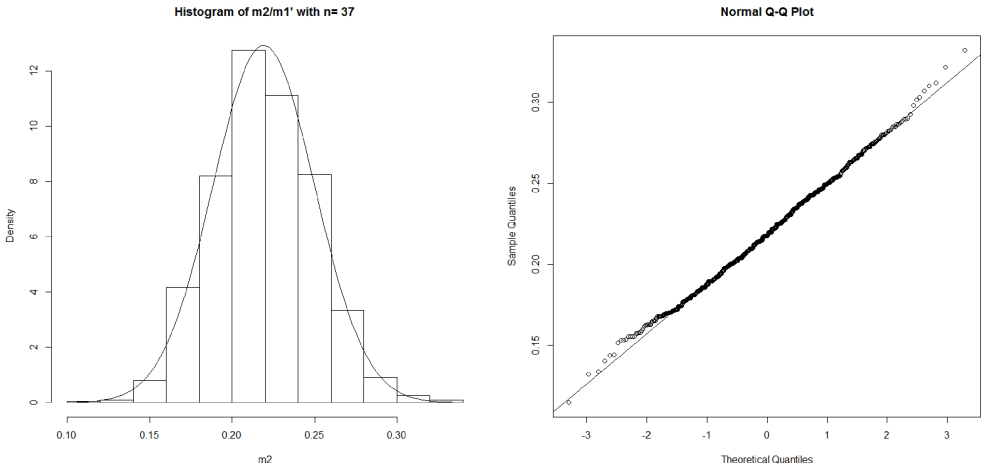
The histogram of the distribution of the statistic $\mathbf{T_4}$ along with the normal density curve is given below for each p. The quantile –quantile plot is also given.



p=0.2



p=0.5



p=0.8

2. Let us consider the parent distribution as **Poisson** with varying values of the parameter λ.

i.e. $\mathbf{X_i \sim P(\lambda)\ i = 1, 2, \dots, n}$

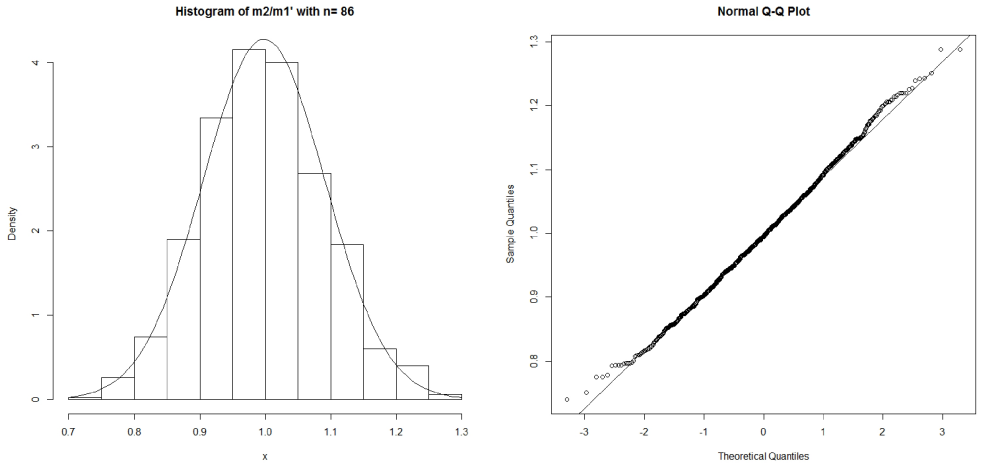
(i) λ = 1 n=86 p-value = 0.1291669

(ii) λ = 2 n=35 p-value = 0.376595

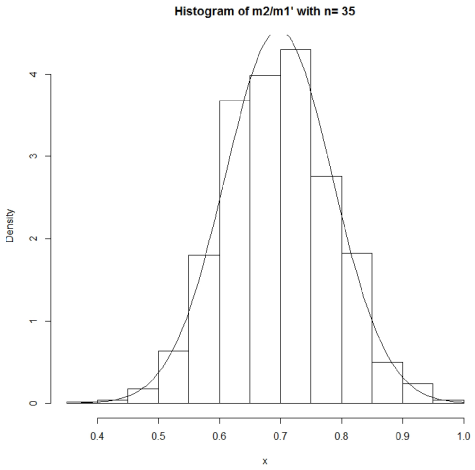
(iii) λ = 5 n=25 p-value = 0.2072368

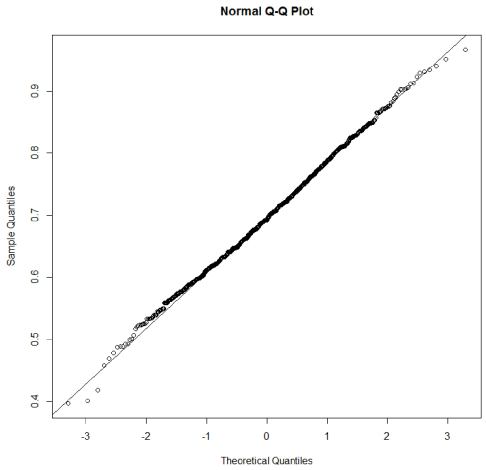
(iv) λ = 10 n=16 p-value = 0.0825968

The histogram of the distribution of the statistic T_4along with the normal density curve is given below for each λ. The quantile –quantile plot is also given.

