Title: *PIMA Indian Diabetes Data Analysis*

Submitted by, *Mitul Shah*
Siddharth Muthe
Yuvraj Singh Tomar

IE6200 Section: 7 Group: 5
Data Analytics Engineering, College of Engineering, Fall'19
Northeastern University, Boston, Massachusetts, United States of America

----

Introduction

Diabetes Mellitus is affecting 382 million people around the world. Hence, there is increase in the number of people with type 2 diabetes worldwide. Only in United states of America, approximately 30.3 million people were identified as suffering from diabetes and 1.5 million Americans are being diagnosed with diabetes each year. The population studied was the PIMA Indian tribe women population near Phoenix, Arizona, United States. The tribe has been under continuous study since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases due to it's high incidence and prediabetes rate of diabetes.

Since in women pregnancy seems to be a factor. According to World Health Organization Criteria, which stated that "if the 2-hour post-load glucose was at least 200 mg/dl at any survey exam or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during routine medical care". Given the data about PIMA, we will be trying to make predictions on how likely a PIMA Indian women is to suffer from diabetes, and therefore, act appropriately towards it. We can start analyzing statistical data that will help us study the onset of diabetes in Pima Indian women.

```
library(tidyverse)

## -- Attaching packages -------------------------------------------------------
---- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -------------------------------------------------------------- t
idyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(gridExtra)

##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine

library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(reshape2)

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths

library(corrplot)

## corrplot 0.84 loaded

library(ggpubr)

## Loading required package: magrittr

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract

library(e1071)
library(caTools)
library(fitdistrplus)

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
## Loading required package: survival

## Loading required package: npsurv

## Loading required package: lsei

library(BivRegBLS)

## Loading required package: ellipse

##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs

library(kableExtra)

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows

#df <- read.csv("Y:/R/Prob & Stats Project/ps_dataframe.csv", header = TRUE,
sep=',')
df <- read.csv("Y:/R/Prob & Stats Project/ps_dataframe_new.csv", header = T,
stringsAsFactors = F, sep=",")
diabetes <- df
```

Data Description

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to statistically predict whether a patient has diabetes, based on certain statistical measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage.

The data consists of two categories viz. tested positive and tested negative. It has 8 features as: number of times pregnant, plasma glucose concentration at 2-hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-hr serum insulin (mu U/ml), BMI (kg/m^2) , diabetes pedigree function and age (years).

```
variable.type <- lapply(df, class)
variable.description <- c("Number of times pregnant",
                          "Plasma glucose concentration at 2 hours in an oral
glucose tolerance test",
                          "Diastolic blood pressure", "Triceps skin fold thic
kness", "2-hour serum insulin (μU/ml)",
                          "Body Mass Index",
                          "Synthesis of the history of Diabetes Mellitus in r
elatives, generic relationship of those relatives to the subject",
                          "Age of the individual",
                          "Occurrence of Diabetes")
variable.name <- colnames(df)
datadesc <- as.data.frame(cbind(variable.name, variable.type, variable.descri
ption))
rownames(datadesc) <- (1:length(datadesc$variable.name))
colnames(datadesc) <- c("Variable Name","Data Type","Variable Description")
datadesc
```

```
##              Variable Name Data Type
## 1              Pregnancies   integer
## 2                  Glucose   integer
## 3            BloodPressure   integer
## 4            SkinThickness   numeric
## 5                  Insulin   numeric
## 6                      BMI   numeric
## 7 DiabetesPedigreeFunction   numeric
## 8                      Age   integer
## 9                  Outcome   integer
##
Variable Description
## 1
Number of times pregnant
## 2                                          Plasma glucose concentration at
2 hours in an oral glucose tolerance test
## 3
Diastolic blood pressure
## 4
Triceps skin fold thickness
## 5
2-hour serum insulin (µU/ml)
## 6
Body Mass Index
## 7 Synthesis of the history of Diabetes Mellitus in relatives, generic rela
tionship of those relatives to the subject
## 8
Age of the individual
## 9
Occurrence of Diabetes
```

Data Preprocessing

Each of the columns had some zero values, e.g.- age of some of the women was zero or their blood pressure level was zero. We had to replace those values, as in real life those values clearly cannot be zero. We calculated the two-separate means of the column based on the value of the outcome and then replaced those values in place of zeros in that column. We replaced the zeros with means for glucose, blood pressure, skin thickness, insulin and BMI columns. Furthermore, the outcome value was in numeric format. We had to convert it in factor format in order to treat it as two categories of 'yes' and 'no'. Also, the variables were not distributed normally. We took log of the variables in order to bring them in normal distribution.

```r
# imputing 0 values
df$Glucose <- ifelse(df$Outcome==0,replace(df$Glucose,df$Glucose==0,value=rou
nd(mean(df[df$Glucose>0 & df$Outcome ==0,]$Glucose))),
                     replace(df$Glucose,df$Glucose==0,value=round(mean(df[df$
Glucose>0 & df$Outcome ==1,]$Glucose))))
df$BloodPressure <- ifelse(df$BloodPressure==0,replace(df$BloodPressure,df$Bl
oodPressure==0,value=round(mean(df[df$BloodPressure>0 & df$Outcome ==0,]$Bloo
dPressure))),
                           replace(df$BloodPressure,df$BloodPressure==0,value
=round(mean(df[df$BloodPressure>0 & df$Outcome ==1,]$BloodPressure))))
df$BMI <- ifelse(df$Outcome==0,replace(df$BMI,df$BMI==0,value=round(mean(df[d
f$BMI>0 & df$Outcome ==0,]$BMI))),
                 replace(df$BMI,df$BMI==0,value=round(mean(df[df$BMI>0 & df$O
utcome ==1,]$BMI))))
```

Converting distribution to normal

It was found that most of the variables did not have normal distribution. So, in order to convert them into normal distribution and to carry out statistical analysis, log of those variables was taken.

