# Survey Sampling

## Survey sampling

a technique of estimating population parameters using the sample data. Sample that we are considering is random.

Advantages of Random sampling: Random selection excludes the chance of bias.

Cost of sample collection is less than the complete population enumeration cost.

Sample data will be of higher quality as compared to complete enumerated data. Sample data collection can be monitored, uses less trained staff etc.

Random sampling methodology allows calculation of error due to sampling.

Random sampling methodology allows determination of a sample size in order to obtain a predetermined error level.

Population Parameters The most common parameters are mean ?? and standard deviation ??. Note: These are not random characteristics/attributes but fixed unknown ones.

## Simple Random Sampling

Characteristics and Assumptions: Sampling is done without replacement. The sample data is random therefore sample mean is random.

The analysis of accuracy with which approximation of population mean is done by the sample mean is probabilistic

Simulation to show as the sample size increases the distribution becomes tighter. -> Implementation of CLT

```
setwd("/Users/cheongsookim/Desktop/Stat/Stat135/data")
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.4.4
```

```
HeartAttack <- read_excel("HeartAttack.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/Los_Angeles'
```

```
BloodPressure <- HeartAttack$Resting_Blood_Pressure
```

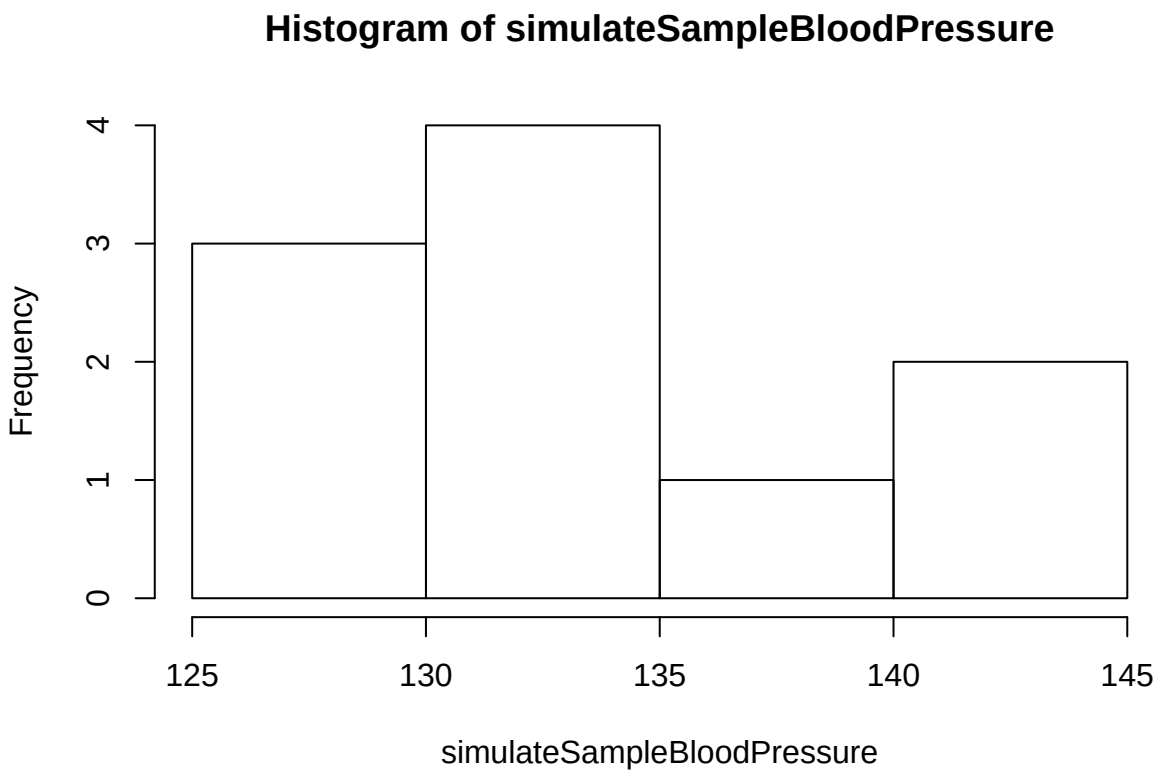
Take a sample of size 8 and find the mean. Repeate this 10 times.

```
simulateSampleBloodPressure <- replicate(10,mean(sample(BloodPressure,8)))
simulateSampleBloodPressure
```

```
##  [1] 140.750 125.250 130.750 127.875 140.125 136.250 134.125 132.500
##  [9] 131.875 129.375
```

