Analysis of Malignant and Benign Cancer Tumors

Probability Modelling

Yashwanth Kumar Yelamanchili

MS Business Analytics

Department of Operations, Business Analytics & Information Systems

Carl H Lindner School of Business: University of Cincinnati

Abstract:

This Wisconsin based cancer dataset gives us the information of characteristics of 569 cell nuclei obtained from the breast mass through Fine Needle Aspiration procedure. The tumors are classified into malignant and benign. There are 30 features (radius, perimeter, texture, smoothness etc.) describing these tumors.

In this case study, we will deploy the techniques learnt in the Probability class and understand the distribution of smoothness of benign and malignant tumors. Also, we will derive important statistical parameters which will help us understand the smoothness variable much better.

Introduction:

a) Data source

The dataset is taken from Kaggle.

Data Source: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

- Diagnosis- Type of Tumor (Malignant or Benign)
- Smoothness_Mean – For each observation, this column measures the mean of local variation in radius lengths

There are several other features that are available, but this report will focus on the feature – **"Smoothness_mean"** - mean of local variation in radius lengths for each observation

The remaining features available in the data are: area_mean, radius_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst

b) Summary of the data

Summary statistics of the variable "Smoothness Mean" broken down by the type of diagnosis is shown below -

Cell Type	Min smoothness mean	Median	Mean	Max	Count
Malignant	0.07371	0.10220	0.10290	0.14470	212
Benign	0.05263	0.09076	0.09248	0.16340	357

Distribution:

We will take a sample of 300 observations from the entire population for our inferences. Before we proceed, we will compare the distributions of both the sample and population to compare if they are similar.

