Computer Assignment

Probability Theory and Statistical Inference 1

Claes Kock
Yuchong Wu

19/9/2019

Contents

1 Assignment 1

Table 1: Distributions for statistics(reference to the text book)

Z	Distribution	Reference
Z_1	Normal Distribution	Theorem 7.1
Z_2	χ^2 Distribution (n-1 df)	Theorem 7.3
Z_3	χ^2 Distribution (n df)	Theorem 7.2
Z_4	t Distribution (n-1 df)	Definition 7.2
Z_5	F Distribution (n-1 df and m-1 df)	Definition 7.3

Table 2: Parameters for statistics Z_1

Z_1	$E(Z_1)$	$var(Z_1)$	$sd(Z_1)$
theory	μ	σ^2/n	$\sqrt{\sigma^2/n}$
n=5	0	1/5	$\sqrt{1/5}$
n=20	0	1/20	$\sqrt{1/20}$

Table 3: Parameters for statistics Z_2

Z_2	$E(Z_2)$	$var(Z_2)$	$sd(Z_2)$
theory	n-1	2n-2	$\sqrt{2n-2}$
n=5	4	8	$\sqrt{8}$
n=20	19	38	$\sqrt{38}$

Table 4: Parameters for statistics Z_3

Z_3	$E(Z_3)$	$var(Z_3)$	$sd(Z_3)$
theory	n	2n	$\sqrt{2n}$
n=5	5	10	$\sqrt{10}$
n=20	20	40	$\sqrt{40}$

Table 5: Parameters for statistics Z_4

Z_4	$E(Z_4)$	$var(Z_4)$	$sd(Z_4)$
theory	0	(n-1)/(n-3)	$\sqrt{(n-1)/(n-3)}$
n=5	0	2	$\sqrt{2}$
n=20	0	19/17	$\sqrt{19/17}$

Table 6: Parameters for statistics Z_5

Z_5	$E(Z_5)$	$var(Z_5)$	$sd(Z_5)$
theory	$\frac{d_2}{d_2-2}$	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$	$\sqrt{\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}}$
n=5, m=20	19/17	$2527/2890 = 0.87$	$\sqrt{2527/2890} = 0.94$

Notes: $d_1 = n - 1$, $d_2 = m - 1$

2 Assignment 2

2.1 Simulation study for n = 5

2.1.1 Sampling

```
n = 5
ma = matrix(rnorm(1000*n), 1000, n)
mu = 0
sigma = 1
```

2.1.2 Z1

```
ve11 = apply(ma, 1, mean)

# E(Z_1)
mean(ve11)
```

```
## [1] -0.01802356
```

```
# var(Z_1)
var(ve11)
```

```
## [1] 0.2050547
```

```
# sd(Z_1)
sd(ve11)
```

```
## [1] 0.4528297
```

2.1.3 Z2

```
ve12 = c()
ma_2 = ma^2
for(i in 1:nrow(ma))
{
  new = sum(ma_2[i,]) - n * (mean(ma[i,])^2)
  ve12 = c(ve12, new)
}
```

```r
# E(Z_2)
mean(ve12)
```

```
## [1] 3.979371
```

```r
# var(Z_2)
var(ve12)
```

```
## [1] 7.85107
```

```r
# sd(Z_2)
sd(ve12)
```

```
## [1] 2.801976
```

2.1.4 Z3

```r
ve13 = c()
for(i in 1:nrow(ma))
{
  new = 0
  for(j in 1:ncol(ma))
  {
    new = new + (ma[i,j] - mu)^2
  }
  new = new / (sigma^2)
  ve13 = c(ve13, new)
}

# E(Z_3)
mean(ve13)
```

```
## [1] 5.005244
```

```r
# var(Z_3)
var(ve13)
```

```
## [1] 10.04894
```

```r
# sd(Z_3)
sd(ve13)
```

```
## [1] 3.170007
```

2.1.5 Z4

```r
ve14 = c()

for(i in 1:nrow(ma))
{
  numerator = (mean(ma[i,]) - mu) / (sigma / sqrt(n))
  denumerator = (sum(ma[i,]^2) - n*mean(ma[i,])^2) / ((n-1)*(sigma^2))
  denumerator = sqrt(denumerator)
  ve14 = c(ve14, (numerator/denumerator))
}
```

```r
# E(Z_4)
mean(ve14)
```

```
## [1] -0.05144921
```

```r
# var(Z_4)
var(ve14)
```

```
## [1] 2.159759
```

```r
# sd(Z_4)
sd(ve14)
```

```
## [1] 1.469612
```

