Project 2

Danny Ceron

November 5, 2018

Central Limit Theorem Simulations

(CLT)

if n is sufficientky large, X has approximately normal distribution with $\mu_x = \mu$ and $\sigma_x^2 = \frac{\sigma^2}{n}$

In addition, even when individual variables themselves are not normally distributed, sums and averages of the variables will under suitable conditions have approximately a normal distribution; this is the content of the Central Limit Theorem

```
leVec <- vector(length = 500)

for(i in 1:length(leVec))#----------------------------uniform--size-5-----------------
{
  result = runif(5, min = 0, max = 5)
  leVec[i]<-(mean(result))
}
#everage
mean(leVec)
```

```
## [1] 2.546178
```

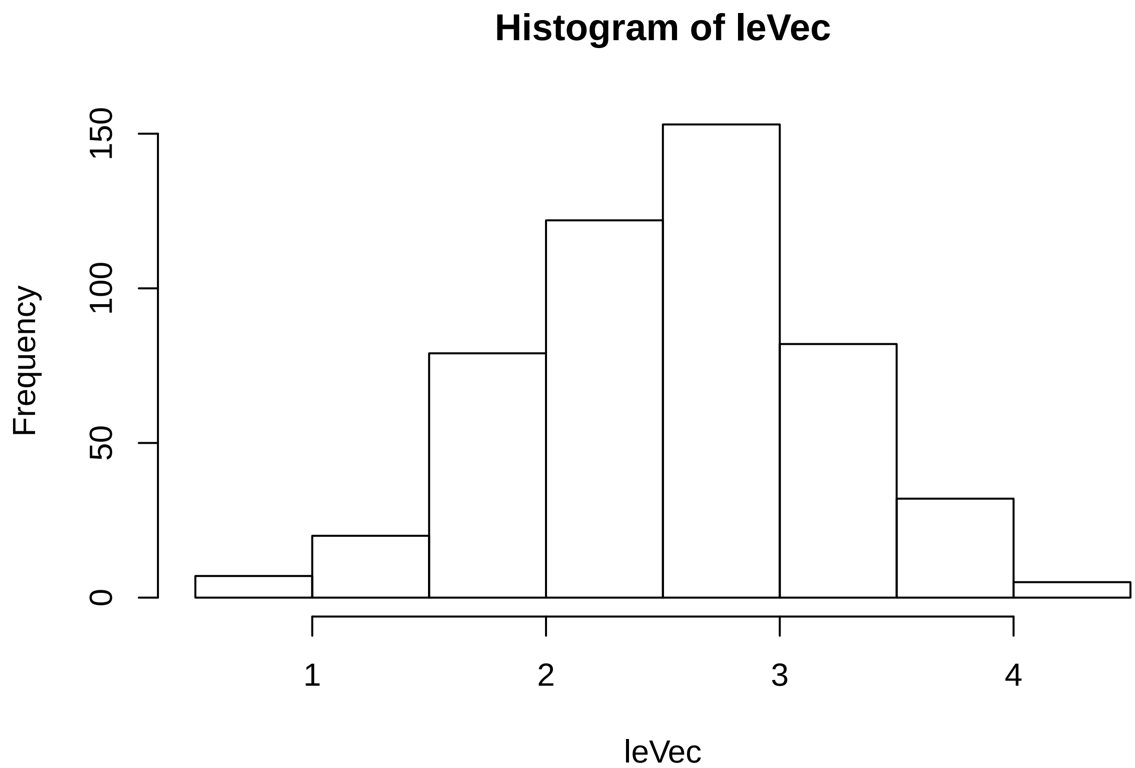
```
#variace
var(leVec)
```

```
## [1] 0.4162332
```

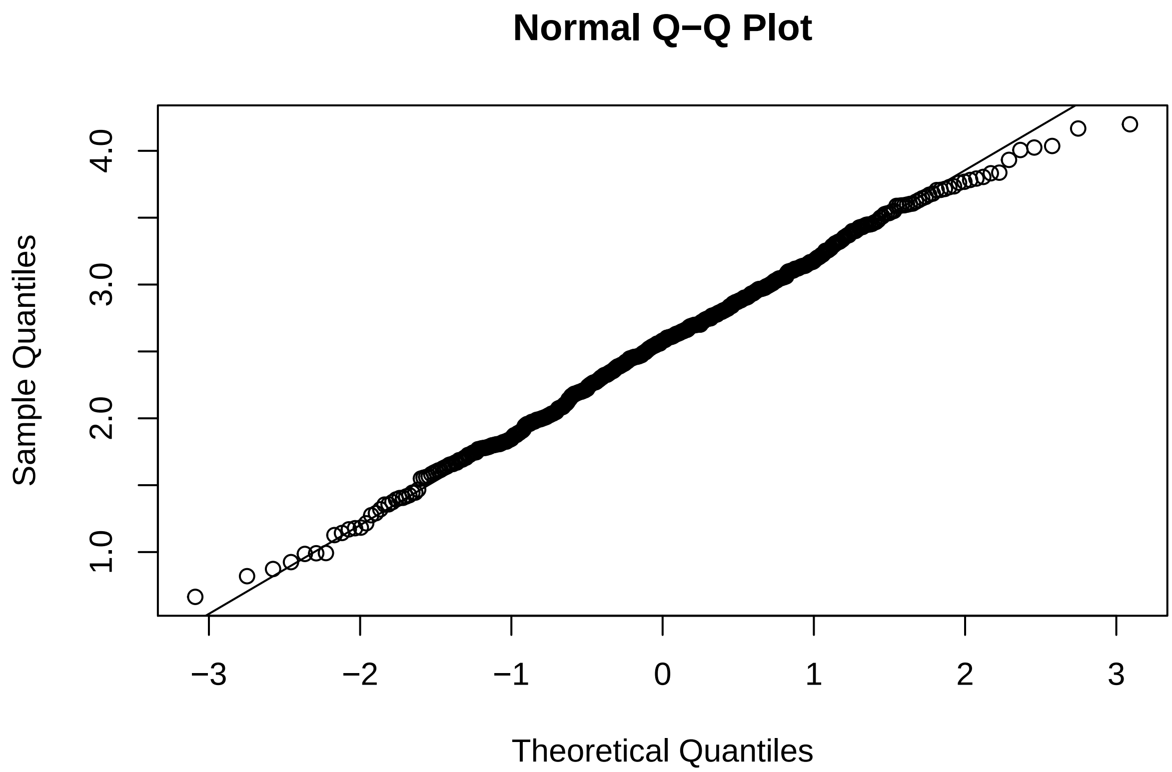
a. What is the average of the 500 sample means when the sample size is n = 5? What is the average of the 500 sample means when the sample size is n = 50? What are the theoretical expected values of sample means, respectively? the theoretical expected values of the uniform distribution: $Ex = \mu x = \frac{A+B}{2}$ in this case we have $Ex = \mu x = \frac{5-0}{2} = 2.5$
b. For n = 5 and n = 50, what are the variances of the 500 sample means, respectively? What are the theoretical variances of sample means, respectively? the theoretical varian values of the uniform distribution with samples size 5: $\sigma^2 = \frac{(B-A)^2}{(12)(5)}$

now for sample size 50 $\sigma^2 = \frac{(B-A)^2}{(12)(50)} = 0.041$ c. Construct histogram for sample means for n = 5 and n = 50, respectively. d. Construct normal probability plot of sample means for n = 5 and n = 50, respectively.

```
hist(leVec)
```




```r
qqnorm(leVec)
qqline(leVec)
```



