STA261: Assignment

Question 1

(a) population mean is 24

```
X=c(21, 22, 23, 24, 25, 26, 27)
mean(X)
```

```
## [1] 24
```

(b) population variance is 4

```
mean((X - 24)^2)
```

```
## [1] 4
```

(c)

```
d=expand.grid(X,X,X,X)
write.csv(d,file="Question1.csv",row.names = F)
```

(d)

```
X_bar = double(2401)
for(i in 1:2401){
  X_bar[i] = sum(d[i, ])/4
}
head(X_bar)
```

```
## [1] 21.00 21.25 21.50 21.75 22.00 22.25
```

(e)

```
# frequencies
table(X_bar)
```

```
## X_bar
##    21 21.25  21.5 21.75    22 22.25  22.5 22.75    23 23.25  23.5 23.75    24
##     1     4    10    20    35    56    84   116   149   180   206   224   231
## 24.25  24.5 24.75    25 25.25  25.5 25.75    26 26.25  26.5 26.75    27
##   224   206   180   149   116    84    56    35    20    10     4     1
```

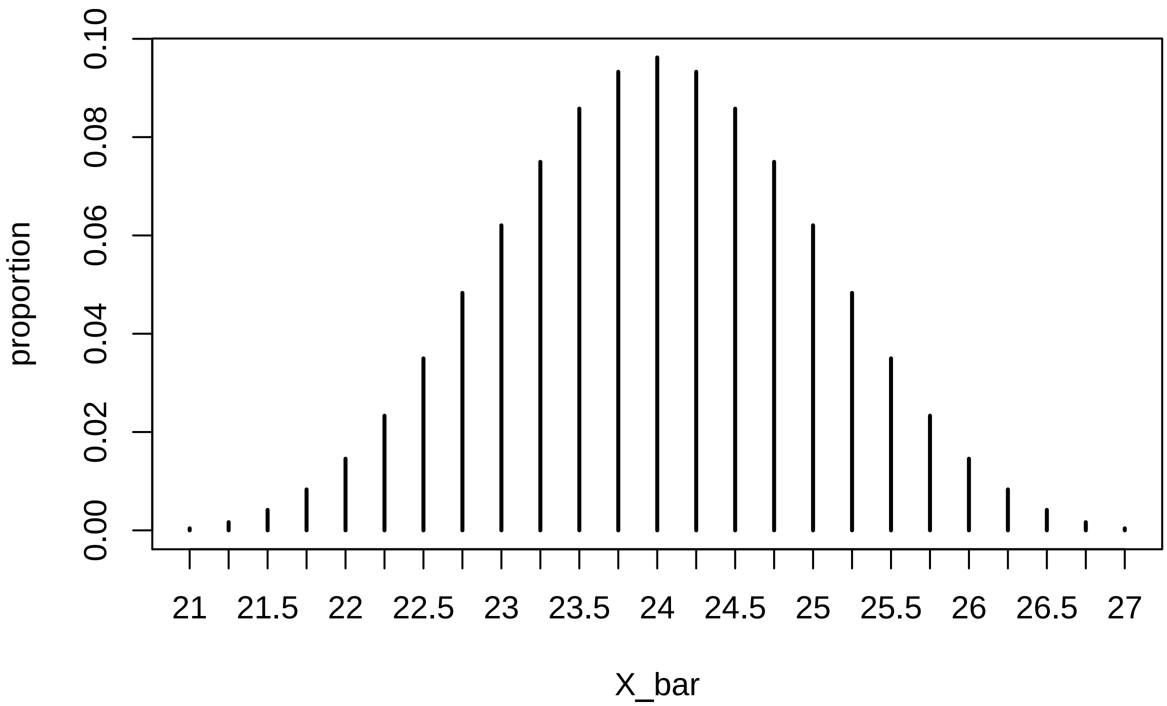
```
# proportion
table(X_bar)/2401
```

```
## X_bar
##           21        21.25         21.5        21.75           22        22.25
## 0.0004164931 0.0016659725 0.0041649313 0.0083298626 0.0145772595 0.0233236152
##         22.5        22.75           23        23.25         23.5        23.75
## 0.0349854227 0.0483132028 0.0620574761 0.0749687630 0.0857975843 0.0932944606
##           24        24.25         24.5        24.75           25        25.25
## 0.0962099125 0.0932944606 0.0857975843 0.0749687630 0.0620574761 0.0483132028
##         25.5        25.75           26        26.25         26.5        26.75
## 0.0349854227 0.0233236152 0.0145772595 0.0083298626 0.0041649313 0.0016659725
```

```
##            27
## 0.0004164931
```

(f) The shape of this plot look like normal distribution.

```r
plot(table(X_bar)/2401, ylab = "proportion")
```



(g) the mean of these 2401 numbers is 24, it is the same as the question in 1(a).

```r
mean(X_bar)
```

```
## [1] 24
```

(h) The variance of these 2401 numbers is 1, it is 1/4 of the previous variance in 1(b).

```r
sum((X_bar - 24)^2)/2401
```

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

(i) Central limit theorem tells us the mean of the sample X1,X2,…X4 follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Question 2

(a)

Bias[S^2] = E[S^2] - 4 = 4 - 4 = 0.