2.2 Simulation study for n = 20

2.2.1 Sampling

```r
n = 20
ma = matrix(rnorm(1000*n), 1000, n)
mu = 0
sigma = 1
```

2.2.2 Z1

```r
ve21 = apply(ma, 1, mean)

# E(Z_1)
mean(ve21)
```

```
## [1] -4.640395e-05
```

```r
# var(Z_1)
var(ve21)
```

```
## [1] 0.0473953
```

```r
# sd(Z_1)
sd(ve21)
```

```
## [1] 0.2177046
```

2.2.3 Z2

```r
ve22 = c()
ma_2 = ma^2
for(i in 1:nrow(ma))
{
  new = sum(ma_2[i,]) - n * (mean(ma[i,])^2)
  ve22 = c(ve22, new)
}
```

```r
# E(Z_2)
mean(ve22)
```

```
## [1] 19.34395
```

```r
# var(Z_2)
var(ve22)
```

```
## [1] 36.044
```

```r
# sd(Z_2)
sd(ve22)
```

```
## [1] 6.003666
```

2.2.4 Z3

```r
ve23 = c()
for(i in 1:nrow(ma))
{
  new = 0
  for(j in 1:ncol(ma))
  {
    new = new + (ma[i,j] - mu)^2
  }
  new = new / (sigma^2)
  ve23 = c(ve23, new)
}

# E(Z_3)
mean(ve23)
```

```
## [1] 20.29091
```

```r
# var(Z_3)
var(ve23)
```

```
## [1] 37.18617
```

```r
# sd(Z_3)
sd(ve23)
```

```
## [1] 6.098046
```

2.2.5 Z4

```r
ve24 = c()

for(i in 1:nrow(ma))
{
  numerator = (mean(ma[i,]) - mu) / (sigma / sqrt(n))
  denumerator = (sum(ma[i,]^2) - n*mean(ma[i,])^2) / ((n-1)*(sigma^2))
  denumerator = sqrt(denumerator)
  ve24 = c(ve24, (numerator/denumerator))
}
```

```r
# E(Z_4)
mean(ve24)
```

```
## [1] -0.002708441
```

```r
# var(Z_4)
var(ve24)
```

```
## [1] 1.046567
```

```r
# sd(Z_4)
sd(ve24)
```

```
## [1] 1.023019
```

2.3 Simulation study for n = 5 and m = 20

2.3.1 Sampling

```r
ma = matrix(rnorm(5000), 1000, 5)
ma_y = matrix(rnorm(20000), 1000, 20)
n = 5
m = 20
mu = 0
sigma = 1
```

2.3.2 Z5

```r
ve5 = c()

for(i in 1:nrow(ma))
{
  numerator = (sum(ma[i,]^2) - n*mean(ma[i,])^2) / ((n-1)*(sigma^2))
  denumerator = (sum(ma_y[i,]^2) - n*mean(ma_y[i,])^2) / ((m-1)*(sigma^2))
  ve5 = c(ve5, (numerator/denumerator))
}

# E(Z_5)
mean(ve5)
```

```
## [1] 1.047995
```

```r
# var(Z_5)
var(ve5)
```

```
## [1] 0.7390057
```

```r
# sd(Z_5)
sd(ve5)
```

```
## [1] 0.8596544
```

2.4 Comparison between theoretical values and simulated values

Table 7: Comparison between theoretical values and simulation values for Z1

Df	Theo_E	Sim_E	Theo_Var	Sim_Var	Theo_Sd	Sim_Sd
n=5	0	-0.02	0.20	0.21	0.45	0.45
n=20	0	0.00	0.05	0.05	0.22	0.22

Table 8: Comparison between theoretical values and simulation values for Z2

Df	Theo_E	Sim_E	Theo_Var	Sim_Var	Theo_Sd	Sim_Sd
n=5	4	3.98	8	7.85	2.83	2.8
n=20	19	19.34	38	36.04	6.16	6.0

Table 9: Comparison between theoretical values and simulation values for Z3

Df	Theo_E	Sim_E	Theo_Var	Sim_Var	Theo_Sd	Sim_Sd
n=5	5	5.01	10	10.05	3.16	3.17
n=20	20	20.29	40	37.19	6.32	6.10

Table 10: Comparison between theoretical values and simulation values for Z4

Df	Theo_E	Sim_E	Theo_Var	Sim_Var	Theo_Sd	Sim_Sd
n=5	0	-0.05	2.00	2.16	1.41	1.47
n=20	0	0.00	1.12	1.05	1.06	1.02