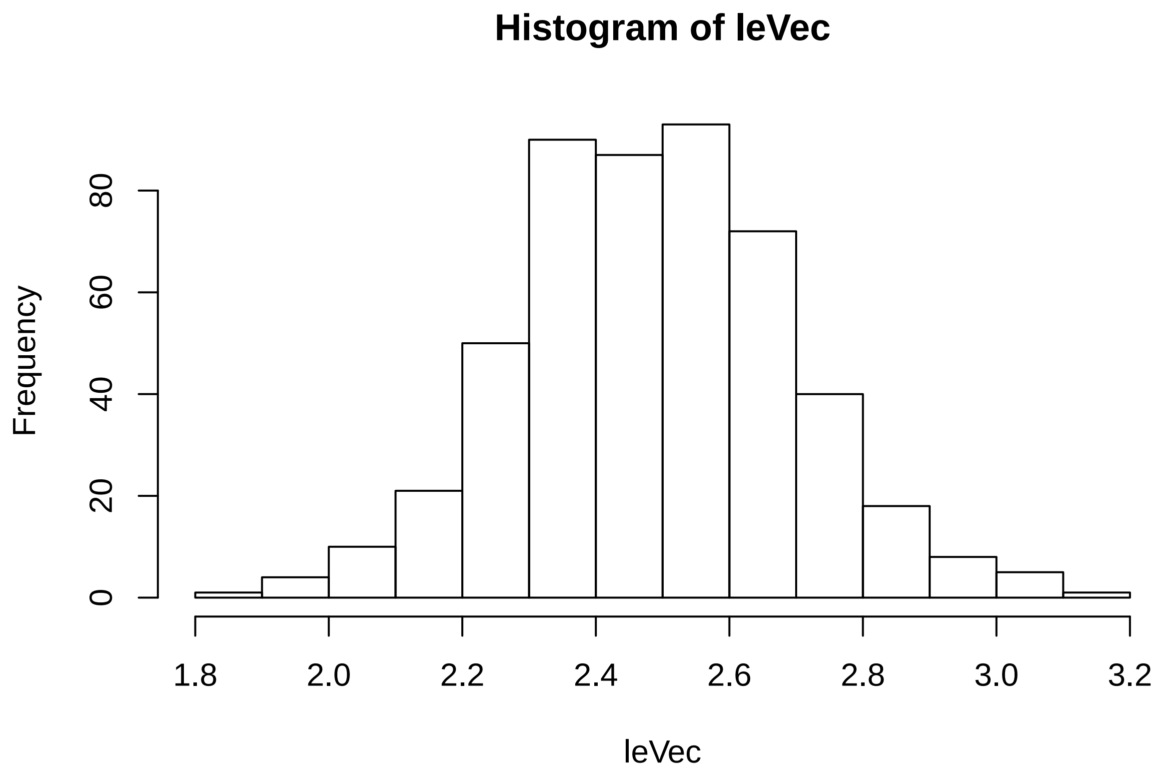

```r
leVec <- vector(length = 500)

for(i in 1:length(leVec))#---------------------------uniform--size-50----------------
{
  result = runif(50, min = 0, max = 5)
  leVec[i]<-(mean(result))
}
#everage
mean(leVec)
```

```
## [1] 2.485561
```

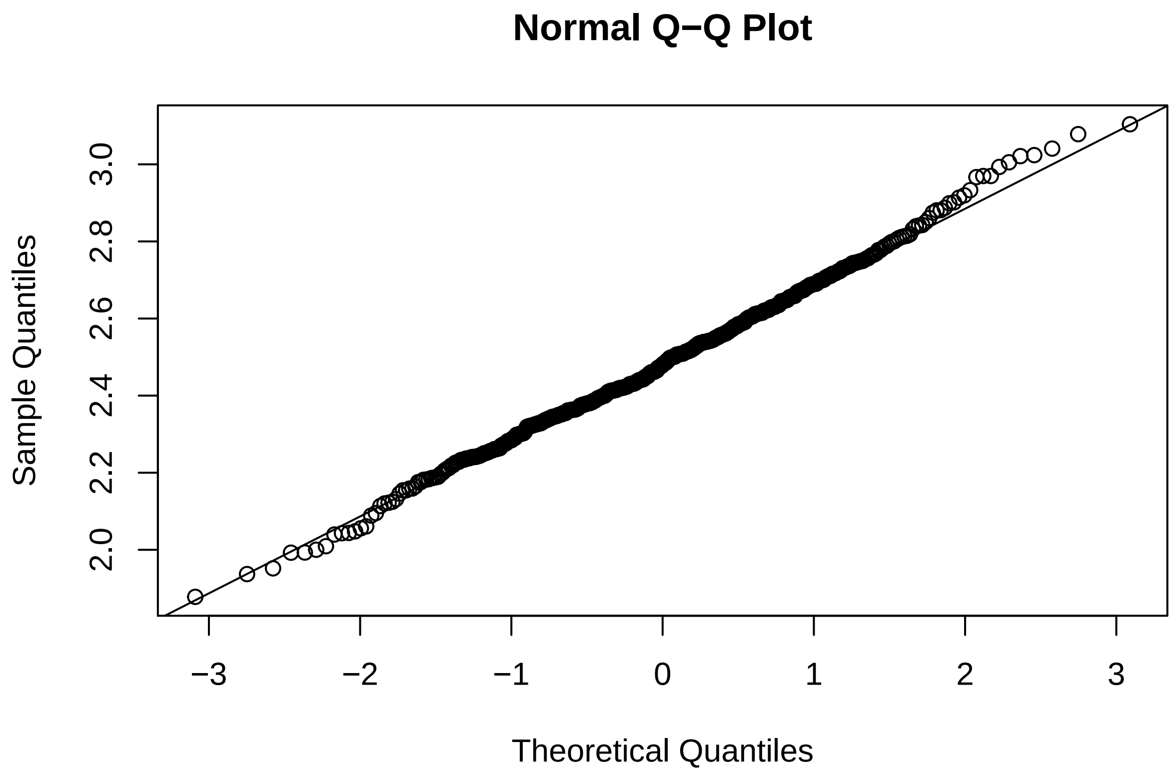
```r
#variace
var(leVec)
```

```
## [1] 0.04203988
```

```r
hist(leVec)
```




```r
qqnorm(leVec)
qqline(leVec)
```



e. After answering parts a-d for all distributions I-IV, you should see some common trends across distributions. Summarize these findings for each of a-d, and use the central limit theorem to explain your findings.

central limit theorem gives us the distribution from a sample popultion. it also gives us the theoretical values of the mean and the variance of that sample. The algarthem is picking ramdom samples with in specifict ranges. then it takes the averages of those values inputting them into a vector and then taking mean and variance of that vector.central limit theorem is something similar.

One thing we can notices is that as the sample size increases the more accurate the teorethical values are with the calculated values. also the higher the samples the more it will start to look like a normal disrtibution. the larges the samples, the more their mean will be towards center. we can even see in the probability-plots, when the samples size was 5 the distribution was not normal. when the sampels was of 50 the distribution was starting to look normal. when the plot does not follow a the linear line, it means that the distribution is either ritght or left skewed whish is what happen with the sample size of 5. ##*****************************End of uniform***************************