Bias[$\hat{\sigma}^2$] = E[$\hat{\sigma}^2$] - 4 = 3 - 4 = -1.

```
S = double(2401)
sigmahat = double(2401)
for(i in 1:2401){
  S[i] = sum((d[i,] - sum(d[i,])/4)^2)/3
  sigmahat[i] = sum((d[i,] - sum(d[i,])/4)^2)/4
}
mean(S) - 4
```

```
## [1] 0
```

```
mean(sigmahat) - 4
```

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

(b)

MSE[$\hat{\sigma}^2$] = E[([$\hat{\sigma}^2$] - 4)^2] = $\frac{1}{n}\sum(\hat{\sigma}^2 - 4)^2$ = 4.1875.

```
mean((sigmahat - 4)^2)
```

```
## [1] 4.1875
```

var[$\hat{\sigma}^2$] = 3.1875.

```
mean((sigmahat - mean(sigmahat))^2)
```

```
## [1] 3.1875
```

$(Bias[\hat{\sigma}^2])^2 = (-1)^2 = 1$.

so $MSE(\hat{\sigma}^2) = var(\hat{\sigma}^2) + (Bias(\hat{\sigma}^2))^2 = 3.1875 + 1 = 4.1875$.

Question 3

(a) 0.941691 of these interval contains $\mu = 24$.

```
CI = double(2401)
for(i in 1:2401){
  upper = sum(d[i,])/4 + 1.96*2/sqrt(4)
  lower = sum(d[i,])/4 - 1.96*2/sqrt(4)
  if(24>=lower & 24<=upper){
    CI[i] = 1
  }
}
mean(CI)
```

```
## [1] 0.941691
```

(b)

Test statistic is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{25.5-24}{2/\sqrt{4}} = 1.5$.

P-value = 2*P(Z>1.5) = 0.1336144, larger than 0.05, so we fail to reject the null hypothesis, we accept $H_0 : \mu = 24$.

```
samp = c(24,25,26,27)
# test statistic
(mean(samp) - 24)/(2/sqrt(4))
```

```
## [1] 1.5
```

```
# p-value
2*(1-pnorm(1.5))
```

```
## [1] 0.1336144
```

(c) The p-value based on the 2401 $\bar{X}$ is the proportion that $\bar{X}$ at least as extreme as the sample, it is 0.1749271, larger than 0.05, so we fail to reject the null hypothesis, we accept $H_0 : \mu = 24$.

```
mean(abs(X_bar - 24) >= abs(mean(samp) - 24))
```