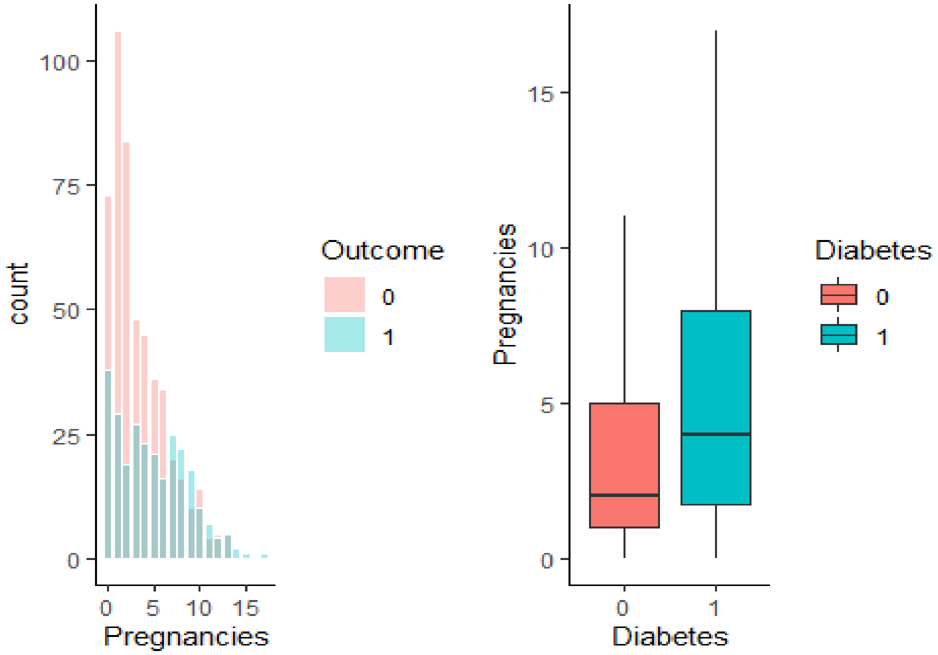
```r
#as factor
df$Outcome <- as.factor(df$Outcome)
df1 <- df
#taking log
df[,c(2,3,6,8)] <- log(df[,c(2,3,6,8)])
```

Population and sample creation

We created two independent populations, one for non-diabetic women and other for diabetic women. pop0 has population of non-diabetic women and pop1 has population of diabetic women. After that we took one sample from each of the populations.

```r
#Creating two independent populations
pop0<-df[df$Outcome==0,]
pop1<-df[df$Outcome==1,]

#Creating samples
sample0<-sample_n(pop0, 450)
sample1<-sample_n(pop1, 250)
```

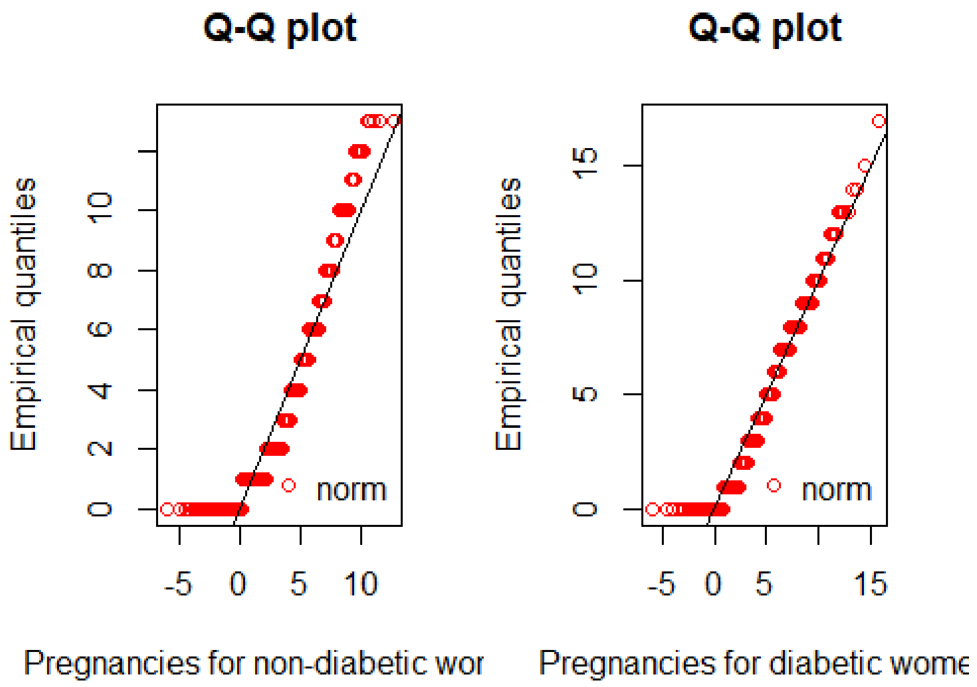
```
#1 Number of Times Pregnant
#Normality Testing
kurtosis(df$Pregnancies) #Calculate Kurtosis

## [1] 0.142184

skewness(df$Pregnancies) #Calculate Skewness

## [1] 0.8981549

shapiro.test(df$Pregnancies) #Significance Testing "Shapiro-Wilk's test" stat
istical analysis

##
##  Shapiro-Wilk normality test
##
## data:  df$Pregnancies
## W = 0.90428, p-value < 2.2e-16

#Plots
p1 <- ggplot(df, aes(x=df$Pregnancies,fill=df$Outcome))+
  geom_bar(alpha=0.35, position="identity",colour="white")+
  xlab("Pregnancies")+labs(fill="Outcome")+theme_classic()
p2 <- ggplot(df, aes(y=df$Pregnancies,x=df$Outcome,fill=df$Outcome))+
  geom_boxplot(outlier.shape=NA)+
  ylab("Pregnancies")+xlab("Diabetes")+labs(fill="Diabetes")+theme_classic()
grid.arrange(p1, p2, ncol = 2,nrow=1)
```