Table 11: Comparison between theoretical values and simulation values for Z5

Df	Theo_E	Sim_E	Theo_Var	Sim_Var	Theo_Sd	Sim_Sd
n=5, m=20	1.12	1.05	0.87	0.74	0.94	0.86

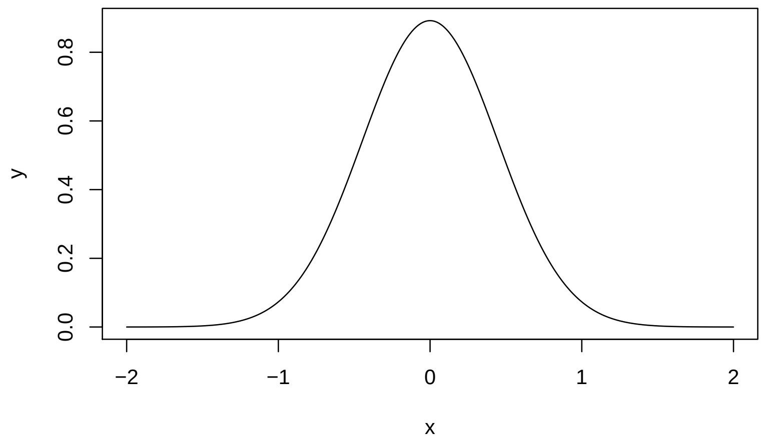
2.5 Histograms for statistics

2.5.1 Z1

2.5.1.1 Histograms for simulation study (n=5)

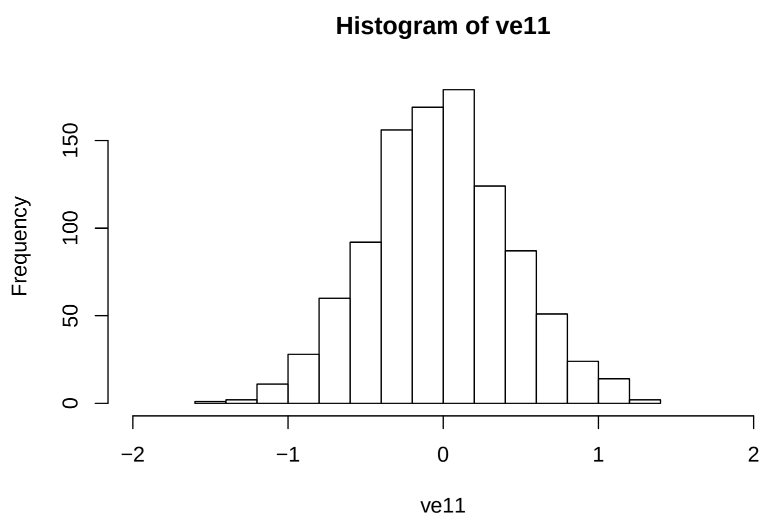
2.5.1.1.1 Theoretical

```
x <- seq(-2, 2, length=1000)
y <- dnorm(x, mean=0, sd=sqrt(1/5))
plot(x, y, type="l", lwd=1)
```



