Coursework 2-Shizhi Chen-10307389

Assume we have a discrete random variable with probability function given by

x	p(x)
0	0.2
3	0.1
5	0.1
10	0.6

Perform the following tasks:

1. Calculate the expected value and variance

x=c (0,3,5,10)

px=c (0.2,0.1,0.1,0.6)

Ex=sum(x*px)

varx=sum((x^2)*px)-(Ex)^2

```
> Ex
[1] 6.8
> varx
[1] 17.16
```

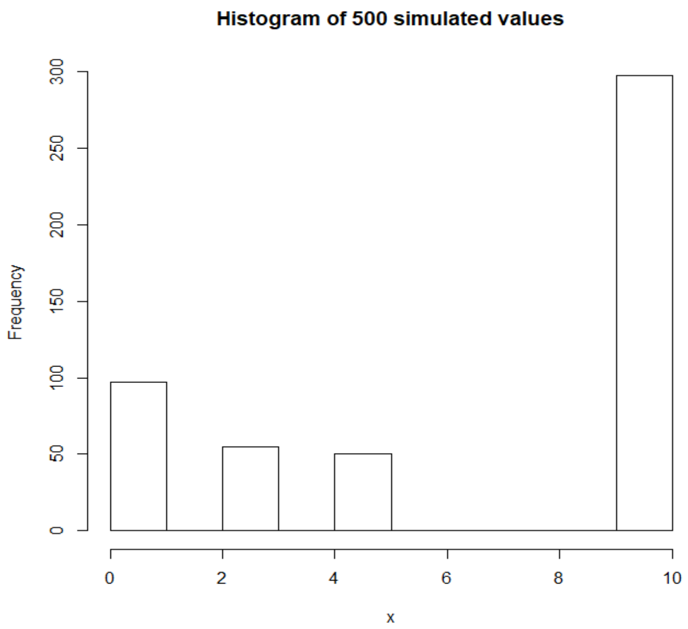
2. Simulate 500 values from this distribution

data=sample (x, prob=px,500, replace=TRUE)

```
> data
  [1]  0 10  3 10 10 10 10 10  0  0 10 10 10  0 10  0  3  0 10  5 10 10  0 10 10
 [26]  3 10  0 10 10 10 10  0  5 10  5  0  0 10  3 10  0 10 10  3  5 10 10 10 10
 [51]  0 10 10  0  5  5 10 10 10  0  0 10  5 10  3 10  3 10 10 10 10 10  0 10  0
 [76] 10 10  0  0  0 10  3 10  5  3 10 10  3 10 10  0  0 10  0 10 10  3 10 10 10
[101]  0 10  0 10 10 10 10 10  0 10  0 10  5 10 10 10 10 10 10 10 10  3 10 10 10
[126] 10  3 10  0 10  5 10 10 10  0  0 10  3  0 10 10 10 10  3  0 10 10 10 10 10
[151] 10 10 10  3 10 10 10 10  0 10  3  3 10 10 10  0  0  5 10 10  0 10  3 10 10
[176]  5 10  5 10 10 10 10 10 10 10 10 10  3 10 10 10  5 10 10 10  5 10  3 10  0
[201] 10  0  3 10 10  5 10  0 10 10 10 10 10  3 10 10 10 10 10  0 10 10 10  0 10
[226] 10 10  3  0 10  5 10 10  3  0 10  0 10  3  0  5 10  5  0 10 10  0 10 10  3
[251]  0 10 10 10 10  3 10  0 10 10 10  0  0 10 10 10 10  5  3 10  5 10 10 10 10
[276] 10 10 10 10 10  3  0 10  5 10 10  0  0 10 10 10 10 10 10 10 10 10 10  5 10
[301]  3  3  0 10  0 10 10  3  0 10 10 10 10  0 10  5 10  0 10  5  0  5 10 10 10
[326]  0 10 10  3 10  0  5  0 10 10  0  3 10  0 10  3  0  3  3  5 10  0 10 10  0
[351] 10 10  5 10  0 10  5  0  3  5 10  3 10  5  5  0 10  5  5 10 10  3  3  5 10
[376]  5 10  0  3 10 10 10 10  5  0 10 10 10  0 10  0 10 10 10  0 10 10  3  5  0
[401] 10  0  3 10 10  0  0  0 10 10 10 10  3 10 10 10 10 10 10 10  3 10  3  0 10
[426] 10  5 10 10  0 10 10  5 10 10 10  0 10 10  5 10 10 10  3 10 10  0 10  5 10
[451] 10  3 10 10  3 10 10  0  3  5 10  5  5  0  0 10 10  0 10  3 10  5 10  0 10
[476] 10  5 10 10  0 10  3 10  0 10  5 10 10  3 10  0 10  5 10  0  0 10 10 10  0
```

3.Draw a histogram of the 500 simulated values. Compare the histogram to the specified probability distribution. Does it meet our expectation?

