MET CS544 A1 Foundation of Analytics
Homework#5

Problem#1

Central Limit Theorem

The input data consists of the sequence 1 to 25. Show the following three plots in a single row.

a) Show the histogram of the densities of this distribution.
b) Using all samples of this data of size 2, show the histogram of the densities of the sample means.
c) Using all samples of this data of size 5, show the histogram of the densities of the sample means.
d) Compare of means and standard deviations of the above three distributions.

Solution:

a) The data in the input is shown here

```
> input
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
>
```

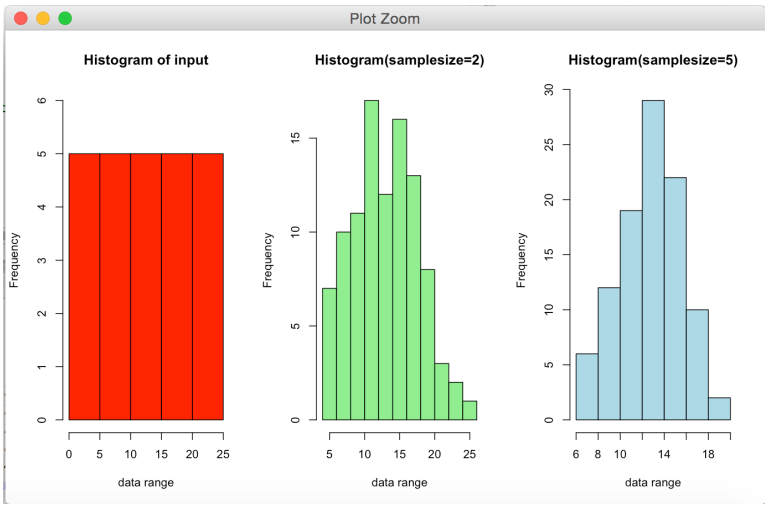
b) With the given sample size of 2, 100 samples are taken and their mean is calculated

```
> x1.mean
  [1] 11.5  5.5 22.5  8.5 15.5 14.5 17.5 13.0 11.5 11.0 14.5 16.5 15.0 18.5 17.5 12.0  6.0 18.0 19.0 14.0 12.0 13.0
 [23] 13.0 10.0  4.5 16.5 20.0 12.0 14.5 14.0  7.5  6.5 11.0 19.5 24.5  7.5 16.5 15.5 17.0  6.5 14.5  8.5  7.5  9.0
 [45] 16.5 11.0 16.0 21.0  9.0 17.5 13.5 10.5  7.5 10.0  8.0 19.5 16.5  8.5 18.0 10.5  4.5 15.5 14.5 12.5 15.0 12.0
 [67] 15.5 10.5 15.5 20.0 13.5 21.5 13.5 12.0 11.0  8.5 17.5 19.5 13.0 15.5  9.0 14.0  6.0  6.5 15.0 21.5 10.5 19.5
 [89] 10.5 15.0 11.0  4.0  9.5 18.0  7.0  9.5 22.5  5.5  6.5 12.5
```

c) With the given sample size of 5, 100 samples are taken and their mean is calculated

```
> x2.mean
  [1] 17.0  9.4 16.4  9.4  8.4 16.0 16.2 13.2 12.4 11.4 13.2 14.2 13.8 16.0  6.6 13.0 18.2 14.0 12.2 13.4 10.8 12.0
 [23] 15.2 12.8 12.6 15.2 13.6 14.2  9.8 10.6 11.8 12.2 12.2 11.0 13.8 15.8 14.6 14.8  7.6 13.8 13.0 12.8 16.0  8.8
 [45]  8.2 10.8 17.4  9.4 12.8 16.0 12.2  8.0 19.0 12.6 15.2 14.6 16.2 15.2 15.0 17.8 16.2 13.0 11.2 13.2 12.2 16.8
 [67] 14.6 12.2 10.2  9.8 12.0 17.8 15.0  7.8 11.6 11.8  7.8 11.8  7.8 10.0  8.4  9.0 14.0 15.8 10.6 12.6 15.4 15.0
 [89]  8.4 10.2 12.4 11.6 12.8 13.6 17.2 10.6 12.0 15.8 10.4 14.4
```

The plot of the histograms is given here



d) The means and the standard deviations of the above three cases are being discussed here

```
> x.mean<-c(mean(input), mean(x1.mean), mean(x2.mean))
> x.mean
[1] 13.00 13.11 12.83
> x.sd<-c(sd(input), sd(x1.mean), sd(x2.mean))
> x.sd
[1] 7.360 4.697 2.795
```

It shows that the no matter what the distribution the data comes from, as we do the average of mean of the samples, the distribution of sample means getting closer to very nearly normal distribution as the sample size is increased.

Problem#2

Central Limit Theorem - Binomial distribution

Suppose the input data follows the negative binomial distribution with the parameters size = 5 and prob = 0.5.