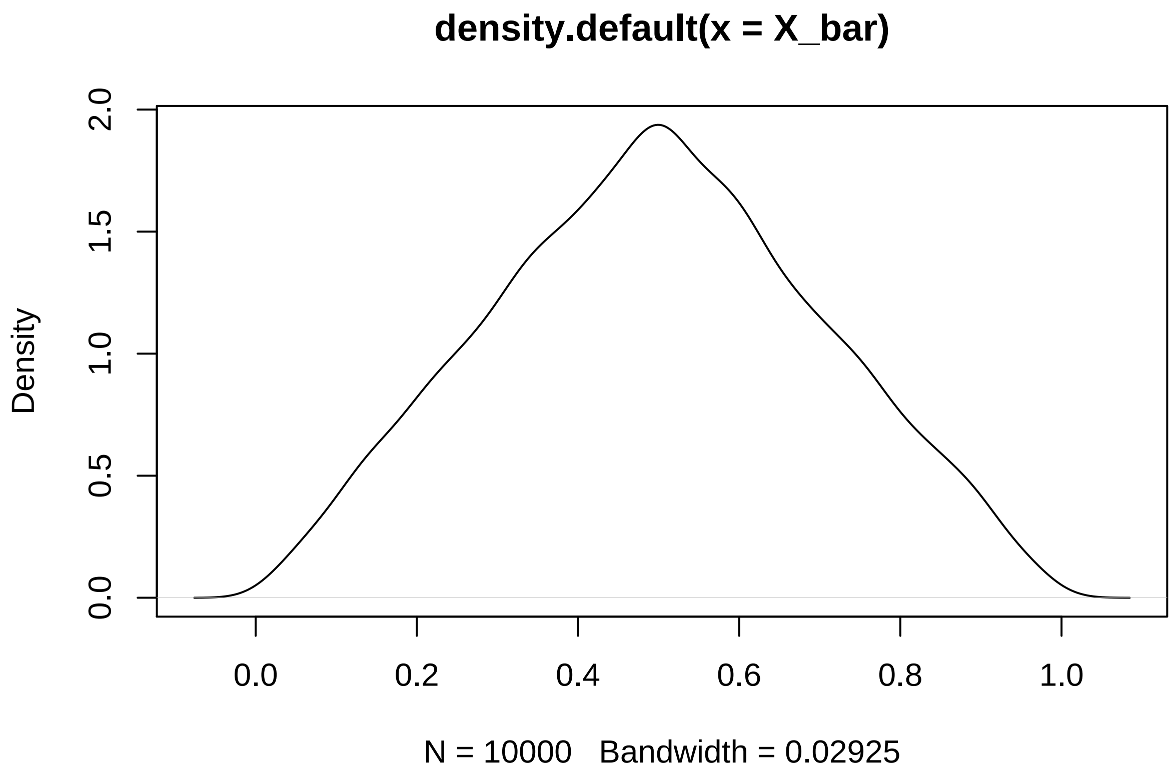
```
## [1] 0.1749271
```

(d) The conclusion is the same, but the p-value is slightly different, since in (b), we assume the data is normal, while it is not normal. If the true distribution is normal, these two numbers will be similar.

Question 4

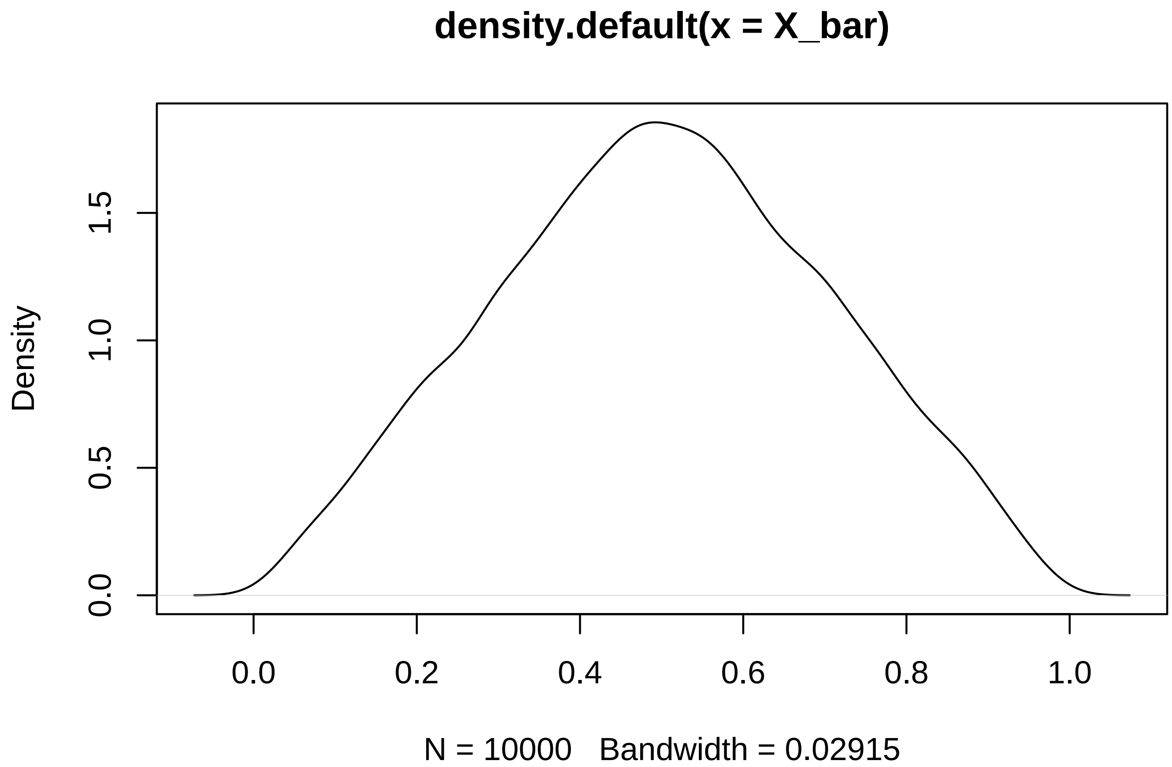
(a)

```
sample_2m_unif=function(){
  s=runif(2,0,1)
  return(mean(s))
}
X_bar=replicate(10000,sample_2m_unif())
plot(density(X_bar))
```