Inference on Pregnancies

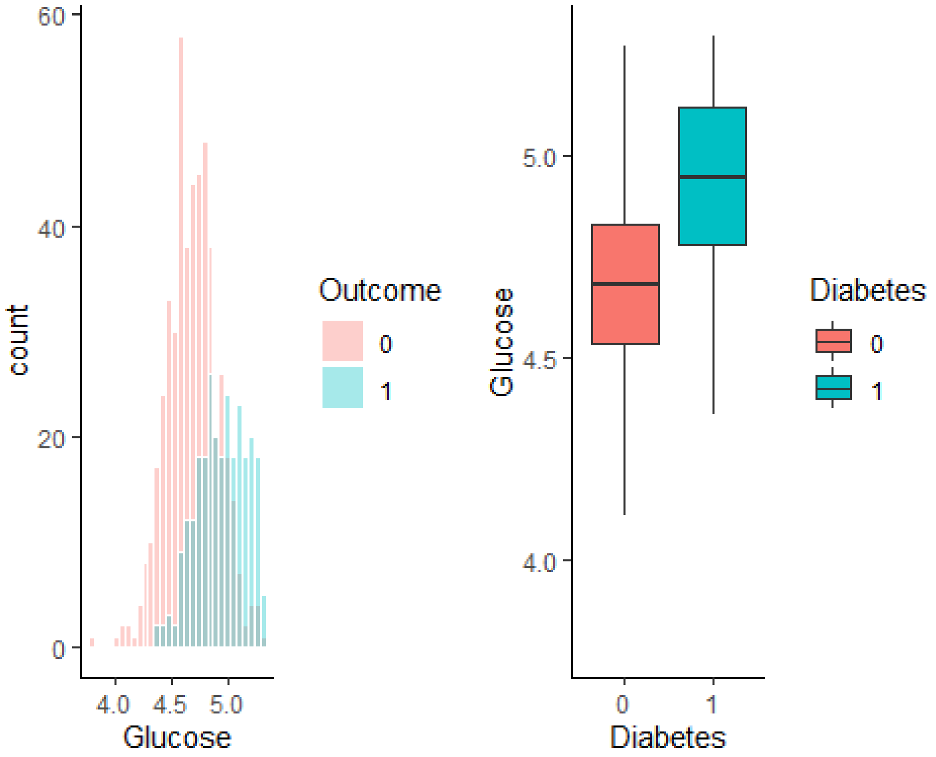
- The first histogram plot is for how many times a woman has been pregnant and from the plot we can see that the greatest number of times a woman has been pregnant is 1. 135 women were pregnant only once, while 103 women were pregnant for two times. 111 women had never been pregnant. The histogram for number of times a woman has been pregnant is right skewed.
- From the segmented histograms we can clearly see that as the number of times woman has been pregnant increases, the chances of having diabetes increases.
- The box plot clearly shows that median of number of times a woman has been pregnant is greater in case of women suffering from diabetes. So, we can say that as pregnancies increases, likelihood of diabetes increases.
- By using the qq plot we can see that the distribution isn't normal distribution. The distribution for pregnancies has skewness of 0.898 and kurtosis is 0.142.

```
fit_preg0 <- fitdist(pop0$Pregnancies, "norm")
fit_preg1 <- fitdist(pop1$Pregnancies, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp  (list(fit_preg0), legendtext = plot.legend, xlab = 'Pregnancies for n
on-diabetic women', xlegend = 'bottomright')
qqcomp  (list(fit_preg1), legendtext = plot.legend, xlab = 'Pregnancies for d
iabetic women', xlegend = 'bottomright')
```



```
#2 Glucose
#Normality Test
kurtosis(df$Glucose) #Calculate Kurtosis
```

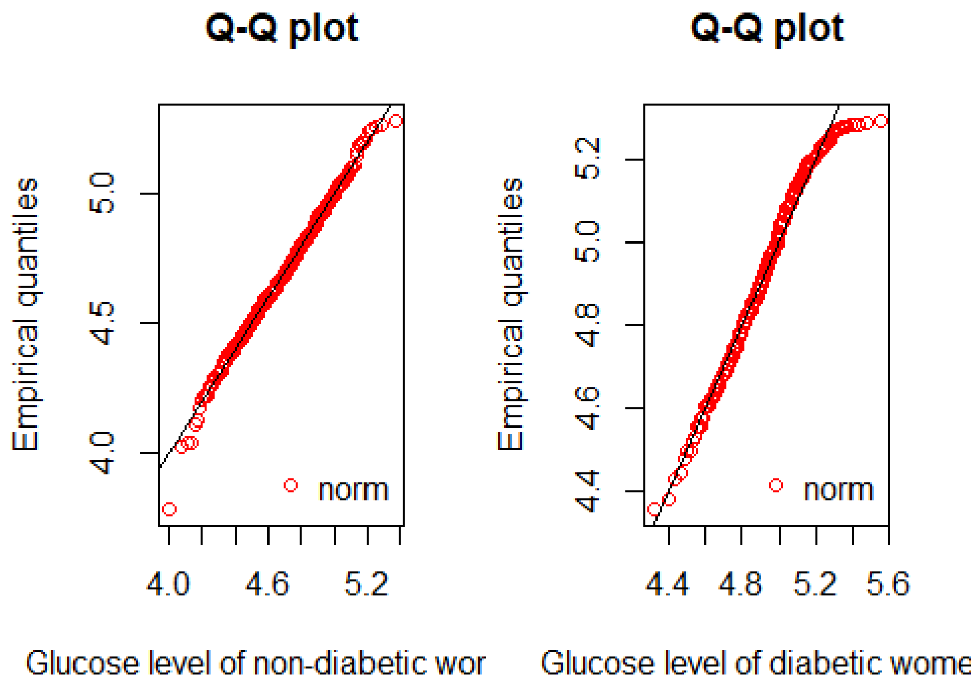
```
## [1] -0.1808438

skewness(df$Glucose) #Calculate Skewness

## [1] -0.06687324

shapiro.test(df$Glucose) #Significance Testing "Shapiro-Wilk's test"

##
##  Shapiro-Wilk normality test
##
## data:  df$Glucose
## W = 0.99155, p-value = 0.0002232

#Plots
p1 <- ggplot(df, aes(x=df$Glucose,fill=df$Outcome))+
  geom_histogram(alpha=0.35, position="identity",colour="white")+
  xlab("Glucose")+labs(fill="Outcome")+theme_classic()
p2 <- ggplot(df, aes(y=df$Glucose,x=df$Outcome,fill=df$Outcome))+
  geom_boxplot(outlier.shape=NA)+
  ylab("Glucose")+xlab("Diabetes")+labs(fill="Diabetes")+theme_classic()
grid.arrange(p1, p2, ncol = 2, nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Inference on Glucose

- In first plot, we have plotted two separate histograms, one for women those have diabetes(blue) and other for non-diabetic(red). From plot it can be clearly seen that the glucose level of women having diabetes is higher than women that do not have diabetes.
- It can be clearly seen from the box plot that the median for glucose is higher in case of diabetic women. From this we can say that as the glucose increases, chances of diabetes increases.
- The glucose distribution has almost normal distribution and this can be clearly seen from qq plot. Skewness for glucose distribution is -0.069 and kurtosis is -0.18.