2.5.1.2 Simulation

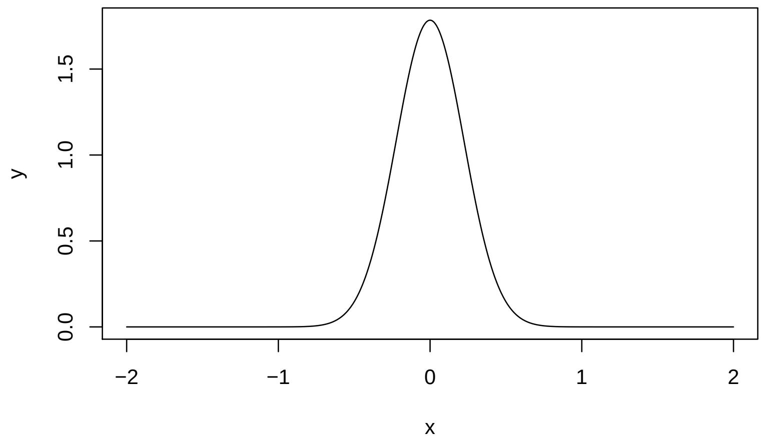
```
hist(ve11, xlim = range(-2,2))
```



2.5.1.3 Histograms for simulation study (n=20)

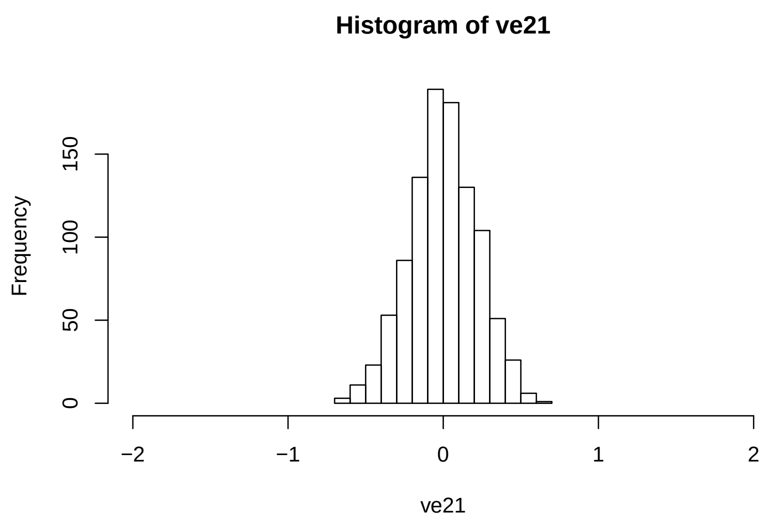
2.5.1.3.1 Theoretical

```
x <- seq(-2, 2, length=1000)
y <- dnorm(x, mean=0, sd=sqrt(1/20))
plot(x, y, type="l", lwd=1)
```



2.5.1.4 Simulation

```
hist(ve21, xlim = range(-2,2))
```



2.5.1.5 Comment for Z1

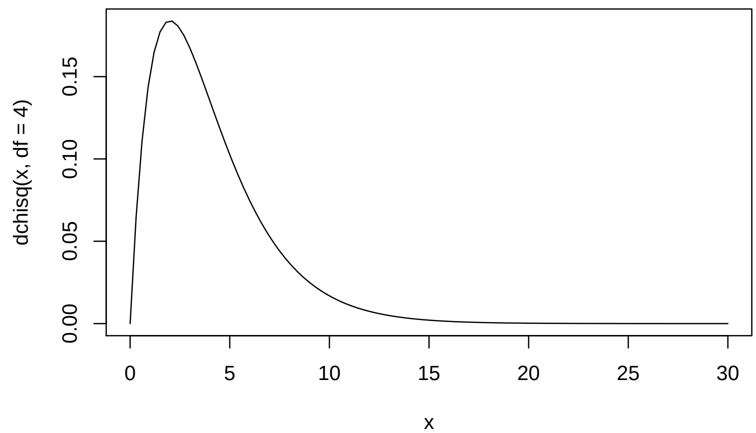
Our 2 Histogram models for n = 5 and n = 20 seem to be normally distributed. They seem to be focused around the theoretical expected value, which is 0. The difference between the 2 models is in the variance, where a larger n leads to a smaller variance, which will make the model more evenly distributed. An increase in the sample size also seems to lead to an increase in height of the model, compared to its width, as n = 20 is mostly distributed between -1 and 1.

2.5.2 Z2

2.5.2.1 Histograms for simulation study (n=5)

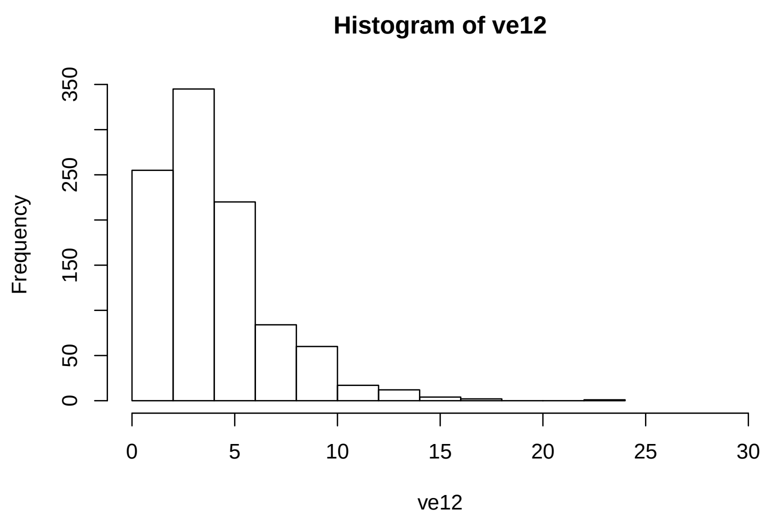
2.5.2.1.1 Theoretical

```
curve(dchisq(x, df=4), xlim = range(0, 30))
```