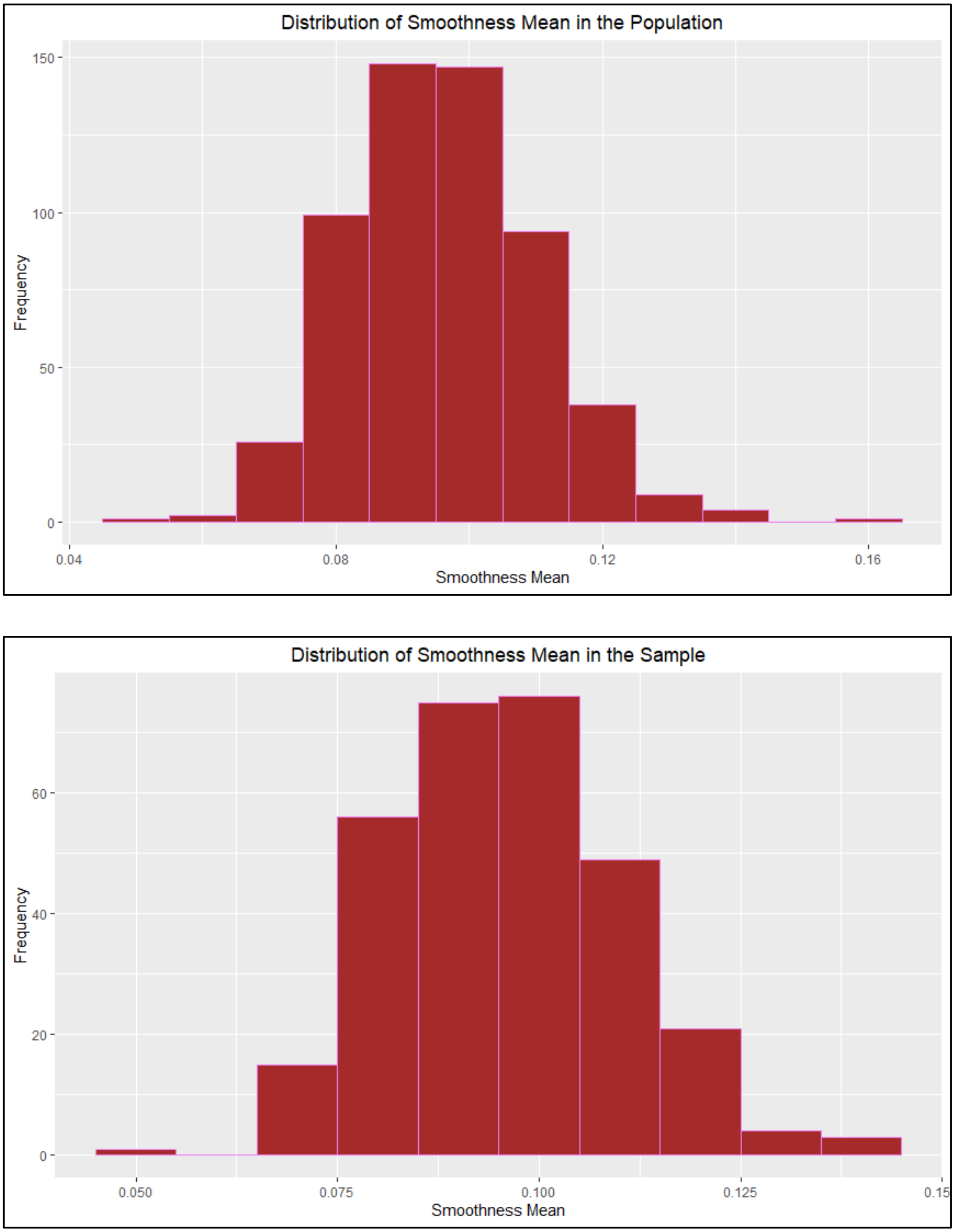


Fig 1.A Distribution of Smoothness of Cancer Cells

The above histograms show that the distribution of the sample is very similar to that of the population. So, we can proceed and estimate the parameters from our sample. Another thing that we can takeaway from the above figures is that they approximately follow a normal distribution.

Data Analysis

a) Empirical Cumulative density function (ECDF)

For the analysis of the smoothness of cancer cells, we will use the ECDF as an estimator for our CDF. We plot this to avoid any data or information loss when we look at the distribution visually.