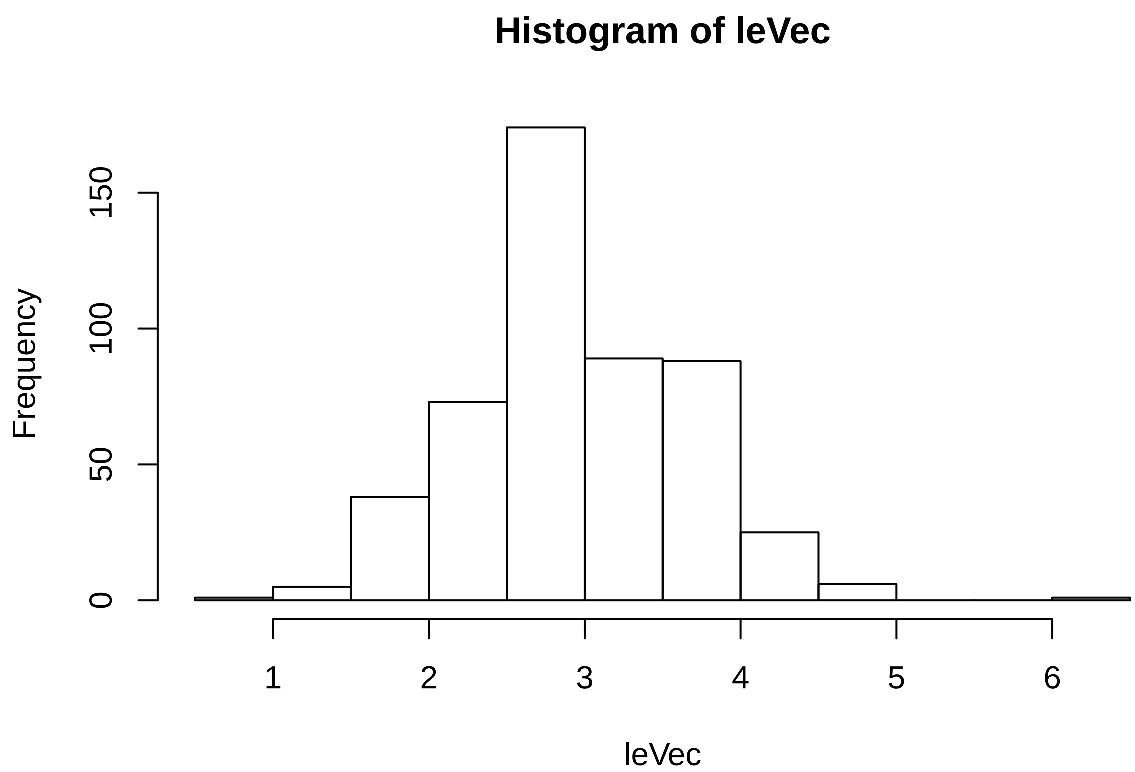
```
for(i in 1:length(leVec)){
  result = rbinom(5,size = 15,prob = .2)#-----------------------------binomial---size-5----------------
  leVec[i]<-(mean(result))
}
#everage
mean(leVec)
```

```
## [1] 3.0064
```

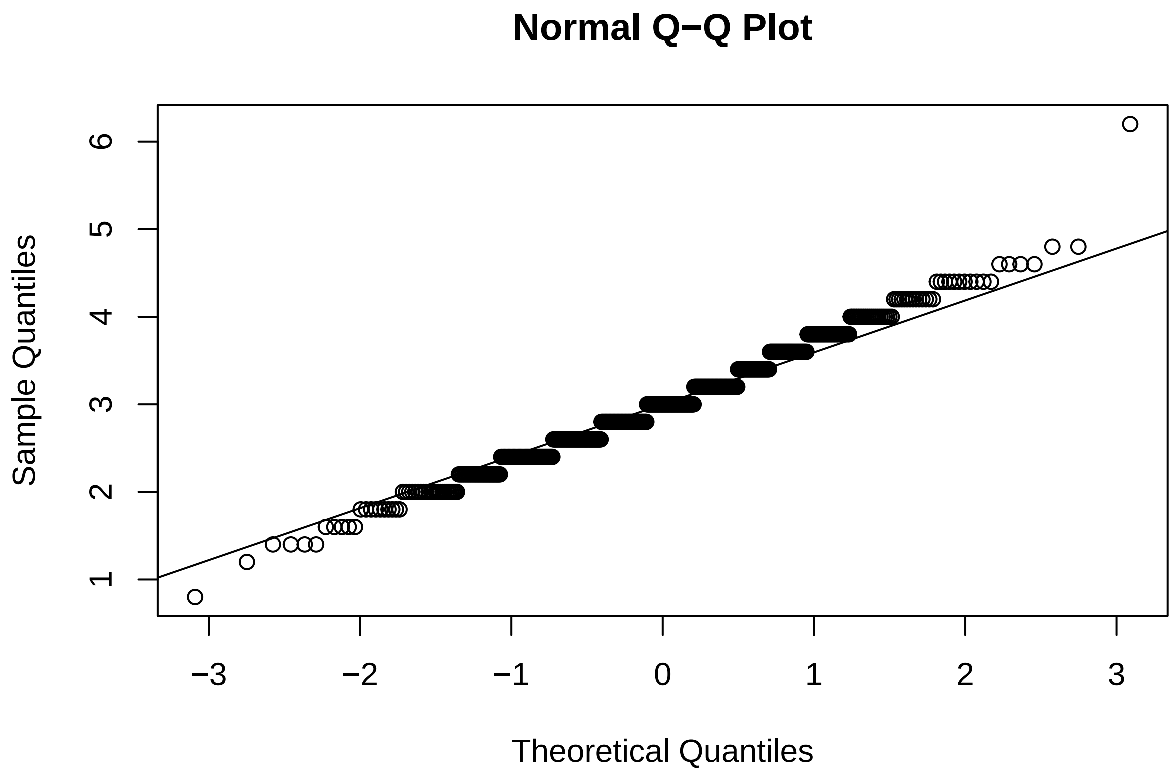
```
#variace
var(leVec)
```

```
## [1] 0.4833257
```

```r
hist(leVec)
```




```r
qqnorm(leVec)
qqline(leVec)
```



a. What is the average of the 500 sample means when the sample size is n = 5? What is the average of the 500 sample means when the sample size is n = 50? What are the theoretical expected values of sample means, respectively? n: number of trials $E(x)=np $ $E(x) = (15)(.2) = 3$

b. For n = 5 and n = 50, what are the variances of the 500 sample means, respectively? What are the theoretical variances of sample means, respectively?

n: number of trials.

with sample of 5 $V(x) = npq = \frac{(15)(.2)(1-.2)}{(5<-samplessize)} = .384$

with samples of 50 $V(x) = npq = \frac{(15)(.2)(1-.2)}{50} = 0.0384$

c. Construct histogram for sample means for n = 5 and n = 50, respectively.
d. Construct normal probability plot of sample means for n = 5 and n = 50, respectively.

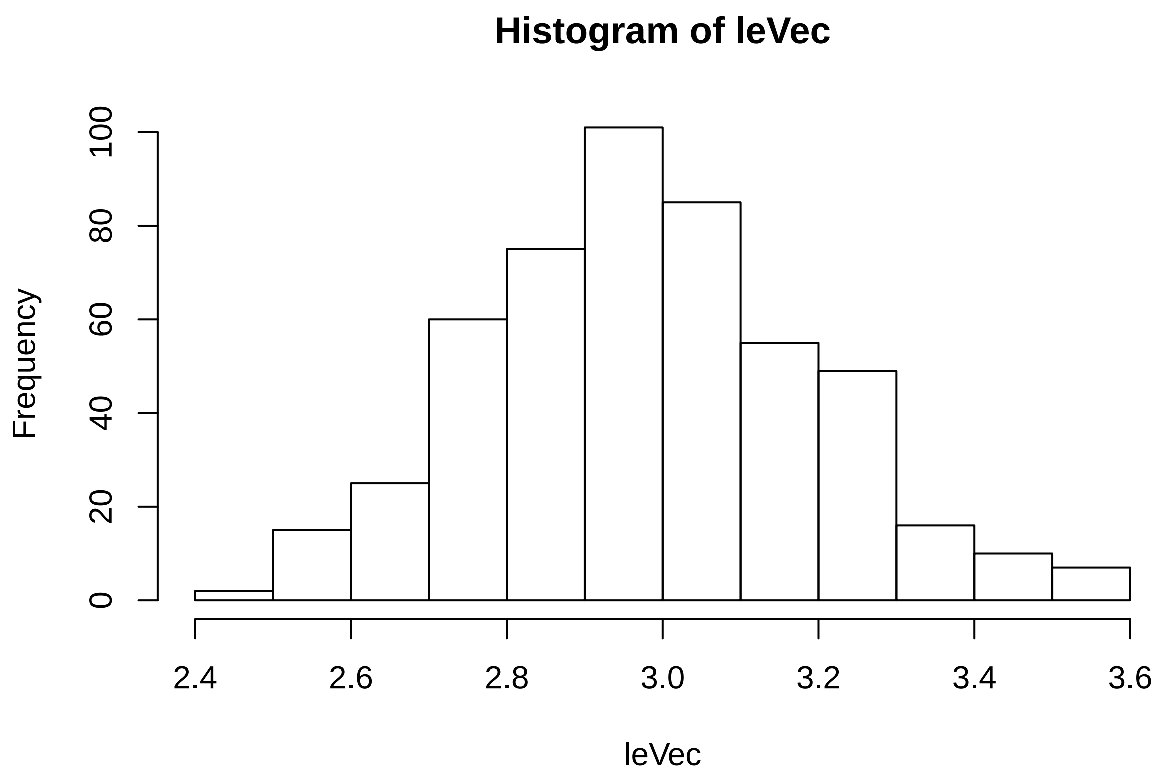
```
  for(i in 1:length(leVec))#-----------------------binomial---size-50-------------
{
  result = rbinom(50,size = 15,prob = .2)
  leVec[i]<-(mean(result))
}
#everage
mean(leVec)
```

```
## [1] 2.9914
```

