Problem Set 7.

Q.2

c) $P(1.8 < X\_\{1\} <= 2.1)$
i.e.
$P(X\_\{1\} = 2) = 0.1$

D) $n = 80$

$$E\bar{X}_{80} = E\left(\sum_{i=1}^{80} \frac{1}{80} \times X_i\right)$$

$$= \frac{1}{80} \sum EX_i = \frac{1}{80} \times 8\cancel{0} \times 2$$

$$\boxed{E\bar{X}_{80} = 2}$$

$$Var\,\bar{X}_{80} = Var\left(\sum_{i=1}^{80} \left(\frac{1}{80}\right) X_i\right)$$

$$= \left(\frac{1}{80}\right)^2 \left(\sum_{i=1}^{80} Var\,X_i\right.$$

$$= \frac{1}{80} \times \frac{1}{80} \times 80 \times 2.4$$

$$\boxed{Var\,\bar{X}_{80.} = 0.03}$$

Problem Set 7

Q.4 one can of coke – 351 gm
one can of pepsi – 350 gm } SD = 1 gm

a) We know weight of 1 can ∴ $X_i \to$ weight , $EX_{40} = 351$ as the mean and $\text{var}X_{40} = 1$ as the variance due to CLT

$$\bar{X}_{40} \sim \text{Normal}\left(351, \frac{1}{40}\right)$$

b) as per we solved (a) $Y_i \to$ weight & $XY_{42} = 350$ as the mean and $\text{Var}X_{42} = 1$ as variance, again due to CLT

$$\bar{X}_{42} \sim \text{Normal}\left(350, \frac{1}{42}\right)$$

c) $P(X_1 > 351.5)$ cannot be found because it has a continous random distribution. X_1 , X_2 , X_3 can be any value.

d) $P(\bar{X}_{40} > 351.5)$

$= 1 - \text{pnorm}(q = 351.5 \text{ , mean} = 351 \text{ , SD} = \text{sqrt}(1/40))$

$= 0.0007$

This can be done because 40 is large value to assume that sample Mean ($\bar{X}_{40}$) is normally distributed.

e) $P(\bar{X}_{40} - \bar{Y}_{42})$

$P(\bar{X}_{40} - \bar{Y}_{42} > 0)$

$= 1 - P(\bar{X}_{40} - \bar{Y}_{42} \leq 0)$

$= 1 - F_{\bar{X}-\bar{Y}}(0)$

$= 1 - \text{pnorm}(0, 351 - 350, \text{sqrt}(1/40 + 1/42))$

$= 0.999$

Note: Rest in R

PS07

Aditya Sanjay Mhaske

2023-02-27

Question 1) Consider an urn that contains 10 tickets, labelled {3,3,3,4,4,7,7,7,10,10}. From this urn, an experiment consist on drawing n = 60 tickets with replacement; let Y and $\bar{X}60$ the random variables that assigns the sum and sample mean of those 60 tickets, respectively; and do the following in R: 1 a.) Create and object called urn that represents the urn with the tickets shown above. Report your R code.

```
urn =  c(3,3,3,4,4,7,7,7,10,10)
```

1b.i.) Run a random seed first using set.seed(520),

```
set.seed(520)
```

1b.ii.) Obtain the sum of a random sample of 60 tickets (with replacement) from the urn, and

```
sample1 = sample(urn, 60, T)
sum(sample1)
```

```
## [1] 336
```

1b.iii.) Obtain the sample mean of another random sample of 60 tickets.

```
sample2 = sample(urn, 60, T)
mean(sample2)
```