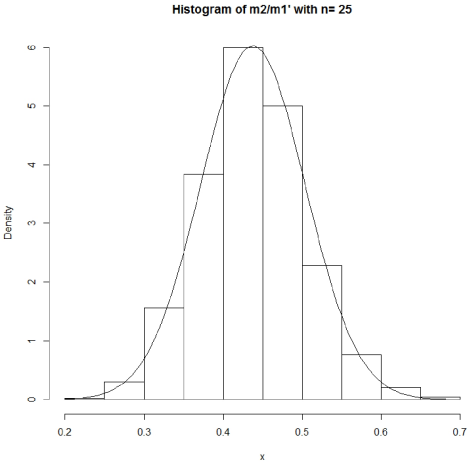


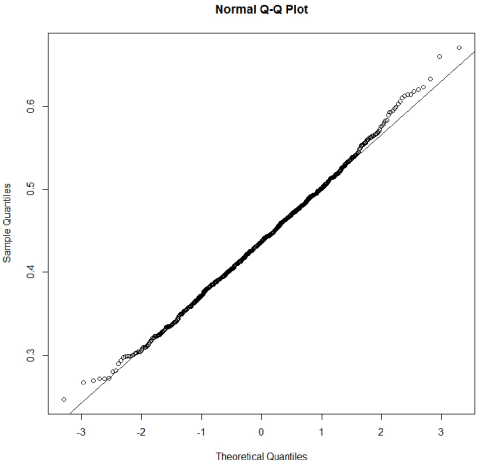
λ = 1



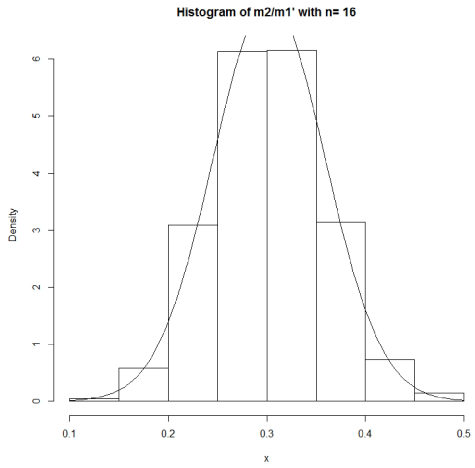


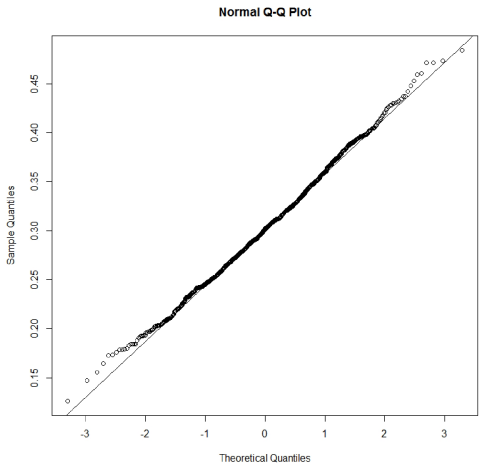
$\lambda = 2$





$\lambda = 5$





$\lambda = 10$

3. Let us consider the parent distribution as **Exponential** with varying values of the parameter λ

i.e. $\mathbf{X_i \sim E\ (\lambda)\ i = 1, 2, \dots, n}$

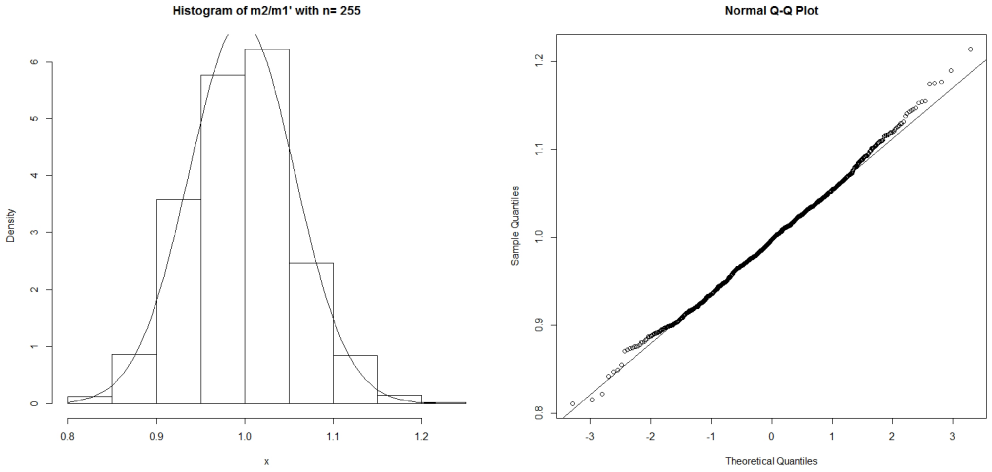
(i) λ = 0.2 n=255 p-value = 0.08316145

(ii) λ = 0.5 n=188 p-value = 0.3484245

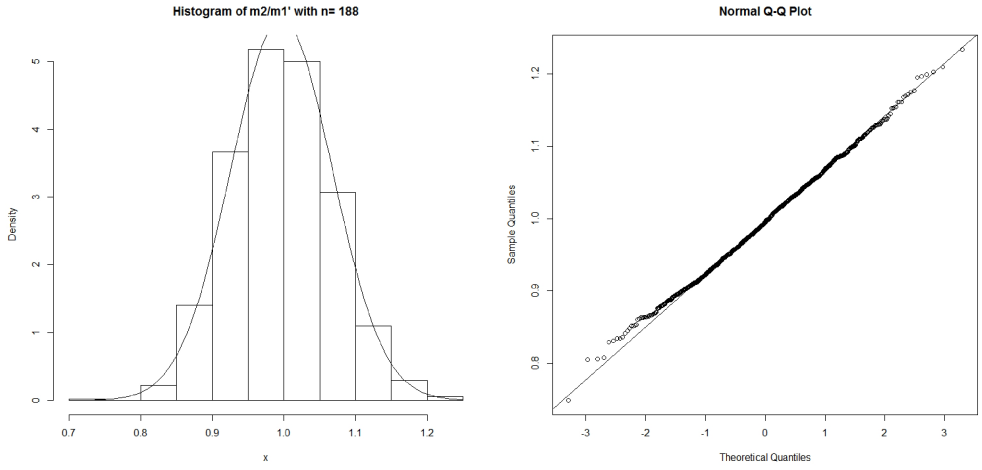
(iii) λ = 1 n=223 p-value = 0.472286

(iv) λ = 3 n=317 p-value = 0.05118578

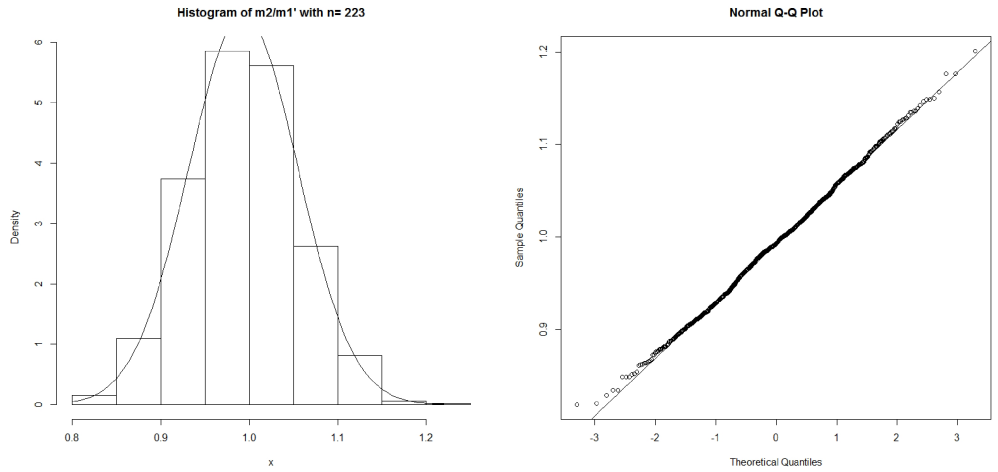
The histogram of the distribution of the statistic T_4 along with the normal density curve is given below for each λ. The quantile –quantile plot is also given.