2.5.2.1.2 Simulation

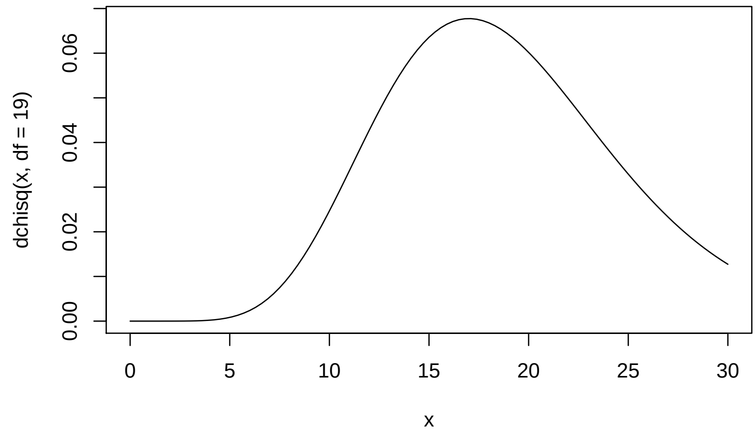
```
hist(ve12, xlim = range(0, 30))
```



2.5.2.2 Histograms for simulation study (n=20)

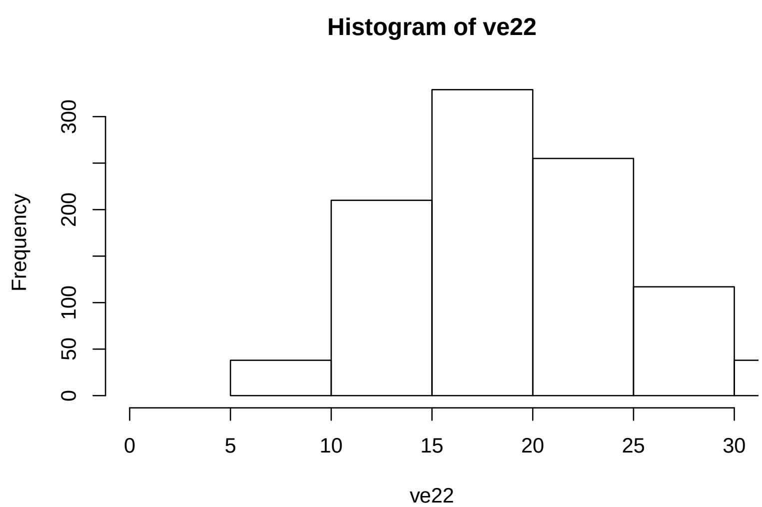
2.5.2.2.1 Theoretical

```
curve(dchisq(x, df=19), xlim = range(0, 30))
```



2.5.2.2.2 Simulation

```
hist(ve22, xlim = range(0, 30))
```



2.5.2.3 Comment for Z2

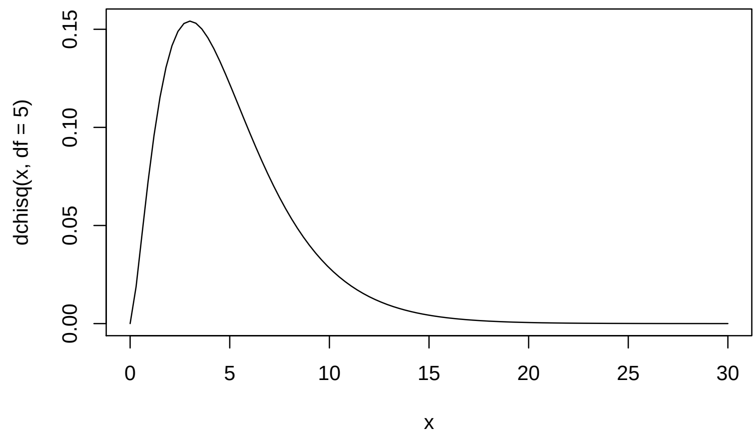
The range of the distribution for both models is potisitive, and the distributions seem to follow their theoretical models. They seem to be focused around the theoretical expected value, which is equal to the sample size minus 1. We can also observe that the higher the sample size, the more the distribution peaks futher away from 0.

