Inferential Statistics for Data Science

DS412

Kotalwar Sangamesh

U101115FCS210

Course in-charge: Prof. Suman Sanyal

Contents

1 Acknowledgement

I'm highly indebted to Prof. Suman Sanyal for his guidance and constant supervision as well as for providing necessary information regarding the assignments.

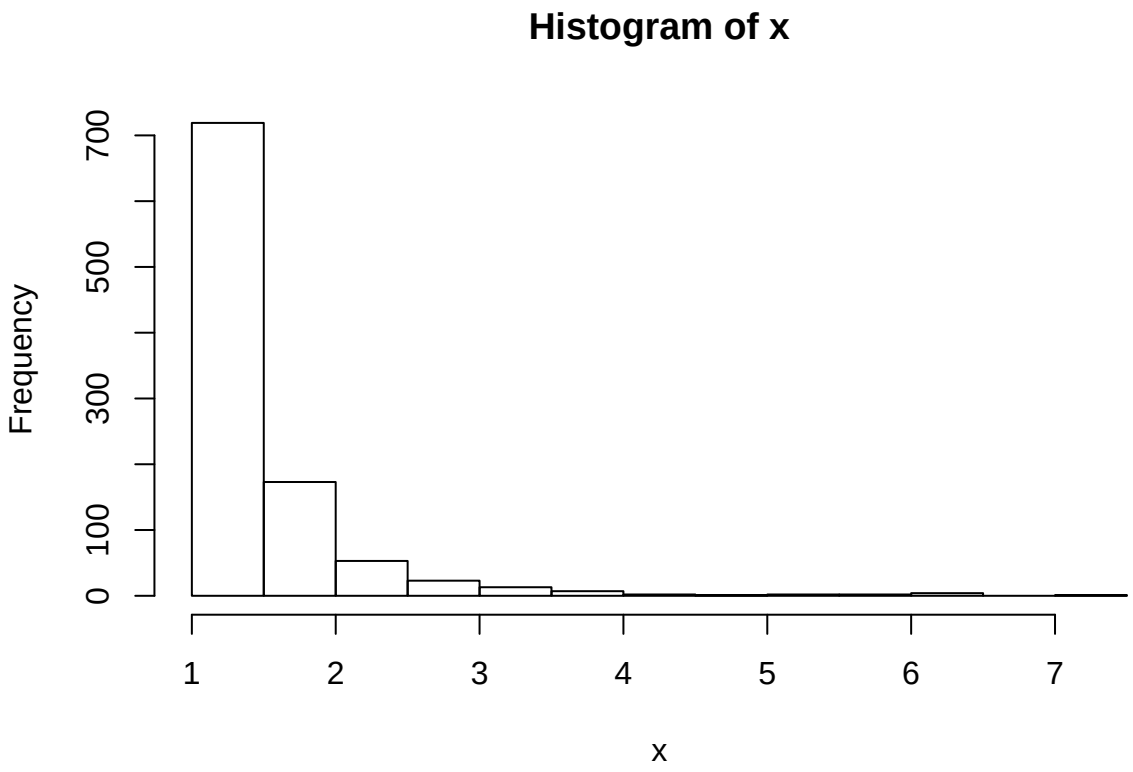
I acknowledge that any work that I submit for assessment at NIIT University:

1. Must be all my own work.
2. Must not have been prepared with the assistance of any other person, except those permitted within University guidelines or the specific assessment guidelines for the piece of work.
3. Has not previously been submitted for assessment at this University or elsewhere.

2 Assignment A

2.1 Q.1] How close is the average of the samples to the expected value of X?

Calculating the Sample mean and draw it's histogram:



Sample mean of X is:

```
## [1] 1.471154
```

Calculating Expected value of the distribution:

Expected value of X is:

