UPPSALA UNIVERSITET

Department of Statistics

Probability Theory and Statistical Inference I, fall 2019

Computer Assignment

Authors: Mayara Latrech & Roua Benammar

1 Theoretical distributions

In this section, we find the distribution and the parameters of five statistics $Z_1, Z_2, ..., Z_5$. Where:

$$Z_1 = \bar{X}$$

$$Z_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$$

$$Z_3 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2}$$

$$Z_4 = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)\sigma^2}}}$$

$$Z_5 = \frac{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)\sigma^2}}{\frac{\sum_{i=1}^{m}(Y_i-\bar{Y})^2}{(m-1)\sigma^2}}$$

Such that, $X_1, ..., X_n$ is a random sample of n independent observations from a population that follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where $\mu = 0$ and $\sigma^2 = 1$.
$Y_1, ..., Y_m$ is also a random sample f m independent observations from a population that follows the same normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Distribution of Z_1

According to THEOREM 7.1, $X_1, ..., X_n$ is a random sample of size n with mean μ and vanriance σ^2 then $\bar{X}$ follows a normal distribution with mean μ and variance σ^2/n.

Hence,

$$Z_1 \sim \mathcal{N}(\mu, \sigma^2/n)$$

According to THEOREM 4.7, Z_1 is a normally distributed random variable with parameters μ and σ^2/n, then
$E[Z_1] = \mu$
$Var(Z_1) = \sigma^2/n$
$Sd(Z_1) = \sqrt{Var(Z_1)} = \sigma/\sqrt{n}$

Distribution of Z_2

According to THEOREM 7.3, $X_1, ..., X_n$ is a random sample of size n with mean μ and vanriance σ^2 then $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$ follows a $\mathcal{X}^2$ distribution with $(n-1)$ degrees of freedom.

Hence,

$$Z_2 \sim \mathcal{X}^2(n-1)$$

$E[Z_2] = n - 1$
$Var(Z_2) = 2(n-1)$
$Sd(Z_2) = \sqrt{2(n-1)}$

Distribution of Z_3

According to THEOREM 7.2, $X_1, ..., X_n$ is a random sample of size n with mean μ and vanriance σ^2 then $\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2}$ follows a $\mathcal{X}^2$ distribution with n degrees of freedom.

Hence,

$$Z_3 \sim \mathcal{X}^2(n)$$

$E[Z_3] = n$
$Var(Z_3) = 2n$
$Sd(Z_3) = \sqrt{2n}$

Distribution of Z_4

Let $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. We know that $Z_1 = \bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ so, Z follows a normal distribution.
$E[Z] = E[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}] = \frac{1}{\sigma/\sqrt{n}}(E[\bar{X}] - \mu) = \frac{1}{\sigma/\sqrt{n}}(\mu - \mu) = 0$
$Var(Z) = \frac{1}{\sigma^2/n}(Var(\bar{X} - 0) = \frac{1}{\sigma^2/n}\sigma^2/n = 1$

Z_4 can be rewritten as $Z_4 = \frac{Z}{\sqrt{Z_2/(n-1)}}$ where Z is a standard normal random variable and $Z_2 \sim \mathcal{X}^2(n-1)$.

According to DEFINITION 7.2, Z_4 has a t distribution with $(n-1)$ degrees of freedom.

$$Z_4 \sim t(n-1)$$

$E[Z_4] = 0$ for $n > 2$
$Var(Z_4) = \frac{n-1}{n-3}$ for $n > 3$
$Sd(Z_4) = \sqrt{\frac{n-1}{n-3}}$ for $n > 3$

Distribution of Z_5

$Z_2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$ follows a $\mathcal{X}^2$ distribution with $(n-1)$ degrees of freedom.
Similarly, $Z_2^* = \frac{\sum_{i=1}^{m}(Y_i-\bar{Y})^2}{\sigma^2}$ follows a $\mathcal{X}^2$ distribution with $(m-1)$ degrees of freedom.
Z_5 can be rewritten as $Z_5 = \frac{Z_2/(n-1)}{Z_2^*/(m-1)}$.

According to DEFINITION 7.3, Z_5 has an F distribution with $(n-1)$ numerator degrees of freedom and $(m-1)$ denominator degrees of freedom.

$$Z_5 \sim F(n-1, m-1)$$

$E[Z_5] = \frac{m-1}{m-3}$ for $m > 3$
$Var(Z_5) = \frac{2(m-1)^2(n+m-4)}{(n-1)(m-3)^2(m-5)}$ for $m > 5$
$Sd(Z_5) = \sqrt{\frac{2(m-1)^2(n+m-4)}{(n-1)(m-3)^2(m-5)}}$ for $m > 5$ & $n > 1$

Numerical Application

Let us calculate the expected value, the variance and the standard deviation of the statistics Z_1 to Z_4 for $n = 5$ and $n = 20$.

	$n = 5$			$n = 20$		
	$E(Z)$	$Var(Z)$	$Sd(Z)$	$E(Z)$	$Var(Z)$	$Sd(Z)$
Z_1	0	0.2	0.447	0	0.05	0.224
Z_2	4	8	2.828	19	38	6.164
Z_3	5	10	3.162	20	40	6.325
Z_4	0	2	1.414	0	1.118	1.057

For $n = 5$ and $m = 20$ we get,
$E[Z_5] = 1.118$
$Var(Z_5) = 0.874$
$Sd(Z_5) = 0.935$

2 Empirical distributions

Statistics Z_1 to Z_4

Case of $n = 5$

In this section, we draw 1000 samples of size $n = 5$ from the population. For each of the four statistics, we calculate the mean, variance, and standard deviation of the 1000 samples and compare the obtained results with the theoretical values.

	$E(Z)$		$Var(Z)$		$Sd(Z)$	
	Computed	Theoretical	Computed	Theoretical	Computed	Theoretical
Z_1	0.0345	0	0.198	0.2	0.445	0.447
Z_2	4.148	4	7.683	8	2.771	2.828
Z_3	4.873	5	9.704	10	3.115	3.162
Z_4	0.0677	0	1.793	2	1.339	1.414

We notice that the results are comparable to the theoretical values in each of the four statistics.
Now let us compare the computed the histograms to the theoretical distribution graphs.