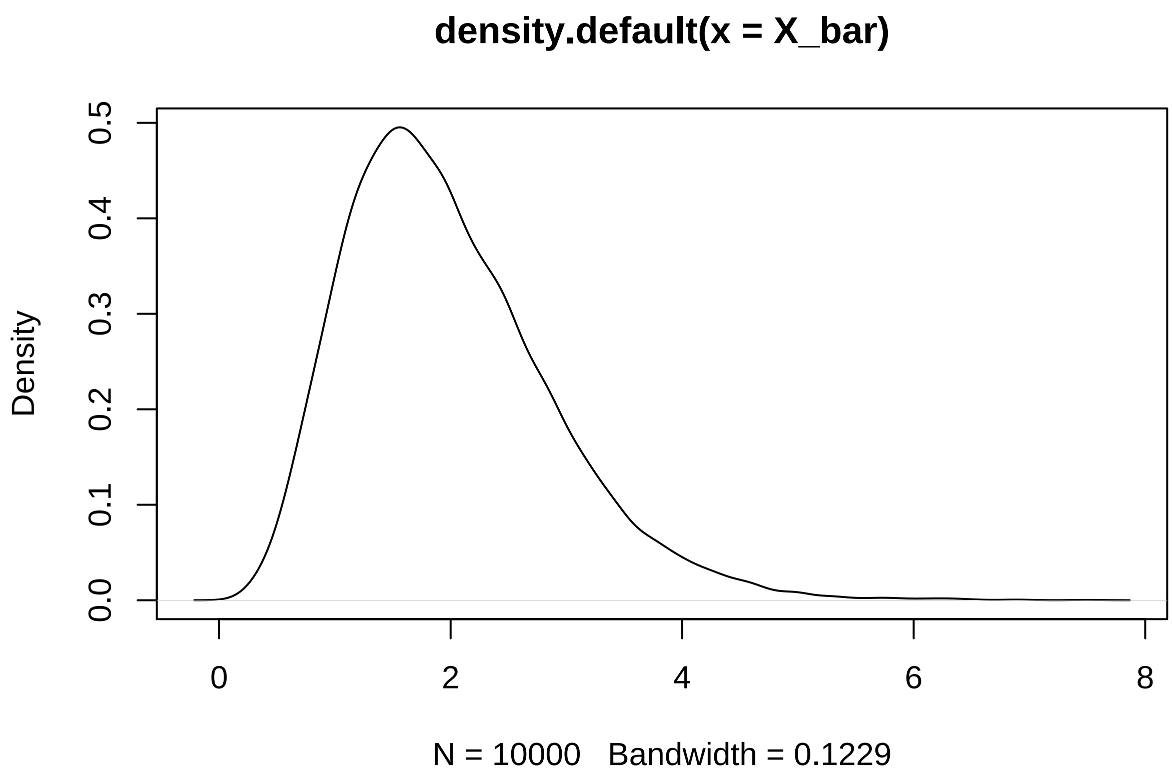
(b)

```
sample_5m_unif=function(){
  s=runif(5,0,1)
  return(mean(s))
}
X_bar=replicate(10000,sample_2m_unif())
plot(density(X_bar))
```