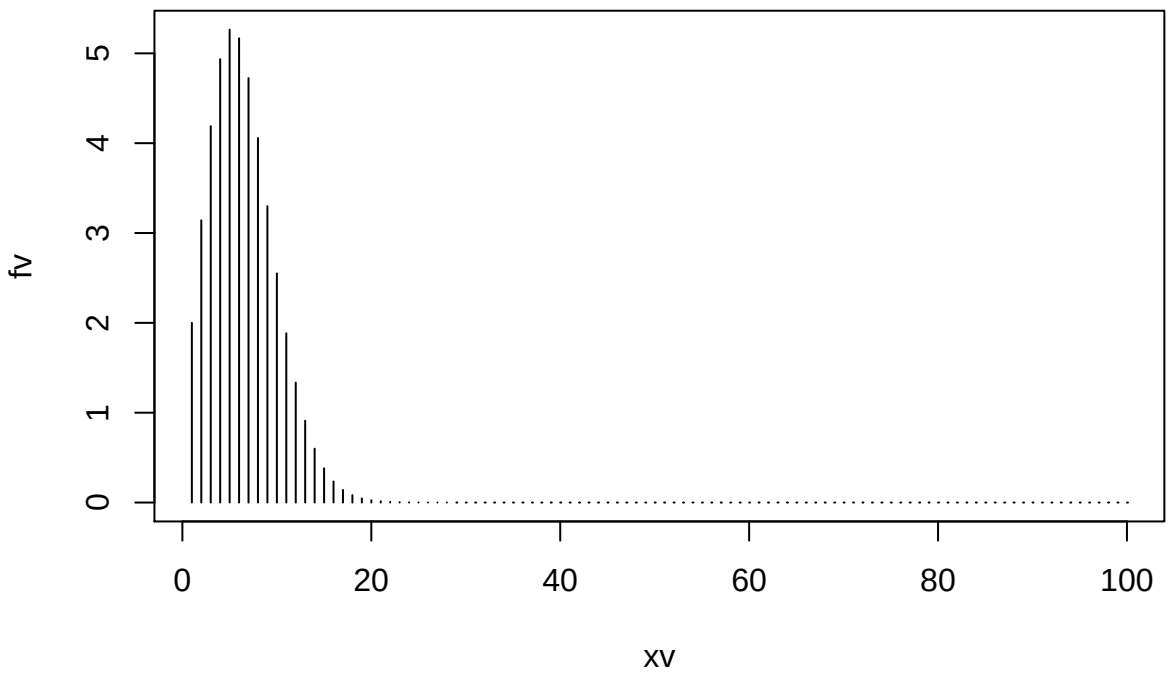
```
## 1.5 with absolute error < 1.7e-14
```

Difference between sample mean and expected value of X is:

```
## [1] -0.0288457
```

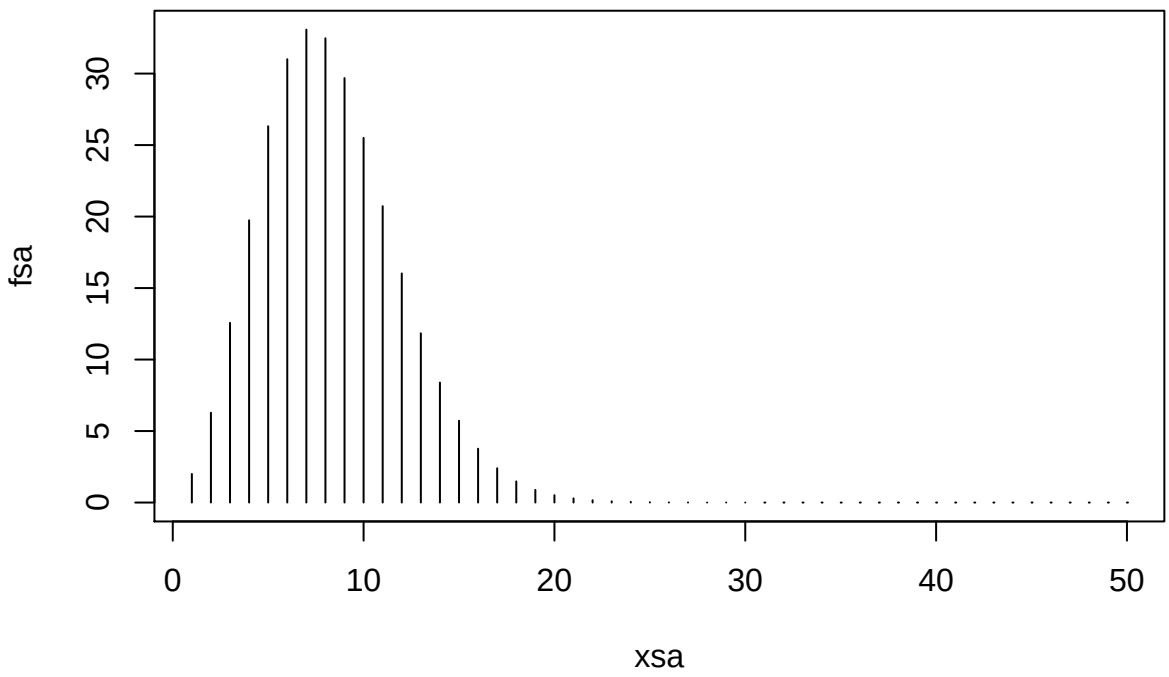
Inference : The difference between the sample mean and expected value is nearly 0.

