CAC 1

Abin Roy, Debolina Chatterjee, Siddhartha Sinha, Swastik Roy

2023-12-11

# AIM: TO FIND SAMPLING DISTRIBUTION AND STANDARD ERROR OF LOAN\_DEFAULT\_PREDICTION.

# IMPORTING THE DATASET:

library(readxl)

## Warning: package 'readxl' was built under R version 4.1.3

CAC1\_DATASET <- read\_excel("C:/Users/DEBOLINA/Downloads/CAC1\_DATASET.xlsx")  
View(CAC1\_DATASET)  
Data <- CAC1\_DATASET

Here we have imported the data set from Microsoft Excel.

# DEFINING THE POPULATION:

population <- Data$`Annual Salary`

Here we have shown the population of our target variable “Annual Salary”.

# MEAN OF THE POPULATION:

mean(population)

## [1] 402203.8

Here we can see the mean of our target variable “Annual Salary” is 402203.8

# SD OF THEPOPULATION:

sd(population)

## [1] 160039.7

Here we can see the sd of our target variable “Annual Salary” is 160039.7.

A standard deviation close to zero indicates that data points are close to the mean. Here,the standard deviation obtained is 160039.7, which indicates that the data points are not so close to mean.

# TAKING SAMPLE SIZE OF 3000

sample\_without <- sample(population,3000,replace = FALSE)  
sample\_with <- sample(population,3000,replace = TRUE)

Here we have taken the sample size of 3000 in both without replacement and with replacement method.

# MEAN OF THE SAMPLE WITHOUT AND WITH REPALCEMENT

mean(sample\_without)

## [1] 402663.9

mean(sample\_with)

## [1] 397105.9

Here we have found the mean of our target variable “Annual Salary” in both without replacement and with replacement procedure.

# SD OF THE SAMPLE WITHOUT AND WITH REPLACEMENT:

sd(sample\_without)

## [1] 161471

sd(sample\_with)

## [1] 161674

Here we have found the sd of our target variable “Annual Salary” in both without replacement and with replacement procedure.

From the 1st sample we have seen that the sample mean varies from the population mean and the standard deviation increases for the 1st sample as compared to the population standard deviation.

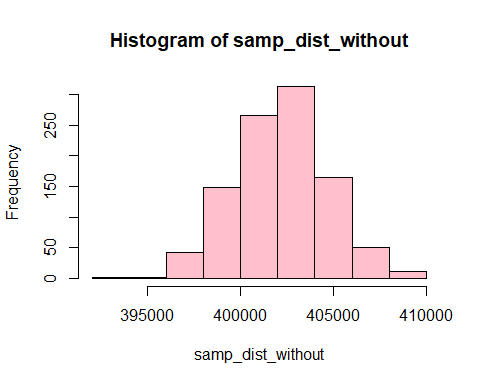
# SAMPLING DISTRIBUTION OF WITHOUT AND WITH REPLACEMENT:

samp\_dist\_without <- replicate(1000,mean(sample(population,3000,replace = FALSE)))  
samp\_dist\_with <- replicate(1000,mean(sample(population,3000,replace = TRUE)))

Here we have the sampling distribution in both without and with replacement procedure in 1000 replications.

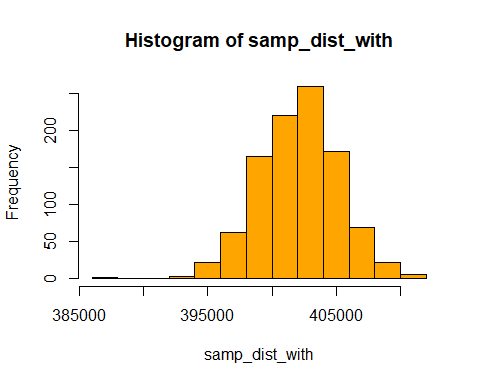
# HISTOGRAM OF SAMPLING DISTRIBUTION WITHOUT REPLACEMENT:

hist(samp\_dist\_without,col = "pink")

 Here we tried to visualize the sampling distribution of without replacement procedure with 1000 replicates in a histogram.

# HISTOGRAM OF SAMPLING DISTRIBUTION WITH REPLACEMENT:

hist(samp\_dist\_with,col ="orange")

 Here we tried to visualize the sampling distribution of with replacement procedure with 1000 replicates in a histogram.

# MEAN OF THE SAMPLING DISTRIBUTION:

mean(samp\_dist\_without)

## [1] 402215.4

mean(samp\_dist\_with)

## [1] 402083.8

# VARIANCE OF THE SAMPLING DISTRIBUTION:

var(samp\_dist\_without)

## [1] 6253759

var(samp\_dist\_with)

## [1] 9187007

Here we have found the variance of sampling distribution of our target variable “Annual Salary” in both without replacement and with replacement procedure.

# SD OF THE SAMPLING DISTRIBUTION:

sd(samp\_dist\_without)

## [1] 2500.752

sd(samp\_dist\_with)

## [1] 3031.008

Here we have found the sd of sampling distribution of our target variable “Annual Salary” in both without replacement and with replacement procedure.

# STANDARD ERROR:

(sd(sample\_without)/sqrt(3000))

## [1] 2948.044

(sd(sample\_with)/sqrt(3000))

## [1] 2951.75

Here we have found the se of our target variable “Annual Salary” in without replacement and with replacement procedure.