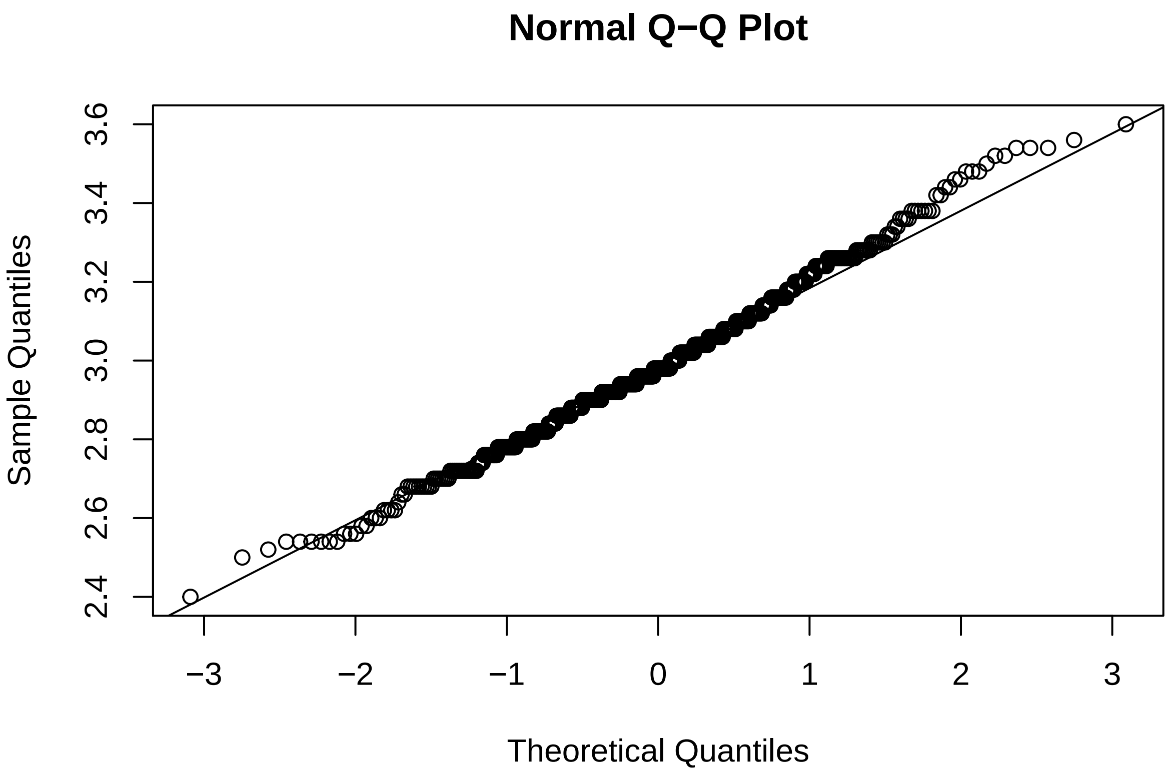
```
#variace
var(leVec)
```

```
## [1] 0.04532309
```

```r
hist(leVec)
```




```r
qqnorm(leVec)
qqline(leVec)
```



e. After answering parts a-d for all distributions I-IV, you should see some common trends across distributions. Summarize these findings for each of a-d, and use the central limit theorem to explain your findings.

looking back at the values of the expecte and theoretical mean, we can see that they are different, not by much but there is quite a difference error.

again central limit theorem tells us the theoretical mean and variece of a samples size. it will always not be sertain because the values taken from the sampels are always random, but as the samples size increases the more normalily distributed.

we can see that the variance values of the samples size 50 are more similar than the ones with samples size 5.

I have noticed that average also changes arcording to its size. again when the size of the calculated value the closer it getrs to the mean of the expexted value which is also the mean of the population.

this means that when the samples sizes of a sample size increased the mean of the sample gets closer to the mean of the population.

for the probability plt with sample population of 5 we can see that distribution is sort of linear but there are some missing values in between. for the plot with samples sixe of 50 the data looks more normally distributed. thats because the variance is determines the spread of sample means. ##********************************End of binomial**********************************

```
  for(i in 1:length(leVec))
{
  result = rexp(5,rate =5)#----------------exponential--size-5---------------
  leVec[i]<-(mean(result))
}
#everage
mean(leVec)
```

```
## [1] 0.2090321
```

```r
#variace
var(leVec)
```

```
## [1] 0.00921812
```

a. What is the average of the 500 sample means when the sample size is n = 5? What is the average of the 500 sample means when the sample size is n = 50? What are the theoretical expected values of sample means, respectively?

$$E(x) = \mu = \frac{1}{(\lambda)(n < -samplesize)}$$

$\mu = \frac{1}{\lambda} = .2$

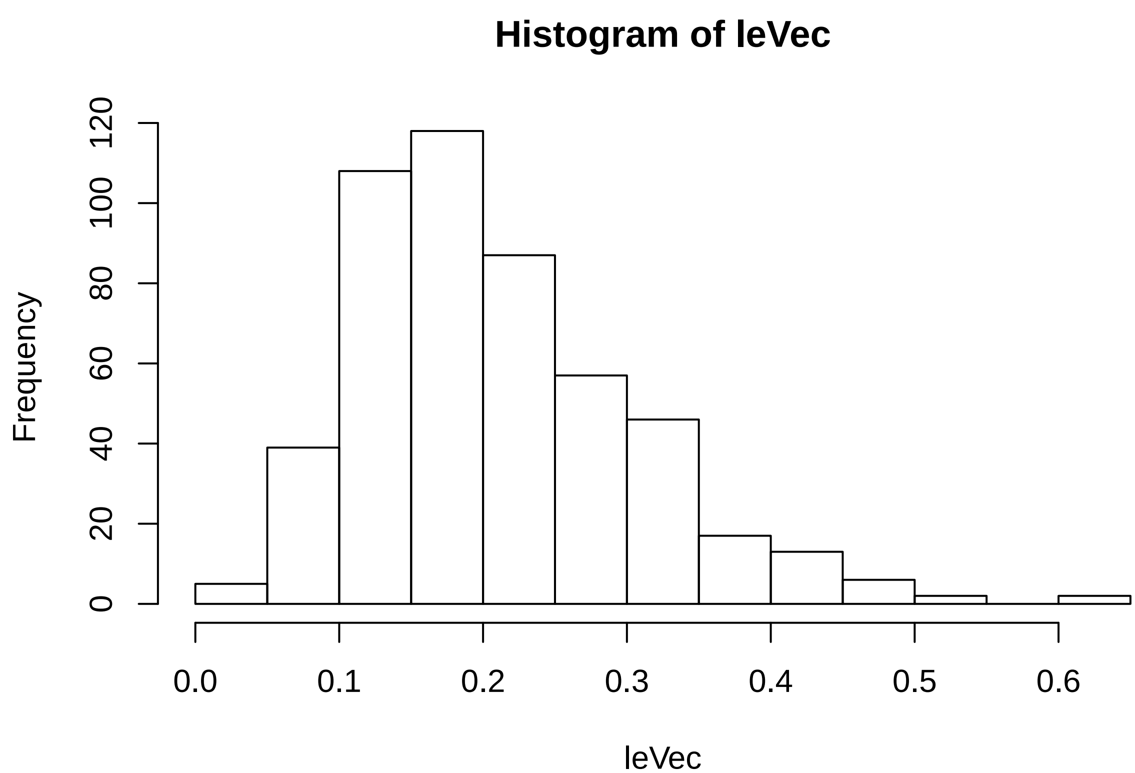
sample size of 5 $\mu_x = \frac{1}{(5)(5)} = .04$ sample size of 50 $\mu_x = \frac{1}{(5)(50)} = .004$

b. For n = 5 and n = 50, what are the variances of the 500 sample means, respectively? What are the theoretical variances of sample means, respectively?

