IST772 Summary Template: Chapter 3 – Probabilities in the Long Run
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Hendi Kushta
Date: 01/29/2023

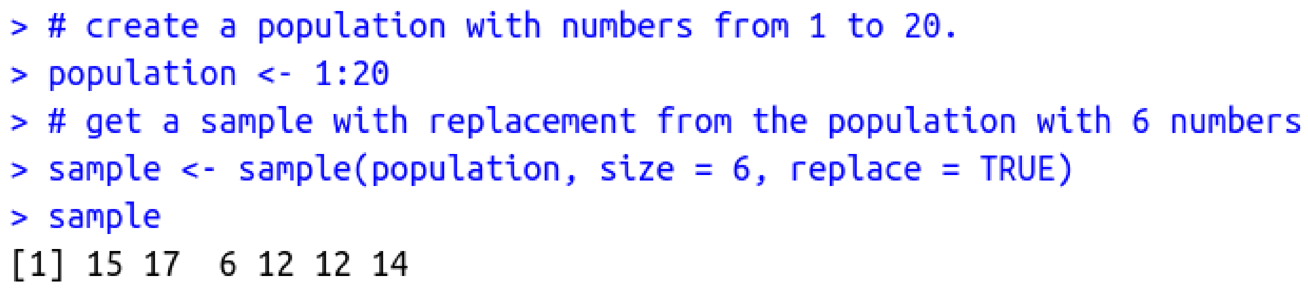
Important **Copying and/or pasting anything from the textbook will not be acceptable for your chapter notes submissions. You must write your notes in your own words and generate your own code, results, and graphs in R. This is what forces your brain to process the material that you read.**

SAMPLING

Population in statistics refers to the total group of people or things that share an interesting trait. The people or things that researchers wish to study or draw conclusions about are the target population. A sample is a subset of the population that is used to investigate and draw conclusions about the population.
Example: If we have a bag with 20 balloons, and there are 10 red and 10 blue, and we want to get 6 balloons from the bag, 20 is the population and 6 is the sample.
N is used for the number of elements in the population → N=20
n is used for the number of elements in the sample → n=6
One of the easiest samplings is known as random sampling with replacement. It states that each person in the population has an equal chance of being chosen, and one person may be chosen more than once. Each participant in this method is picked and added to the sample before being returned to the population for potential selection in subsequent draws. As a result, the sample can have members that are duplicates.

REPETITIOUS SAMPLING WITH R

Sampling is impossible to be performed in real life, because the population size can be huge, and the sampling repetition will be very hard to be conducted.
In R we perform sampling using sample() function.

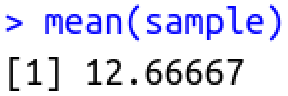


```
> # create a population with numbers from 1 to 20.
> population <- 1:20
> # get a sample with replacement from the population with 6 numbers
> sample <- sample(population, size = 6, replace = TRUE)
> sample
[1] 15 17  6 12 12 14
```

Every time that I will run the sample function the results will be different.

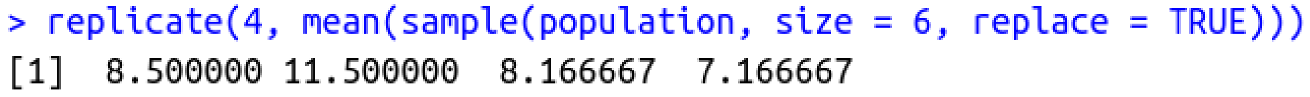
IST772 Summary Template: Chapter 3 – Probabilities in the Long Run
Originality Assertion: By submitting this file you affirm that this writing is your own.

To summarize each of the samples that we choose, we can find the mean of each sample using mean() function.



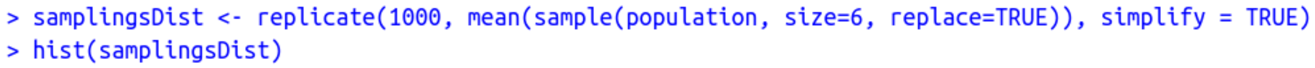
```
> mean(sample)
[1] 12.66667
```

To repeat the process of sampling, there is a function named replicate() in R.