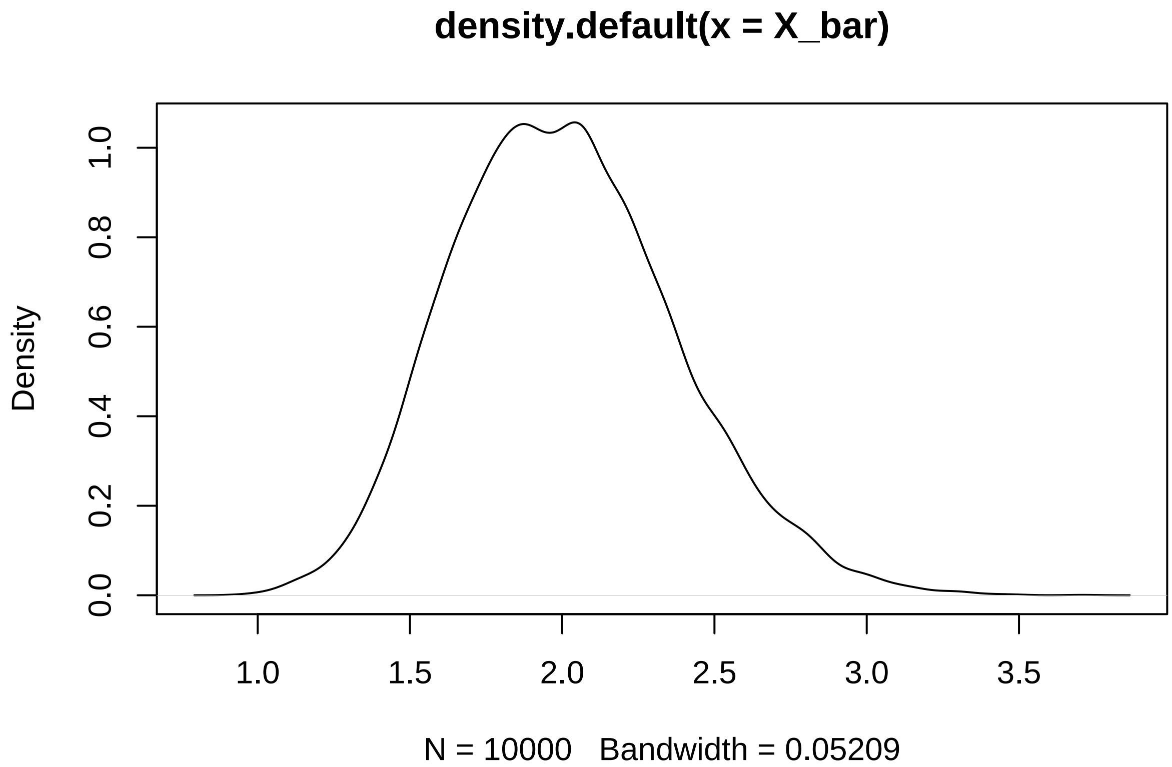
(c)

```
sample_5m_chi=function(){
  s=rchisq(5,2)
  return(mean(s))
}
X_bar=replicate(10000,sample_5m_chi())
plot(density(X_bar))
```



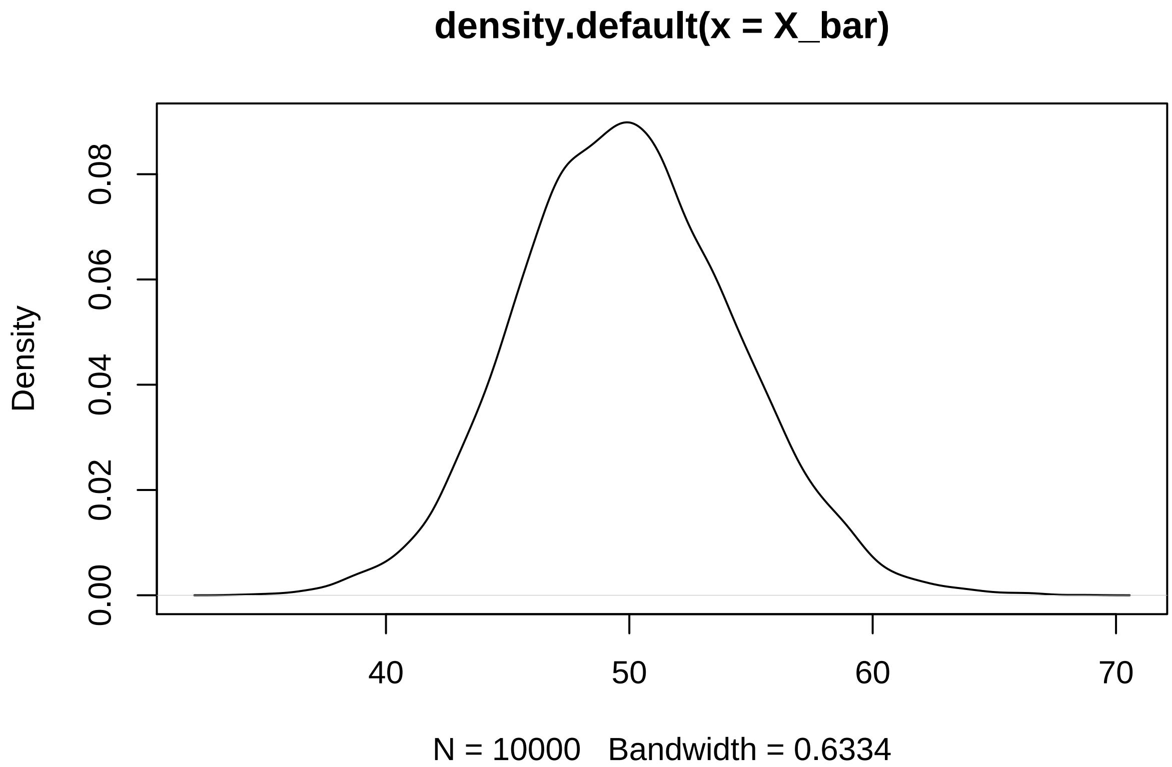
(d)

```
sample_30m_chi=function(){
  s=rchisq(30,2)
  return(mean(s))
}
X_bar=replicate(10000,sample_30m_chi())
plot(density(X_bar))
```