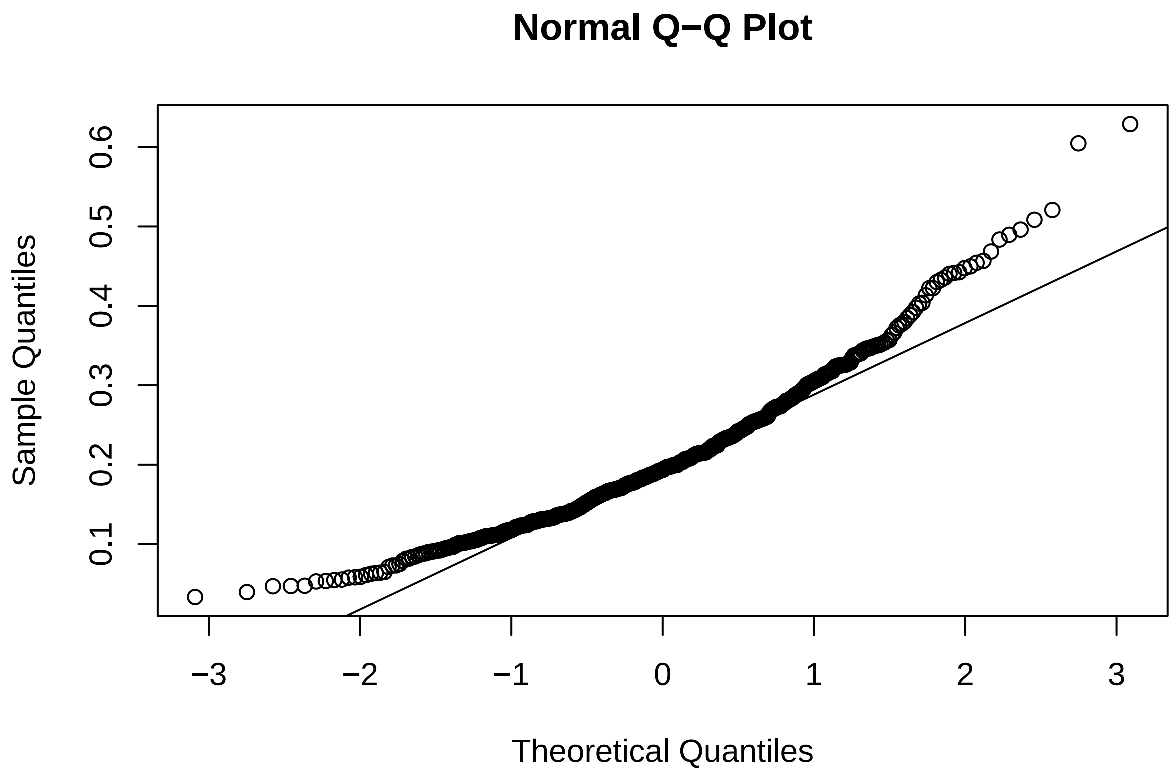
c. Construct histogram for sample means for n = 5 and n = 50, respectively.

d. Construct normal probability plot of sample means for n = 5 and n = 50, respectively.

```r
hist(leVec)
```




```r
qqnorm(leVec)
qqline(leVec)
```



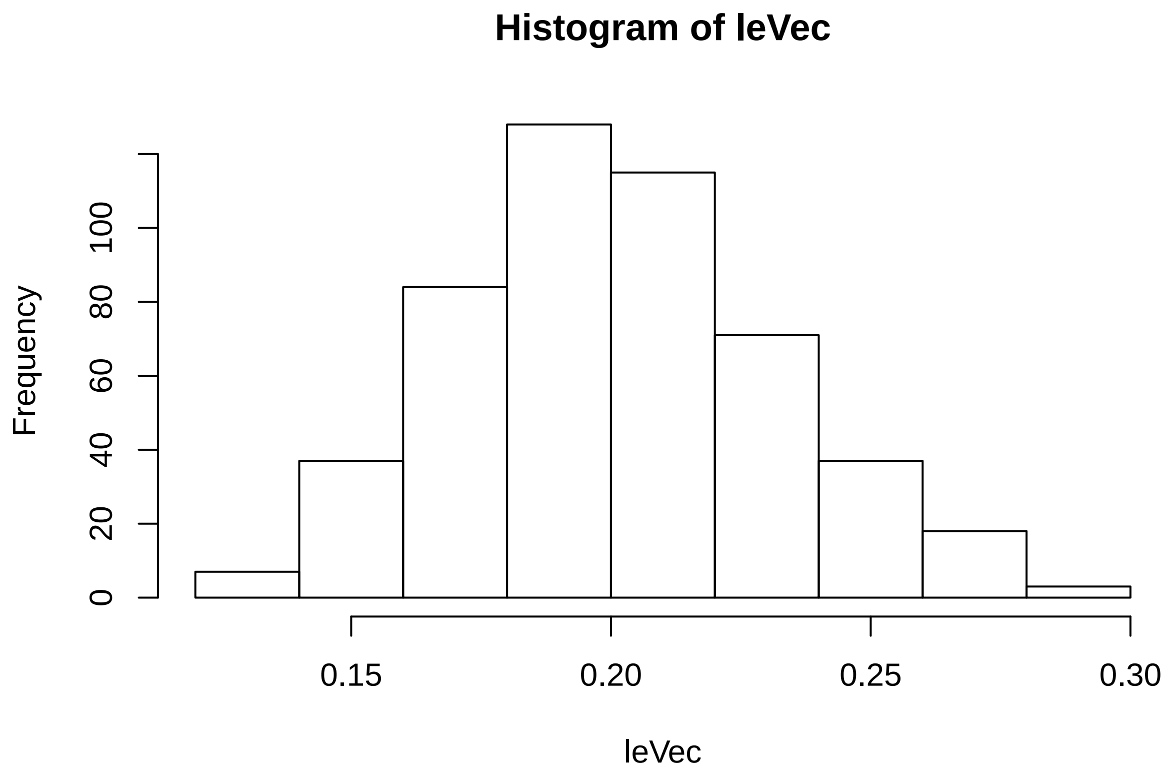

```
  for(i in 1:length(leVec))
{
  result = rexp(50,rate =5)#---------------------exponential--size-50---------------
  leVec[i]<-(mean(result))
}
#everage
mean(leVec)
```

```
## [1] 0.2012114
```