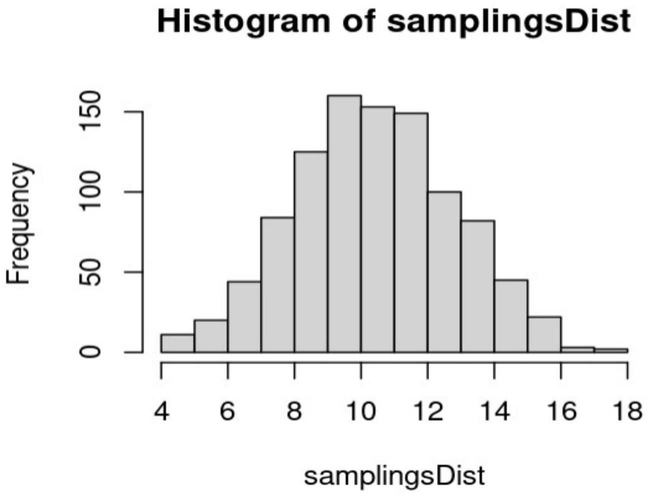


```
> replicate(4, mean(sample(population, size = 6, replace = TRUE)))
[1]  8.500000 11.500000  8.166667  7.166667
```

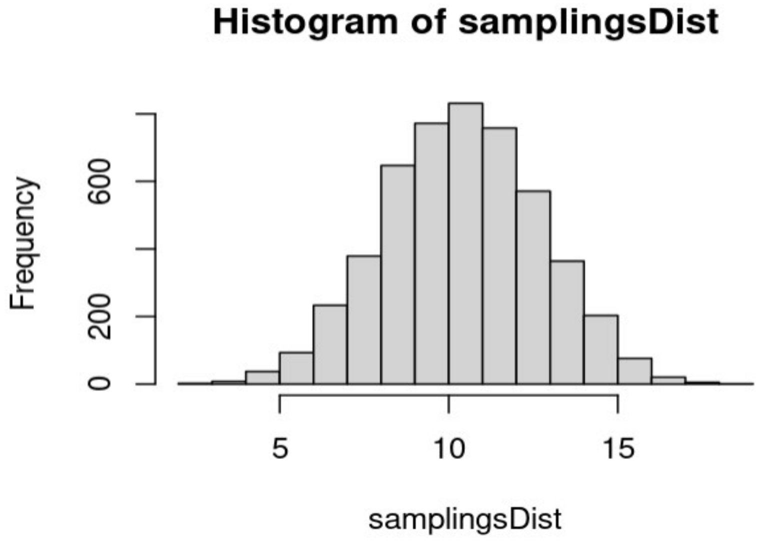
In this line, we have found the mean for 4 different samples because there are 4 replications. The results are different, because every time we run the sample() function, it will select different numbers.
Most of the time, when we deal with huge amount of data, to find the proper mean for the population, there will be needed more than just 4 replications. So below I am showing for 1000 replications, and building the histogram for the distribution of these sampling means.



```
> samplingsDist <- replicate(1000, mean(sample(population, size=6, replace=TRUE)), simplify = TRUE)
> hist(samplingsDist)
```

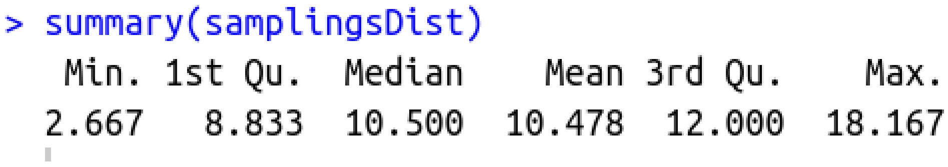


As we see the mean for 1000 replications will be close to 10 and the histogram has a bell shape, normal distribution. And if we repeat the process with more replications, it will be represented even better, for example with 5000 replications.



USING SAMPLING DISTRIBUTIONS AND QUANTILES TO THINK ABOUT PROBABILITIES

To find some statistics about the distribution of the means of the samplings we found with 5000 replications, R has a function named summary()

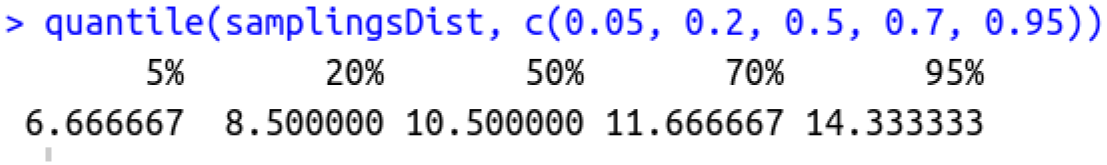


```
> summary(samplingsDist)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.667   8.833  10.500  10.478  12.000  18.167 
```