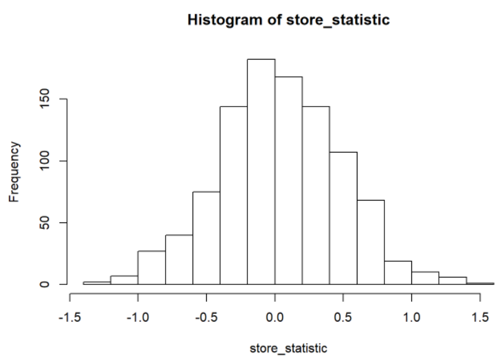
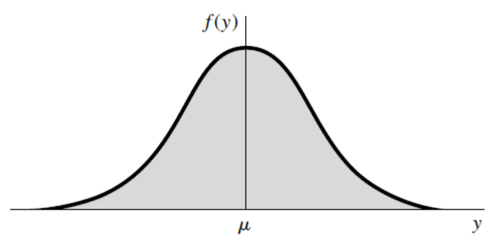


FIGURE **4.10** The normal probability density function



The histogram of Z_1 is indeed comparable to the theoretical graph of the normal distribution.

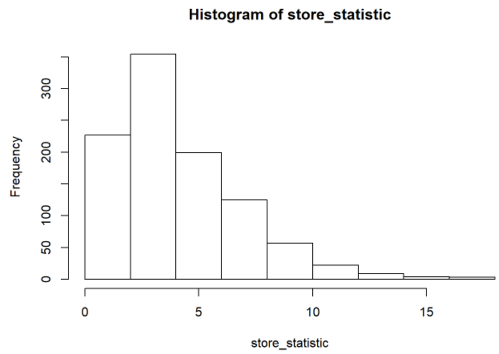
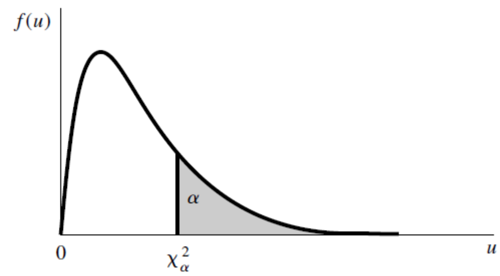


FIGURE 7.2
A χ^2 distribution showing upper-tail area α



The histogram of Z_2 is indeed comparable to the theoretical graph of the $\mathcal{X}^2$ distribution.

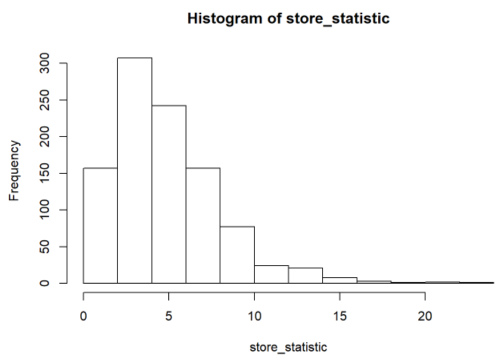
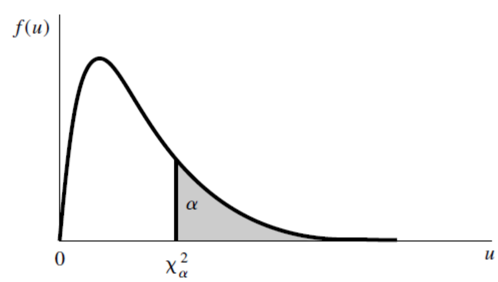


FIGURE 7.2
A χ^2 distribution showing upper-tail area α



The histogram of Z_3 is indeed comparable to the theoretical graph of the $\mathcal{X}^2$ distribution.

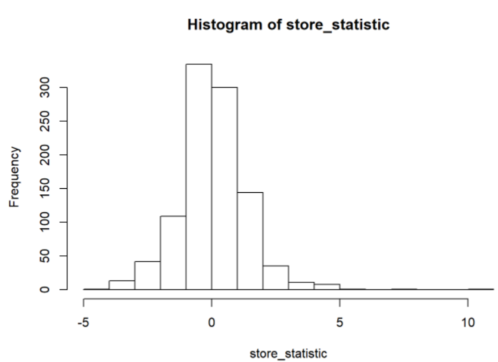
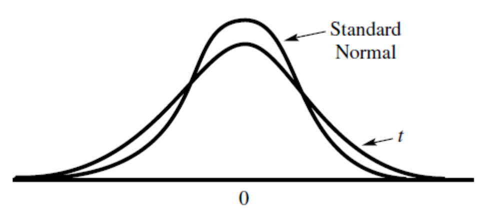


FIGURE 7.3
A comparison of the standard normal and t density functions.



The histogram of Z_4 is indeed comparable to the theoretical graph of the t distribution.

Case of $n = 20$

We repeat the same experiment for $n = 20$ and we obtain the following results.

	$E(Z)$		$Var(Z)$		$Sd(Z)$	
	Computed	Theoretical	Computed	Theoretical	Computed	Theoretical
Z_1	0.001	0	0.0494	0.05	0.222	0.224
Z_2	18.843	19	38.368	38	6.194	6.164
Z_3	19.929	20	42.232	40	6.499	6.325
Z_4	0.001	0	1.153	1.118	1.074	1.057

Now let us compare the computed the histograms to the theoretical distribution graphs.

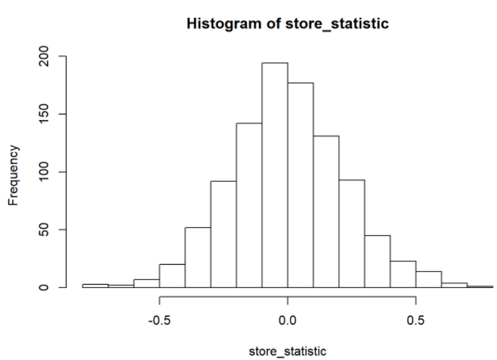
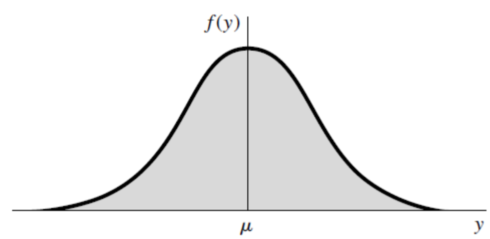


FIGURE 4.10 The normal probability density function



The histogram of Z_1 is indeed comparable to the theoretical graph of the normal distribution.