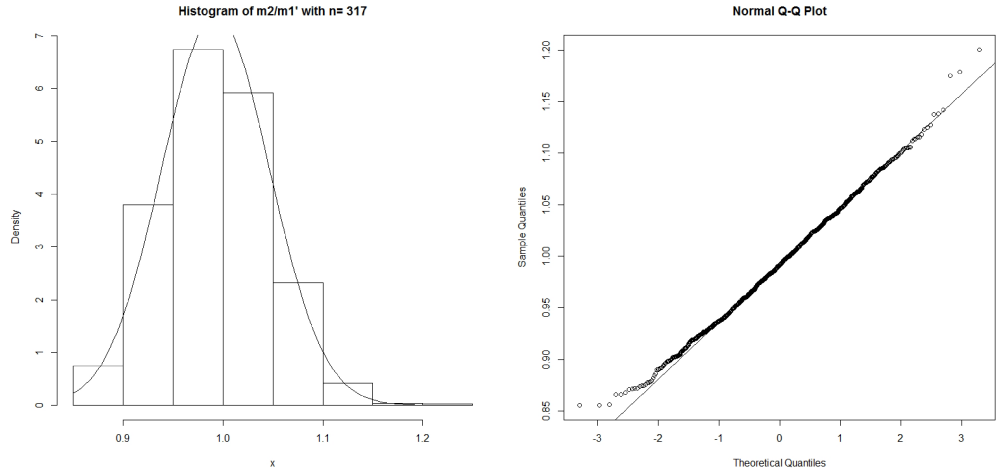
λ = 0.2



λ = 0.5



$\lambda = 1$

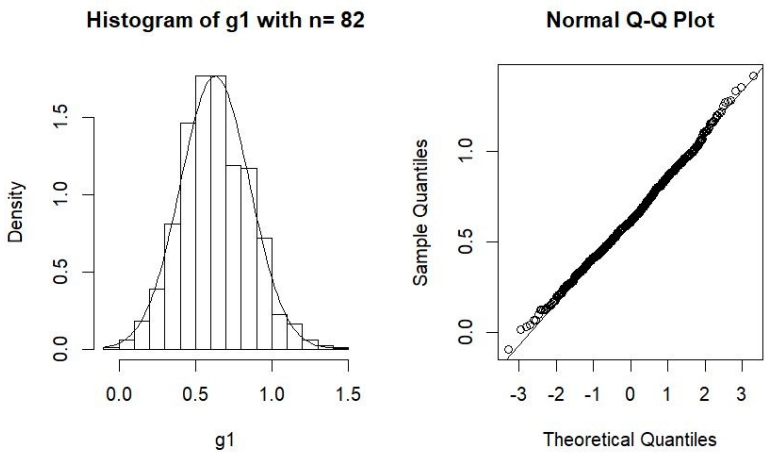


$\lambda = 3$

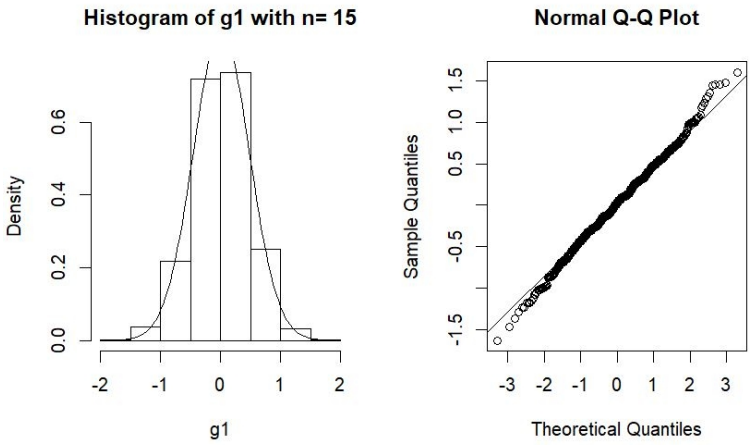
- **<u>Results obtained for the statistic g_1:</u>**

1. Let us consider the parent distribution as **Binomial** with parameters **k=5** and varying values of p i.e. $\mathbf{X_i \sim B\ \ (5, p)\ i = 1, 2, \dots, n}$
   (i) p=0.2 n=82 p-value=0.1020306
   (ii) p=0.5 n=15 p-value=0.07469013
   (iii) p=0.8 n=111p-value=0.1329466

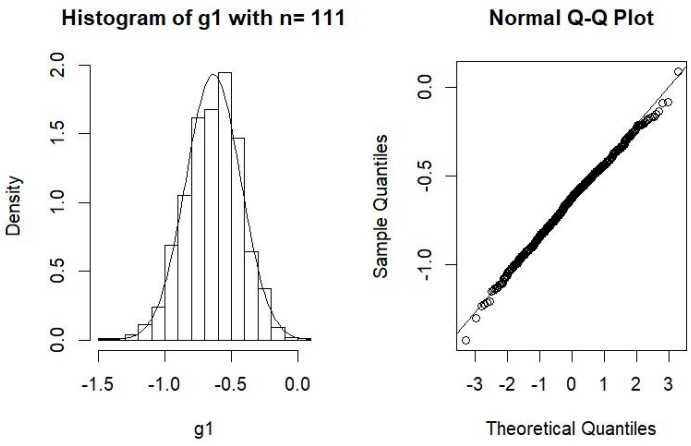
The histogram of the distribution of the statistic $\mathbf{g_1}$ along with the normal density curve is given below for each p. The quantile –quantile plot is also given.