Min stands for the smallest sampling mean from our 5000 replications in this case is 2.667.
Max stands for the largest sampling mean from our 5000 replications in this case is 18.167.
Median is the number in the middle after ordering the means which is 10.5.
Mean is the average of all 5000 means which is 10.478.
A set of observations is divided into four equal halves using values called quartiles. The distribution of a dataset is frequently summarized in statistics using the quartiles.
The quartiles summarize the size and pattern of the data distribution as a whole. They can also be applied to a dataset to find outliers and skewness.
1st quartile, or 25% of all the means less than or equal to 8.833.
3rd quartile or 25% of all the means have values greater than or equal to 12.
25% of the data falls between 1st quartile and median and last 25% falls between Median and 3rd quartile.

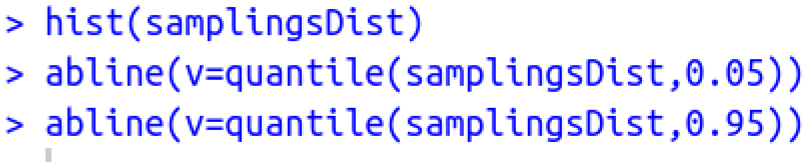
IST772 Summary Template: Chapter 3 – Probabilities in the Long Run
Originality Assertion: By submitting this file you affirm that this writing is your own.

To make better cuts of our distribution, R has a function named quantile() which divide the distribution into portions that we want.

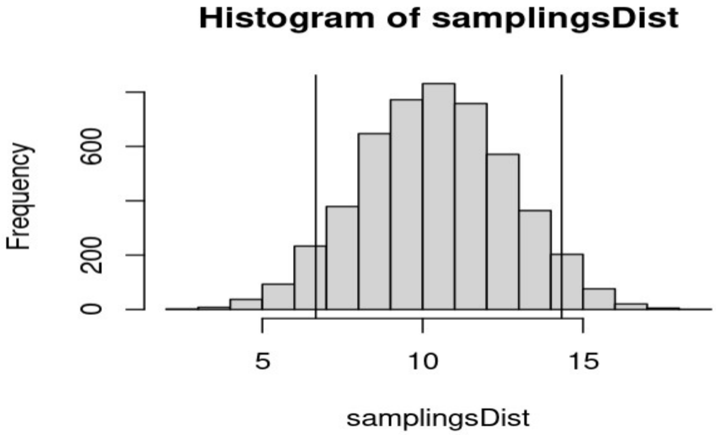


```
> quantile(samplingsDist, c(0.05, 0.2, 0.5, 0.7, 0.95))
       5%       20%       50%       70%       95% 
 6.666667  8.500000 10.500000 11.666667 14.333333 
```

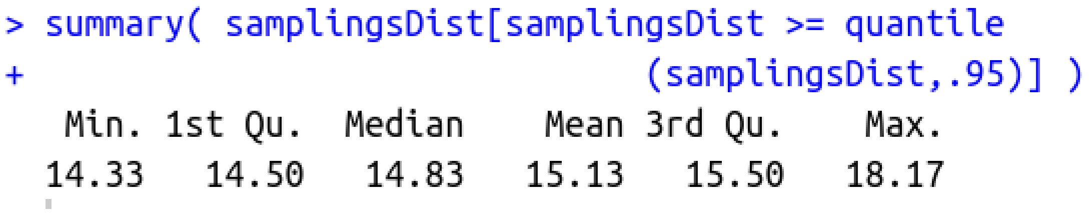
The 5% and 95% values are the limits of the distribution, known also as tails. To draw the lines on the tails, or in any other point that we want, we use the function abline() and v for vertical line.



```
> hist(samplingsDist)
> abline(v=quantile(samplingsDist,0.05))
> abline(v=quantile(samplingsDist,0.95))
```



To go more in depth with the statical values we can find in right tail of the mean distribution, we sample the samplingDist only for values greater or equal to 95%.



```
> summary( samplingsDist[samplingsDist >= quantile
+                                (samplingsDist,.95)] )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  14.33   14.50   14.83   15.13   15.50   18.17 
```