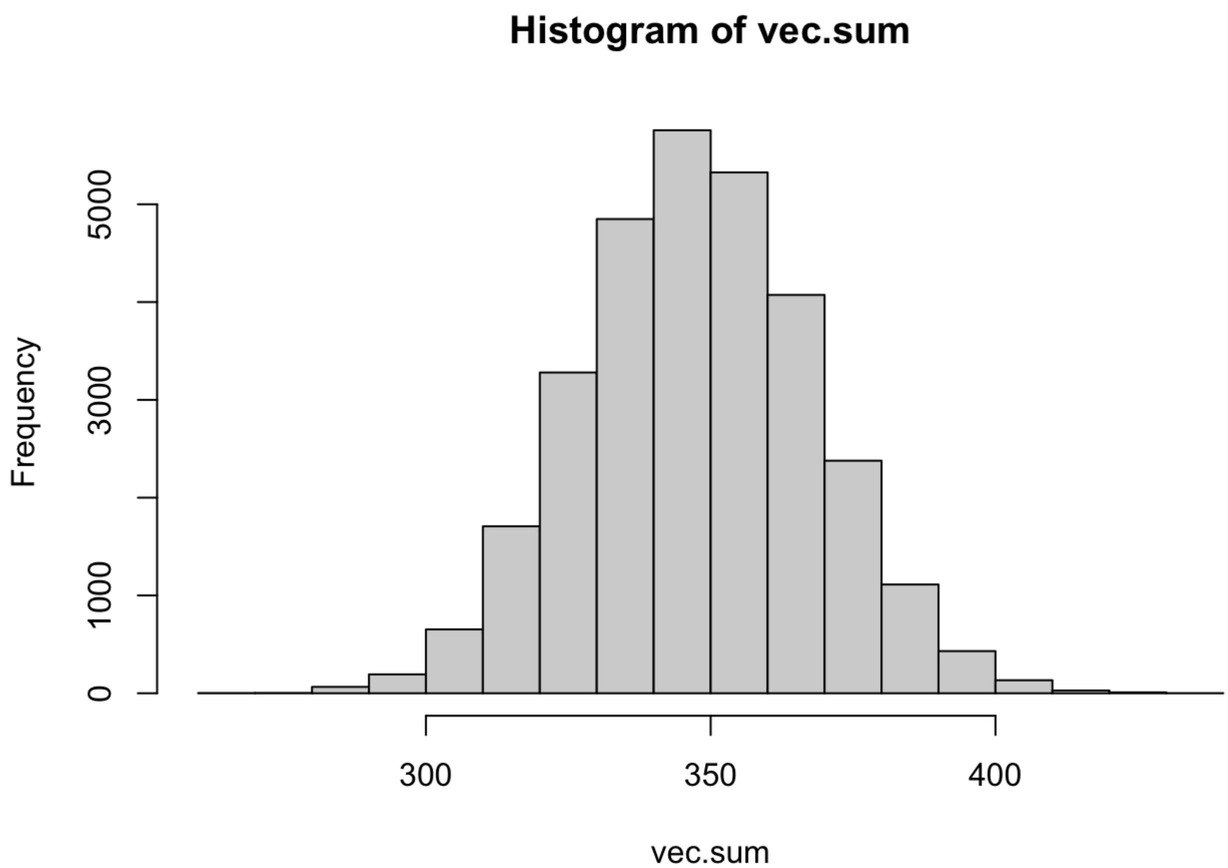
```
## [1] 5.666667
```

1c.) Obtain a big vector of 30000 sums of 60 tickets each. Call this vector vec.sum

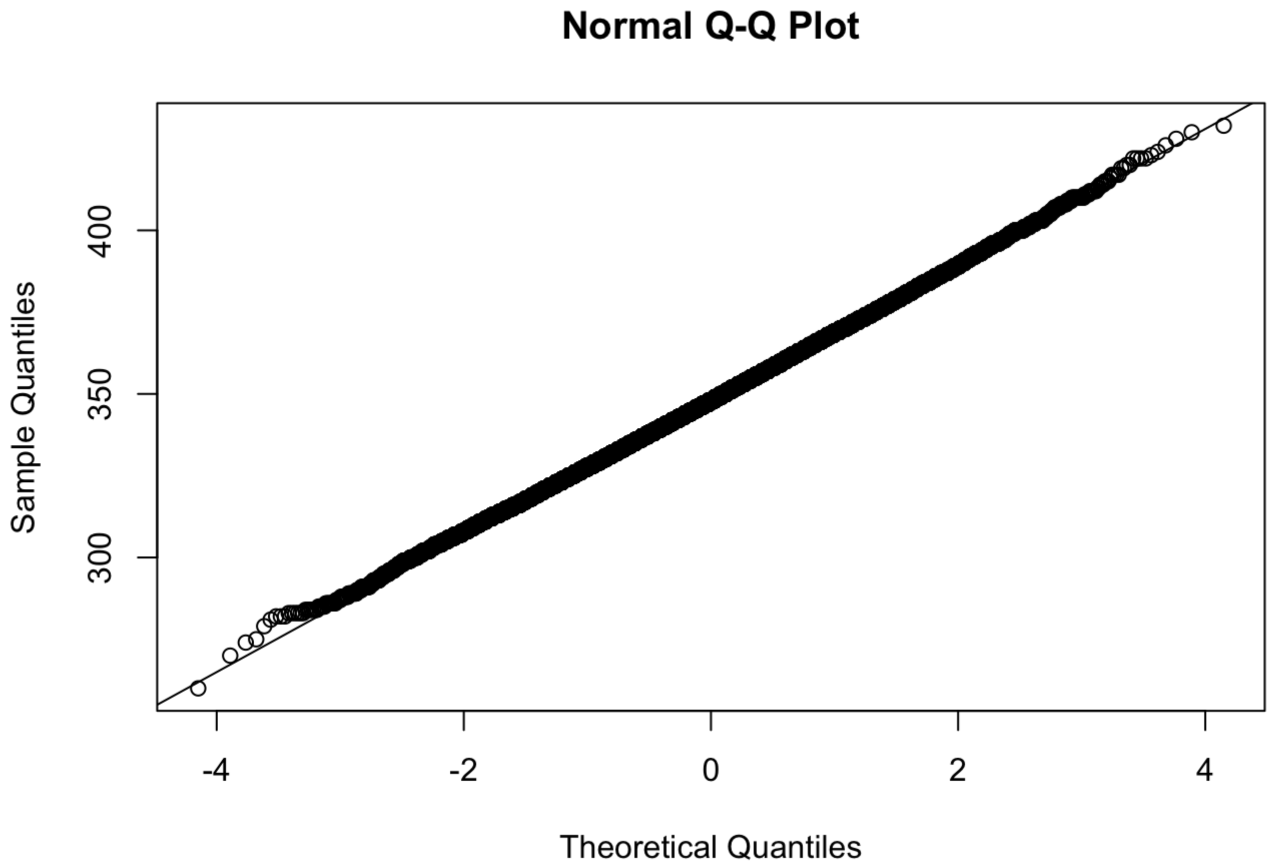
```
vec.sum = replicate(30000, sum(sample(urn, 60, T)))
```