Histogram of g1 with n= 82
Density
g1
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

p=0.2


Histogram of g1 with n= 15
Density
g1
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

p=0.5


Histogram of g1 with n= 111
Density
g1
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

p=0.8

2. Let us consider the parent distribution as **Poisson** with varying values of the parameter λ.

i.e. $\mathbf{X_i \sim P(\lambda)\ i = 1, 2, \dots, n}$

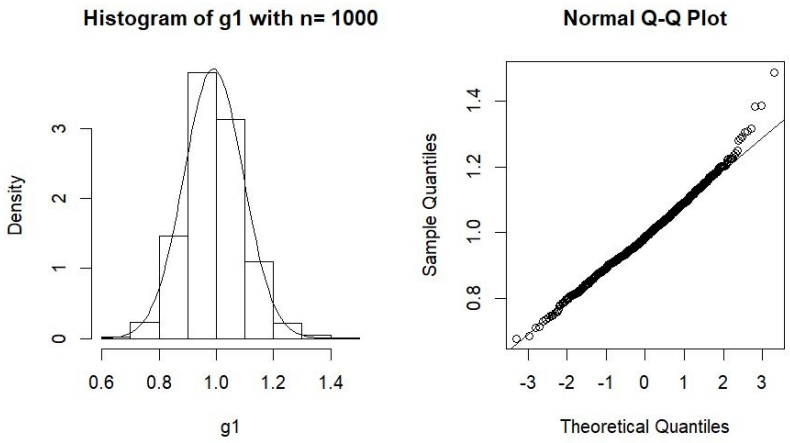
(i) λ = 1 n=1000 p-value = $7.448405e^{-5}$

(ii) λ = 2 n=8 p-value = 0.05817784

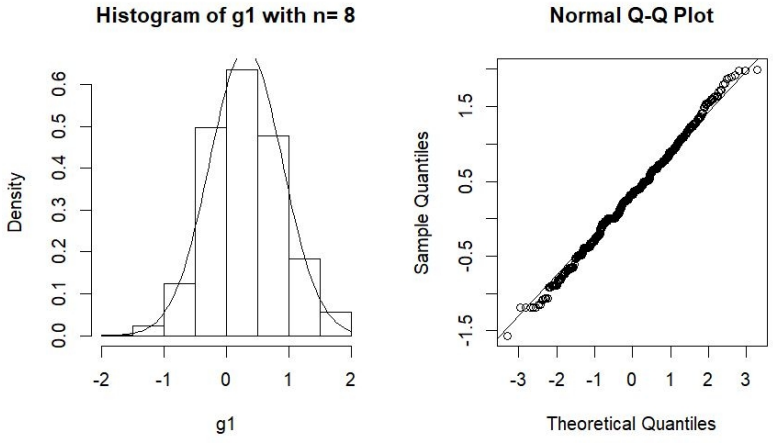
(iii) λ = 5 n=7 p-value = 0.1251024

(iv) λ = 10 n=8 p-value = 0.9319101

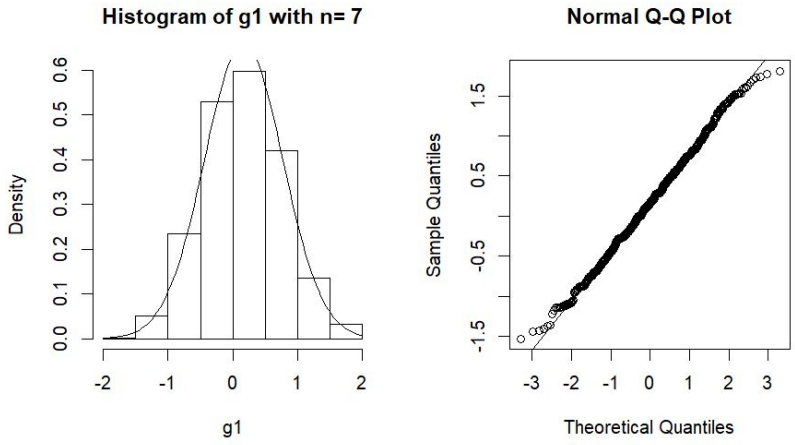
The histogram of the distribution of the statistic $\mathbf{g_1}$ along with the normal density curve is given below for each λ. The quantile –quantile plot is also given.



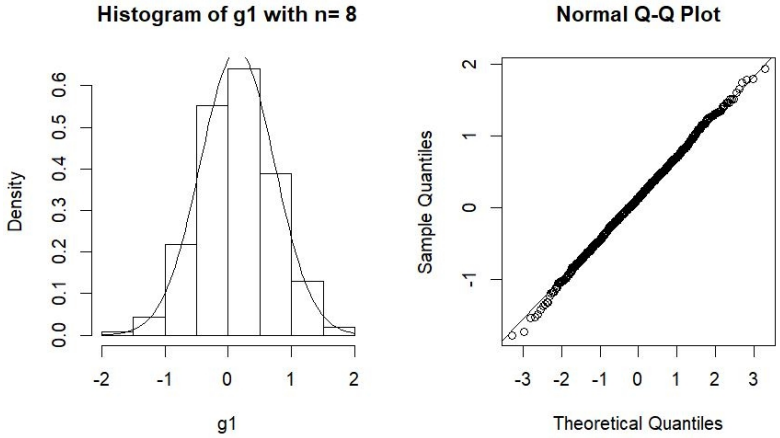
λ = 1



λ = 2



$\lambda = 5$



$\lambda = 10$

<u>Comments:</u> Here the sample skewness g1 does not converge asymptotically to the standard normal distribution for $\lambda = 1$ (the value of n comes out to be 1000 since we consider sample sizes up to 1000 in the R code).