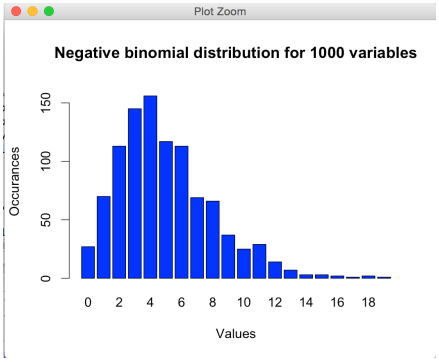
a) Generate 1000 random numbers from this distribution. Show the barplot with the proportions of the distinct values of this distribution.

b) With samples sizes of 10, 20, 30, and 40, generate the data for 5000 samples using the same distribution. Show the histograms of the densities of the sample means. Use a 2 x 2 layout.

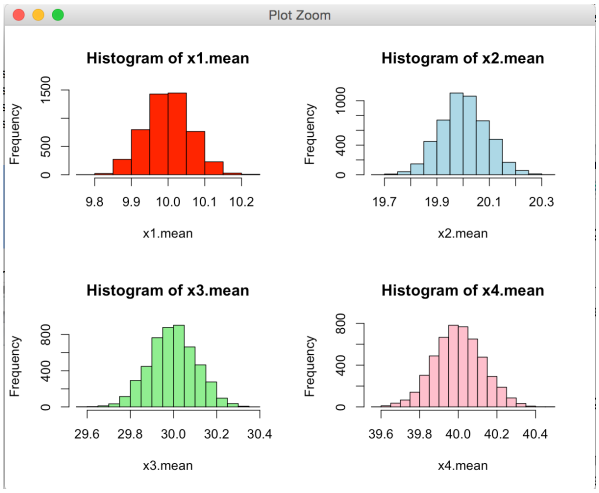
c) Compare of means and standard deviations of the data from a) with the four sequences generated in b).

Solution:

a) 1000 random numbers are picked from the negative binomial distribution and the distribution is plotted



b) The histograms are plotted for the different sample sizes



c) The means and the standard deviations are computed and plotted

```
> xmean.mean<-c(mean(x1.mean),mean(x2.mean),mean(x3.mean),mean(x4.mean))
> xmean.sd<-c(sd(x1.mean),sd(x2.mean),sd(x3.mean),sd(x4.mean))
> xmean.mean
[1]  9.999 20.002 29.997 40.002
> xmean.sd
[1] 0.06225 0.08935 0.10827 0.12650
```

The mean of the data is closer to the sample size of the data and the standard deviation of the increases as the sample size increases.

Problem#3

Sampling

Use the MU284 dataset from the sampling package. Use a sample size of 20 for each of the following.

a) Show the sample drawn using simple random sampling without replacement. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.
b) Show the sample drawn using systematic sampling. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.
c) Calculate the inclusion probabilities using the S82 variable. Using these values, show the sample drawn using systematic sampling. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.
d) Order the data using the REG variable. Draw a stratified sample using proportional sizes based on the REG variable. Show the frequencies for each region (REG). Show the percentages of these with respect to the entire dataset.
e) Compare the means of RMT85 variable for these four samples with the entire data.

Solution:

The proportions of the REG in the original dataset are

```
> x.count<-(input3.reg$Freq/sum(input3.reg$Freq))*100
> names(x.count)<-input3.reg$Var1
> x.count
     1      2      3      4      5      6      7      8
 8.803 16.901 11.268 13.380 19.718 14.437  5.282 10.211
> sum(x.count)
[1] 100
>
```

a) The proportions of the REG using the simple random sampling are

```
> xa.count<-(x1.reg.count$Freq/sum(x1.reg.count$Freq))*100
> names(xa.count)<-x1.reg.count$x1.reg
> xa.count
 1  2  3  4  5  6  7  8
15 20 10 10 25  5  5 10
> sum(xa.count)
[1] 100
```

b) The proportions of the REG using the systematic sampling are

```
> xb.count<-(x2.reg.count$Freq/sum(x2.reg.count$Freq))*100
> names(xb.count)<-x2.reg.count$input3b.sampled.reg
> xb.count
[1]  5.263 15.789 10.526 15.789 21.053 15.789  5.263 10.526
> sum(xb.count)
[1] 100
>
```

c) The proportions of the REG using the systematic sampling with inclusive probabilities are

```
> xc.count<-(x3.reg.count$Freq/sum(x3.reg.count$Freq))*100
> names(xc.count)<-x3.reg.count$Var1
> xc.count
 1  2  3  4  5  6  7  8
10 15 10 15 20 15  5 10
> sum(xc.count)
[1] 100
>
```

d) The proportions of the REG using the stratified sampling are

```
> xd.count<-(x4.reg.count$Freq/sum(x4.reg.count$Freq))*100
> names(xd.count)<-x4.reg.count$Var1
> xd.count
    1     2     3     4     5     6     7     8
 6.25 18.75 12.50 12.50 18.75 12.50  6.25 12.50
> sum(xd.count)
[1] 100
>
```

e) The means of RMT85 variable for these four samples with the entire data are compared

```
> input3.mean<-c(input3a.mean,input3b.mean,input3c.mean,input3d.mean)
> names(input3.mean)<-c(1,2,3,4)
> input3.mean
    1     2     3     4
114.4 209.5 227.6 150.5
>
```

The mean of the attribute is 245.1 before any sampling is done. The means of the RMT85 are compared and it shows the mean calculating through the systematic sampling with inclusive probabilities is closer to the original mean of the attribute.