Chapter 1: Introduction

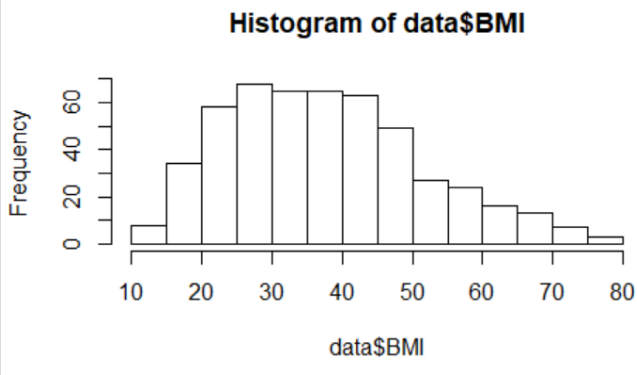
The Kaggle dataset having body mass index was used for analysis.

https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex/kernels

The dataset had 4 columns

1) Gender: Values- Categorical variable with values Male, Female
2) Height – Numerical variable describing the height of individual in cm
3) Weight – Numerical variable describing the weight of individual in kgs
4) Index - Categorical variable describing the condition of a variable
5) BMI - A new numerical variable BMI is created to hold the actual value of the body mass index using the formula weight/(height)^2

Distribution of variable – From below histogram, we see that the distribution of variable BMI is approximately normal.

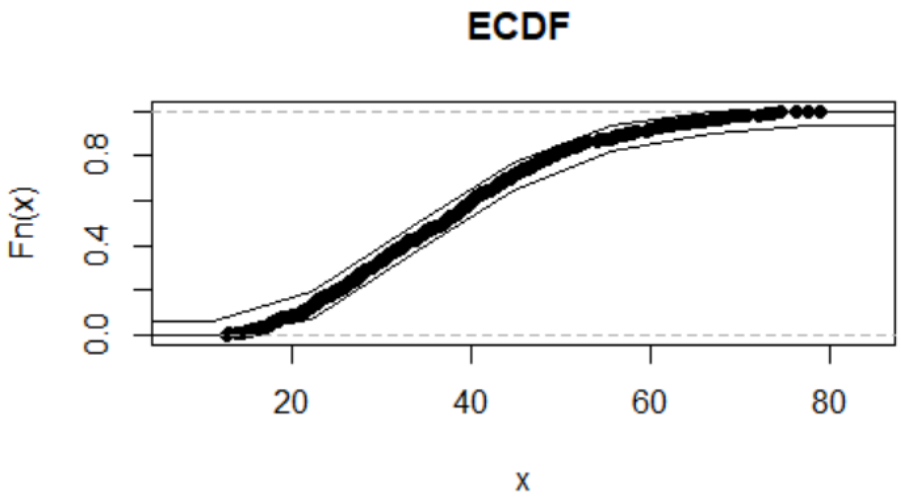


Chapter 2: Empirical Distribution Function

Empirical Distribution function is used in estimating the actual population distribution function. The estimation of distribution function will give a picture of how the is distributed. Here we are trying to estimate the population distribution function of the variable BMI

Code:

```
bmi_ecdf<-ecdf(data$BMI)
plot(bmi_ecdf,main="ECDF ")
Alpha=0.05
n=length(data$BMI)
Eps=sqrt(log(2/Alpha)/(2*n))
grid<-seq(0,100, length.out = 10)
lines(grid, pmin(bmi_ecdf(grid)+Eps,1))
lines(grid, pmax(bmi_ecdf(grid)-Eps,0))
bmi_ecdf(100)-bmi_ecdf(25)
```



Observations:

1) The empirical cdf function of BMI is represented above along with its 95% confidence interval.
2) 80 percent of the people in this dataset are overweight or obese.

Chapter 3: Bootstrap and Confidence Intervals

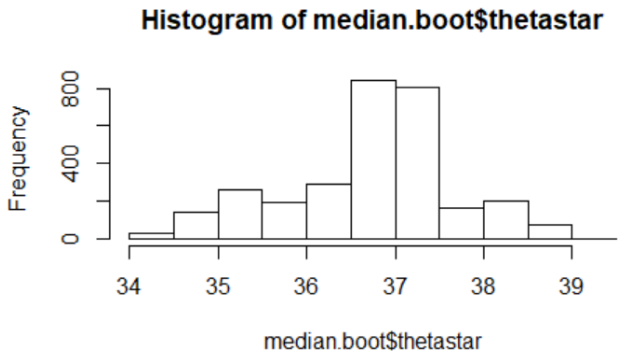
Bootstrap is a test that uses random sampling with replacement. In the context of statistical inference, bootsraping is used for estimating variance of the sampling distribution.

In this analysis, the chosen parameter of interest is median of variable BMI.

Code:

```
n<-length(bmi)
B=3000
library(bootstrap)
median.boot<-bootstrap(bmi,B,median)
hist(median.boot$thetastar)

se_boot <- var(median.boot$thetastar)
theta_hat <- median(bmi)

normal.ci<-c(theta_hat-2*se_boot,theta_hat+2*se_boot)
pivotal.ci<-c(2*theta_hat - quantile(median.boot$thetastar,0.975),2*theta_hat
- quantile(median.boot$thetastar,0.0275))

quantile.ci<-quantile(median.boot$thetastar,c(0.025,0.975))
```

Output:



Observations:

1) The estimated standard error of sampling distribution of median from bootstrap is 0.896
2) The normal confidence interval is [35.165,38.749]
3) The pivotal confidence interval is [35.396,39.076]
4) The quantile confidence interval is [34.8,38.518]
5) The point estimation for the median of the BMI is 36.597
6) Here we can state that 95% of the times, the confidence intervals would consist the actual population median of BMI.