2.5.3 Z3

2.5.3.1 Histograms for simulation study (n=5)

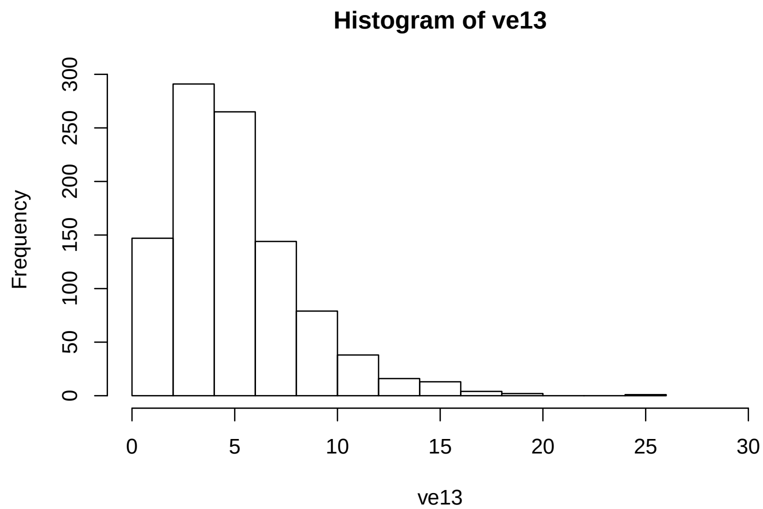
2.5.3.1.1 Theoretical

```
curve(dchisq(x, df=5), xlim = range(0, 30))
```



2.5.3.1.2 Simulation

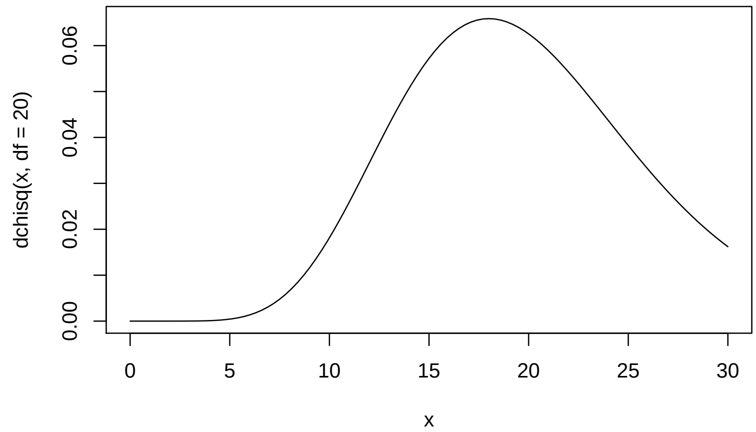
```
hist(ve13, xlim = range(0, 30))
```



2.5.3.2 Histograms for simulation study (n=20)

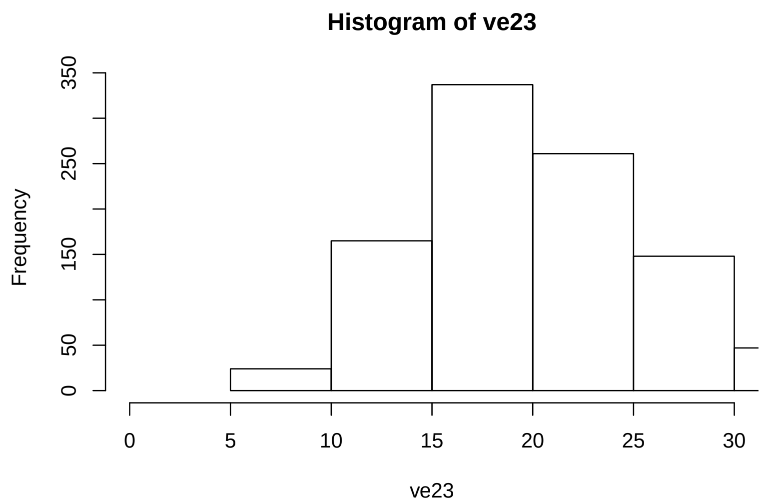
2.5.3.2.1 Theoretical

```
curve(dchisq(x, df=20), xlim = range(0, 30))
```