3. Let us consider the parent distribution as **Exponential** with varying values of the parameter λ

i.e. $\mathbf{X_i \sim E \quad (\lambda)\ i = 1, 2, \dots, n}$

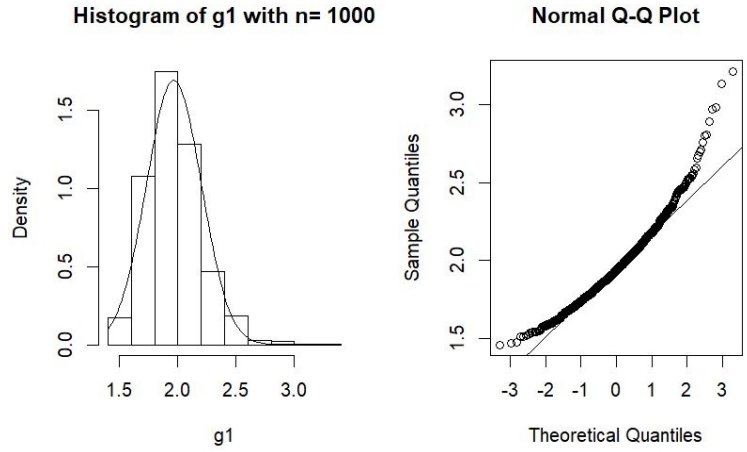
(i) $\lambda = 0.2$ n=1000 p-value = $4.920692e^{-1}$

(ii) $\lambda = 0.5$ n=1000p-value = $2.864848e^{-1}$

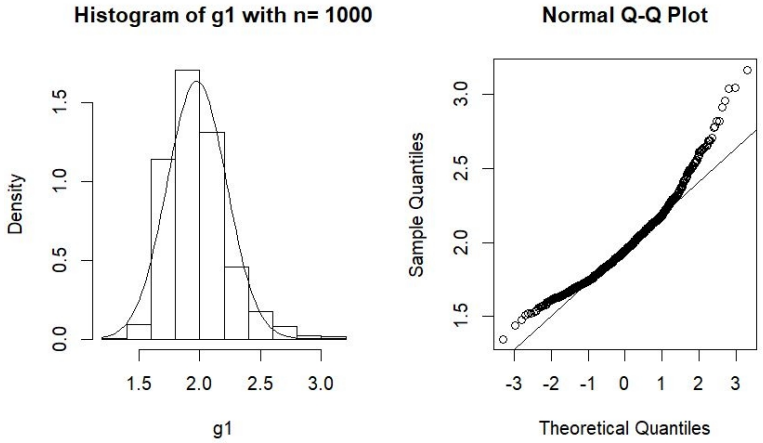
(iii) $\lambda = 1$ n=1000 p-value = $4.162683e^{-2}$

(iv) $\lambda = 3$ n=1000 p-value = $4.31622e^{-2}$

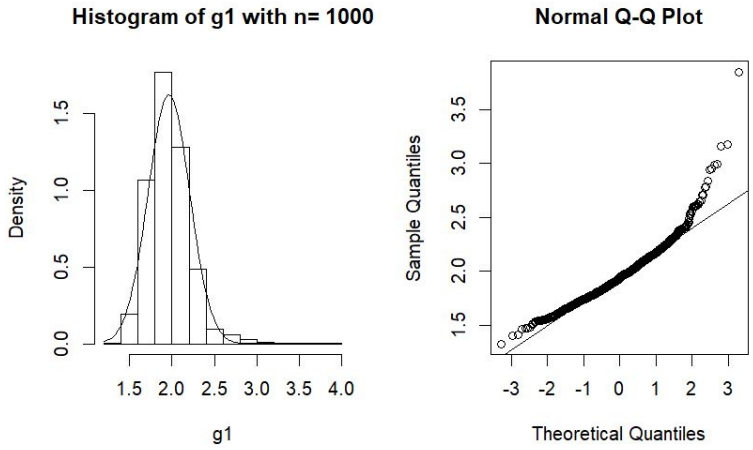
The histogram of the distribution of the statistic $\mathbf{g_1}$ along with the normal density curve is given below for each λ. The quantile –quantile plot is also given.



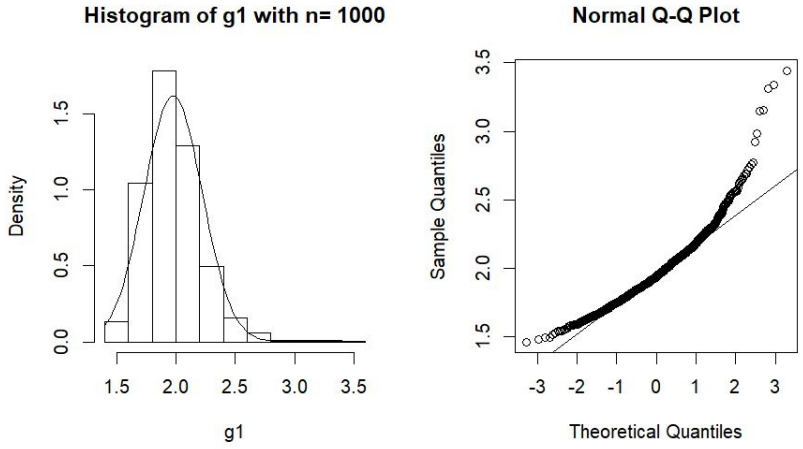
λ = 0.2



λ = 0.5



λ = 1



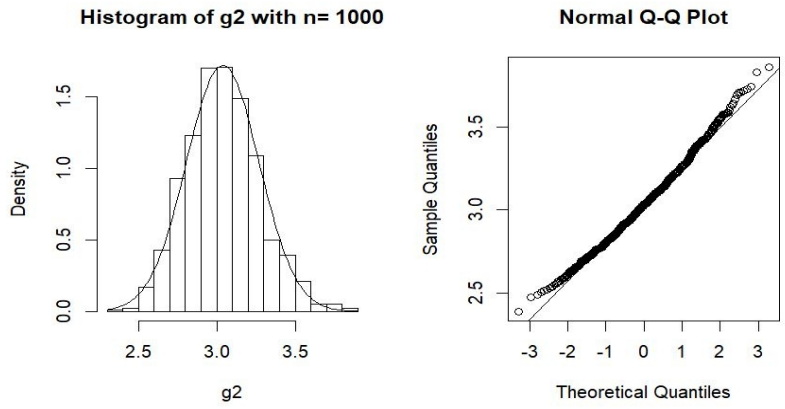
$\lambda = 3$

<u>Comments:</u> Here the sample skewness g1 does not converge asymptotically to the standard normal distribution for λ = 0.2,0.5,1 and 3 (the value of n comes out to be 1000 since we consider sample sizes up to 1000 in the R code).