2.2 Q.2] 2-D The volume of a d-dimensional unit ball



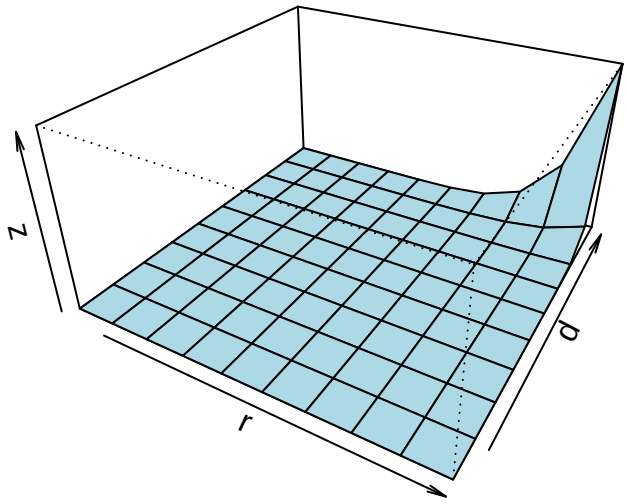
Observation: if the limit as d goes to infinity, the volume of the ball goes to zero.

2.3 Q.3] 2-D The surface area of a d-dimensional unit ball



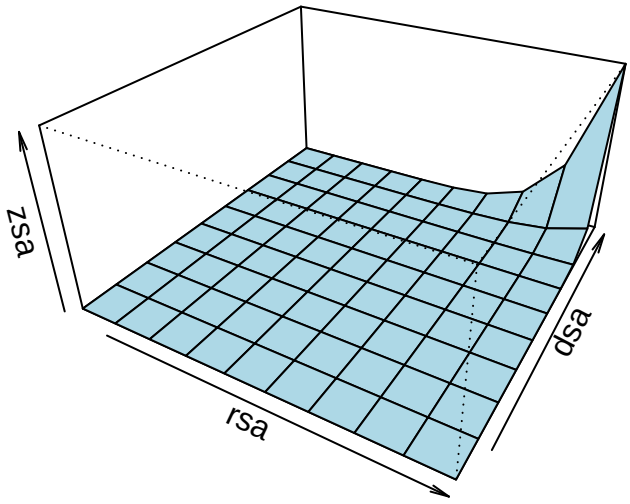
Observation: if the limit as d goes to infinity, the surface area of the ball goes to zero.

2.4 Q.4] 3-D The volume of a d-dimensional unit ball



Observation: if the limit as d and r goes to z, the volume of the ball goes to zero.

2.5 Q.5] 3-D The surface area of a d-dimensional unit ball



Observation: if the limit as d and r goes to infinity, the surface area of the ball goes to zero.

2.6 Q.6] Calculate distance

The differences in euclidean distances for the subspace projection is as follows:

```
## [1] "Subspace of dimension: 1"
## [1] 43.48805
```

```
## [1] "Subspace of dimension: 2"
## [1] 60.15291
```

```
## [1] "Subspace of dimension: 3"
## [1] 68.35696
```

```
## [1] "Subspace of dimension: 4"
## [1] 76.92791
```

```
## [1] "Subspace of dimension: 5"
## [1] 86.46117
```

```
## [1] "Subspace of dimension: 10"
## [1] 109.5695
```

Observation : As we decrease the dimension, the differences in euclidean distances for the subspace projection decreases.