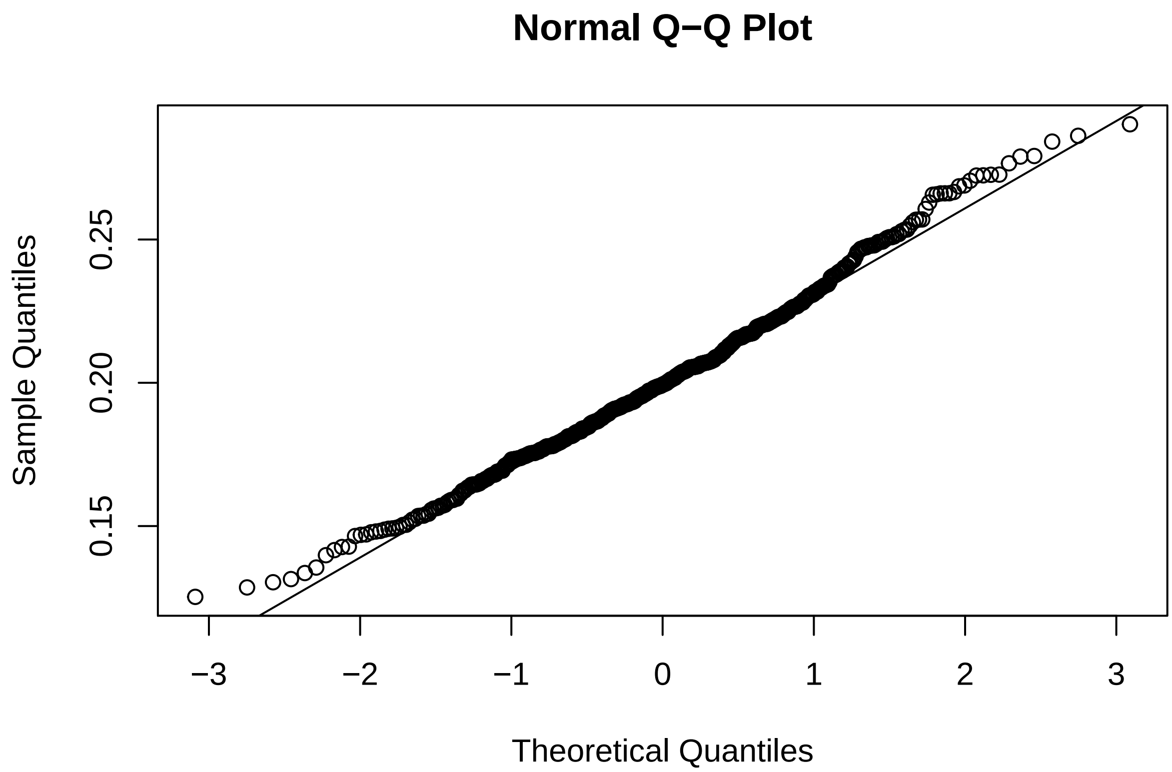
```
#variace
var(leVec)
```

```
## [1] 0.0009394677
```

```
hist(leVec)
```




```
qqnorm(leVec)
qqline(leVec)
```



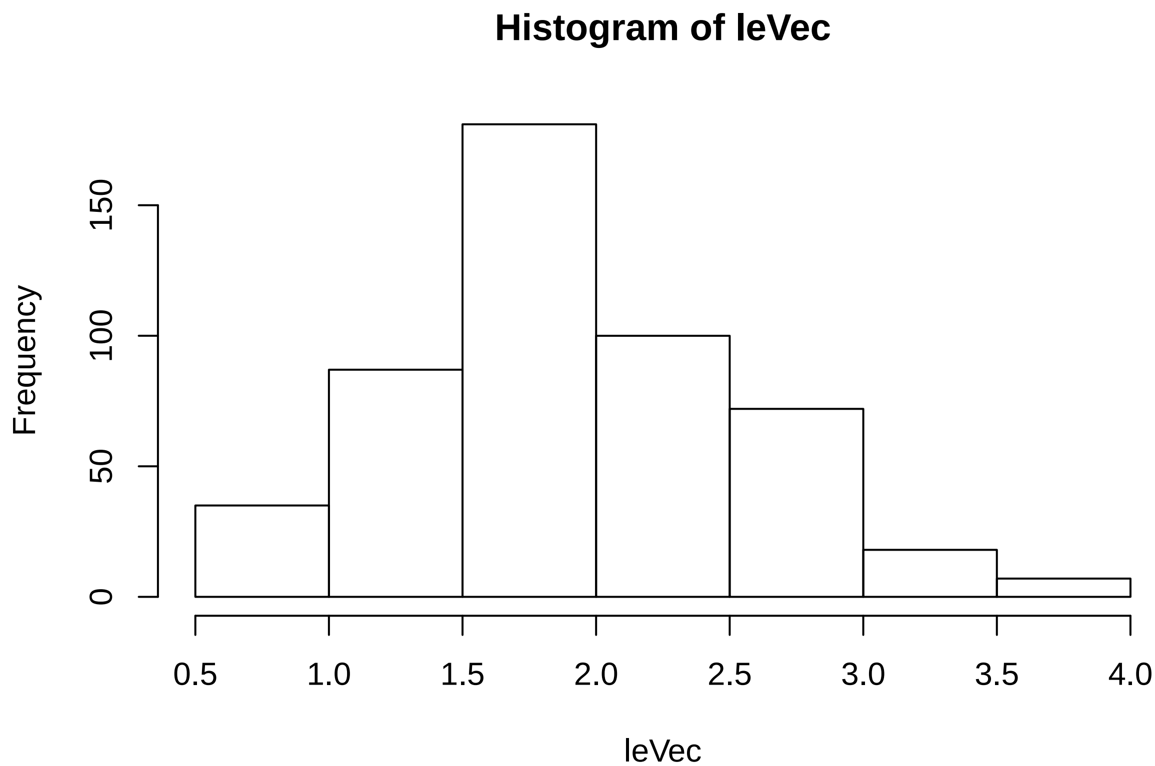
e. After answering parts a-d for all distributions I-IV, you should see some common trends across distributions. Summarize these findings for each of a-d, and use the central limit theorem to explain your findings.

for the mean and variance of the sample size 5, the theoretical value is less closely to the calculated value. by looking at the hitogram we can see that the data is very scewed. when the samples size is 50 the calculated mean and the variance is more closely to the theoretical.

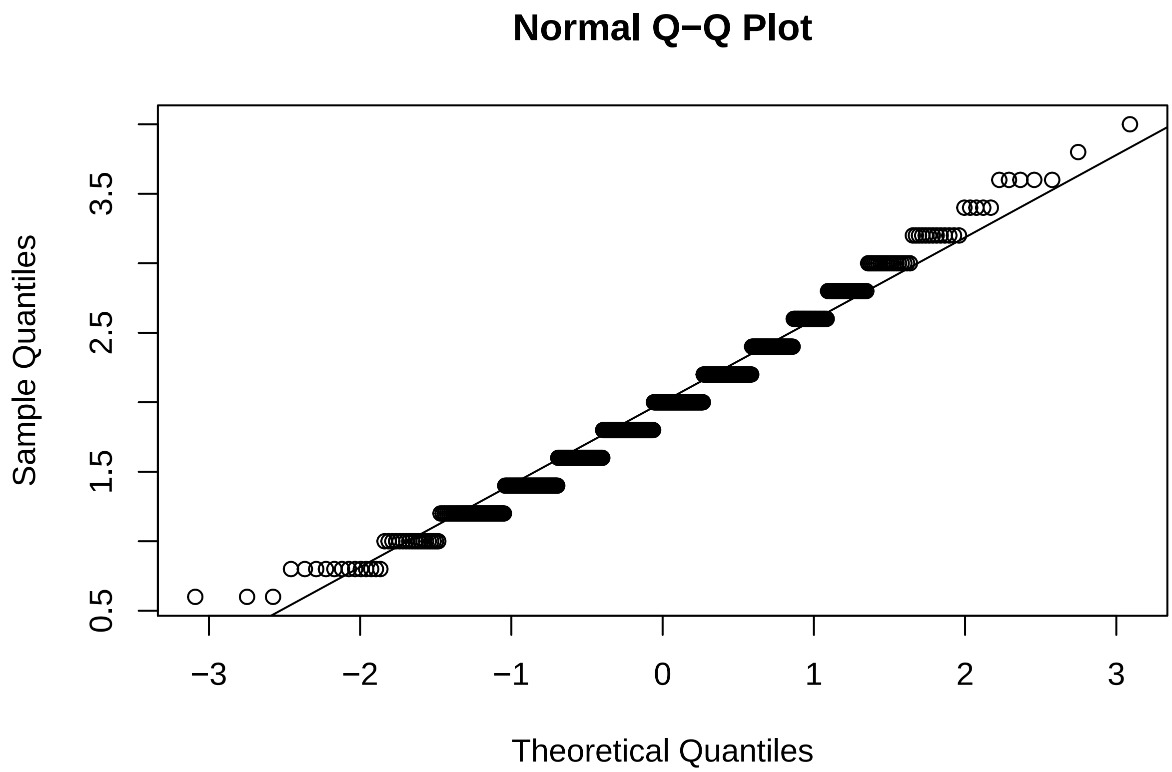
if we compare the graph of the 5 sampels compared to the graph with 50 samples, we can see that the one with 5 is highly spread towards the middle, and for the 50 samples the data is more normal. for the smaller samples is less predictable on how scewed the data will be.