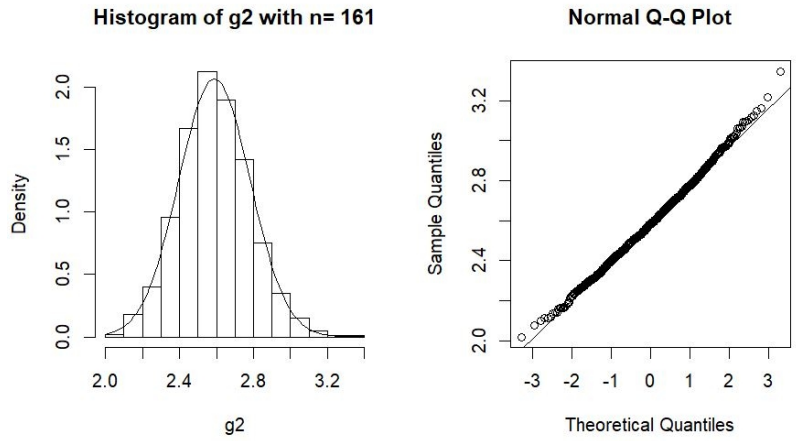
- **<u>Results obtained for the statistic g_2:</u>**

1. Let us consider the parent distribution as **Binomial** with parameters **k=5**and varying values of p i.e. $\mathbf{X_i \sim B\ \ (5, p)\ i = 1, 2, \dots, n}$
   (i) p=0.2 n=1000 p-value=0.0001310632
   (ii) p=0.5 n=161 p-value=0.9317819
   (iii) p=0.8 n=1000 p-value=$5.81909e^{-9}$

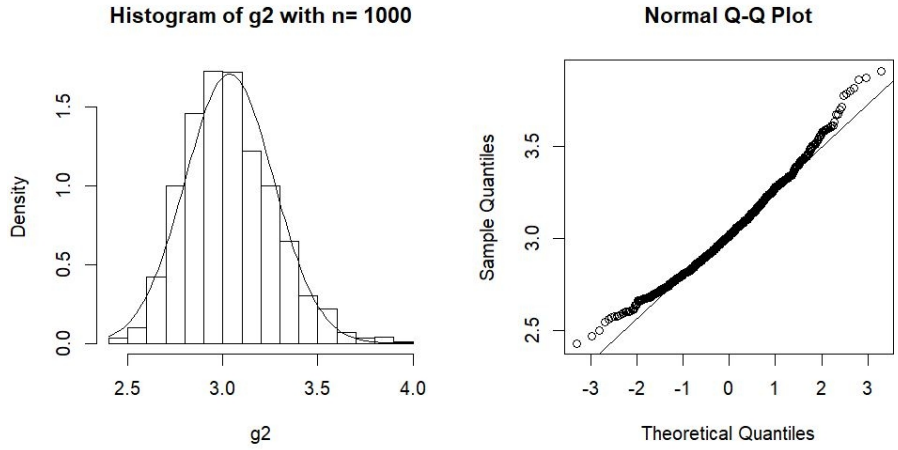
The histogram of the distribution of the statistic $\mathbf{g_2}$ along with the normal density curve is given below for each p. The quantile –quantile plot is also given.



p=0.2



p=0.5



p=0.8

<u>Comments:</u> Here the sample skewness g1 does not converge asymptotically to the standard normal distribution for p= 0.2 and 0.8 (the value of n comes out to be 1000 since we consider sample sizes up to 1000 in the R code).

2. Let us consider the parent distribution as **Poisson** with varying values of the parameter λ.
i.e. $\mathbf{X_i \sim P(\lambda)\ i = 1, 2, \dots, n}$

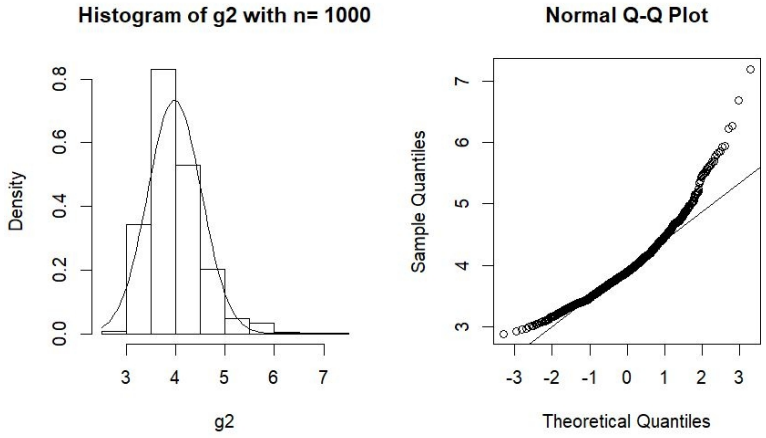
(i) λ = 1 n=1000 p-value=$1.615391e^{-2}$

(ii) λ = 2 n=1000 p-value=$4.447445e^{-1}$

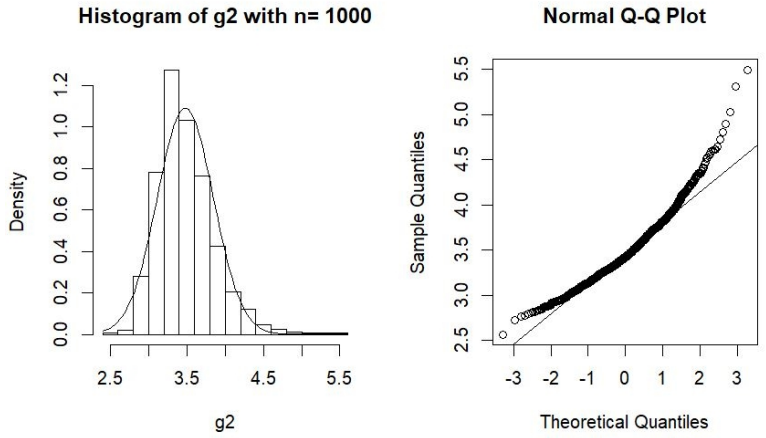
(iii) λ = 5 n=1000 p-value = $1.928747e^{-1}$

(iv) λ = 10 n=1000p-value = $3.0183862e^{-1}$

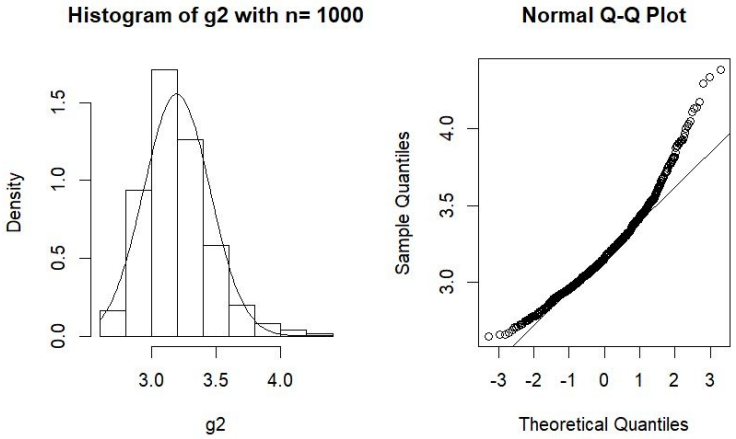
The histogram of the distribution of the statistic $\mathbf{g_2}$ along with the normal density curve is given below for each λ. The quantile –quantile plot is also given.