Attached File for code

3 Assignment B

The app can be run on cloud here!

```
library(shiny)
# Define UI for application that draws a histogram
ui <- fluidPage(
  tags$head(
    tags$style(HTML("
                    .mg-histogram .mg-bar rect {
                    fill: #75AADB;
                    shape-rendering: auto;
                    }
                    .mg-histogram .mg-bar  {
                    fill: #75AADB;
                    shape-rendering: auto;
                    }


                    .mg-histogram .mg-bar rect.active {
                    fill: #ffa500;
                    }"))),

  # Sidebar with a slider input for number of bins
  sidebarLayout(
    sidebarPanel(   selectInput("typee","Select the type of distribution",c("Uniform Distribution"="uni
                    selectInput("sele","Replacement",c("Yes"="yess","No"="noo")),

                    sliderInput("numb",
                                "Number of Samples:",
                                min = 20,
                                max = 100,
                                value = 20),
                    sliderInput("unimin",
                                "Enter the Minimum Value for for Uniform Distibution: ",
                                min = 1,
                                max = 99,
                                value = 1),
                    sliderInput("unimax",
                                "Enter the Maximum Value for Uniform Distibution: ",
                                min = 2,
                                max = 100,
                                value = 100),
                    sliderInput("meann",
                                "Enter the Mean for Normal Distibution: ",
                                min = 0,
                                max = 50,
                                value = 10),
                    sliderInput("stddev",
                                "Enter the Standard Deviation for Normal Distibution: ",
                                min = 0,
                                max = 50,
                                value = 10),
```

```
                    sliderInput("exprate",
                                "Enter the Rate for Exponential Distibution: ",
                                min = 1,
                                max = 100,
                                value = 1),
                    tags$div(class="header", checked=NA,
                             tags$p("Check the deployed version here:"),
                             tags$a(href="https://sangamkotalwar.shinyapps.io/Assignmnet_2Final/", "Cli
                    )
    ),


    # Show a plot of the generated distribution
    mainPanel(
      plotOutput("plot",brush = brushOpts(id = "plot_brush"),hover = hoverOpts(id = "plot_hover"))
    )
  )
    )

# Define server logic required to draw a histogram
server <- function(input, output) {
  d <- reactive({

    sele <- switch(input$sele,
                   noo = TRUE,
                   yess = FALSE,
                   TRUE)
    disttt<-switch(input$typee,
                   unii=sample(input$unimin:input$unimax,input$numb,replace = sele),
                   normm=rnorm(input$numb,mean=input$meann,sd=input$stddev),
                   expp=rexp(input$numb,input$exprate),
                   sample(input$unimin:input$unimax,input$numb,replace = sele)
    )

  })
  faa<- reactive({
    disttt<-switch(input$typee,
                   unii="Uniform Distribution of ",
                   normm="Normal Distribution of ",
                   expp="Exponential Distribution of ",
                   "Uniform Distribution of "
    )
  })
  dkk<-reactive({

    if( is.null(input$plot_brush$xmax) && is.null(input$plot_hover$x))
      color="blue"
    else if(!is.null(input$plot_hover$x))
    {
      color=dkkb2()
    }
    else if( !is.null(input$plot_brush$xmax) && is.null(input$plot_hover$x))
    {
```

```
    color=dkkb()
  }

  else color=dkkb()
})

dkkb<-reactive({
  color="blue"
  flag=1
  i=1
  differe =((max(d())-min(d()))/10)
  check=min(d())
  while(i<11)
  {
    if (check>(input$plot_brush$xmax))
    {
      flag=2
    }
    if(((input$plot_brush$xmin-differe)<check) && (flag==1))
    {
      color[[i]]<-"orange"
    }
    else{
      color[[i]]<-"blue"
    }
    i=i+1
    check=check+differe
  }
  check=min(d())


  return(color)
})
dkk2<-reactive({

  if( is.null(input$plot_hover$x) )
  {
    color="blue"
  }
  else
    color=dkkb2()

})
dkkb2<-reactive({
  color=c("blue","blue","blue","blue","blue","blue","blue","blue","blue","blue")
  abcc=(as.integer((input$plot_hover$x-  min(d()))*10  / (max(d())-min(d()) ))+1)
  color[[abcc]]="orange"
  return(color)
  #print(color)
})

output$plot <- renderPlot({
  dist <- input$dist
```

```
    n <- input$numb
    minv=min(d())
    maxv=max(d())
    hist(d(),breaks=seq(minv,maxv,l=11),main = paste(faa(),n, " Random Variables", sep = ""),col = dkk()
  })
}

# Run the application
shinyApp(ui = ui, server = server)


#The app can be run on cloud at https://sangamkotalwar.shinyapps.io/Assignmmet_2Final/
```

The app can be run on cloud here!