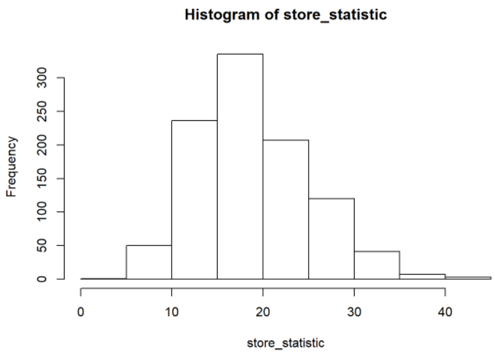
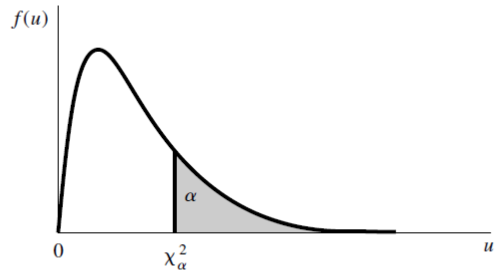


FIGURE 7.2
A χ^2 distribution showing upper-tail area α



The histogram of Z_2 is indeed comparable to the theoretical graph of the $\mathcal{X}^2$ distribution.

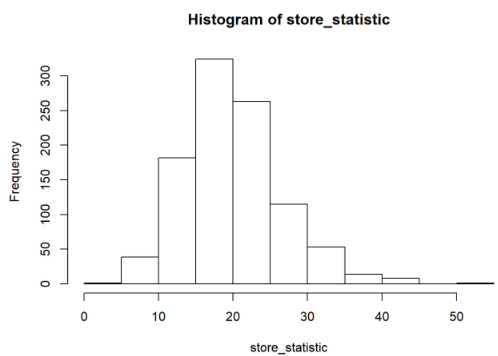
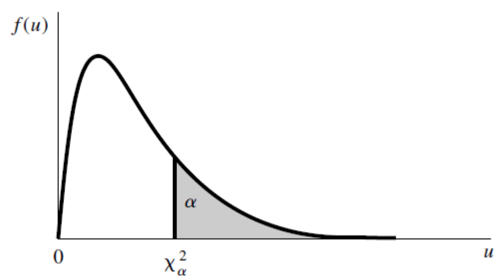


FIGURE 7.2
A χ^2 distribution showing upper-tail area α



The histogram of Z_3 is indeed comparable to the theoretical graph of the $\mathcal{X}^2$ distribution.

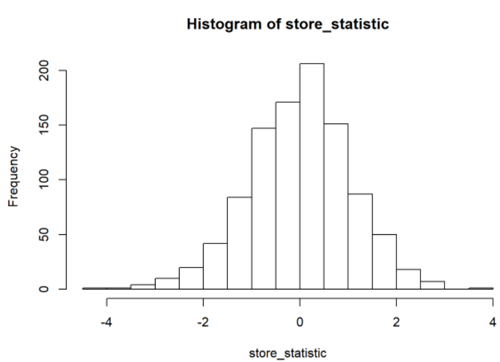
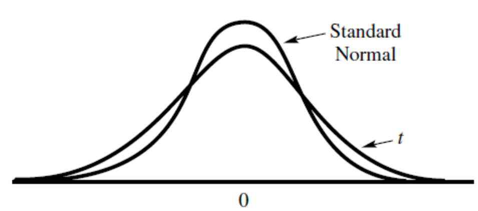


FIGURE 7.3
A comparison of the standard normal and t density functions.



The histogram of Z_4 is indeed comparable to the theoretical graph of the t distribution.

Statistic Z_5

In this section, we draw 1000 samples of Z_5 with $n = 5$ and $m = 20$ and calculate the mean, the variance and the standard deviation.

	Computed value	Theoretical value
$E(Z)$	1.274	1.118
$Var(Z)$	0.873	0.874
$Sd(Z)$	0.934	0.935

The computed values are comparable to the theoretical values for the Z_5 statistic.
Now let us compare the histogram of Z_5 to its theoretical distribution graph.

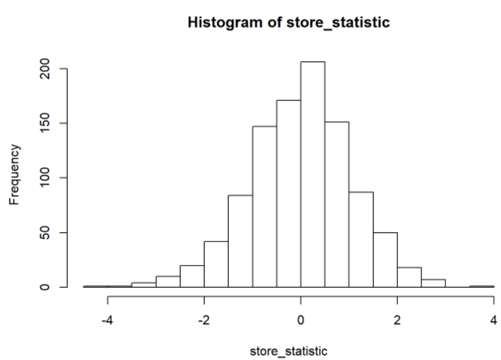
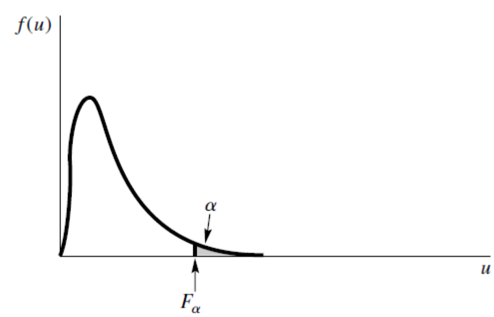


FIGURE 7.4
A typical F probability density function



As expected, the histogram of Z_5 is comparable to the theoretical graph of the F distribution.
In the index we put the detailed R code we used.