```r
fit_glucose0 <- fitdist(pop0$Glucose, "norm")
fit_glucose1 <- fitdist(pop1$Glucose, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp(list(fit_glucose0), legendtext = plot.legend, xlab = 'Glucose level of 
non-diabetic women', xlegend = 'bottomright')
qqcomp(list(fit_glucose1), legendtext = plot.legend, xlab = 'Glucose level of 
diabetic women', xlegend = 'bottomright')
```



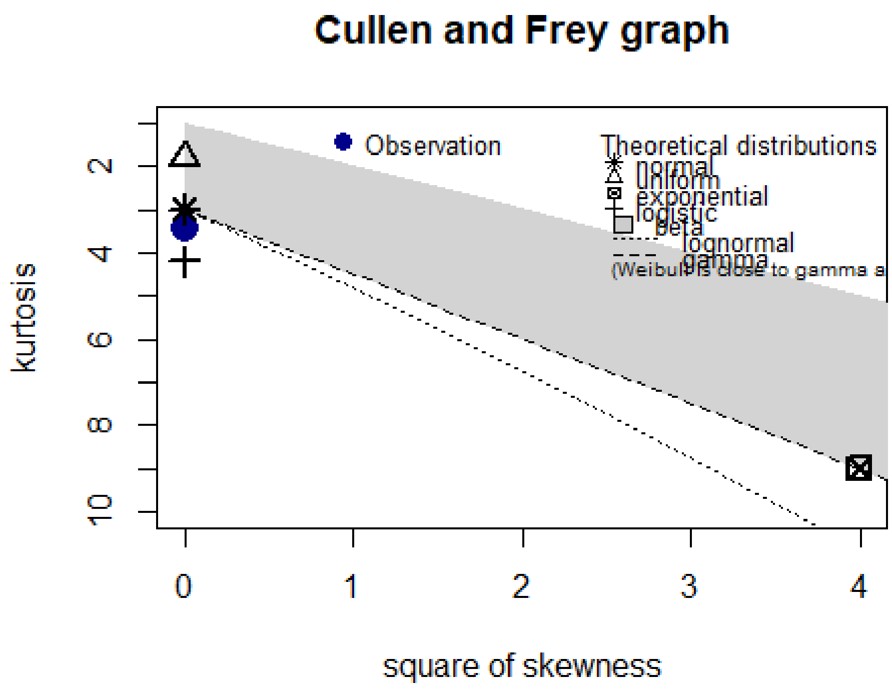
Normality testing

From the Q-Q probability plots for non-diabetic & diabetic PIMA India Women, it can be inferred that the observed values againt normally distributed data(represented by the line). Normally distributed data fall along the line.

Goodness-of-fit test

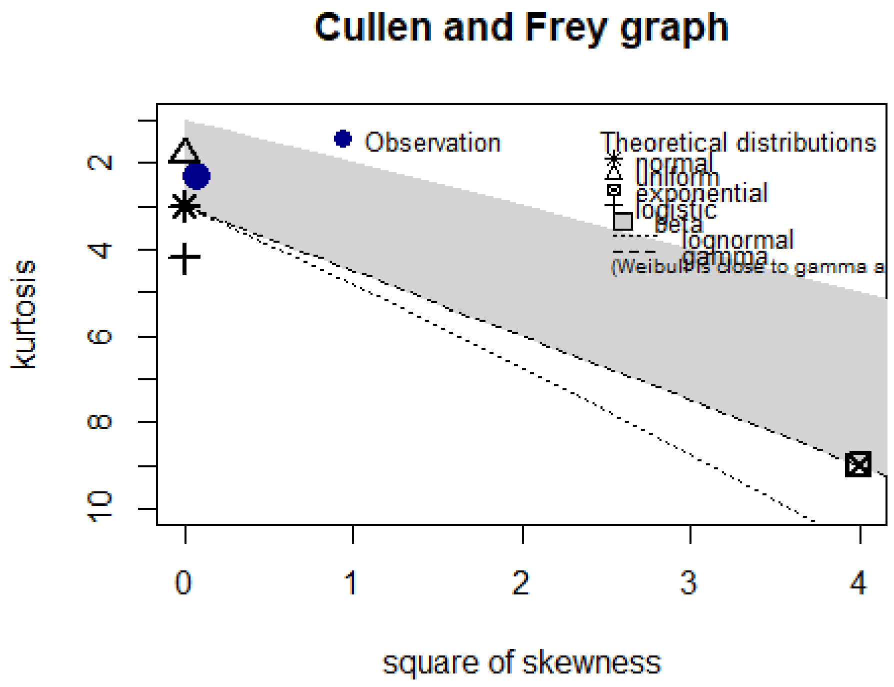
The goodness-of-fit tests can be used to determine whether a certain distribution is a good fit. Calculating the goodness-of-fit statistics also helps to order the fitted distributions accordingly to how good they fit to your data. This feature is very helpful for comparing the fitted models.