(e)

```
sample_5m_chi=function(){
  s=rchisq(5,50)
  return(mean(s))
}
X_bar=replicate(10000,sample_5m_chi())
plot(density(X_bar))
```



(f) In (a)(b), we could not see any normal shape from the plot, in (c), we could roughly see the bell curve, while it is seriously skewed, in (d) and (e), we can see a good normal curve, so for around n = 30, $\bar{X}$ can converge to normal distribution, that is CLT tells us. The skewness of the original distributions will lead to the shift of the normal curve away from the original point.

Question 5

(a)

(i)

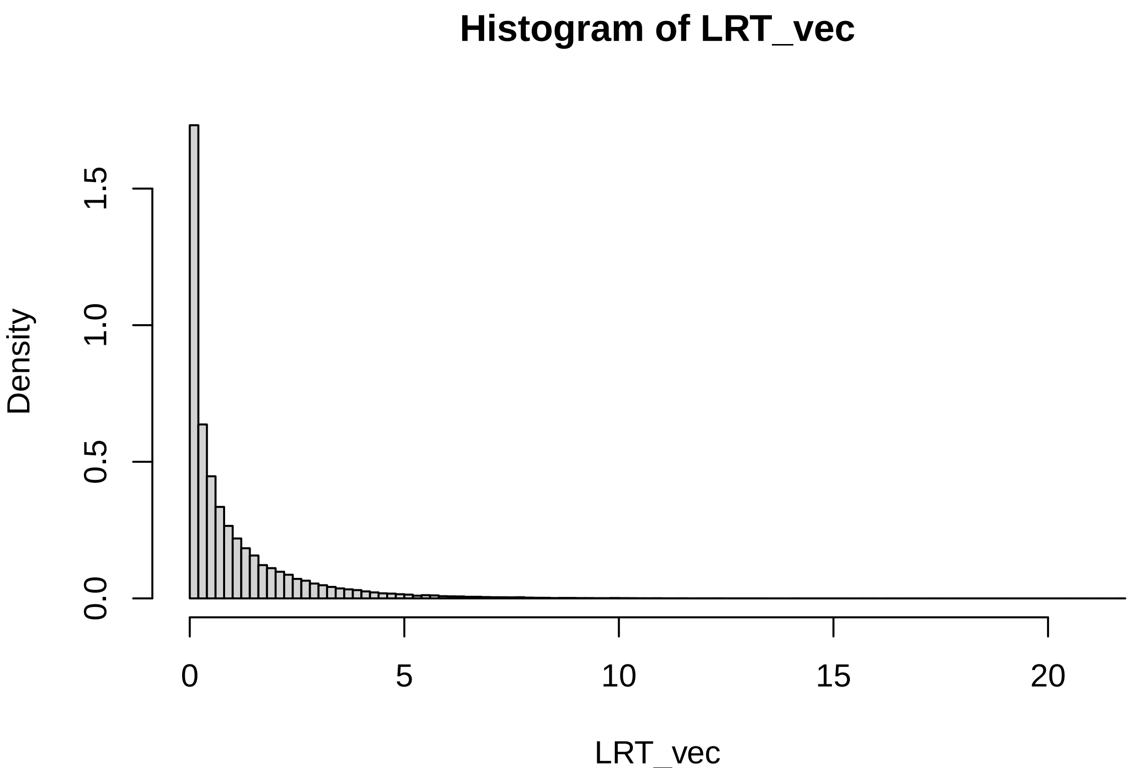
```
sample20_normal <-function(){
  s = rnorm(20, 10, 4)
  L_theta0 = prod(dnorm(s, mean = 10, sd = 4))
  L_theta1 = prod(dnorm(s, mean = mean(s), sd = 4))
  return(-2*log(L_theta0/L_theta1))
}
```