1d.) Using vec.sum, construct a histogram, a normal probability plot, and a kernel density estimate. Does the data seem to be drawn from a normal distribution? Explain.

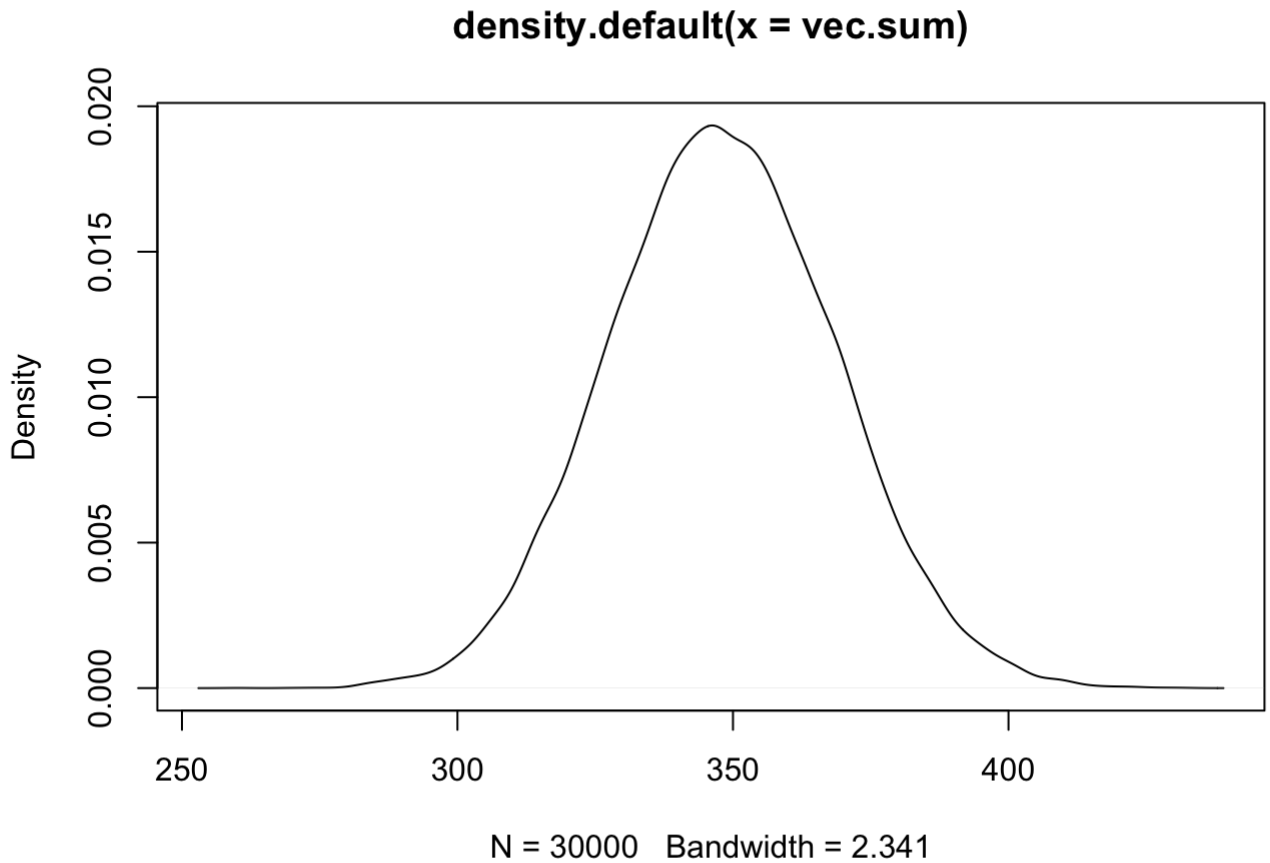
```
hist(vec.sum)
```




```
qqnorm(vec.sum)
qqline(vec.sum)
```




```
plot(density(vec.sum))
```