Make the histogram of sample means

```
hist(simulateSampleBloodPressure)
```




Calculate the mean of the sample means(Sampling Distribution Mean)

```
mean(simulateSampleBloodPressure)
```

```
## [1] 132.8875
```

Calculate the Standard Error(Standard Deviation of the Sampling Distribution)

```
sd(simulateSampleBloodPressure)
```

```
## [1] 5.039583
```

If the same code is implemented with a sample size of 32,64 and the simulation is done 1000 times we get the following histograms with the following mean and standard errors:

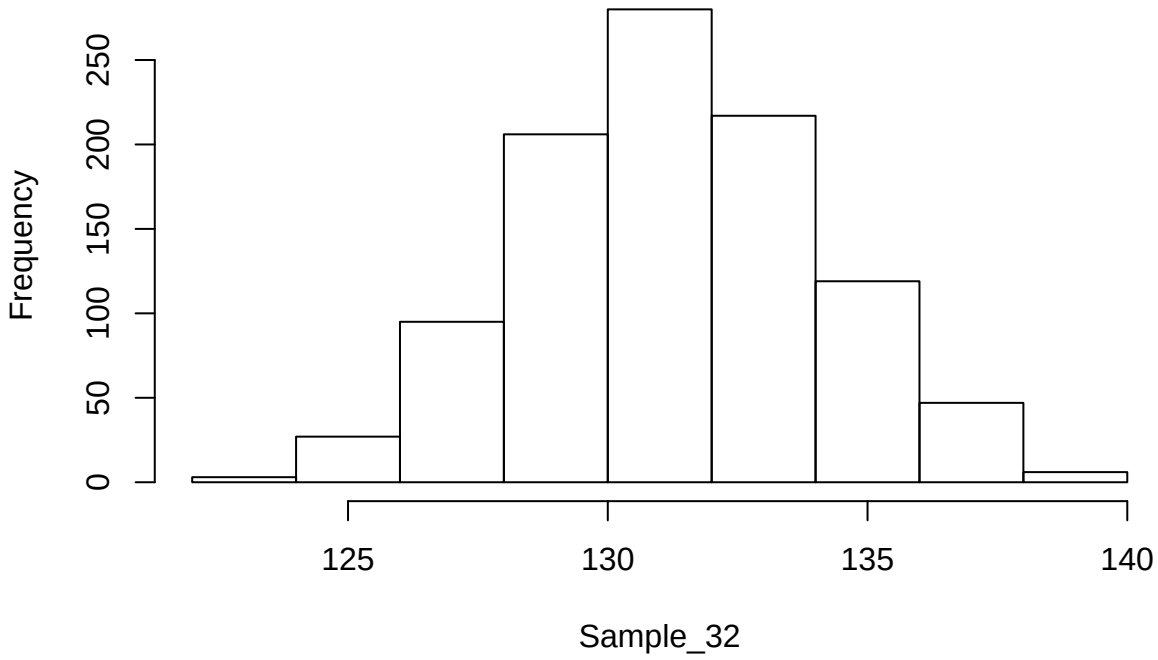
```
Sample_32 <- replicate(1000,mean(sample(BloodPressure,32)))
mean(Sample_32)
```

```
## [1] 131.2821
```

```
sd(Sample_32)
```

```
## [1] 2.851537
```

```
hist(Sample_32)
```