```
#Goodness of fit
descdist(pop0$Glucose)
```



```
## summary statistics
## ------
## min:  3.78419   max:  5.283204
## median:  4.67748
## mean:  4.681925
## estimated sd:  0.2217067
## estimated skewness:  -0.08460002
## estimated kurtosis:  3.429065
```

```
descdist(pop1$Glucose)
```



```
## summary statistics
## ------
## min:  4.356709   max:  5.293305
## median:  4.945201
## mean:  4.935938
## estimated sd:  0.2130465
## estimated skewness:  -0.2720811
## estimated kurtosis:  2.312458
```

As it can be observed from the plot, normal distribution as well as lognormal distribution are the closest to the distribution for Glucose for both the cases.

Hypothesis testing

We want to test if the outcome is dependent on the selected variable. So, for testing the hypothesis we would be creating two independent populations based on the outcome and we will take one sample from each of the populations. Then we will confirm if the selected variable contributes towards diabetes by carrying out hypothesis testing for difference of means of population. There won't be any difference in means of two populations if the variable does not contribute towards diabetes and on the other hand, if the variable is contributing towards diabetes then there would be difference in means of the two populations. As we know the variance of both the populations and we want to measure the difference of means, we would be using two sample z-test. Furthermore, we just want to check if the difference is zero or not, so we would be performing a two-sided test. We have selected the confidence level of 95%. If the z-value is between -1.96 and 1.96, we would fail to reject the null hypothesis and if it lies in rejection region then we would reject the null hypothesis.

zcalc = ((x1-x2)-(mu1-mu2)) / sqrt((sqr(σ1)/n1) + (sqr(σ2)/n2))

where,
x1 is the mean of sample 1 (non-diabetic women),
x2 is the mean of sample 2 (diabetic women),
mu1-mu2 is the hypothesized difference between population means which is 0,
σ1 is the standard deviation of population 1 (non-diabetic women),
σ2 is the standard deviation of population 2 (diabetic women),
n1 is the size of sample 1 (non-diabetic women),
n2 is the size of sample 2 (diabetic women)

```
#hypothesis testing
#function for z-statistic
z_test = function(a, b, var_a, var_b){
  n.a = length(a)
  n.b = length(b)
  z = ((mean(a, na.rm=TRUE) - mean(b, na.rm=TRUE)) / (sqrt((var_a)/n.a + (var
_b)/n.b)))
  return(z)
}
z_test(sample0$Glucose, sample1$Glucose, var(pop0$Glucose), var(pop1$Glucose)
)

## [1] -14.86188
```

Glucose hypothesis

We will check whether outcome is dependent on glucose.

H0: μ1 - μ2 = 0
H1: μ1 - μ2 != 0
z-value = -14.95246
Thus, for a significance level of α = 0.05, we reject the null hypothesis since the z-value lies outside the range [−1.96, 1.96] and conclude that there is significant difference between the mean of glucose of two population. So, we can say that glucose has effect on diabetes.

Confidence interval of mean

From the hypothesis testing we found out that the factors contributing towards diabetes are glucose, diabetes pedigree function, BMI and blood pressure. After finding out the factors that are responsible for diabetes, we will now calculate the interval in which these factors lie for diabetic women as well as non-diabetic women. We will use the confidence interval of means formula for this, which is given by,

x-(z(alpha/2)*(sd_pop)/sqrt(sample_size)) < mu
x+(z(alpha/2)(sd_pop)/sqrt(sample_size))

where,
x is mean of sample,
sd_pop is standard deviation of population,
sample_size is size of sample

```r
#Confidence interval of mean for glucose level of diabetic women
mean_sample<-exp(mean(sample1$Glucose))
sd_pop<-exp(sd(pop1$Glucose))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 138.8634

upper_limit

## [1] 139.8335
```

Confidence interval of mean for glucose level of diabetic women

Sample mean of diabetic women = 139.5339 mg/dL
Standard deviation of population of diabetic women = 1.237442 mg/dL
Sample size = 25
α = 0.05
After calculation we find that the glucose level of diabetic women lies between 139.05 and 140.01. So, if a woman is diabetic, we can say with 95% confidence that her glucose level lies in this range.

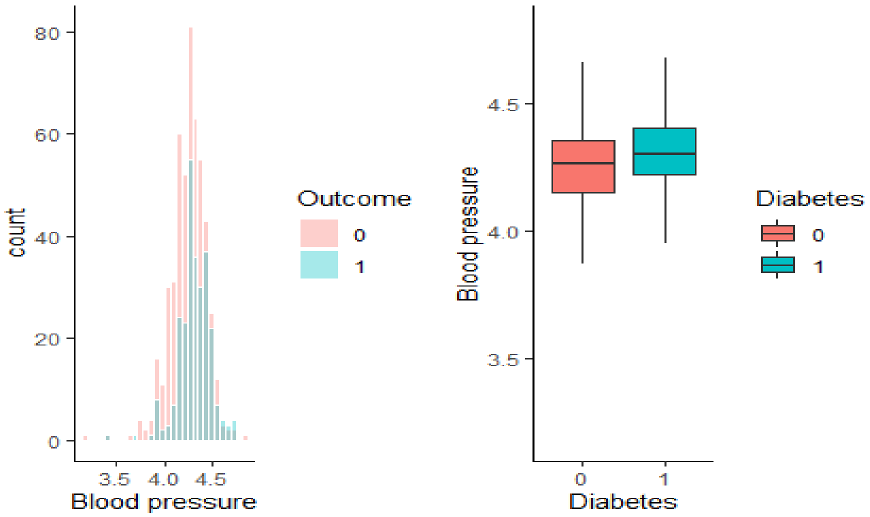
```r
#Confidence interval of mean for glucose level of non-diabetic women
mean_sample<-exp(mean(sample0$Glucose))
sd_pop<-exp(sd(pop0$Glucose))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 107.6636