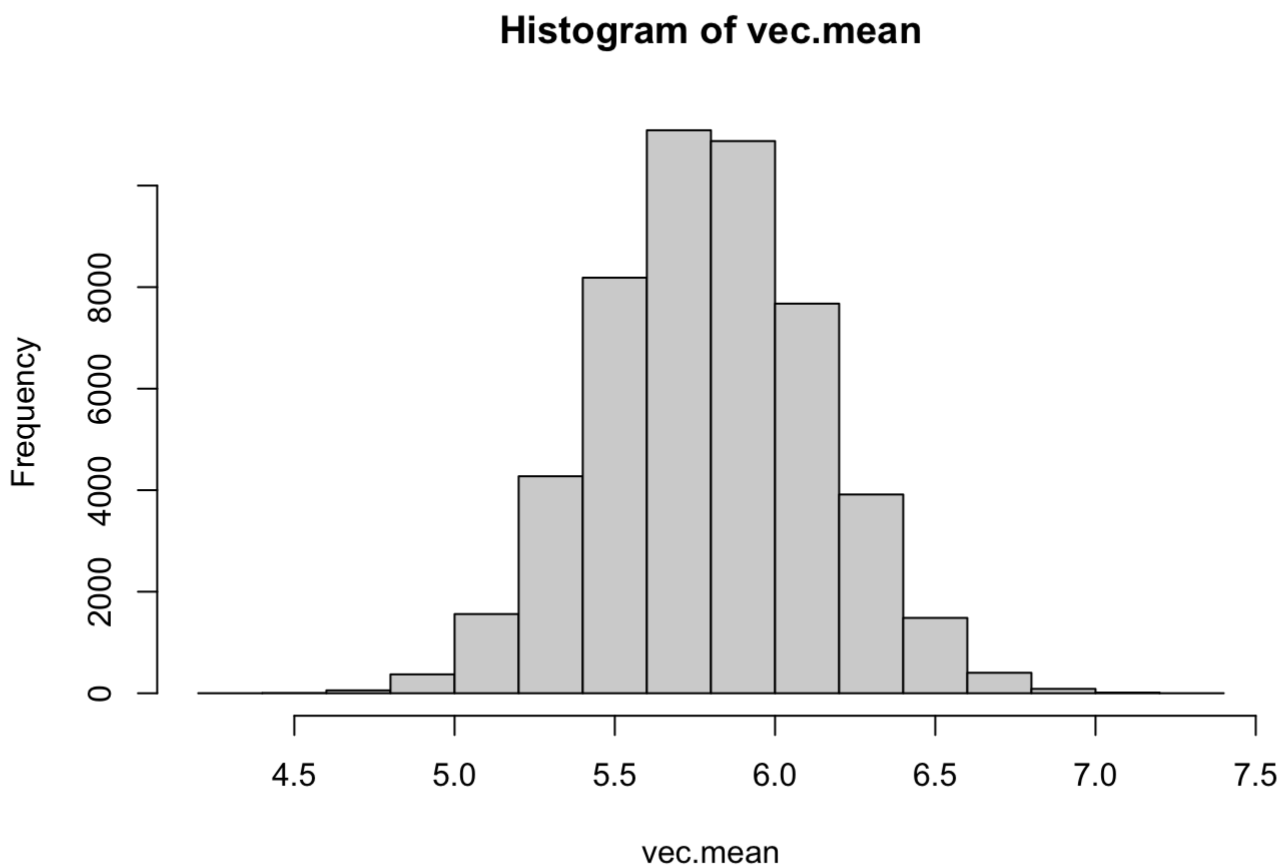
Yes, the data seems to be drawn from a normal distribution. In the histogram and Kernel density plot, we can see close resemblance to the bell curve and in the Normal probability plot, we see a major overlap between the line and data points, altough there is some deviation at the ends.

1e.) Obtain a big vector of 50000 sample means of 60 tickets each. Call this vector vec.mean.