Chapter 4: MLE and its asymptotic distributions

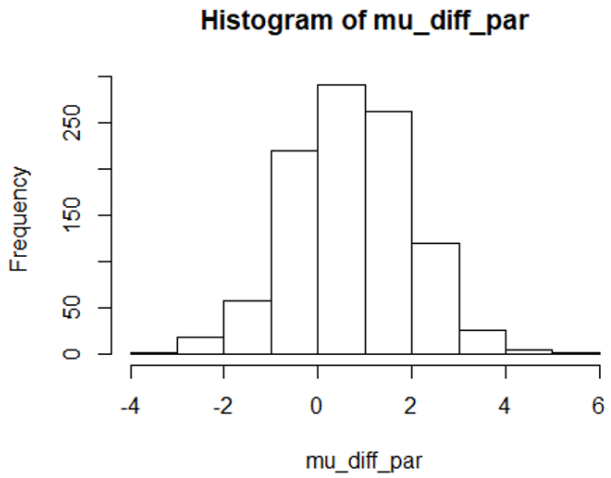
In statistics, maximum likelihood estimation is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

Now there are two gender here in the data. We calculated the mean BMI of each gender and researched if it differs for male and female. #MLE of mu1 - mu2

Code:

```
male_bmi <- data[which(data$Gender=='Male'),"BMI"]
female_bmi <- data[which(data$Gender=='Female'),"BMI"]

mu1<-mean(male_bmi) #male bmi mean
mu2<-mean(female_bmi) #female bmi mean
mudiff_hat <- mu1-mu2
#Paramteric bootstrap to calculate CI for mu1-mu2

var1<-var(male_bmi)
var2<-var(female_bmi)
sd1<-sqrt(var1)
sd2<-sqrt(var2)
n_male <- length(male_bmi)
n_female <- length(female_bmi)

mu_diff_par<-c()
for (i in 1:1000){
  x <- rnorm(n_male,mu1,sd1)
  y <- rnorm(n_female,mu2,sd2)
  mu_diff_par[i]<-mean(x)-mean(y)
}

mu_diff_par.mean <- mean(mu_diff_par)
mu_diff_par.sd<-sd(mu_diff_par)
CI<-c(mu_diff_par.mean-1.96*mu_diff_par.sd,
mu_diff_par.mean+1.96*mu_diff_par.sd)

hist(mu_diff_par)
```

Output:



Observations:

1) The Maximum likelihood estimate is given by the value `0.7575097`
2) The estimated standard error was found using parametric bootstrap
3) Then using the MLE and standard error, a confidence interval was built – (`-1.751231, 3.141853`)

Chapter 5: Hypothesis Testing (Wald Test)

In statistics, the Wald test assesses constraints on statistical parameters based on the weighted distance between the unrestricted estimate and its hypothesized value under the null hypothesis, where the weight is the precision of the estimate.

Here we are performing wald test for null hypothesis if difference of mean BMI between male and female is zero

$H_0 : \mu_m - \mu_f = 0$

$H_a : \mu_m - \mu_f \neq 0$

Code:

```
z <- mudiff_hat - 0/(sqrt((var1+var2)/n_male+n_female))
p_value <- 2*(1-pnorm(z))
p_value
```

The observed p-value is 0.4487.

Since p-value is significant we cannot reject the null hypothesis. We cannot make any conclusions about the difference of mean of male BMI and female BMI.

Chapter 6: Bayesian Analysis

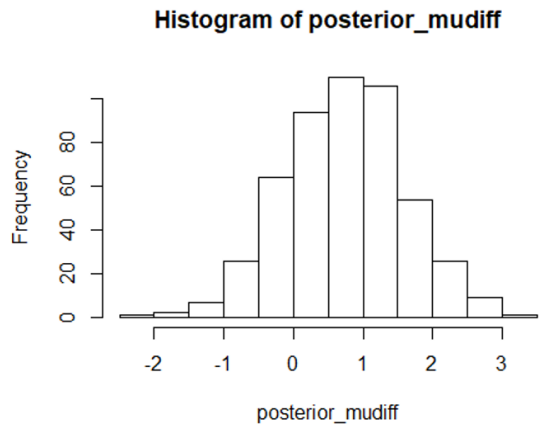
Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference is an important technique in statistics, and especially in mathematical statistics.

We continue to work with the estimate of difference of mean between male and female. Considering flat prior for the mean difference we generate posterior of mu1-mu2.

Code:

```
mu_diff_var <- var1+var2
posterior_mudiff<-rnorm(n_male+n_female,mudiff_hat,sqrt(mu_diff_var/(n_male+n
_female)))
hist(posterior_mudiff)
```

Output:



Observation:

We see that the distribution of difference of mean using Bayesian analysis is similar to the distribution of difference of mean obtained in frequentist approach.

Chapter 7: Conclusion

We have thus conducted various statistical tests and inferences on real world dataset as shown in previous chapters. This project could further be extended to do tests and analysis on this data to obtain other inferences on population. For instance, research can be done on proportion of unhealthy male to unhealthy female.