Index

Assignment 2:

#Evaluating Z1 for n=5

```
#Drawing 1000 samples with 5 observations each from a normally distributed population
sample<-matrix(rnorm(n=5000,m=0,sd=1),nrow = 1000, ncol=5 )
store_statistic<-numeric(1000)
#Evaluating Z1 (Z1: Empirical mean)
for(i in 1:1000){
  row<-sample[i,]
  store_statistic[i]<-mean(row)
}
#Calculating , the mean , variance and standard deviation:
#The thoritical Expected value=0,
mean(store_statistic)
```

```
## [1] 0.03457911
```

```
#The theoritical variance is 1/5(=0.2)
var(store_statistic)
```

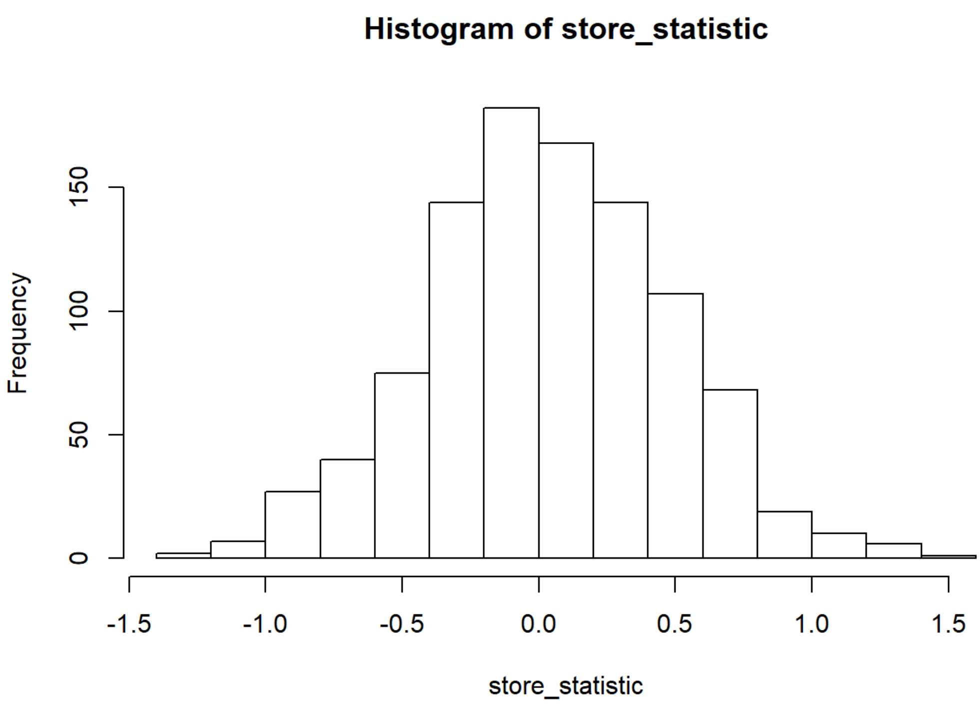
```
## [1] 0.1983152
```

```
#The theoritical standard deviation is sqd(1/5)= 0.447
sd(store_statistic)
```

```
## [1] 0.445326
```

#Graphical Illustration of the distribution of Z1 , (n=5)

```
hist(store_statistic)
```



#For Z1 when n=5 , The empirical mean is equal to 0.02470494 which is slightly larger than the theoritical expected value of 0 but still somewhat close.The empirical values of the variance (0.1867699) and standard deviation (0.4321688) are consistent with the theoritical values of 1/5 and sqr(1/5) respectively.

#Evaluating Z1 for n=20

```
#Drawing 1000 samples from the population with n=20
sample <- matrix(rnorm(n=20000, m=0, sd= 1), nrow = 1000, ncol = 20)
store_statistic <- numeric(1000)
# Z1
for(i in 1:1000){
  row <- sample[i,]
  store_statistic[i] <- mean(row)
}
#Calculating , the mean , variance and standard deviation:
mean(store_statistic)
```

```
## [1] 0.0007333152
```

```
var(store_statistic)
```

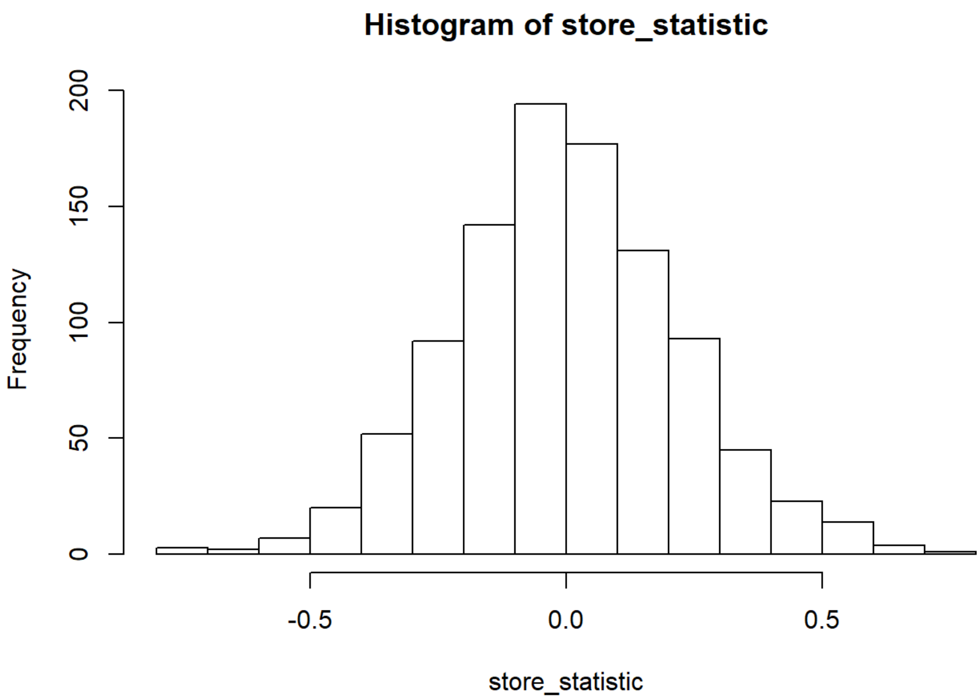
```
## [1] 0.04937548
```

```
sd(store_statistic)
```

```
## [1] 0.2222059
```

```
#Graphical Illustration of the distribution of Z1, (n=20)

hist(store_statistic)
```