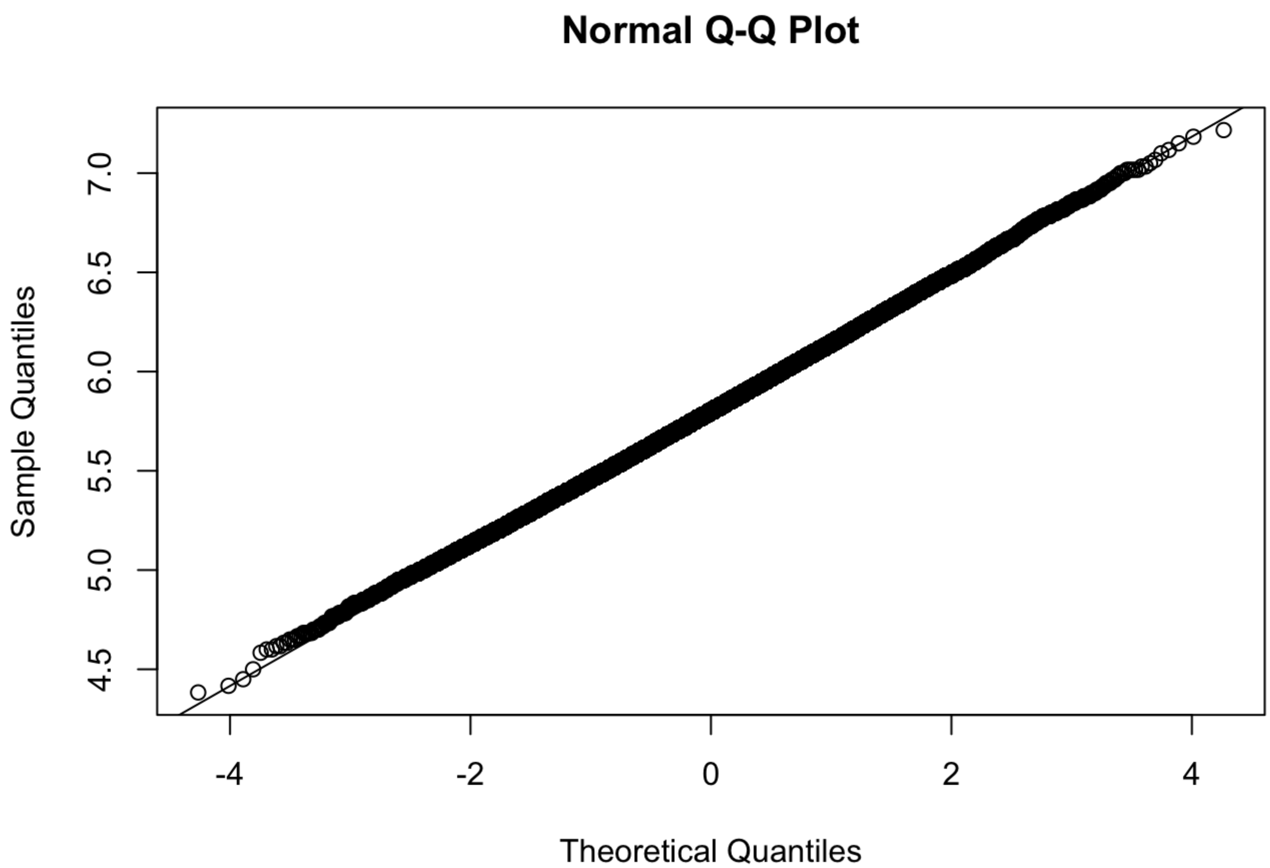
```
vec.mean = replicate(50000, mean(sample(urn, 60, T)))
```

1f.) Using vec.mean, construct a histogram, a normal probability plot, and a kernel density estimate. Does the data seem to be drawn from a normal distribution? Explain.