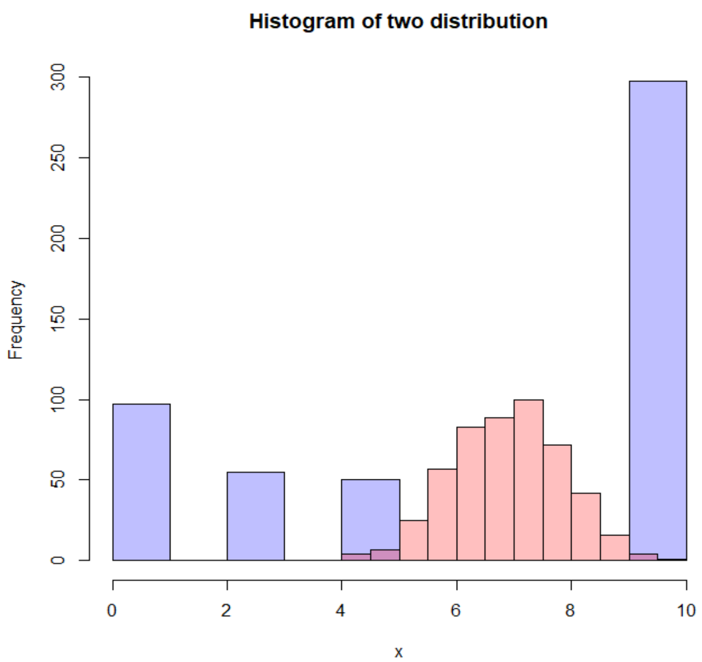
hist (data, xlab = "x", main="Histogram of 500 simulated values")



p1<-hist (data, xlab = "x", main="Histogram of 500 simulated values")
p2<-hist (rnorm (500, mean = 6.8, sd = 1))

plot (p1, col=rgb (0,0,1,1/4), xlim=c (0,10), xlab="x", main="Histogram of two distribution")

plot (p2, col=rgb (1,0,0,1/4), xlim=c (0,10), add=T)



Draw the histogram of these 500 simulated values distributions and a normal distribution (mean=6.8, sd=1). Compared to the normal distribution, the peak value of this distribution is not at the mean 6.8 but at x=10.

The distribution of these 500 values is discrete, and the frequency at which each value appears obeys its probability. The frequency of each value (0,3,5,10) is close to 2:1:1:6, so I consider it meets my previous expectation.

4.Use variance of estimator $\bar{X}$ given by Var ($\bar{X}$) =σ^2/n. Use R to compute this variance.

var_xbar=varx/4

```
> var_xbar
[1] 4.29
```

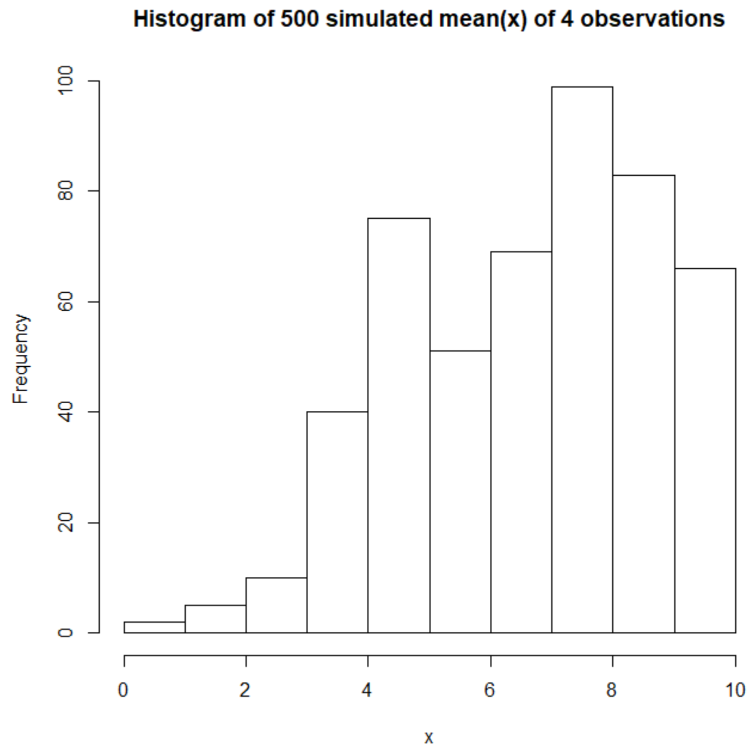
5. Generate 500 simulated values of $\bar{x}$ where $\bar{x}$ is the mean value of 4 observations drawn from the specified discrete distribution.