4 Assignment C

4.1 Q.1] 1. Perform the following steps and comment on the observation.

4.1.1 Step I. Generate one U(-100,100) random number. Call it m

```
## [1] "m = -47"
```

###Step II. Generate one U(10,50) random number. Call it s.

```
## [1] "s = 20"
```

###Step III. Generate one U(10,25) random number. Call it n.

```
## [1] "n = 13"
```

4.1.2 Step IV. Generate 1000 N(m,s) random numbers. Call this the population.

```
## [1] -59.52908 -43.32713 -63.71257 -15.09438 -40.40984 -63.40937
```

4.1.3 Step V. Sample n numbers without replacement from the population.

```
##  [1] "Head of sample:  -41.5989019812554"
##  [2] "Head of sample:  -41.4417173509891"
##  [3] "Head of sample:  -30.0149922823928"
##  [4] "Head of sample:  -67.6580047716401"
##  [5] "Head of sample:  -38.8119632069813"
##  [6] "Head of sample:  -60.1782399955601"
##  [7] "Head of sample:  -65.6903515831049"
##  [8] "Head of sample:  -46.9516838211305"
##  [9] "Head of sample:  -33.2079955613181"
## [10] "Head of sample:  -47.7848000546634"
## [11] "Head of sample:  -53.6181560136553"
## [12] "Head of sample:  -53.6800168473309"
## [13] "Head of sample:  -42.781854714974"
```

4.1.4 Step VI. Construct 90%, 95%, and 99% confidence intervals for the population mean.

```
## [1] "For 90% interval: "
```

```
##
##  One-sample z-Test
##
## data:  population
```

```
## z = -74.682, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  -48.27326 -46.19267
## sample estimates:
## mean of x
## -47.23296
```

```
## [1] "For 95% interval: "
```

```
##
##  One-sample z-Test
##
## data:  population
## z = -74.682, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -48.47255 -45.99337
## sample estimates:
## mean of x
## -47.23296
```

```
## [1] "For 99% interval: "
```

```
##
##  One-sample z-Test
##
## data:  population
## z = -74.682, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  -48.86206 -45.60387
## sample estimates:
## mean of x
## -47.23296
```

4.1.5 Step VII. Construct 90%, 95%, and 99% confidence intervals for the population variance.

```
## [1] "degree of freedom =  999"
```

```
## [1] "Population variance = 428.420318187078"
```

```
## [1] "For 90% interval: "
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 20
##
## Alternative Hypothesis:          True variance is not equal to 20
##
## Test Name:                       Chi-Squared Test on Variance
```

```
##
## Estimated Parameter(s):          variance = 428.4203
##
## Data:                            population
##
## Test Statistic:                  Chi-Squared = 21399.59
##
## Test Statistic Parameter:        df = 999
##
## P-value:                         0
##
## 90% Confidence Interval:         LCL = 398.6353
##                                  UCL = 461.8795
```

```
## [1] "For 95% interval: "
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 20
##
## Alternative Hypothesis:          True variance is not equal to 20
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 428.4203
##
## Data:                            population
##
## Test Statistic:                  Chi-Squared = 21399.59
##
## Test Statistic Parameter:        df = 999
##
## P-value:                         0
##
## 95% Confidence Interval:         LCL = 393.1989
##                                  UCL = 468.6209
```

```
## [1] "For 99% interval: "
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 20
##
## Alternative Hypothesis:          True variance is not equal to 20
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 428.4203
##
## Data:                            population
##
## Test Statistic:                  Chi-Squared = 21399.59
```

```
##
## Test Statistic Parameter:        df = 999
##
## P-value:                         0
##
## 99% Confidence Interval:         LCL = 382.8567
##                                  UCL = 482.1786
```