#For Z1 , when n=20, the theoritical mean is equal to 0.01247727 which is closer to the theoritical mean of 0 compared to the previous simulation where n= 5. The distribution is more centralized compared to the theoritical results and the results of the previous simulation as the empirical variance obtained is 0.01247727 and the standard deviation is 0.2299652

#Evaluating Z2 for n=5:

```r
sample<-matrix(rnorm(n=5000,m=0,sd=1),nrow = 1000, ncol=5 )
store_statistic<-numeric(1000)
for(i in 1:1000){
  row<-sample[i,]
  k<-1
  m=mean(row)
  store_statistic[i]<-sum((row-m)^2)*k
}

#Calculating , the mean , variance and standard deviation for Z2:
#Theoritical mean = 4
mean(store_statistic)
```

```
## [1] 4.148241
```

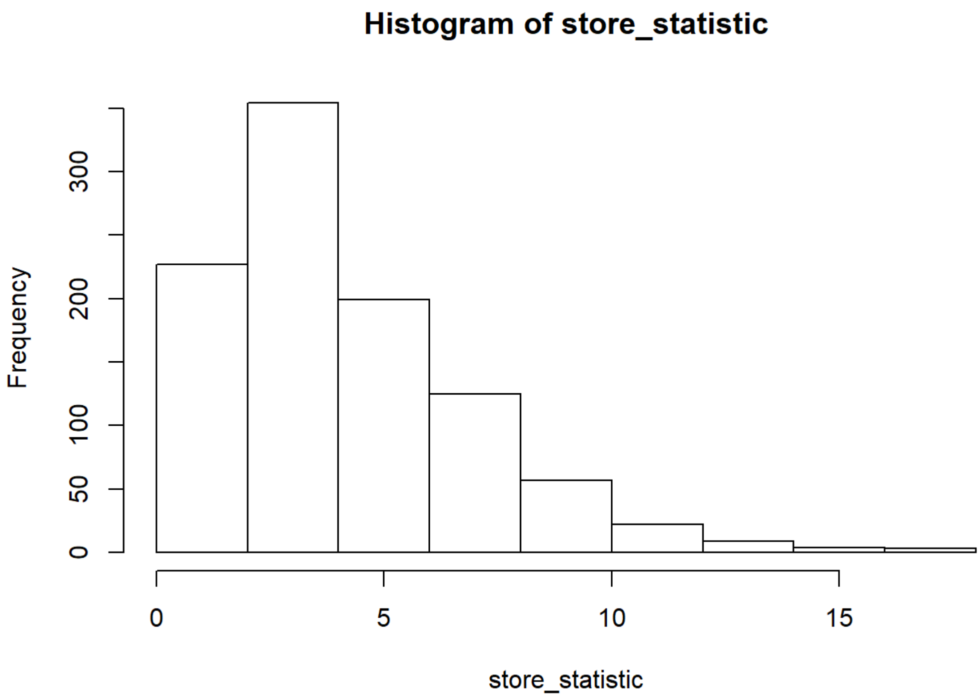
```r
#Theoritical variance =8
var(store_statistic)
```

```
## [1] 7.683631
```

```r
#the theoritical sd is sqr(8)=2.828427
sd(store_statistic)
```

```
## [1] 2.771936
```

```r
#Graphical Illustration of the distribution of Z2, (n=5):
hist(store_statistic)
```



#For Z2 when n=5, the empirical mean is equal to 3.860413 which is slightly lower than the theoritical value of 4. The empirical variance is equal to 8.37643 and the sd is 2.894206 , both of these values are slightly higher than the theoritical values

#Evaluating Z2 for n=20

```r
sample <- matrix(rnorm(n=20000, m=0, sd= 1), nrow = 1000, ncol = 20)
store_statistic <- numeric(1000)

for(i in 1:1000){
  row <- sample[i,]
  k <- 1
  m= mean(row)
  store_statistic[i]<- sum((row-m)^2)*k
}
#Calculating , the mean , variance and standard deviation for Z2:
#Theoritical mean = 19
mean(store_statistic)
```

```
## [1] 18.84303
```

```r
#Theoritical variance = 38
var(store_statistic)
```

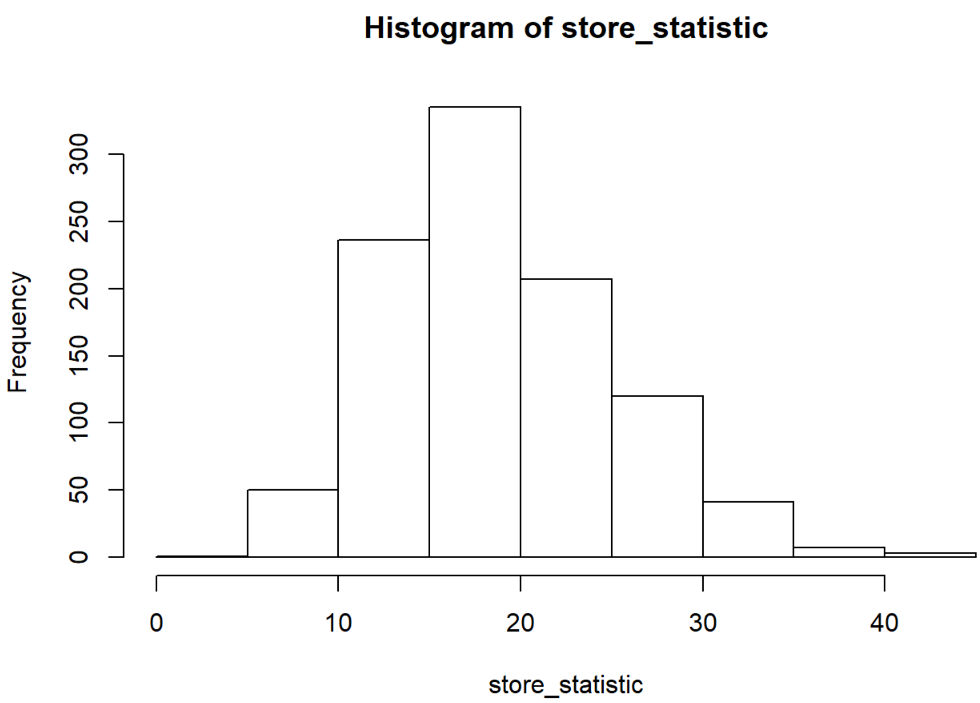
```
## [1] 38.36808
```

```r
#theoritic sd= 6.164414
sd(store_statistic)
```

```
## [1] 6.194198
```