2.5.3.2.2 Simulation

```
hist(ve23, xlim = range(0, 30))
```



2.5.3.3 Comment for Z3

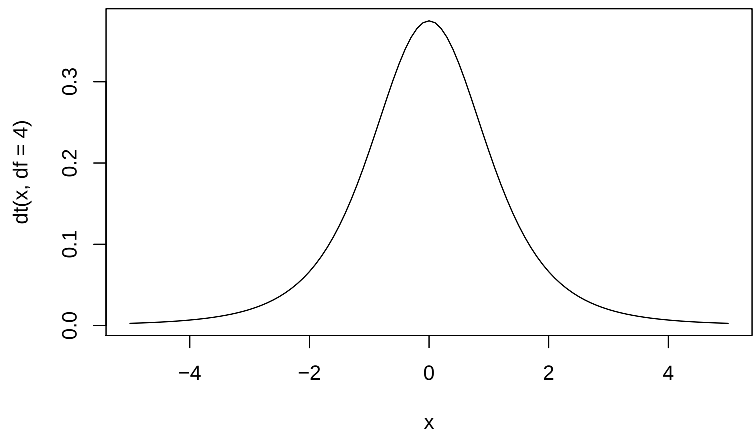
The distributions for Z3 is quite similiar to those for Z2, for the reason that they are the same distribution with different degrees of freedom. The mean, variance and standard deviation are sightly larger from those of Z2, due to one extra degree.

2.5.4 Z4

2.5.4.1 Histograms for simulation study (n=5)

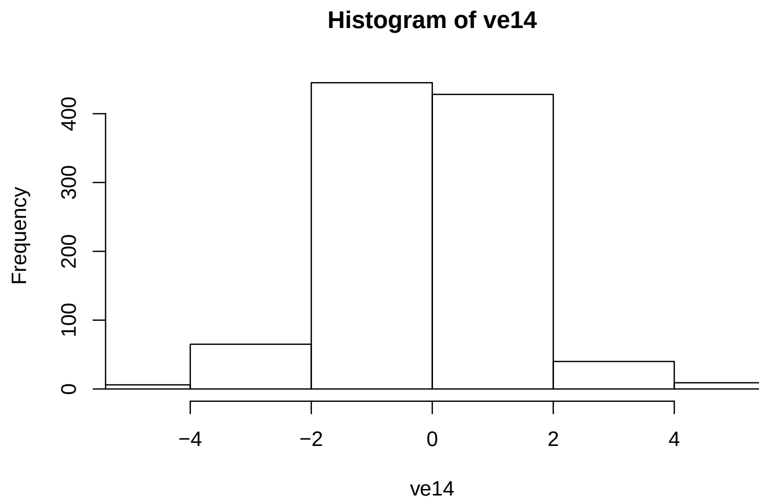
2.5.4.1.1 Theoretical

```r
curve(dt(x, df=4), xlim = range(-5, 5))
```



2.5.4.1.2 Simulation

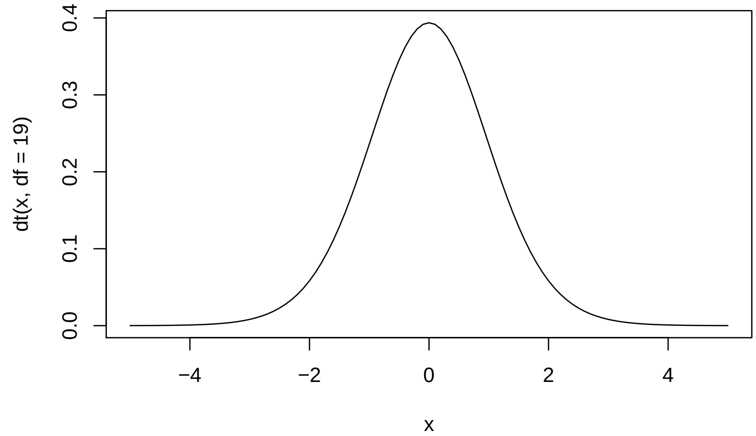
```r
hist(ve14, xlim = range(-5, 5))
```