simu_xbar= apply (matrix (sample (x, size = 4 * 500, replace = TRUE, prob = px), 4), 2, mean)

```
> simu_xbar
  [1]  7.50  7.50 10.00 10.00  6.25  7.00  5.00  5.75  5.00 10.00  6.25  7.50
 [13]  5.75  5.75  3.25  8.25 10.00  7.50  6.25 10.00  7.50  4.00 10.00  8.75
 [25] 10.00  7.50 10.00  8.25  3.75  5.00  5.00  7.50  3.25  5.75  6.25  7.00
 [37]  0.00  3.25  3.25  5.00  3.75  3.75  5.00  5.00  8.25  8.75 10.00  7.50
 [49]  7.50  7.50  3.75  4.50  5.00  6.25  7.50 10.00  7.50  8.25  7.50 10.00
 [61]  2.50  8.75  6.25  8.25  3.25  7.00  7.50  3.75  2.50  6.25  5.00  5.00
 [73] 10.00  5.00  7.50  6.25  7.00  6.50  3.75  7.00  5.00  8.25 10.00  5.00
 [85]  5.00  6.25  6.50  7.50  5.75  6.25  7.50  7.50  3.25  7.50  8.25 10.00
 [97] 10.00  3.75 10.00  7.00  7.50  7.50  5.00  7.50  7.50  7.50  7.50  5.75
[109]  8.75  5.75  7.50  7.50  3.25  5.00  8.75  7.50  5.25  6.25  5.75  8.25
[121]  7.50  5.75 10.00  3.25  7.50 10.00 10.00  5.75 10.00  6.25  3.75  7.50
[133] 10.00  7.50  5.75  5.00  5.00  2.50  3.75  8.25 10.00  7.50  8.25  8.75
[145]  7.50  6.25  2.50  6.25  7.50  5.75  8.25  3.25  8.75  7.50  7.50 10.00
[157]  8.25  6.50  8.75  5.00  5.75  5.75  8.25  6.25  5.00  7.50  7.00  7.00
[169]  3.75  7.50  5.00  7.50  4.50  7.00  7.50  8.25  4.00  7.50  8.75  5.75
[181]  7.50  5.75  8.25  7.00  8.75  5.75  5.25  8.75  7.50  5.75  6.25  5.75
[193]  5.75  7.50  7.50  1.25  7.00  5.00  5.00  5.75  5.00  7.50  5.00  5.00
[205] 10.00  8.75  2.50  2.50  3.25  3.25  7.50 10.00  8.25  8.75  7.50 10.00
[217]  7.50  8.25  7.50  7.50  5.00 10.00  5.00  6.25  8.25  7.50  5.00 10.00
[229]  7.50  8.25  8.25  3.75  8.25 10.00  7.50  7.50  6.25  7.50  5.75  6.50
[241]  7.50 10.00  8.75  5.00  6.50  5.00  7.50  3.75  6.25  5.00  5.00 10.00
[253]  7.50  5.00 10.00  6.25  4.50  8.25  6.25  7.50 10.00  8.25  3.75 10.00
[265]  7.50  4.00  4.75  5.75  5.00  4.50  5.00  7.00  5.00  5.75  5.00  5.00
[277]  6.25  7.50  6.50  8.75  5.75  7.50  2.50  7.50  4.75  8.25  6.25  7.00
[289]  7.50  5.00  5.00  5.00  7.00  5.75  6.25  8.25  8.25  5.75  5.75  5.75
[301]  2.50  7.50  4.50  7.50 10.00  5.00  5.75  5.00  8.25  2.50 10.00  8.75
[313]  3.75  5.75  5.25  8.75  5.00  3.25  5.00  3.75 10.00 10.00  2.50  5.00
[325]  6.25  7.50  5.75 10.00  4.50  8.75  8.25  7.00  6.25  8.25 10.00 10.00
[337]  4.50  5.75  5.25  7.50  7.50  7.00  5.00  7.50  5.75  5.75 10.00  4.50
[349] 10.00  8.25  8.25  5.00  8.75 10.00  7.50 10.00  5.00 10.00  8.75  4.50
[361]  8.75  7.50  8.75  4.00 10.00  5.00  8.75  5.75  7.50  7.50  8.75 10.00
[373]  6.50  5.75 10.00  8.25  3.75  7.00 10.00  2.00  1.25  4.50  5.00  5.00
[385]  8.75  8.75  8.25  5.75  6.25  8.75  3.25  7.50  8.75  6.50  6.25  5.00
[397]  8.75  8.25 10.00  3.25 10.00  8.75  8.25  7.00  5.25 10.00 10.00  5.75
[409]  8.75  3.75  8.25  5.75 10.00  6.25  6.25  8.75  7.50  8.25  6.25  5.00
[421] 10.00  6.25  5.00  7.50  7.50  3.75 10.00 10.00  4.50  8.75  8.75  7.50
[433]  7.50  6.50  8.25  8.75  6.25 10.00  5.00  8.75  5.00  4.00  5.00 10.00
[445]  7.50  1.25  7.50  6.25  5.75  5.00 10.00  7.00 10.00  7.50  7.50  7.50
[457]  7.50  5.00  7.50  8.75  6.50  8.25  8.75  8.75  7.50  5.75 10.00  3.75
[469]  7.50  5.00 10.00  1.25  5.25  8.25  7.50  7.50 10.00 10.00  4.00  3.75
[481]  7.50  8.25  8.75  6.25  7.50  8.75  6.25  7.50  7.50  8.25  5.75  7.50
[493]  7.00  6.25  6.50  7.50  5.75  6.25  0.75  5.75
```