#Graphical Illustration of the distribution of Z2, (n=20):

```r
hist(store_statistic)
```



#The empirical mean is 19.00542 which very close to the theoritical value of 19.The variance is 38.06069 and the standard deviation is 6.169334 , both of these empirical values are consistent with the theoritical values of 38 and 6.164414, although they are slightly larger

#Evaluating Z3 for n=5,

```
sample<-matrix(rnorm(n=5000,m=0,sd=1),nrow = 1000, ncol=5 )
store_statistic<-numeric(1000)
for(i in 1:1000){
  row<-sample[i,]
  k<-1
  mu=0
  store_statistic[i]<-sum((row-mu)^2)*k
}

#Calculating , the mean , variance and standard deviation for Z3:
#Theoritical mean = 5
mean(store_statistic)
```

```
## [1] 4.873743
```

```
#Theoritical variance=10
var(store_statistic)
```

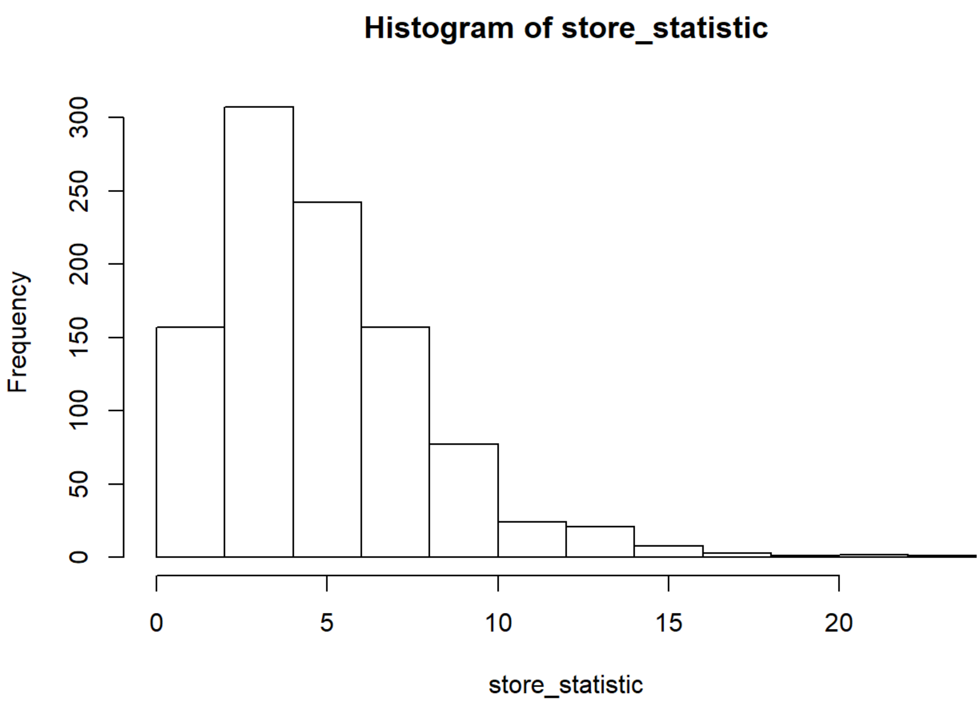
```
## [1] 9.704561
```

```
#theoritical sd= sqr(10)=3.162277
sd(store_statistic)
```

```
## [1] 3.115214
```

#Graphical Illustration of the distribution of Z3 , (n=5):

```
hist(store_statistic)
```



#For Z3 when n= 5, the empirical mean = 4.948427 , the variance is 9.6804 and the sd is 3.111334. These values are slightly below the theoritical values of 5 , 10 and sqr of 10 respectively

#Evaluating Z3 for n=20

```
sample <- matrix(rnorm(n=20000, m=0, sd= 1), nrow = 1000, ncol = 20)
store_statistic <- numeric(1000)

for(i in 1:1000){
  row <- sample[i,]
  k <- 1
  store_statistic[i]<- sum((row-0)^2)*k
}
#Calculating , the mean , variance and standard deviation for Z3:
#theoritical mean = 20
mean(store_statistic)
```

```
## [1] 19.92902
```

```
#theoritical variance = 40
var(store_statistic)
```