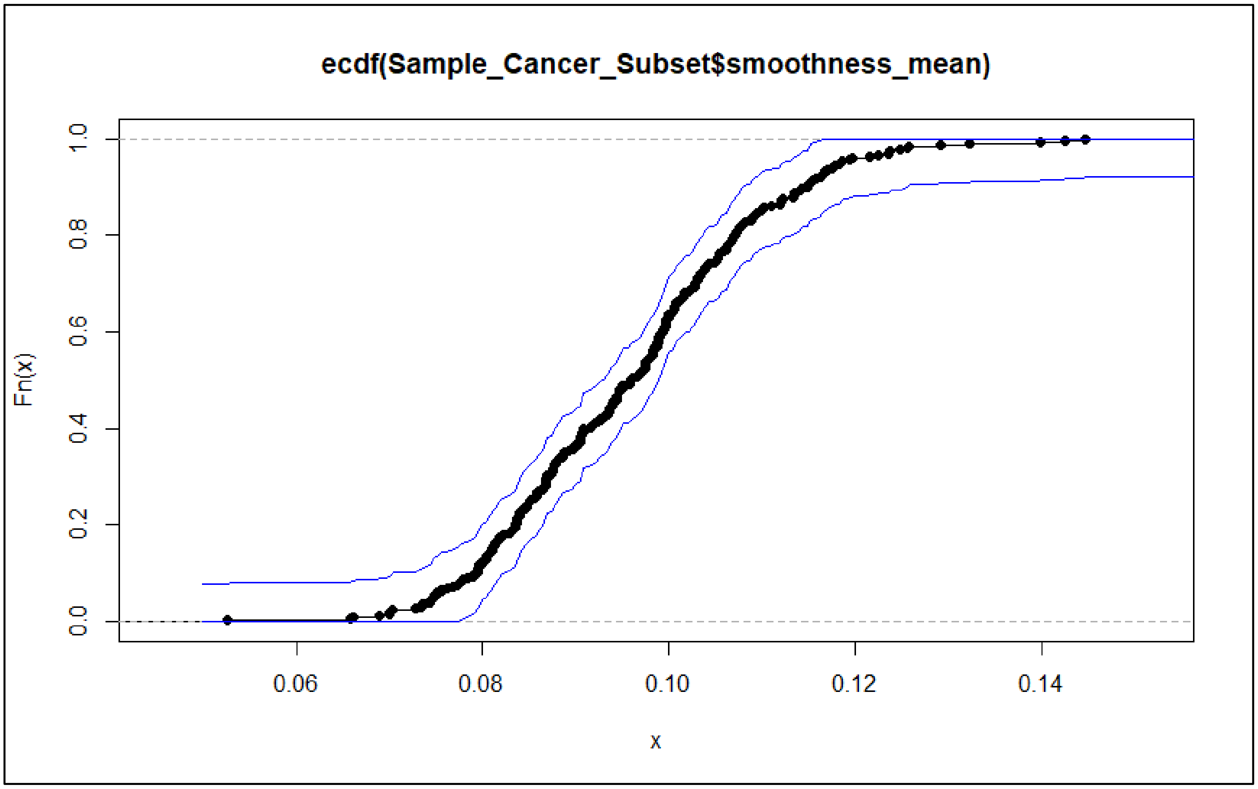


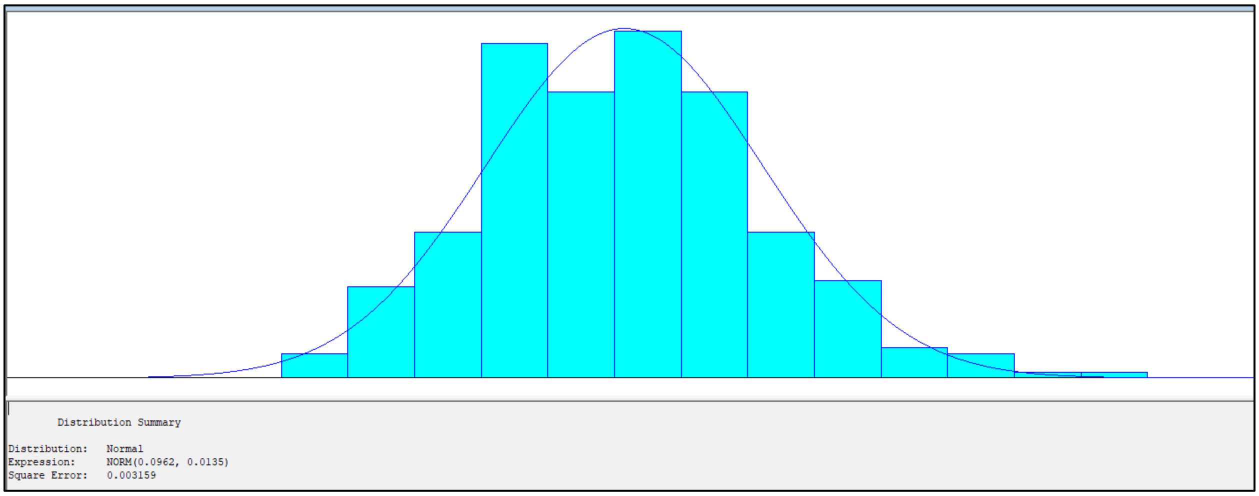
Fig 2. ECDF For smoothness of Cancer Cells

The above fig is the empirical CDF for the smoothness of cancer cells along with a 95% confidence interval band.

The cases of cancer have mean smoothness of tumor generally in the range of 0.078 to 0.118 after which we can again see the datapoints scattered and sparse for this sample.

Using ARENA to estimate the distribution of our sample:

We also used ARENA simulation to double check the distribution and it indeed follows normal distribution. Results shown below –



```
        Distribution Summary

Distribution:   Normal
Expression:     NORM(0.0962, 0.0135)
Square Error:   0.003159
```

Distribution: Normal

Expression: NORM (0.0962, 0.0135)

Square Error: 0.003159

- This distribution also had the least mean square error for the data in this particular sample.