2.5.4.2 Histograms for simulation study (n=20)

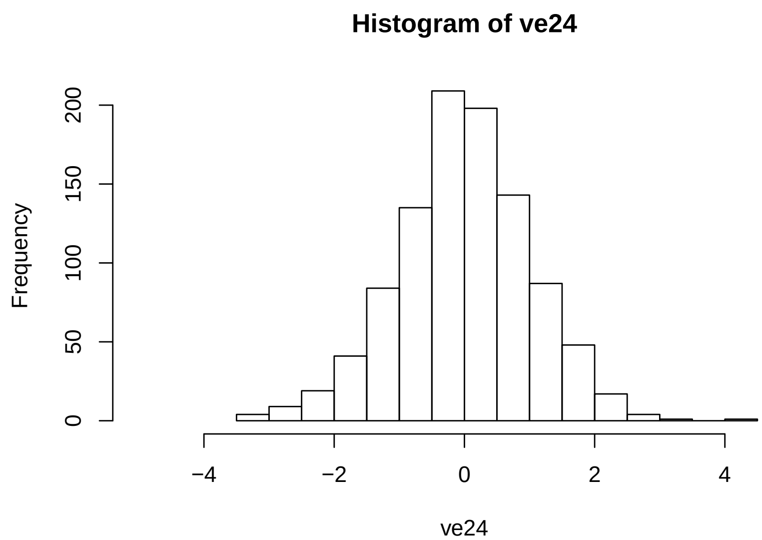
2.5.4.2.1 Theoretical

```
curve(dt(x, df=19), xlim = range(-5, 5))
```



2.5.4.2.2 Simulation

```
hist(ve24, xlim = range(-5, 5))
```



2.5.4.3 Comment for Z4

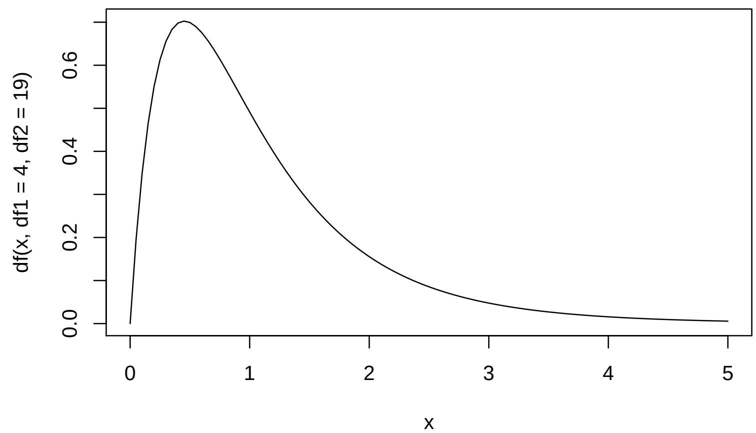
The two models of the T-distribution *($n = 5$ and $n = 20$)* are quite the same, whose expected values are both equal to 0. Since the variance of T-distribution is $\frac{n}{n-2}$, a larger sample size *($n = 20$)* leads to the variance being closer to 1. As can be seen in the plot, the model with the larger sample size is more concentrated around the expected value, which is 0. If the sample size goes really large, it can be imagined that the whole distribution will be almost completely centered around 0.

2.5.5 Z5

2.5.5.1 Histograms for simulation study (n=5, m=20)

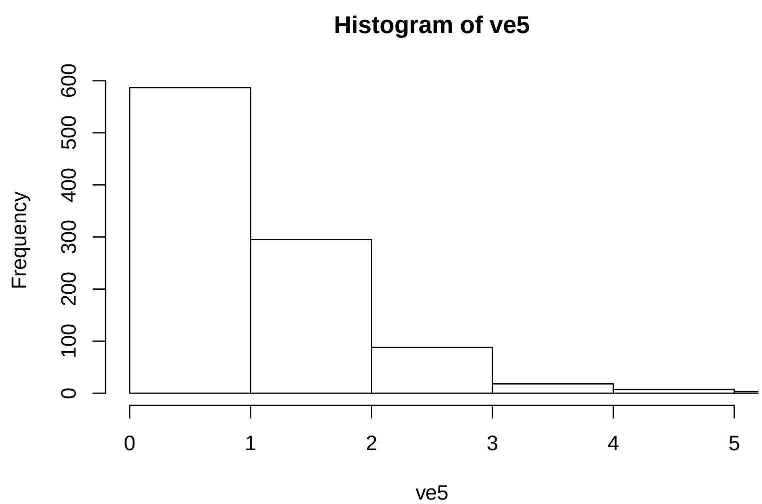
2.5.5.1.1 Theoretical

```
curve(df(x, df1 = 4, df2 = 19), xlim = range(0, 5))
```



2.5.5.1.2 Simulation

```
hist(ve5, xlim = range(0, 5))
```



2.5.5.2 Comment for Z5

The range of the distribution for the model is always potisitive, and the distributions seem to follow its theoretical models. As can be seen in the plot, the F-distribution is similiar to chi-square distribution in some particular situations. Compared to the chi-square distribution, the peak of the F-distribution will never exceed 1. The model seems to be focused around the theoretical expected value, which is almost equal to 1 when the sample size is really large. With the increase of the randon variable*(x)*, the probability density will gradually approach 0.