**4.1 Sampling**

Samples: A subset of the total population that is supposed to represent the population. The R syntax to draw samples is:

* sample( dataSetToSample, size = m, replace=TRUE)  
  Here m is the number of times a value is drawn from the sample set and replace is set to TRUE when after every draw the value goes back in the data set with a possibility of it being drawn again.

Forces of randomness drive uncertainty and to decrease the error we do multiple draws or the long run test. It increases the confidence level. In R, functions can be used on the sample set to use descriptive statistics.

To get multiple runs, we replicate the single run multiple times. In R, the syntax for replicating a single run ‘n’ number of times is:

* replicate(n, function(sample( dataSetToSample, size = m, replace=TRUE)), simplify = TRUE)

We can also use a function on the replicate function like this:

* mean(replicate(n, function(sample( dataSetToSample, size = m, replace=TRUE)), simplify = TRUE))

Law of large numbers: If a statistical process is run a large number of times, it converges on a stable result.

Central Limit Theorem: The sampling means start to create a bell-shaped curve and the mean of the means is close to the actual mean of the population.

The sampling information can be stored in a single vector and multiple functions can be then run on the vector.

**4.2 Roundtable on Sampling**

When is it impossible to sample the whole population?

Currently, for the covid-19 vaccine trials, a sample of population is given the vaccine during trials to determine whether the vaccine will be sustainable. In such a case, it is expensive and also unwise to determine the vaccine’s viability by putting the whole population on the trial.

**4.3 Mystery Samples**

We use it to determine whether a sample is part of a given population or not.  
We check the mean of the mystery sample and do quantile analysis of the samples drawn from the population. From the quantile analysis it can be inferred with good statistical evidence, whether the mystery sample is a sample of the population. We can infer whether it is statistically likely for a sample to have the mean presented by the mystery sample. There are “extreme” zones, which are unlikely to occur, hence when a mystery sample falls in those zones it can be tentatively inferred that since such an event is unlikely to occur, the mystery sample is not a part of the population.

Standard deviation of the sample is also the standard error of the mean.

**4.4 Sampling in R**

Some R commands with syntax are as follows:

1. rnorm( n, mean, sd)  
   Creates a normal distribution with population n, mean and standard deviation as the arguments in the rnorm function.
2. library(moments)  
   skewness(data)  
   Gives the value of how much the data is skewed using the package moments.
3. sample( dataSetToSample, size = m, replace=TRUE)  
   Here m is the number of times a value is drawn from the sample set and replace is set to TRUE when after every draw the value goes back in the data set with a possibility of it being drawn again.
4. replicate(n, function(sample( dataSetToSample, size = m, replace=TRUE)), simplify = TRUE)

Where the sampling is done n number of times and the function is called on each sample.

**3.3 Data Science in the Real World**

Insights from coke machines have been integrated in the supply chains for the coke bottle production. Actionable insights are important. Customisation to make the user experience easier to navigate can be done using the analysis. Quality control in a machine can also be done using analysis.

**Questions from the videos:**

* Why is sampling from a population important?
* In real world data science problems, the population is very huge and it takes a lot of computing power to run operations on the entire population. Hence, we choose a sample of the population to run descriptive statistics to save computing power.
* What are some key things to think about when sampling?
* The important thing in sampling is making sure the sample is as closely representative of the entire population as possible. The sample should not be biased and presents the whole picture of the population. It is also important that the sample size is large enough to accommodate the variations in the data set.
* Why is it useful (or when is it useful) to compare two samples?
* It can be used to compare whether the two samples behave similarly or according to the trend we projected them to. It can be used to assess the performance of a machine or a part of a machine. The context is also important here.

**Questions for Professor:**

* How could we ensure that a sample population is representative of the population when the population data is unknown?