upper_limit

## [1] 108.6422
```

Confidence interval of mean for Glucose level of non-diabetic women

Sample mean of non-diabetic women = 108.1993 mg/dL
Standard deviation of population of non-diabetic women = 1.248205 mg/dL

Sample size = 25
α = 0.05
After calculation we find that the glucose level of non-diabetic women lies between 107.71 and 108.6886. So, if a woman is non-diabetic, we can say with 95% confidence that her glucose level lies in this range.

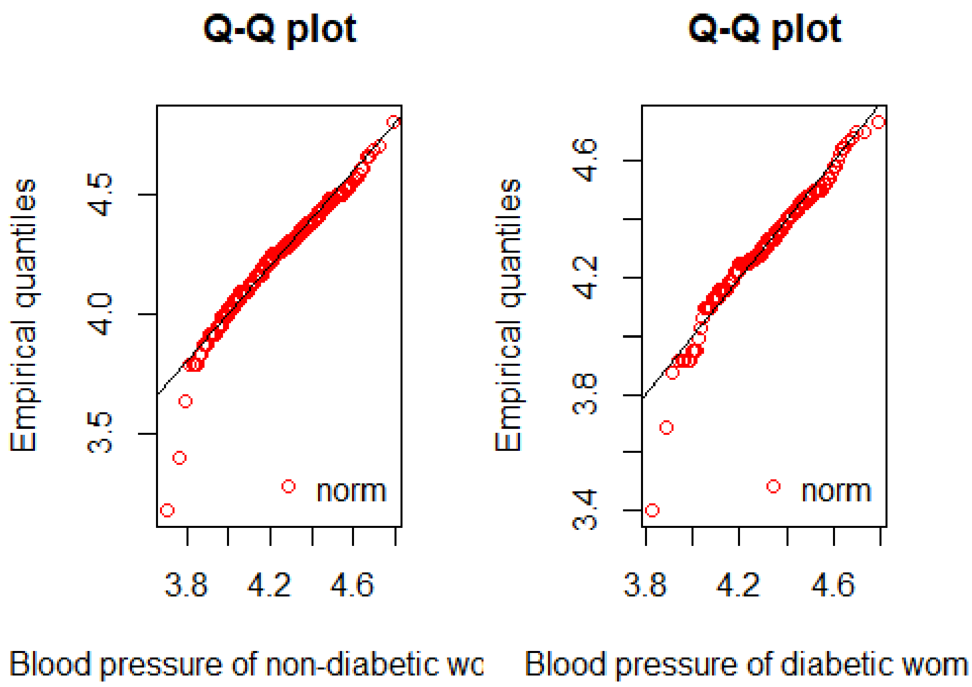
```
#3 Blood pressure
#Normality Test
kurtosis(df$BloodPressure) #Calculate Kurtosis

## [1] 3.315423

skewness(df$BloodPressure) #Calculate Skewness

## [1] -0.806832

shapiro.test(df$BloodPressure) #Significance Testing "Shapiro-Wilk's test"

## 
##  Shapiro-Wilk normality test
## 
## data:  df$BloodPressure
## W = 0.96227, p-value = 3.607e-13

#Plots
p1 <- ggplot(df, aes(x=df$BloodPressure,fill=df$Outcome))+
  geom_histogram(alpha=0.35, position="identity",colour="white")+
  xlab("Blood pressure")+labs(fill="Outcome")+theme_classic()
p2 <- ggplot(df, aes(y=df$BloodPressure,x=df$Outcome,fill=df$Outcome))+
  geom_boxplot(outlier.shape=NA)+
  ylab("Blood pressure")+xlab("Diabetes")+labs(fill="Diabetes")+theme_classic
()
grid.arrange(p1, p2, ncol = 2,nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Inference on Blood Pressure

- From first plot we can see that most of the women those who don't have diabetes have a blood pressure level of around 69 and on the other hand majority of diabetic women have slightly higher blood pressure as compared to them. From the segmented his
- Median of blood pressure level is high in case of women suffering from diabetes. Although there is not too much of difference in median, we can say that with increase in blood pressure level, chances of diabetes increases.
- Skewness of blood pressure distribution is -0.8 and kurtosis is 3.32

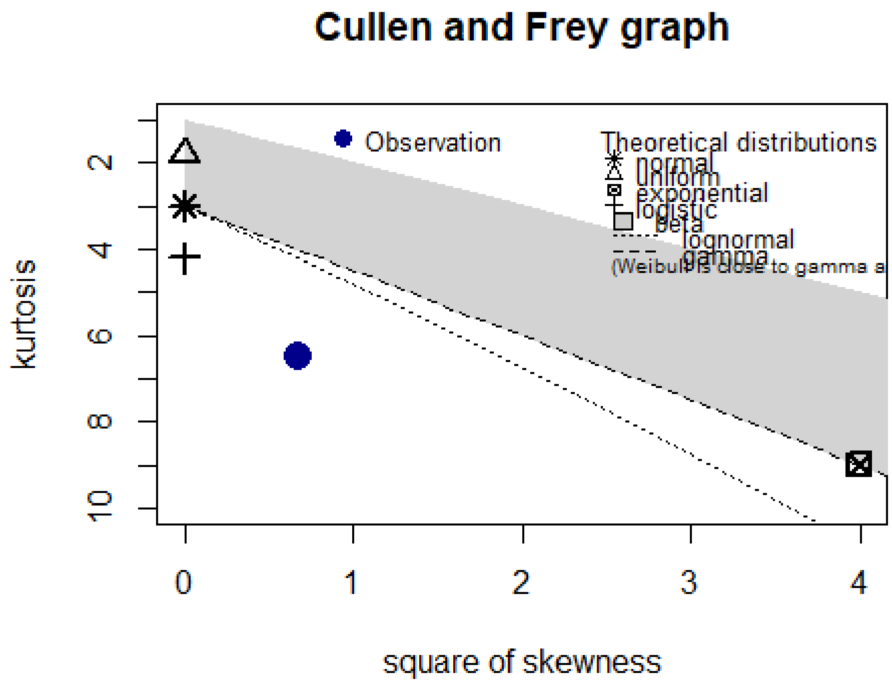
```
fit_bp0 <- fitdist(pop0$BloodPressure, "norm")
fit_bp1 <- fitdist(pop1$BloodPressure, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp(list(fit_bp0), legendtext = plot.legend, xlab = 'Blood pressure of non
-diabetic women', xlegend = 'bottomright')
qqcomp(list(fit_bp1), legendtext = plot.legend, xlab = 'Blood pressure of dia
betic women', xlegend = 'bottomright')
```