- Summary statistics of the entire column – "Smoothness Mean"

```
> summary(Sample_Cancer_Subset$smoothness_mean)
•     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
•  0.05263 0.08520 0.09598 0.09595 0.10530 0.14470
```

b) Maximum Likelihood Estimator

To estimate the difference in the "Smoothness mean" of the cancer diagnosis – "Malignant" and "Benign", we will use the maximum likelihood estimator.

Parameter of interest = $\mu_{M\,-}\,\mu_{B,}$ where

$\mu_{M\,=}$ Mean smoothness of all the Malignant cancer observations

$\mu_{B=}$ Mean smoothness of all the Benign cancer observations

MLE ($\mu_{M\,-}\,\mu_B$) = MLE (μ_M)-MLE (μ_B).

MLE (μ_M - μ_B) = Sample smoothness mean of Malignant cases - Sample smoothness mean of Benign cases = 0.01104583

c) Bootstrap and Confidence Intervals

Since we are done with the estimation of MLE, we will initially proceed with parametric bootstrap as we are going ahead with the assumption of a normal distribution i.e. the bootstrap method followed when we know the distribution. The mean, standard error and the parametric confidence interval is mentioned below –

Parametric Bootstrapping:

Mean:

```
(mean(theta_hat))
[1] 0.01104001
```

Standard Error:

```
(parametric_se <- sd(theta_hat))
[1] 0.001518125
```

Parametric CI:

```
(parametric_CI <- c(mean(theta_hat)-2*sd(theta_hat),mean(theta_h
at)+2*sd(theta_hat)))
[1] 0.008003757 0.014076259
```

Now, we will perform non – parametric bootstrapping to estimate the confidence interval, mean and standard error to compare the 2 methods and see how different the results will be –

Parametric Bootstrapping:

Mean:

```
(mean(theta_hat2))
[1] 0.01102862
```

Standard Error:

```
(non_parametric_se <- sd(theta_hat2))
[1] 0.001563499
```

Parametric CI:

```
(non_parametric_CI <- c(mean(theta_hat2)-2*sd(theta_hat2),mean(t
heta_hat2)+2*sd(theta_hat2)))
[1] 0.007901626 0.014155621
```

We can clearly see that – both parametric bootstrap and non – parametric bootstrap are providing approximately identical results. This shows the normal assumption for parametric bootstrap is not a wrong assumption.

Now, let us look at the histograms to look at distributions using parametric and non – parametric bootstrapping.