(ii)

```
LRT_vec = replicate(100000, sample20_normal())
```

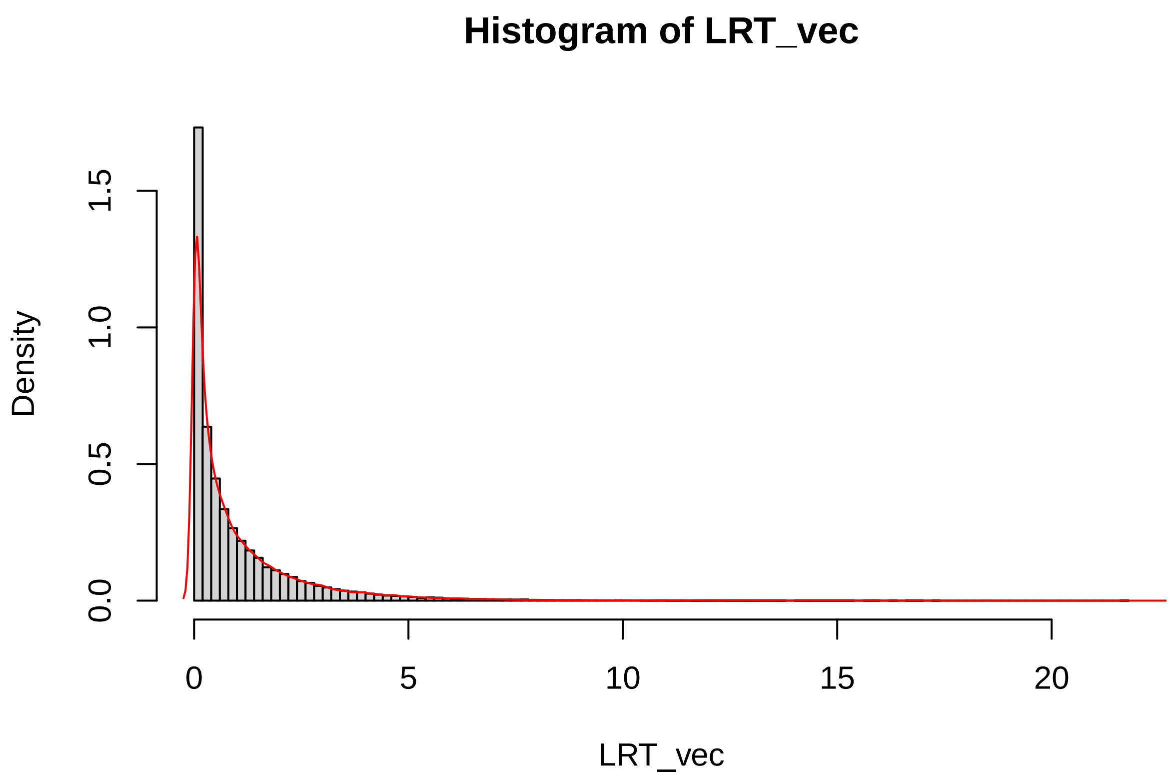
(iii)

```
hist(LRT_vec, freq=FALSE, breaks=100)
```



(iv)

```
schi = rchisq(100000, df = 1)
hist(LRT_vec, freq=FALSE, breaks=100)
lines(density(schi), col = "red")
```



(b)

(i)

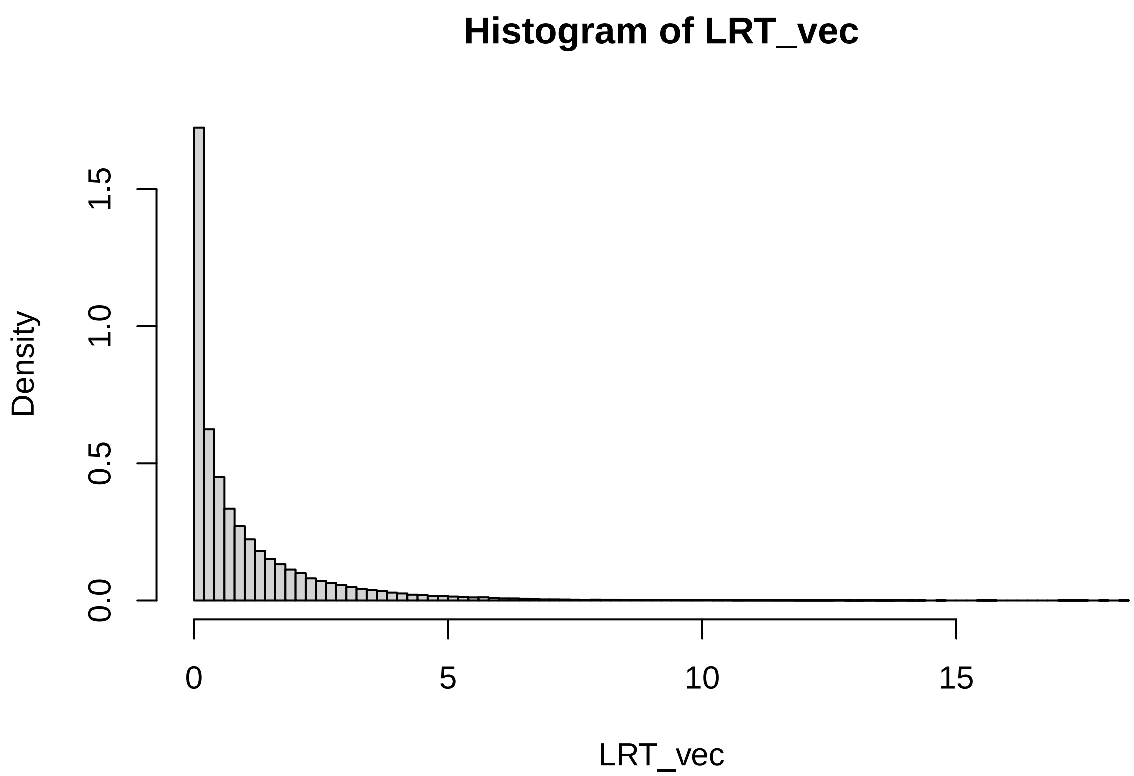
```
sample20_exp <- function(){
  s = rexp(20, 0.1)
  L_theta0 = prod(dexp(s, 0.1))
  L_theta1 = prod(dexp(s, 1/mean(s)))
  return(-2*log(L_theta0/L_theta1))
}
```