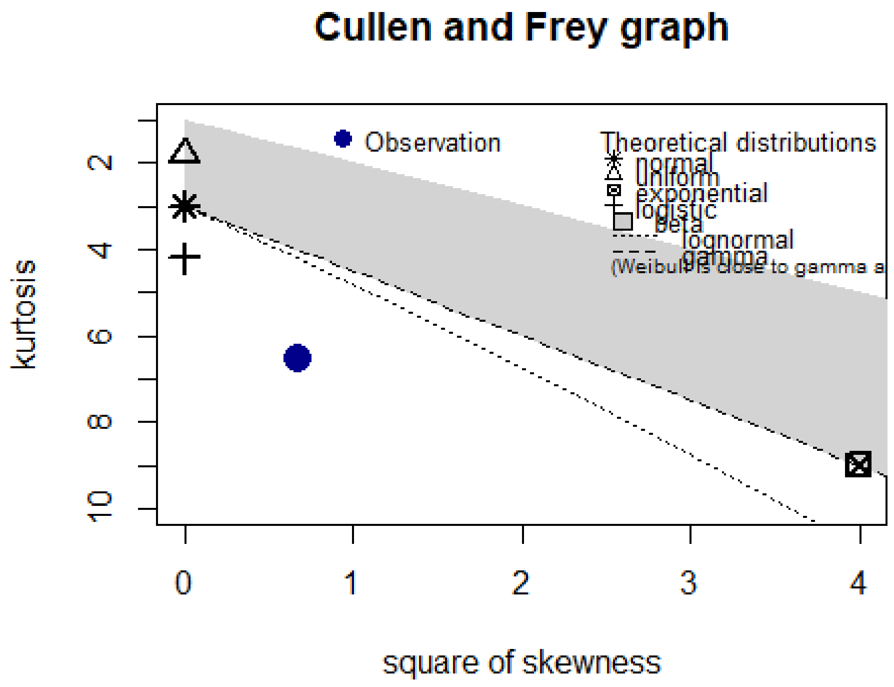
Normality testing

From the Q-Q probability plots for non-diabetic & diabetic PIMA India Women, it can be inferred that the observed values againt normally distributed data(represented by the line). Normally distributed data fall along the line.

```
#Goodness of fit
descdist(pop0$BloodPressure)
```



```
## summary statistics
## ------
## min:  3.178054   max:  4.804021
## median:  4.26268
## mean:  4.246209
## estimated sd:  0.1758183
## estimated skewness:  -0.8184393
## estimated kurtosis:  6.509338

descdist(pop1$BloodPressure)
```



```
## summary statistics
## ------
## min:  3.401197   max:  4.736198
## median:  4.304065
## mean:  4.305076
## estimated sd:  0.1663548
## estimated skewness:  -0.823675
## estimated kurtosis:  6.53796
```

As can be observed from the plot, normal distribution as well as lognormal distribution are the closest to the distribution for BloodPressure for both the cases and also, due to existing outliers the data is skewed.

```
#hypothesis testing
z_test(sample0$BloodPressure, sample1$BloodPressure, var(pop0$BloodPressure),
var(pop1$BloodPressure))

## [1] -5.015484
```

Blood pressure hypothesis

We will check whether outcome is dependent on blood pressure.

H0: µ1 - µ2 = 0
H1: µ1 - µ2 != 0
z-value = -4.587262

Thus, for a significance level of α = 0.05, we reject the null hypothesis since the z-value lies outside the range [−1.96, 1.96] and conclude that there is significant difference between the mean of blood pressure of two population. So, we can say that blood pressure has effect on diabetes.

```
#Confidence interval of mean for blood pressure of diabetic women
mean_sample<-exp(mean(sample1$BloodPressure))
sd_pop<-exp(sd(pop1$BloodPressure))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 73.98471

upper_limit

## [1] 74.9106
```

Confidence interval of mean for blood pressure of diabetic women

Sample mean of diabetic women = 74.14762 mm Hg
Standard deviation of population of diabetic women = 1.180992 mm Hg
Sample size = 25
α = 0.05
After calculation we find that the blood pressure of diabetic women lies between 73.68 and 74.61. So, if a woman is diabetic, we can say with 95% confidence that her blood pressure lies in this range.

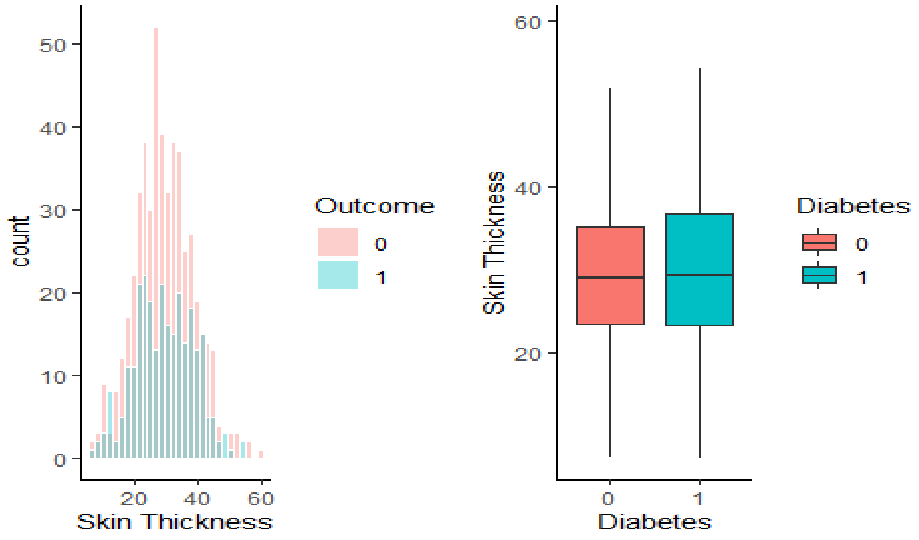
```
#Confidence interval of mean for blood pressure of non-diabetic women
mean_sample<-exp(mean(sample0$BloodPressure))
sd_pop<-exp(sd(pop0$BloodPressure))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 69.14352

upper_limit

## [1] 70.07822
```

Confidence interval of mean for blood pressure of non-diabetic women

Sample mean of non-diabetic women = 69.72911 mm Hg
Standard deviation of population of non-diabetic women = 1.192221 mm

Hg Sample size = 25
α = 0.05 After calculation we find that the blood pressure of non-diabetic women lies between 69.26176 and 70.19646. So, if a woman is non-diabetic, we can say with 95% confidence that her blood pressure lies in this range.