6. Visualize the results from (5) in a histogram and interpret the results with respect to the following: What can you say about the variance of the 500 simulated means? Are they close to the theoretical value?

hist (simu_xbar, xlab="x", main="Histogram of 500 simulated mean(x) of 4 observations")



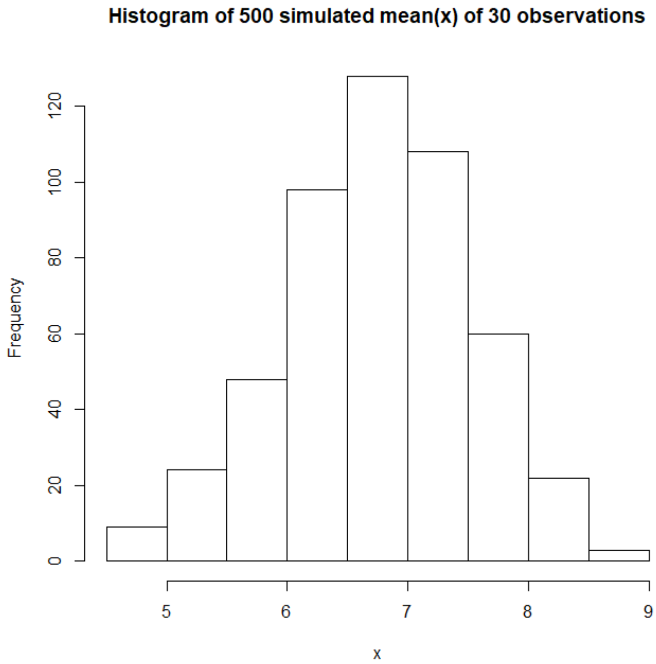

```
> var(simu_xbar)
[1] 4.282107
```

Absolute Error=0.007893.

The variance of the mean of 500 samples is approximately one quarter of the population variance. The variance of the 500 simulated means is very close to the theoretical value with only 0.007893 absolute error.

7. Repeat step (4) and (5) but use averages over 16 observations instead of 4. Compare and comment on the results with respect to the following: unbiasedness, consistency and the central limit theorem.

simu_xbar 2 = apply (matrix (sample (x, size = 30 * 500, replace = TRUE, prob = px), 30), 2, mean) hist (simu_xbar 2, xlab="x", main="Histogram of 500 simulated mean(x) of 30 observations")




```
> mean(simu_xbar2)
[1] 6.772533
> var(simu_xbar2)
[1] 0.5963093
```

Use 30 observations instead of 4. With the increase of our observations (sample size), the 'central' value is closer to the true mean of 6.8. We can see that, on average, the sample average is equal to the population mean. The expected value of the sample average is equal to the population mean. In other words, the sample average is an unbiased estimator of the population mean.

The spread of sample averages is smaller than the spread of the total population (Var ($\bar{X}$) =σ^2/n, here n=30). Compared to 4 observations, we note that the values are less spread out. This means that we have better consistency than 4 observations' samples. It shows us that the spread tends to decrease as the sample size increases.

Moreover, here you can see that the shape of the distribution of 500 simulated means of 30 observations is approximately normal. This is a clear example of the central limit theorem. Regardless of the distribution of the original population, when the sample size extracted from the population is large enough, the sample distribution of the sample mean approximates a normal distribution.