***********************************End of exponetial**************

```
  for(i in 1:length(leVec))
{
  result = rpois(sample(5),2)#---------------------Poisson--size-5----------------
  leVec[i]<-(mean(result))
}
hist(leVec)
```




```
qqnorm(leVec)
qqline(leVec)
```




```
mean(leVec)
```

```
## [1] 1.9728
```

```
var(leVec)
```

```
## [1] 0.3986174
```

```
  for(i in 1:length(leVec))
{
    mu <- 2
  result = rpois(sample(50),mu)#--------------------Poisson-size-50-------------------
  leVec[i]<-(mean(result))
}
#everage
mean(leVec)
```

```
## [1] 2.00028
```

```
#variace
var(leVec)
```

```
## [1] 0.04104601
```

a. What is the average of the 500 sample means when the sample size is n = 5? What is the average of the 500 sample means when the sample size is n = 50? What are the theoretical expected values of sample means, respectively? $E(x) = \mu = V(x)$ $\mu_x = 2$
b. For n = 5 and n = 50, what are the variances of the 500 sample means, respectively? What are the theoretical variances of sample means, respectively?

$v(x) = \mu = \lambda = \alpha t$

expected value for sampels sieze 5

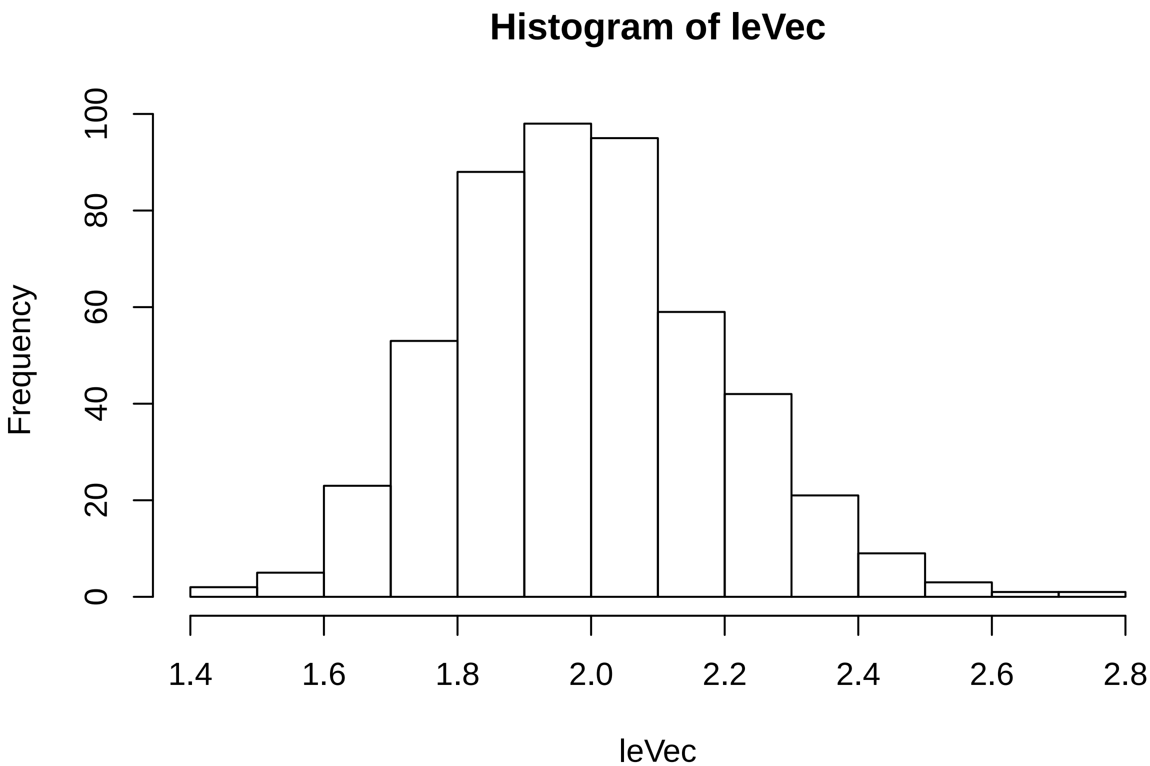
$\sigma_x^2 = \frac{2}{n} = \frac{2}{5} = 0.4$

expected value for sampels sieze 50

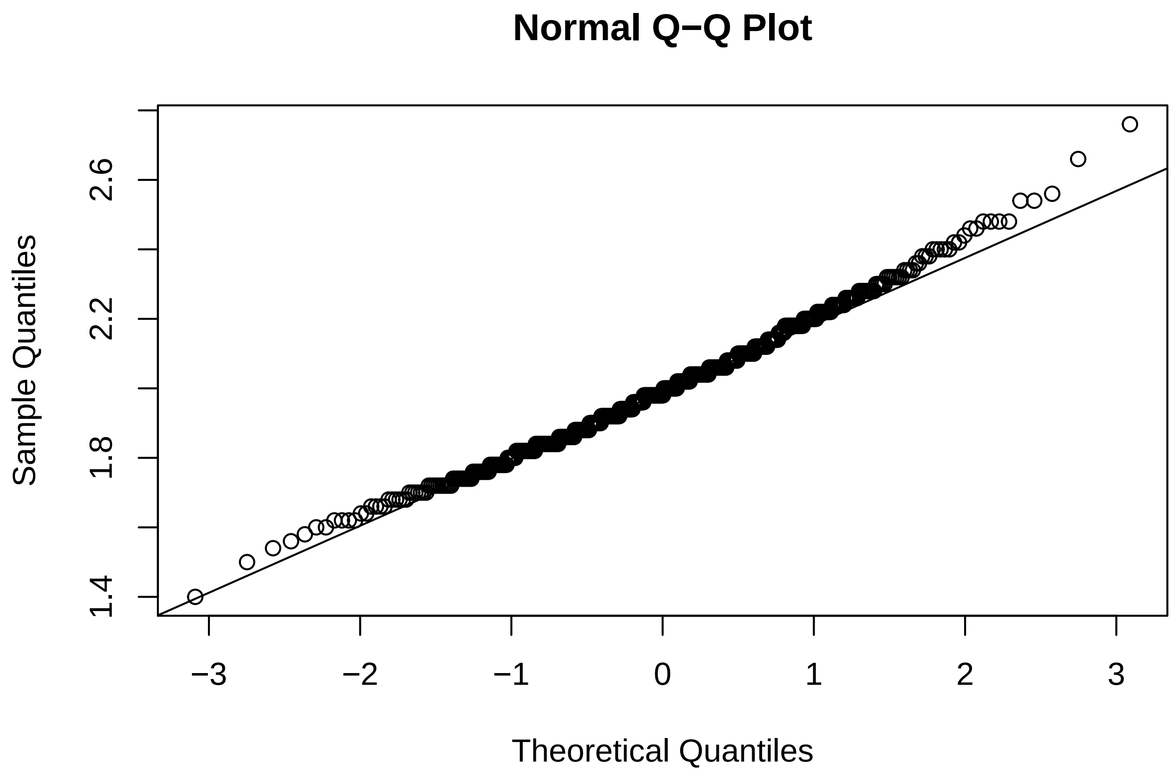
$\sigma_x^2 = \frac{2}{n} = \frac{2}{50} = 0.04$

c. Construct histogram for sample means for n = 5 and n = 50, respectively.
d. Construct normal probability plot of sample means for n = 5 and n = 50, respectively.

```
hist(leVec)
```




```
qqnorm(leVec)
qqline(leVec)
```



e. After answering parts a-d for all distributions I-IV, you should see some common trends across distributions. Summarize these findings for each of a-d, and use the central limit theorem to explain your findings.