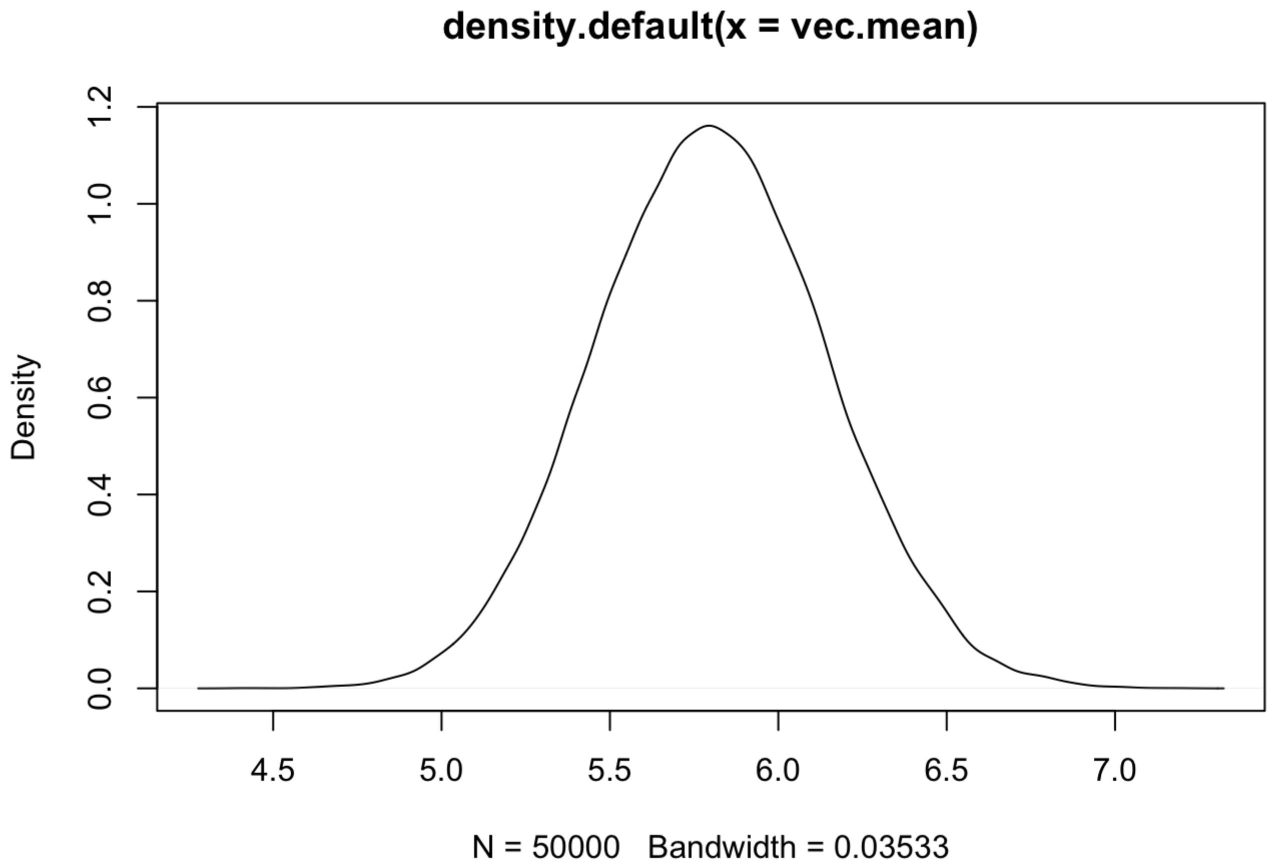
```
hist(vec.mean)
```




```
qqnorm(vec.mean)
qqline(vec.mean)
```




```
plot(density(vec.mean))
```



Yes, the data seems to be drawn from a normal distribution. In the histogram and Kernel density plot, we can see close resemblence to the bell curve and in the Normal probability plot, we see a major overlap between the line and data points.

Ans 2a.) Find E(X1).

```
x = c(1,2,3,6)
p = c (0.6,0.1,0.2,0.1)
mean_x1 = sum(x*p)
mean_x1
```

```
## [1] 2
```

2b.) Find Var(X1).

```
variance_x1 = sum(((x-mean_x1)^2)*p)
variance_x1
```

```
## [1] 2.4
```

2c.) P(1.8 < X1 ≤2.1) = P(X1 = 2)

```
C_2 = pnorm(2.1, 2, sqrt(2.4)) - pnorm(1.8, 2, sqrt(2.4))
C_2
```

```
## [1] 0.07709426
```

2d.) Let n = 80. Find E-X80 and V ar-X80

```
x80 = sample(x,80, prob = p,T)
mean_x80 = mean(x80)
mean_x80
```

```
## [1] 2.075
```

```
var_x80 = mean(x80^2) - mean_x80^2
var_x80
```

```
## [1] 1.994375
```

2e.) Let n = 80. Based on the CLT, approximate P(1.8 <-X80 ≤2.1)

```
sol = pnorm(2.1, 2, sqrt(0.03)) - pnorm(1.8, 2, sqrt(0.03))
sol
```

```
## [1] 0.594042
```

2f.) Construct a simulation of 40000 replications, each replication results in the observed sample mean. Use your simulation to obtain the approximate probability that P(1.8 <-X80 ≤2.1) and compare the result to part (e).

```
X2 = c(10,20,20,20,30,30,40,40,40,50)
xbar.vec = replicate(40000, mean(sample(X2, 80, replace = T)))
mean(xbar.vec > 1.8) - mean(xbar.vec <= 2.1)
```

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

Question 3 3.a) Write in R the proposed code, evaluate urn.model a total of 105 times, share your code, and based on that answer the questions.

```
urn.model <- c(1, 1, 1, 2, 2, 5, 10, 10, 10, 10)
n_draws <- 40

big_vec = replicate(n = 10^5, expr = sum(sample(urn.model, n_draws, replace = T)))
# big_vec

# Define the interval of interest
a <- 170.5
b <- 199.5

mean(big_vec < b) - mean(big_vec <= a)
```

```
## [1] 0.30078
```

3.b

```
EY = n_draws* (mean(urn.model))
EY
```

```
## [1] 208
```

```
VarY = n_draws* (sum((urn.model - mean(urn.model))^2*0.1))
VarY
```

```
## [1] 662.4
```

```
## Using the given formula and inserting the values obtained

se <- sqrt(VarY)
pnorm(199.5, mean= EY, sd = se) - pnorm(170.5, mean= EY, sd = se)
```

```
## [1] 0.2980481
```