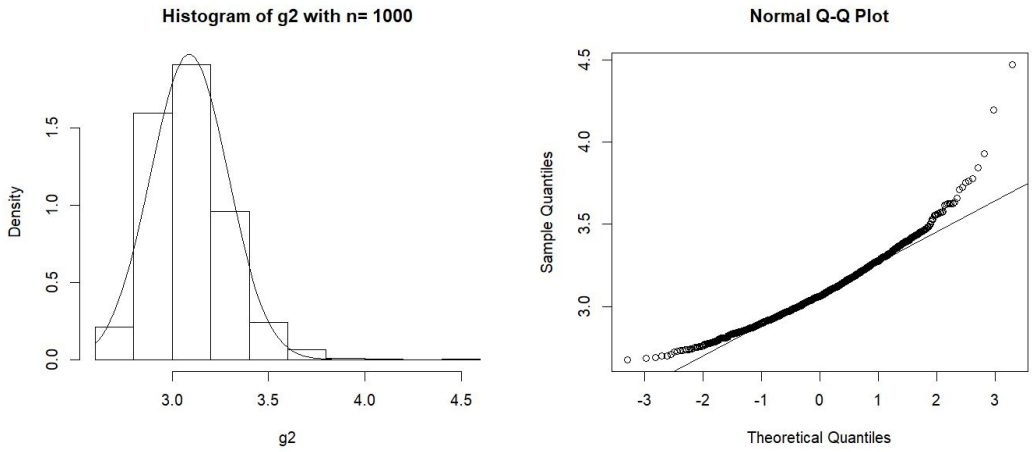
λ = 1



λ = 2



λ = 5



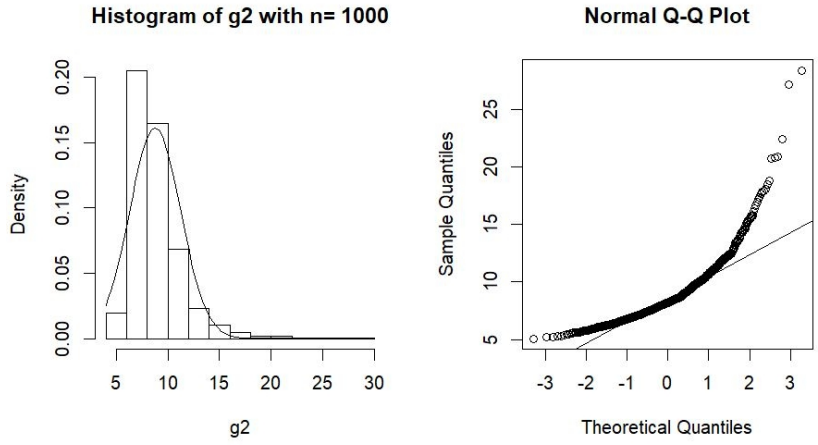
$\lambda = 10$

<u>Comments:</u> Here the sample skewness g1 does not converge asymptotically to the standard normal distribution for λ = 1,2,5 and 10 (the value of n comes out to be 1000 since we consider sample sizes upto 1000 in the R code).

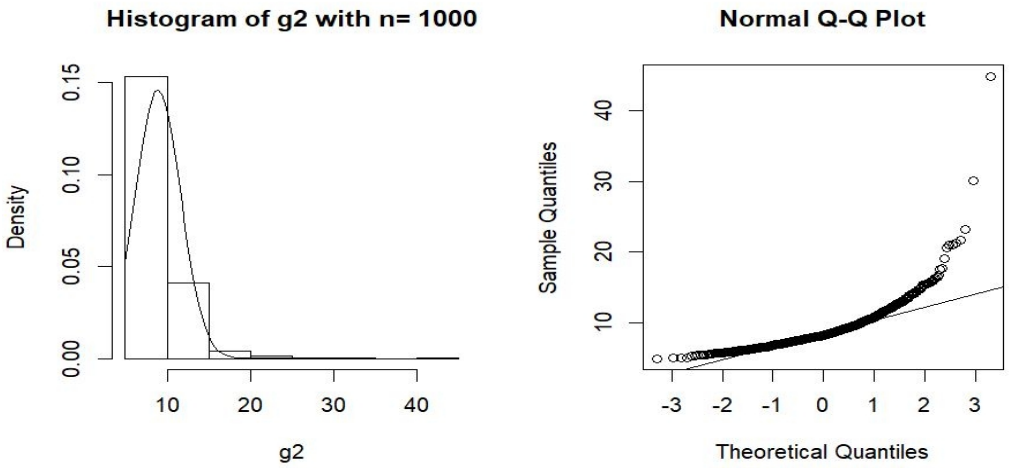
3. Let us consider the parent distribution as **Exponential** with varying values of the parameter λ
i.e. $\mathbf{X_i \sim E\ (\lambda)\ i = 1, 2, \dots, n}$

(i) λ = 0.2 n=1000 p-value = $2.637047e^{-3}$
(ii) λ = 0.5 n=1000 p-value = $1.2154e^{-3}$
(iii) λ = 1 n=1000 p-value = $3.648004e^{-3}$
(iv) λ = 3 n=1000 p-value = $7.256362e^{-2}$