4.1.6 Step VIII. Repeat steps V & VI 100/500/1000 times and count the number of times (and percentage) that the population mean is captured by the confidence interval.

```
## [1] "For n=100: Count =  93 , Percentage =  93 %"
## [1] "For n=500: Count =  476 , Percentage =  95.2 %"
## [1] "For n=1000: Count =  954 , Percentage =  95.4 %"
```

4.1.7 Step IX. Repeat steps V & VII 100/500/1000 times and count the number of times (and percentage) that the population variance is captured by the confidence interval.

```
## [1] "For n=100: Count =  98 , Percentage =  98 %"
## [1] "For n=500: Count =  473 , Percentage =  94.6 %"
## [1] "For n=1000: Count =  954 , Percentage =  95.4 %"
```

4.2 Q.2] In a filament cut test, a razor blade was tested six different times with ultimate forces corresponding to 8.5, 13.9, 7.4, 10.3, 15.7, 4.0.

4.2.1 a] find 95% confidence interval on mean using standard t-distribution

```
## [1] "For 95% interval: "
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 mean = 0
##
## Alternative Hypothesis:          True mean is not equal to 0
##
## Test Name:                       One Sample t-test
##
## Estimated Parameter(s):          mean of x = 9.966667
##
## Data:                            forces
```

```
##
## Test Statistic:                  t = 5.666986
##
## Test Statistic Parameter:        df = 5
##
## P-value:                         0.002379959
##
## 95% Confidence Interval:         LCL =  5.445722
##                                  UCL = 14.487611
```

###b] Find a 95% confidence interval on the mean using Efron's percentile method.

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bs, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 6.817, 13.300 )
## Calculations and Intervals on Original Scale
```

4.2.2 c] Find a 95% confidence interval on the mean using the BCa method and the ABC method.

```
## [1] "BCa test"
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bsBCa, conf = 0.95, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 6.633, 13.000 )
## Calculations and Intervals on Original Scale
```

```
## [1] "ABC test"
```

```
## [1]  0.950000  6.863332 13.157697
```

4.2.3 d] Find a 95% confidence interval on the mean using the percentile-t method.

```
## [1] "percentile-t  test:"
```

```
##          2.5%    97.5%
## mean 5.616667 14.31667
```

5 Assignment D

5.1 Q.1]

5.1.1 (a) Estimate an Efron percentile bootstrap 90% confidence interval on the mean aflatoxin residue.

Use B = 1000 resamples

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_obj, conf = 0.9, type = "perc")
##
## Intervals :
## Level     Percentile
## 90%   ( 4.712,  4.959 )
## Calculations and Intervals on Original Scale
```

5.1.2 (b) Compare the alfatoxin level found with the industry average value of 5.7 ppm

Is the upper confidence limit less than 5.7 ppb, or is it equal or above? What does this imply about a hypothesis test of H0 : mu >= 5.7 ppb versus H1 : mu < 5.7 ppb at the alpha = 0.05 significance level?

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_obj, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 4.690,  4.978 )
## Calculations and Intervals on Original Scale
```

5.1.3 (c) Find the P-value for the test in (b)

```
## [1] 6.757152e-30
```

Result:

- Since P-value < 0.05, so reject null hypothesis.

5.2 Q.2]

5.2.1 (a) Find the observed Recall R, Precision P, figure of merit F2.

```
## [1] "Precison:"
```

```
## [1] 0.1644385
## [1] "Recall:"
## [1] 0.82
## [1] "F-score:"
## [1] 0.4562315
```

5.2.2 (b) Resample the 2 × 2 contingency table B = 1000 times. (Hint: Use the multinomialdistribution and rmultinom() in R.)

->

```
Relevant_true <- c(123,27)
Irrelevant_true <- c(625,6703)
total_relevant_doc <- 150
total_irrelevant_doc <- 7328
Relevant_true <- rmultinom(1000, total_relevant_doc, Relevant_true)
Irrelevant_true <- rmultinom(1000, total_irrelevant_doc, Irrelevant_true)
```

5.2.3 (c) Find 90% and 95% confidence intervals for the true F2 for the complete database using Efron's percentile method.

-> The 90% confidence interval for F2 of database is:

```
##        5%       95%
## 0.4258237 0.4855876
```