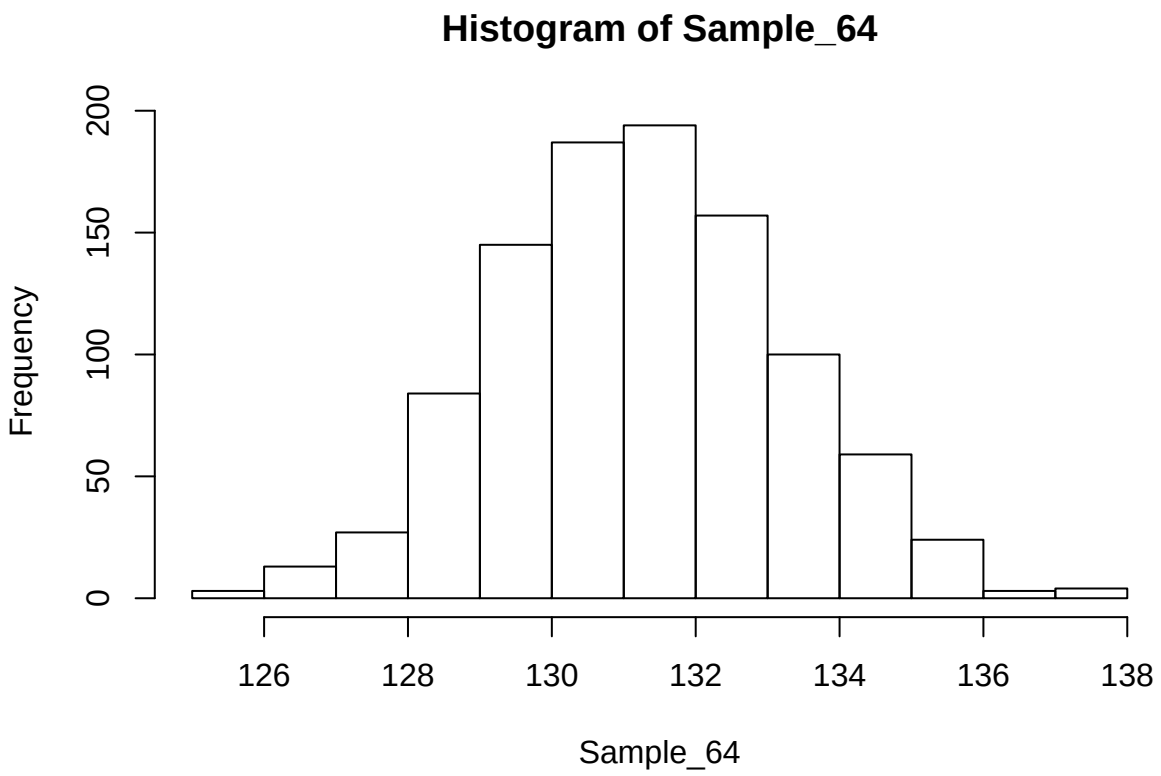

```
Sample_64 <- replicate(1000,mean(sample(BloodPressure,64)))
mean(Sample_64)
```

```
## [1] 131.2605
```

```
sd(Sample_64)
```

```
## [1] 1.985201
```

```
hist(Sample_64)
```




Sample statistics is a unbiased estimator of population parameter. The sample proportion is a unbiased estimator of population proportion. Standard Deviation of sample mean is Standard Error. Standard error is dependent on population standard deviation ?? and sample size n. Standard error is inversely proportional to square root of sample size n.

## ESTIMATION OF POPULATION VARIANCE

Generally we do not know the population standard deviation/variance therefore we cannot compute the standard error. The population variance therefore is estimated using the sample variance.

As the sample size increases the standard error decreases and therefore we get a better estimate of the population mean from this estimated standard error.

We use the Central Limit theorem to approximate the Sampling Distribution with the Normal/Gaussian Distribution.

In the survey sampling scenario (n never approaches infinity as N is fixed) when n is sufficiently large (But small as compared to N) the mean of the simple random sample is approximately normal.

## CONFIDENCE INTERVAL FOR POPULATION PARAMETER

95% Confidence Interval for parameter ?? is a random interval that contains ?? with a probability of .95. If we would take random samples and create Confidence intervals for them then 95% of these would contain ??.

```
# conducting a t test to obtain confidence interval
sampleBloodPressure <- (sample(BloodPressure,50))
t.test(sampleBloodPressure, conf.level = 0.9)$conf.int
```

```
## [1] 127.5481 135.9719
## attr(,"conf.level")
```

```
## [1] 0.9
```

```
#Plotting the Confidence intervals for 100 samples.
numberSamples=100
n=50
#Drawing 100 samples with each sample of size 50.Matrix is of size 100*50.

#Creating a function for generating confidence intervals.

#Applying the function on the 100 samples to find Confidence intervals which get stored in two rows eac

#Summing the number of confidence intervals that include the sample mean in each case.

#Create a plot for the CI s
draws = matrix(rnorm(numberSamples*n,mean(BloodPressure),sd(BloodPressure)), n)

get.conf.int = function(x) t.test(x)$conf.int

conf.int=apply(draws, 2, get.conf.int)

sum(conf.int[1, ] <= mean(BloodPressure) & conf.int[2, ] >= mean(BloodPressure))
```

```
## [1] 94
```

```
plot(range(conf.int), c(0, 1 + numberSamples), type = "n", xlab = "mean tail length",ylab = "sample run

for (i in 1:numberSamples) lines(conf.int[, i], rep(i, 2), lwd = 2)

abline(v = mean(BloodPressure), lwd = 2, lty = 2)
```