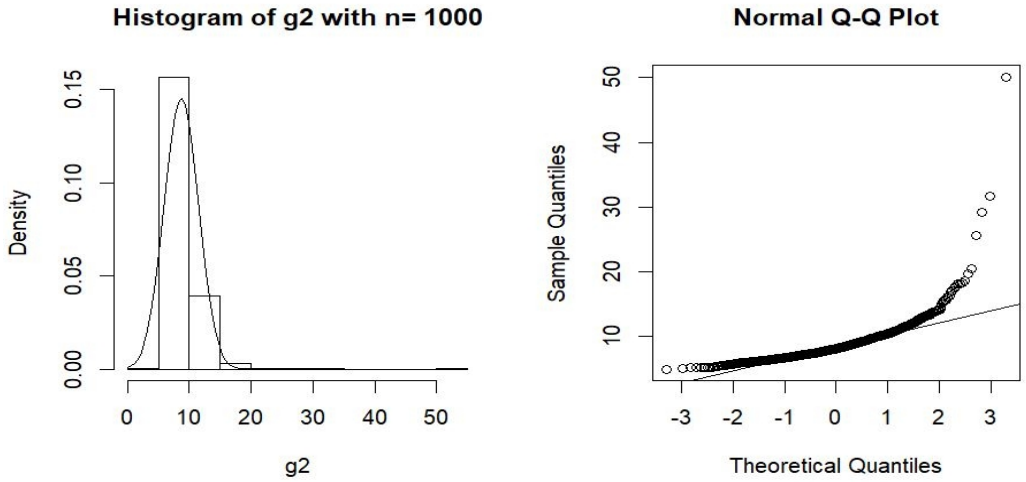
The histogram of the distribution of the statistic $\mathbf{g_2}$ along with the normal density curve is given below for each λ. The quantile –quantile plot is also given.



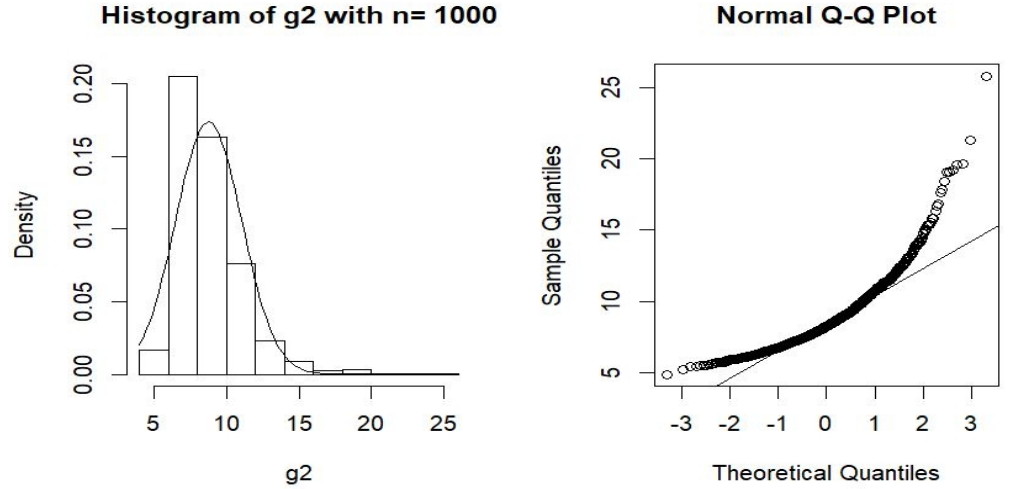
$\lambda = 0.2$



$\lambda = 0.5$



$\lambda = 1$



$\lambda = 3$

<u>Comments:</u> Here the sample skewness g1 does not converge asymptotically to the standard normal distribution for λ = 0.2, 0.5, 1and

3 (the value of n comes out to be 1000 since we consider sample sizes upto 1000 in the R code).

Examples of R code implemented for obtaining the results:

1.For the statistic T_2 in case of Bin(10,0.1) distribution:

```
p=0.1 #value of p
m=10*p #value of mean
r=1000  #no. of samples
los=0.05  #level of significance
for(n in 2:1000) #loop where sample size increases by 1 in each iteration
{
    T2=rep(0,r) #vector of size 1000 initialised to 0 to store 1000 values of T2
    s=rep(0,r)
    for(i in 1:r) #loop to generate 1000 samples
    {
        x=rbinom(n,10,p) #random sample of size n from binom(10,p)
            s[i]=sum((x-mean(x))^2)/n  #sd for each sample
        T2 [i]=(mean(x)-m)/s/sqrt(n) # T2 for each sample
            T2=na.omit(T2)
    }
    if(shapiro.test(T2)$p.value>los) #if Shapiro-Wilk's test is accepted
    {
        break #exits from loop
    }
}
hist(T2,freq=F,main=paste("Histogram of T2 with n=",n)) #histogram of that T3 for which the test was accepted is the first plot
m=mean(T2)
std=sd(T2)
curve(dnorm(x, mean=m, sd=std), add=T) #adds normal curve to histogram
```

```
shapiro.test(T2)$p.value #gives p value for the test which was accepted

n #gives sample size n for which the test was accepted first
```

2.For the statistic T_3 in case of Exp(0.2) distribution:

```
r=1000 #no. of samples

l=0.2 #lambda

los=0.05 #level of significance

for(n in 2:1000) #loop where sample size increases by 1 in each iteration

{

    T3=rep(0,r) #vector of size 1000 initialised to 0 to store 1000 values of T3

    for(i in 1:r) #loop to generate 1000 samples

    {

        x=rexp(n,l) #random sample of size n from exp(lambda), where mean=1/lambda

        T3 [i]=sqrt(sum((x-mean(x))^2)/n) # T3 for each sample

    }

    if(shapiro.test(T3)$p.value>los) #if Shapiro-Wilk's test is accepted

    {

        break #exits from loop

    }

}

par(mfrow=c(1,2)) #to draw 2 plots in a row

hist(T3, freq=F, main=paste("Histogram of T3 with n=",n)) #histogram of that T3 for which the test was accepted is the first plot

m=mean(T3)

std=sd(T3)

curve(dnorm(x, mean=m, sd=std), add=T) #adds normal curve to histogram

qqnorm(T3) #Q-Q plot is the second plot to compare quantiles of T3 with quantiles of the normal distribution

qqline(T3) #adds the diagonal line in the Q-Q plot

shapiro.test(T3)$p.value #gives p value for the test which was accepted

n #gives sample size n for which the test was accepted first
```

REFERENCES:

1. An Introduction to Probability and Statistics, Rohatgi , V.K. and Saleh, A.K.M.E.,John Wiley & Sons,1976

2. An Outline of Statistical Theory Vol.-I , Gun,A.M.,Gupta, M.K., Dasgupta,B.,The World Press Pvt. Limited,2003

3. Laws of Large Numbers, Chandra T.K., Alpha Science International Pvt. Limited, 2012