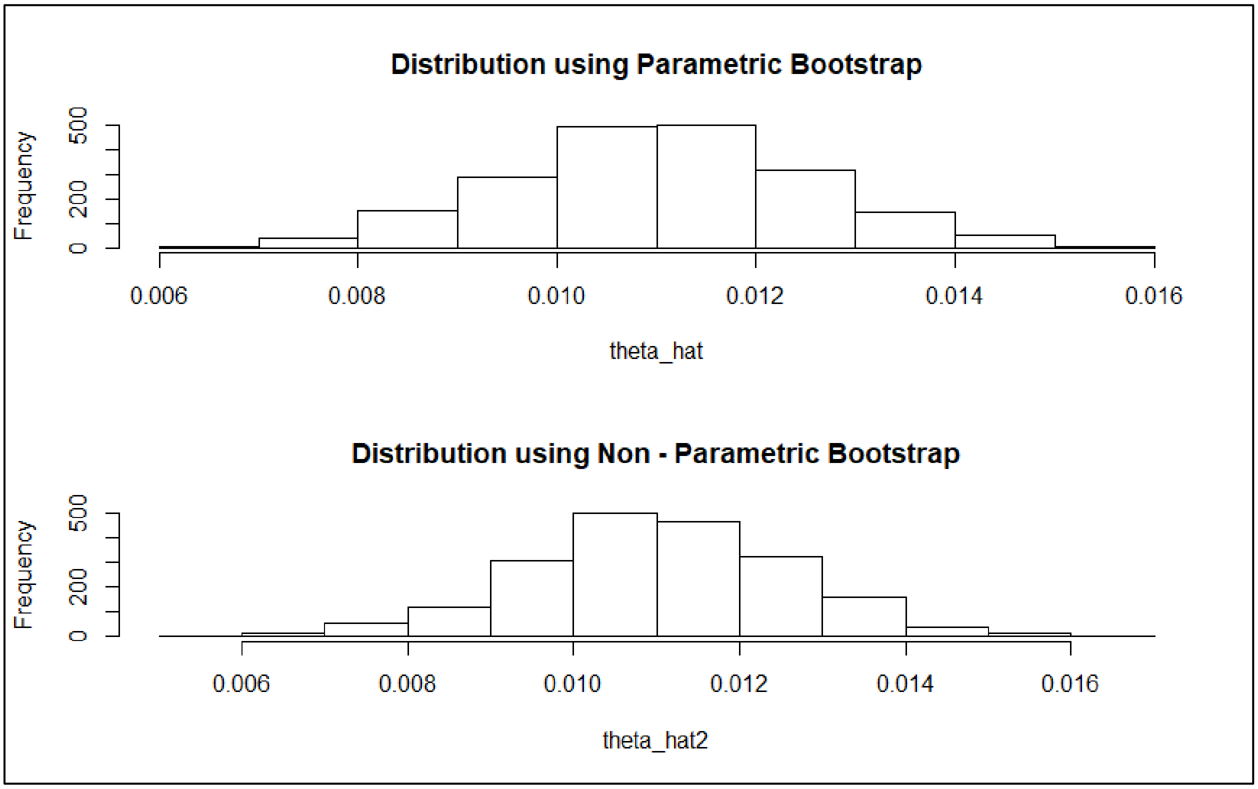


Fig A Distribution of $\widehat{\mu_M} - \widehat{\mu_B}$ using parametric bootstrap

Fig B Distribution of $\widehat{\mu_M} - \widehat{\mu_B}$ using non - parametric bootstrap

We can clearly see an identical distribution using both types of bootstrapping.

d) Hypothesis Test

Given below is the formulation of hypothesis to check if the Setosa and Versicolor have the same sepal widths.

$\mathbf{H_0}$: $\mu_M - \mu_B = 0$ (Mean smoothness of malignant tumor equals mean smoothness of benign tumor)

H_a : $\mu_M - \mu_B \neq 0$ (Mean smoothness of malignant tumor not equals mean smoothness of benign tumor)

Wald test statistic = $\frac{(\widehat{\mu_M} - \widehat{\mu_B} - 0)}{\widehat{s.e}}$ = 7.148396

P-Value = 8.779644e-13

At α = 0.05, since the p-value is less than α, we reject the null hypothesis

Conclusion – Mean smoothness of malignant tumor is significantly different from mean smoothness of benign tumor

Hypothesis Testing (Wilcoxon Rank Sum Test)

Our dataset has the information pertaining to two types of tumors: Malignant and Benign. Now to check if there is a difference in the perimeters of these two tumors, hypothesis testing can be used.

The formulation for our hypothesis is:

H_0 : Median of smoothness of malignant tumor observations = Median of smoothness of benign tumor observations

H_a : Median of perimeters of malignant tumor observations != Median of perimeters of benign tumor observations

The results obtained post performing the Wilcoxon test are:

```
	Wilcoxon rank sum test with continuity correction

data:  Malignant$smoothness_mean and Benign$smoothness_mean
W = 15183, p-value = 3.976e-11
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.007970328 0.014026551
sample estimates:
difference in location
        0.01100833
```

Again, as we see from the results as P value is less than alpha – we reject the null hypothesis and accept the alternative hypothesis that true location shift is not equal to 0 or "Median of perimeters of malignant tumor observations != Median of perimeters of benign tumor observations"

e) Bayesian Analysis

(To be updated)