3.c

```
pnorm((199.5- EY)/sqrt(VarY)) - pnorm((170.5 - EY)/sqrt(VarY))
```

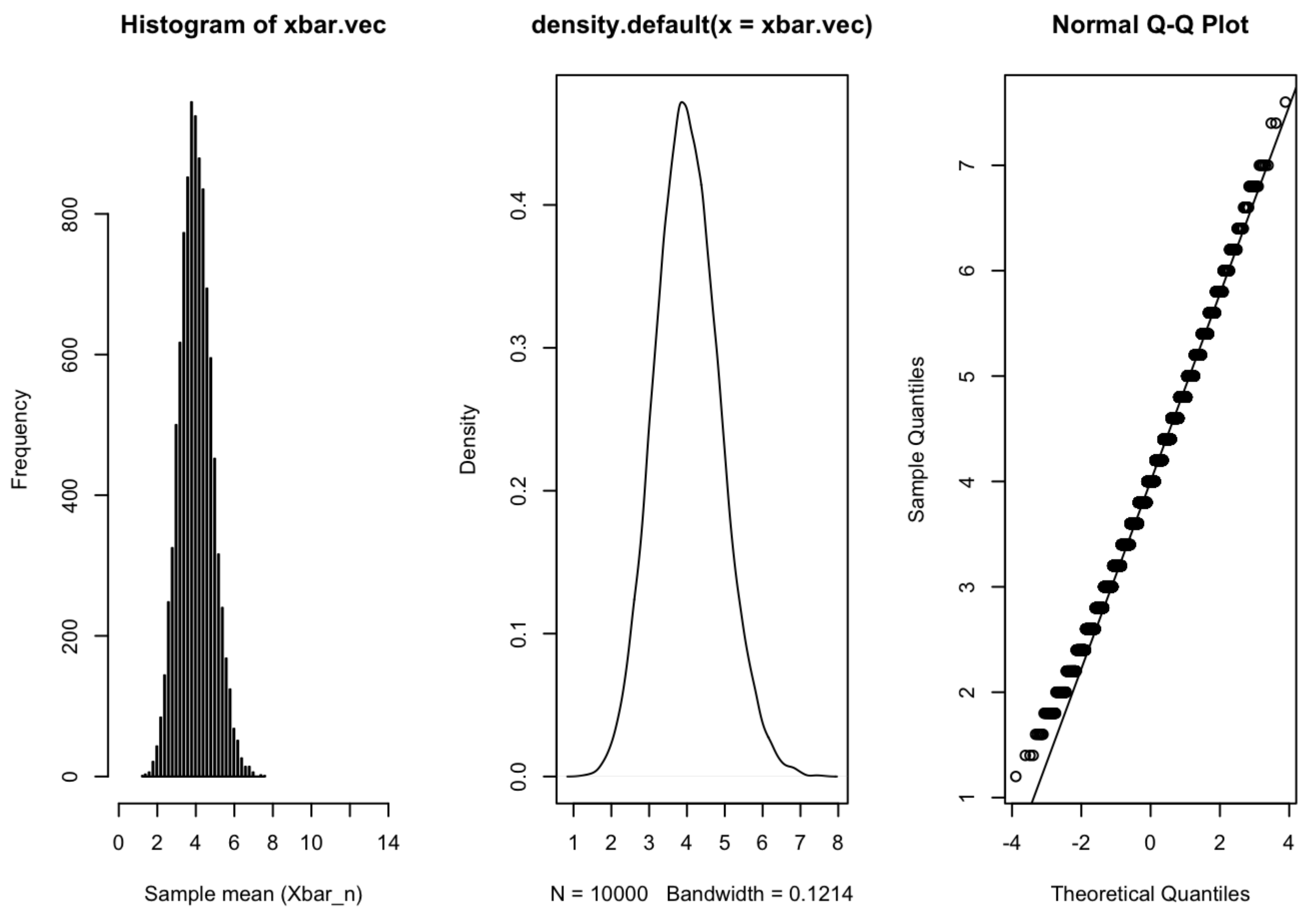
```
## [1] 0.2980481
```

#'The second students approach will provide a better estimate of the probability because when we assume CLT, we assume an infinite sample and the sample error would be less with a greater sample. # Question 5 Recall the heuristics when applying the CLT tell us that when the sample size is $n \geq 30$ the sample mean approximately follows the normal distribution. In this question you are asked to come up with counter-examples, i.e., examples that completely violate this rule of thumb.

```
clt = function(x, n, N = 10^4){
  xbar.vec = replicate(N, mean(sample(x, n, replace = T)))
  op = par(mfrow = c(1,3))
  hist(xbar.vec, breaks = 100,
       xlim = c(min(x), max(x)),
       xlab = paste("Sample mean (Xbar_n)"))
  plot(density(xbar.vec))
  qqnorm(xbar.vec);qqline(xbar.vec)
  par(op)
}
```