(ii)

```
LRT_vec = replicate(100000, sample20_exp())
```

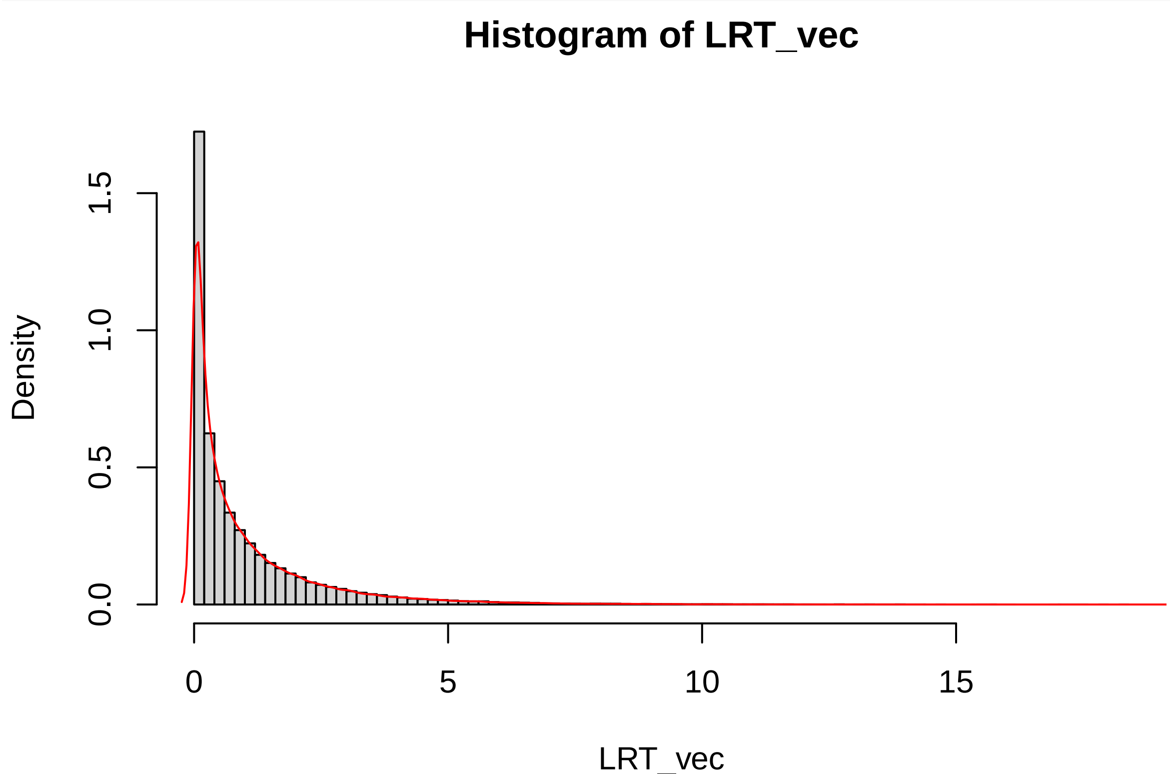
(iii)

```
hist(LRT_vec, freq=FALSE, breaks=100)
```



(iv)

```
schi = rchisq(100000, df = 1)
hist(LRT_vec, freq=FALSE, breaks=100)
lines(density(schi), col = "red")
```



(c)

From the density curve on top of histogram in (a) and (b), we find the histograms match the density very well. What's more, as the sample size n goes larger, the distribution of test statistic will converge to $\chi^2_{(df)}$, that is the histogram and $\chi^2_{(df=1)}$ density are more close in this question.