Conclusion

Appendix

R Code

```
library(dplyr)
library(ggplot2)
Cancer <- read.csv("cancer.csv")
dim(Cancer)
colnames(Cancer)
#View(Cancer)

#Let us focus our analysis on the column - "smoothness mean".
# The distribution will be estimated and we check to see if the mean is same in both Malignant and Benign cancer

#There are a total of 569 observations and we do not have any NA's in the column of interest

#Distribution
#We will consider the distribution for both Malignant and Benign cases
#hist(Cancer$smoothness_mean)

cancer_smoothness_mean_malignant <- Cancer %>%
  filter(diagnosis == "M")
cancer_smoothness_mean_Benign <- Cancer %>%
  filter(diagnosis == "B")

summary(cancer_smoothness_mean_malignant$smoothness_mean)
summary(cancer_smoothness_mean_Benign$smoothness_mean)
```

```
hist(Cancer$smoothness_mean)  #Close to normal - lets check

Sample_Cancer_Subset <- sample_n(Cancer,size = 300,replace = F)

# write.csv(Sample_Cancer_Subset,"sample_subset.csv")



Cancer_subset <- Cancer %>% filter(Cancer$diagnosis == "B")

dim(Cancer_subset)


#Took a sample of 300 observations from the population

#write.csv(Sample_Cancer_Subset,"sample_subset.csv")

#Sample_Cancer_Subset_check <- Sample_Cancer_Subset %>% filter(Sample_Cancer_Subset$diagnosis == "B")


(Sample_Cancer_Subset$smoothness_mean)


#Checking the histogram to see of the distribution of the sample matches with that of the population

par(mfrow = c(2, 1))


ggplot(data = Cancer, aes(smoothness_mean)) +

  geom_histogram(binwidth = 0.01,col = I("violet"), fill = I("brown")) +

  ggtitle("Distribution of Smoothness Mean in the Population") +

  labs(x = "Smoothness Mean", y = "Frequency") +

  theme(plot.title = element_text(hjust = 0.5))


ggplot(data = Sample_Cancer_Subset, aes(smoothness_mean)) +

  geom_histogram(binwidth = 0.01, col = I("violet"), fill = I("brown")) +

  ggtitle("Distribution of Smoothness Mean in the Sample") +

  labs(x = "Smoothness Mean", y = "Frequency") +

  theme(plot.title = element_text(hjust = 0.5))
```

```
# hist(Cancer$smoothness_mean)

# hist(Sample_Cancer_Subset$smoothness_mean)


summary(Sample_Cancer_Subset$smoothness_mean)


#As they look same, we can estimate the parameters of our population from our sample data


#Plotting ECDF

alpha = 0.05

n = length(Sample_Cancer_Subset$smoothness_mean)

eps = sqrt(log(2/alpha)/(2*n))


par(mfrow = c(1,1))

Sample_Cancer_Subset_smoothness_mean.ecdf <- ecdf(Sample_Cancer_Subset$smoothness_mean)

plot(Sample_Cancer_Subset_smoothness_mean.ecdf)

grid <- seq(0.05,0.18, length.out = 300)

lines(grid,pmin(Sample_Cancer_Subset_smoothness_mean.ecdf(grid) + eps,1), col = 'blue')

lines(grid,pmax(Sample_Cancer_Subset_smoothness_mean.ecdf(grid) - eps,0), col = 'blue')


# The cases of cancer has mean smoothness of tumor generally in the range of 0.078 to 0.118

# after which we can again see the datapoints slightly scattered


#MLE Estimate

#Tau = mu1-mu2 (1-Benign) , (2-Malignant)