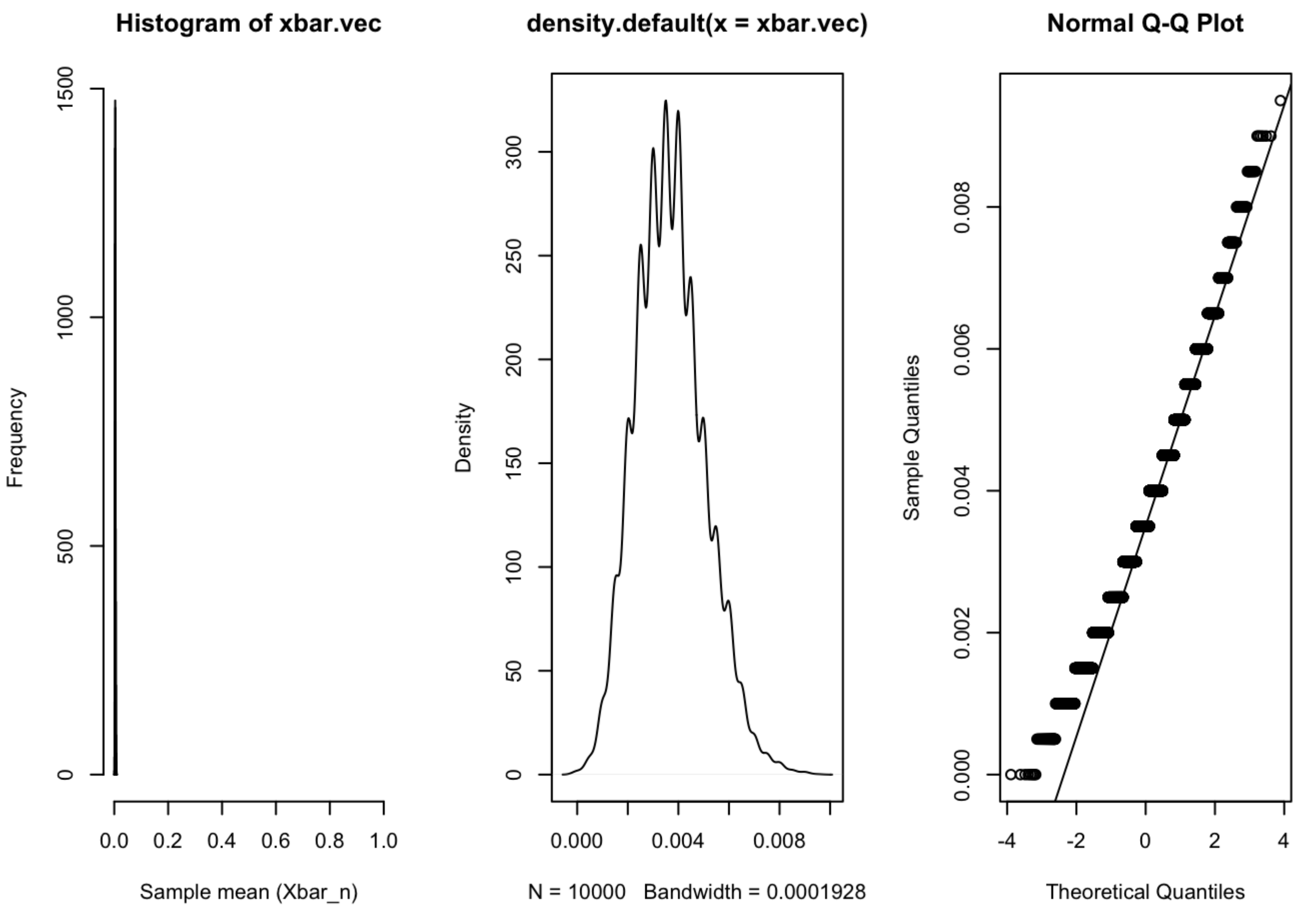
5.a) Constructing a Random Variable where sample mean is close to normal but not the population distribution

```
x1 = rbinom(10^4, 40, .1)
clt(x = x1, n = 5)
```



5.b) Constructing a Random Variable where is not normal

```
x2 = rbinom(10^4, 40, 0.0001)
clt(x = x2, n = 2000)
```


Histogram of xbar.vec
Frequency
0
500
1000
1500
0.0 0.2 0.4 0.6 0.8 1.0
Sample mean (Xbar_n)
density.default(x = xbar.vec)
Density
0
50
100
150
200
250
300
0.000 0.004 0.008
N = 10000 Bandwidth = 0.0001928
Normal Q-Q Plot
Sample Quantiles
0.000
0.002
0.004
0.006
0.008
-4 -2 0 2 4
Theoretical Quantiles