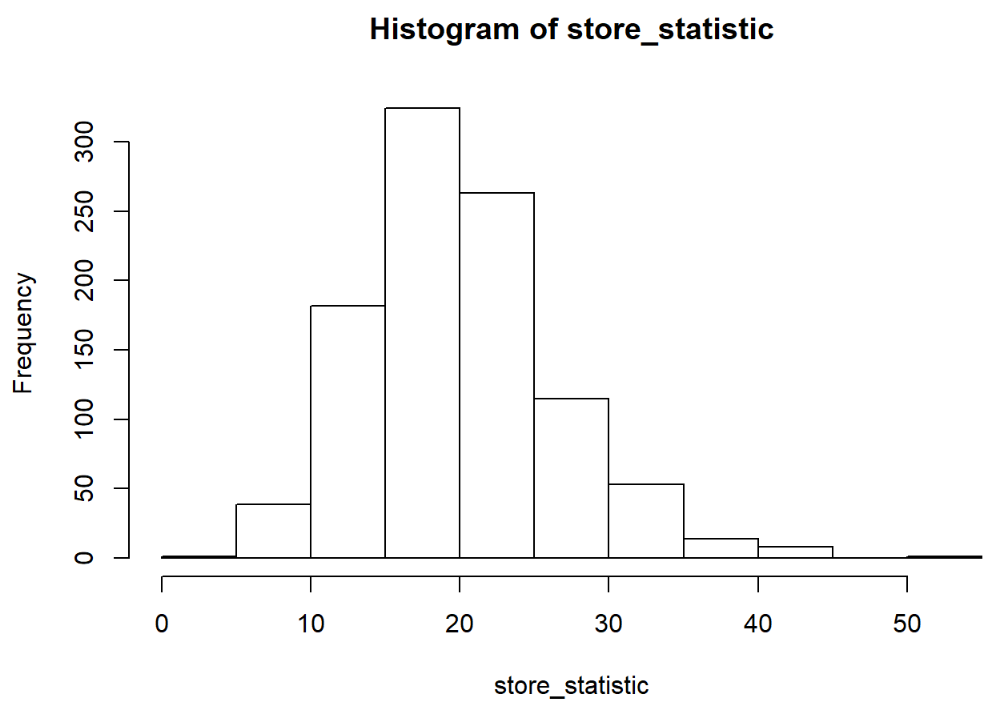
```
## [1] 42.23243
```

```
#Theoritical sd= 6.32455
sd(store_statistic)
```

```
## [1] 6.498648
```

#Graphical Illustration of the distribution of Z3 , (n=20):

```
hist(store_statistic)
```



For z3 , when n= 20 , the empirical mean , variance and sd are equal to 19.93316, 40.81489 and 6.388653 respectively , these empirical mean is slightly below the theoritical mean of 20 , and the variance and the sd are both slightly higher than the theoritical values of 40 and 6.32455

#Evaluating Z4 for n=5

```
sample <- matrix(rnorm(n=5000, m=0, sd= 1), nrow = 1000, ncol = 5)
store_statistic <- numeric(1000)

for(i in 1:1000){
  row <- sample[i,]
  m=  mean(row)
  mu=0
  n=5
  store_statistic[i]<-(n^0.5)*(m-mu)/((sum((row-m)^2)/(n-1))^0.5)
}

#Calculating , the mean , variance and standard deviation for Z4:
#theoritical mean= 0
mean(store_statistic)
```

```
## [1] 0.0677109
```

```
#theoritical var= 2
var(store_statistic)
```

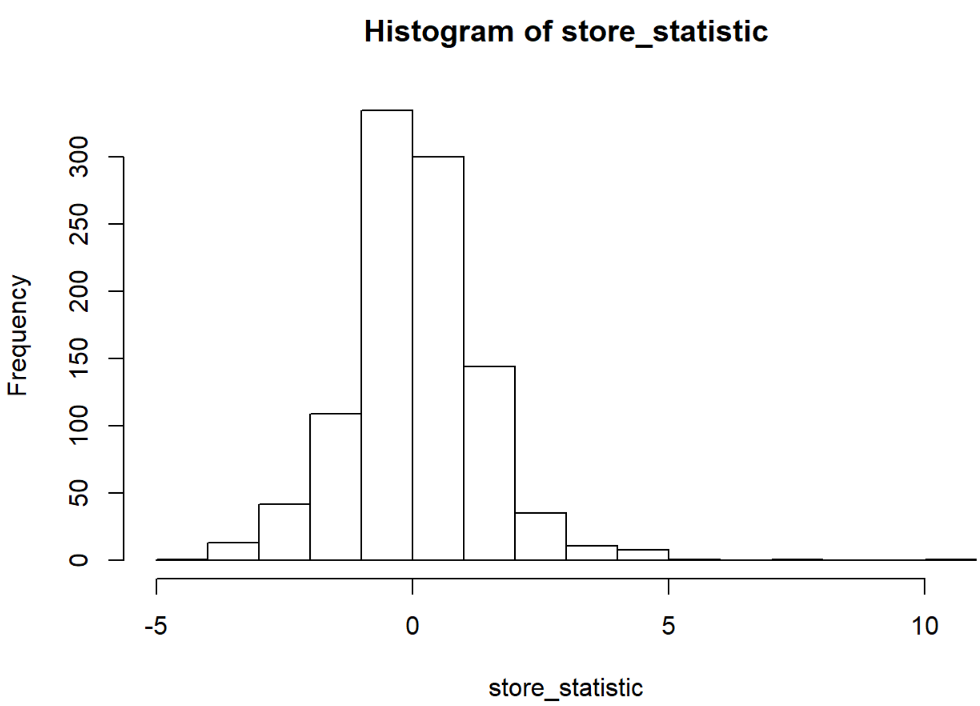
```
## [1] 1.793697
```

```
#theoritical  sd = sqr(2)=1.41
sd(store_statistic)
```

```
## [1] 1.33929
```

#Graphical Illustration of the distribution of Z4, (n=5)

```
hist(store_statistic)
```



#For Z4 when n=5, the mean , variance and sd are : 0.00217372 , 2.42553 and 1.557414 respectively. The results are somewhat comparable to the theoritical values , although , the mean is slightly higher than the theoritical value of 0 and the variance and the standard variation are slightly above the theritical values of 2 and 1.41 respectively

#Evaluating Z4 for n=20

```
sample <- matrix(rnorm(n=20000, m=0, sd= 1), nrow = 1000, ncol = 20)
store_statistic <- numeric(1000)

for(i in 1:1000){
  row <- sample[i,]
  k <- 1
  m=  mean(row)
  mu=0
  n=20
  store_statistic[i]<- (n^0.5)*(m-mu)/((sum((row-m)^2)/(n-1))^0.5)
}
#Calculating , the mean , variance and standard deviation for Z4:
mean(store_statistic)
```

```
## [1] 0.001306036
```

```
var(store_statistic)
```