(mle_est = mean(Sample_Cancer_Subset[Sample_Cancer_Subset$diagnosis == "M",]$smoothness_mean) - mean(Sample_Cancer_Subset[Sample_Cancer_Subset$diagnosis == "B",]$smoothness_mean))

# mle_est = mean(Cancer[Cancer$diagnosis == "M",]$smoothness_mean) - mean(Cancer[Cancer$diagnosis == "B",]$smoothness_mean)


#parametric bootstrap

Malignant <- Sample_Cancer_Subset[Sample_Cancer_Subset$diagnosis == "M",]
```

```
Benign <-Sample_Cancer_Subset[Sample_Cancer_Subset$diagnosis == "B",]
Malignant_mu1.hat = mean(Malignant$smoothness_mean)
Benign_mu2.hat = mean(Benign$smoothness_mean)
n1=length(Malignant$smoothness_mean)
n2=length(Benign$smoothness_mean)
Malignant_sd1.hat = sd(Malignant$smoothness_mean)
Benign_sd2.hat = sd(Benign$smoothness_mean)

theta_hat<-c()
for (i in 1:2000){
  x1<-rnorm(n1,Malignant_mu1.hat,Malignant_sd1.hat)
  x2<-rnorm(n2,Benign_mu2.hat,Benign_sd2.hat)
  theta_hat[i]=mean(x1)-mean(x2)
}

(mean(theta_hat))
(parametric_se <- sd(theta_hat))
(parametric_CI <- c(mean(theta_hat)-2*sd(theta_hat),mean(theta_hat)+2*sd(theta_hat)))

par(mfrow = c(2,1))
hist(theta_hat,main = "Distribution using Parametric Bootstrap")


#Non parametric bootstrap
theta_hat2<-c()
for (i in 1:2000){
  x1<-sample(Malignant$smoothness_mean,n1,replace=T)
  x2<-sample(Benign$smoothness_mean,n2,replace=T)
  theta_hat2[i]=mean(x1)-mean(x2)
}
```

```
(mean(theta_hat2))
(non_parametric_se <- sd(theta_hat2))
(non_parametric_CI <- c(mean(theta_hat2)-2*sd(theta_hat2),mean(theta_hat2)+2*sd(theta_hat2)))
hist(theta_hat2,main = "Distribution using Non - Parametric Bootstrap")


#Hypothesis Testing
# Null hypothesis = mu1=mu2 , Alternate hypothesis = mu1 not equal to mu2
## WALD


(sigma.hat <- sqrt((sd(Malignant$smoothness_mean)^2/n1)+(sd(Benign$smoothness_mean)^2/n2)))




(z.stat<-(mle_est-0)/(sigma.hat))


(p.value=2*(1-pnorm(abs(z.stat))))


#p value is less than 0.05 and hence we reject the null hypothesis


#Wilcox Rank Sum Test


#Wilcoxon Rank Sum Test
wilcox.test(x=Malignant$smoothness_mean , y = Benign$smoothness_mean , conf.int = T)


# As p values is less than alpha -> we reject the null and accept the medians do vary


#Bayesian analysis
#prior of mu1 ~ N(0,1), prior of mu2 ~ N(0,1)
#posterior of mu1 ~ N(n1*xbar/1+n1,1/1+n1), posterior of mu2 ~ N(n2*xbar/1+n2,1/1+n2)


#posterior of mu1
lb.1=1
```

```
Ix.1 = n1/var(Malignant$smoothness_mean)
pos1.mean=((mean(Malignant$smoothness_mean))*Ix.1)/(Ib.1+Ix.1)
pos1.var = 1/(Ib.1+Ix.1)
posterior.mu1 = rnorm(100,pos1.mean,sqrt(pos1.var))

#posterior of mu2
Ib.2=1
Ix.2 = n2/var(Benign$smoothness_mean)
pos2.mean=((mean(Benign$smoothness_mean))*Ix.2)/(Ib.2+Ix.2)
pos2.var = 1/(Ib.2+Ix.2)
posterior.mu2 = rnorm(100,pos2.mean,sqrt(pos2.var))

#posterior of mu1-mu2
pos3.mean = pos1.mean-pos2.mean
pos3.var = posterior.mu1+pos2.var
posterior.mu1.mu2 = rnorm(100,pos3.mean,sqrt(pos3.var))
hist(posterior.mu1.mu2,main="Posterior distribution of mu1-mu2")
mean(posterior.mu1.mu2)

#The mean from Bayesian approach is indeed different from that of obtained from the frequentist approach
```