for Poisson, the is distributes similarly to the distribution of the binomial. but when we increase the sample size the the distribution becomes more normal.

looking at both the histrgrams and probability plot. we can see that the spread is skewed and has missing varibles in between when the sample size is small. *——————————————Algorithm 2——————————————- ##Bootstrap: Estimating the Variance of a Statistic

```
bootstrap_vector <- vector(length = 40000)
set_trials = 15
sample_size = 50
probability = .2
samples <- rbinom(sample_size,set_trials,probability)
for(i in 1:length(bootstrap_vector))
{

samplesx <- sample(x = samples,size = sample_size,replace = TRUE)

mean(samplesx)
bootstrap_vector[i] <- mean(samplesx)
}
```

a. Report the variance from Step 8 and consider the theoretical variance you got in Exercise 1-II-b for n = 50. Is your bootstrapped estimator for the variance on the right scale? (That is, if the theoretical variance is somewhere in 1-2, is the boostrapped variance also somewhere in 1-2?)

the Theoretical variance I got from exercise 1-II-b for sample 50 was 0.0384 it is also very close. in on of the run I got 0.045… considereing that we are taking samples from one sample of that binomial, we are really close to theoretical variance.

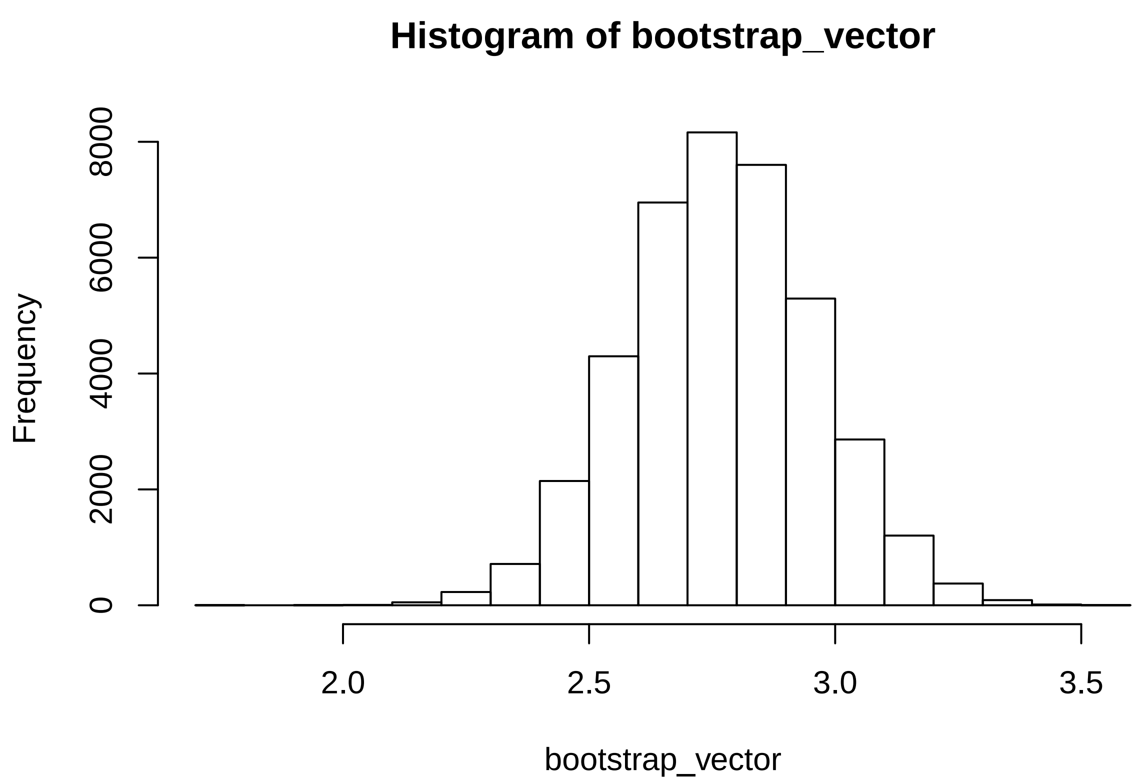
b. If you actually could compute the theoretical variance, as you can for a sample mean of a binomial distribution, would you prefer to bootstrap an estimated variance for that mean as we did here or calculate the theoretical variance as we did in Exercise 1? Why? (Think about computational cost and accuracy.)

they would both work just fine. but i think I would take a lot less sampling then using the bootstrap. in the boots strap we ran a loop 40000 times just to approximate what a loop 500 did. I belive that sampling more of the original data is better then sampling froma sample. As we have seen with this exercise. althou the bootstrap looks more like a normal distribution, I think if we have a much larger loop we wouls also get a more normal distribution.

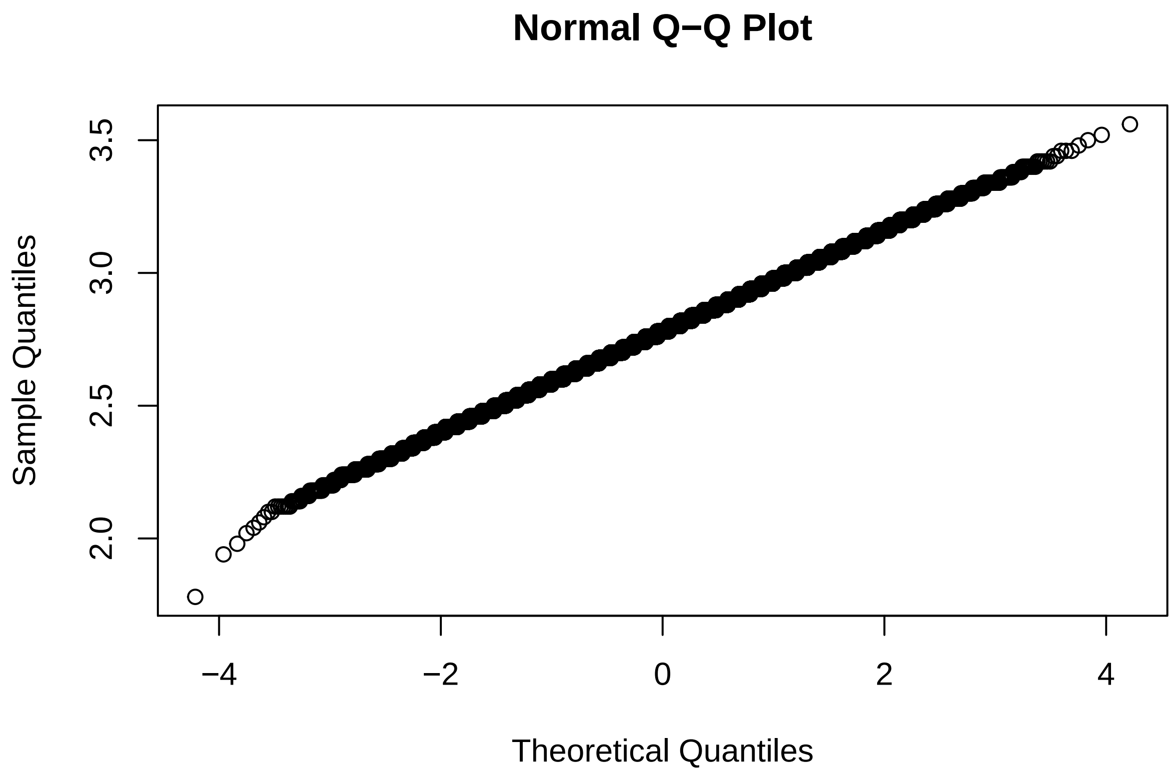
```
var(bootstrap_vector)
```

```
## [1] 0.03633362
```

```
hist(bootstrap_vector)
```




```
qqnorm(bootstrap_vector)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
2.0
2.5
3.0
3.5
−4
−2
0
2
4