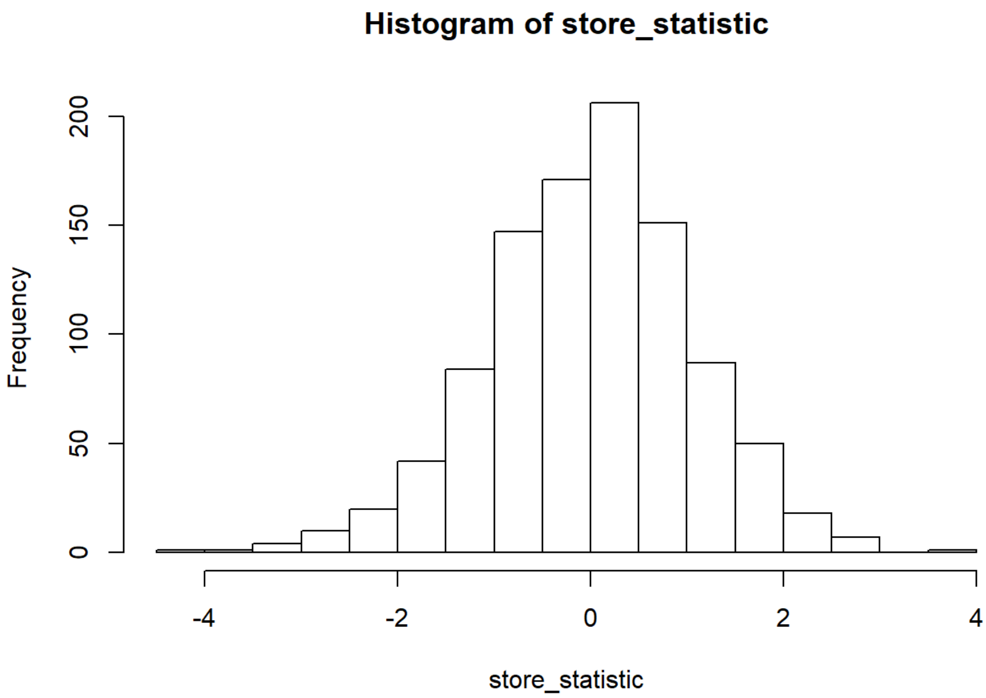
```
## [1] 1.152899
```

```
sd(store_statistic)
```

```
## [1] 1.073731
```

#Graphical Illustration of the distribution of Z4, (n=20)

```
hist(store_statistic)
```



#For Z4 when n=20, the mean , variance and sd are : 0.05047637, 1.03612 and 1.0179 respectively . The theoritical result are : mean = 0 , variance = 1.117 and the sd = 0.9459. The empirical results are overall consistent with the theoritical values. Though the empirical is slightly higher than the theoritical mean and the variance and sd are slightly larger

#Evaluating Z5 for n=5 and m=20

```
sample1 <- matrix(rnorm(n=5000, m=0, sd= 1), nrow = 1000, ncol = 5)
sample2 <- matrix(rnorm(n=20000, m=0, sd= 1), nrow = 1000, ncol = 20)
store_statistic <- numeric(1000)

for(i in 1:1000){
  row1 <- sample1[i,]
  row2 <- sample2[i,]
  k <- 1
  m1=  mean(row1)
  m1=  mean(row1)
  mu=0
  n=5
  m=20
  store_statistic[i]<- (sum((row1-0)^2)/(n-1))/(sum((row2-0)^2)/(m-1))
}

#Calculating , the mean , variance and standard deviation for Z5:
mean(store_statistic)
```

```
## [1] 1.274105
```

```
var(store_statistic)
```

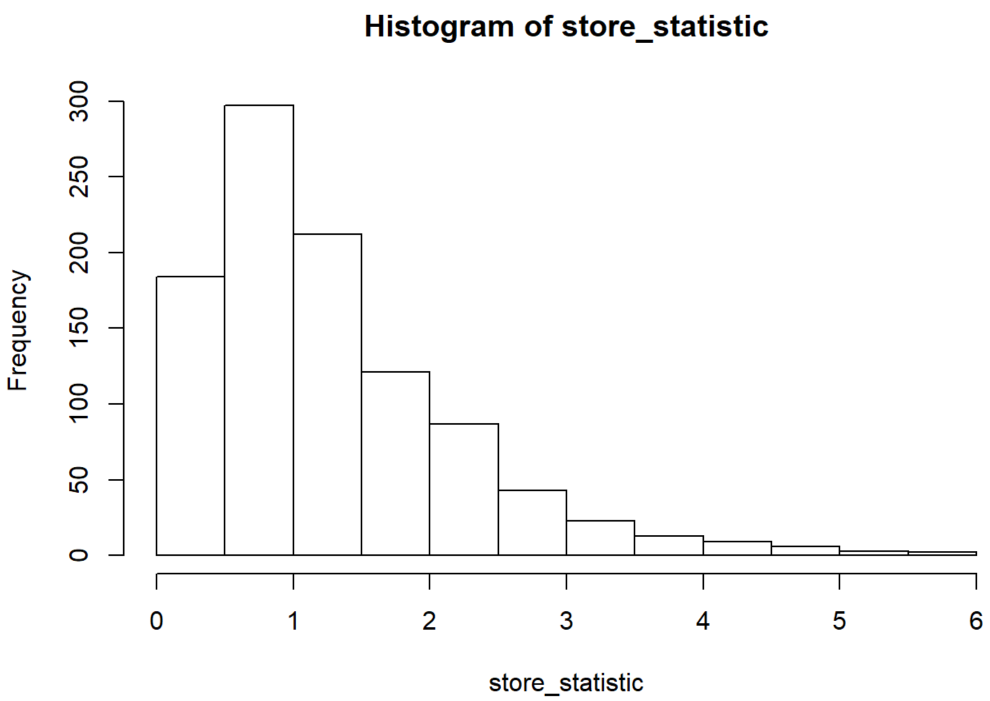
```
## [1] 0.8736793
```

```
sd(store_statistic)
```

```
## [1] 0.9347081
```

#Graphical Illustration of the distribution of Z5:

```
hist(store_statistic)
```



#For Z5 with n=5 and m=20 , the mean = 1.260348 which is close to the theoritical value of 1.11764 , the variance is 0.7821797 which slightly lower but close to the theoritical value of 0.874 and the sd is 0.8844093 , which is close to the theoritical value of 0.934879