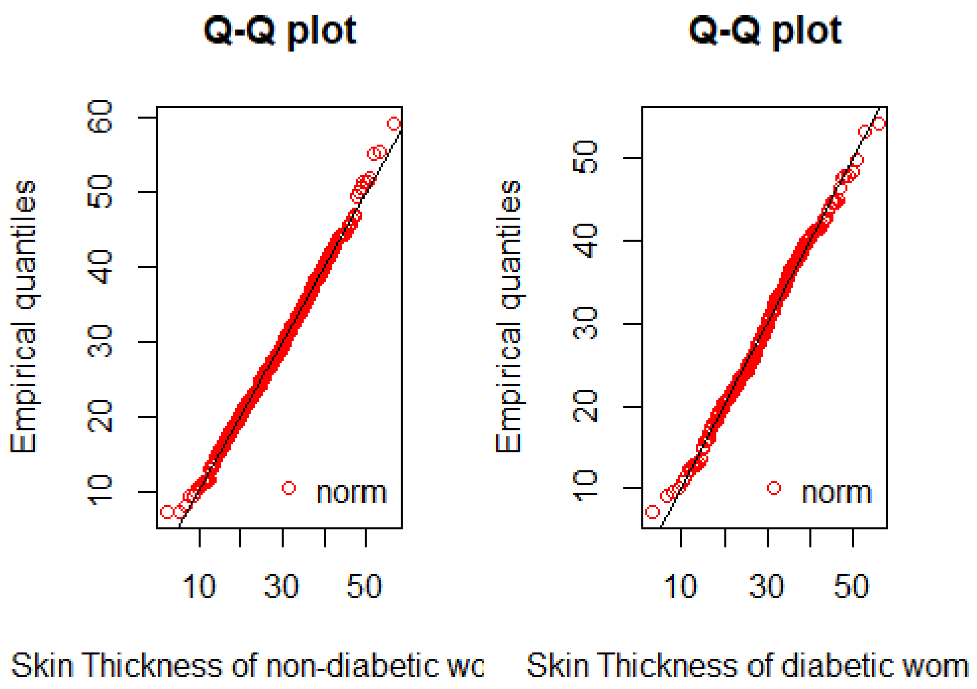
```
#4 Skin Thickness
#Normality Test
kurtosis(df$SkinThickness, na.rm=TRUE) #Calculate Kurtosis

## [1] -0.1704684

skewness(df$SkinThickness, na.rm=TRUE) #Calculate Skewness

## [1] 0.1193945

shapiro.test(df$SkinThickness) #Significance Testing "Shapiro-Wilk's test"

##
##  Shapiro-Wilk normality test
##
## data:  df$SkinThickness
## W = 0.997, p-value = 0.165

#Plots
p1 <- ggplot(df, aes(x=df$SkinThickness,fill=df$Outcome))+
  geom_histogram(alpha=0.35, position="identity",colour="white")+
  xlab("Skin Thickness")+labs(fill="Outcome")+theme_classic()
p2 <- ggplot(df, aes(y=df$SkinThickness,x=df$Outcome,fill=df$Outcome))+
  geom_boxplot(outlier.shape=NA)+
  ylab("Skin Thickness")+xlab("Diabetes")+labs(fill="Diabetes")+theme_classic
()
grid.arrange(p1, p2, ncol = 2,nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Inference on Skin Thickness

- From the first plot, for segmented histograms, it can be seen that there is not much of difference in skin thickness of diabetic and non-diabetic women
- The median of skin thickness is almost same for both diabetic and non-diabetic women. It can be said that skin thickness hardly has any effect on diabetes.
- The skewness of skin thickness distribution is 0.119 and kurtosis is -0.17.

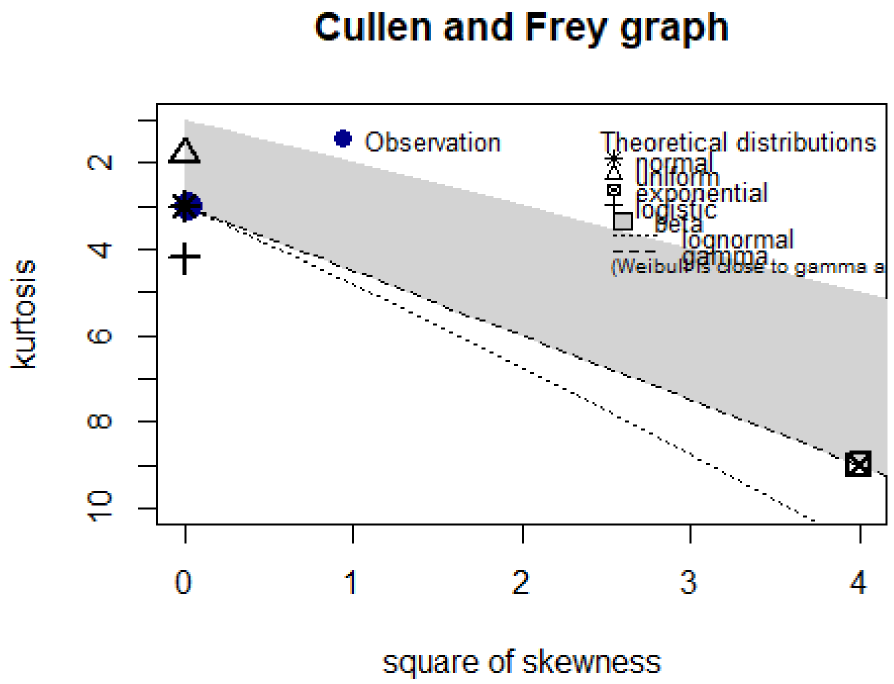
```
fit_st0 <- fitdist(pop0$SkinThickness, "norm")
fit_st1 <- fitdist(pop1$SkinThickness, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp(list(fit_st0), legendtext = plot.legend, xlab = 'Skin Thickness of non
-diabetic women', xlegend = 'bottomright')
qqcomp(list(fit_st1), legendtext = plot.legend, xlab = 'Skin Thickness of dia
betic women', xlegend = 'bottomright')
```



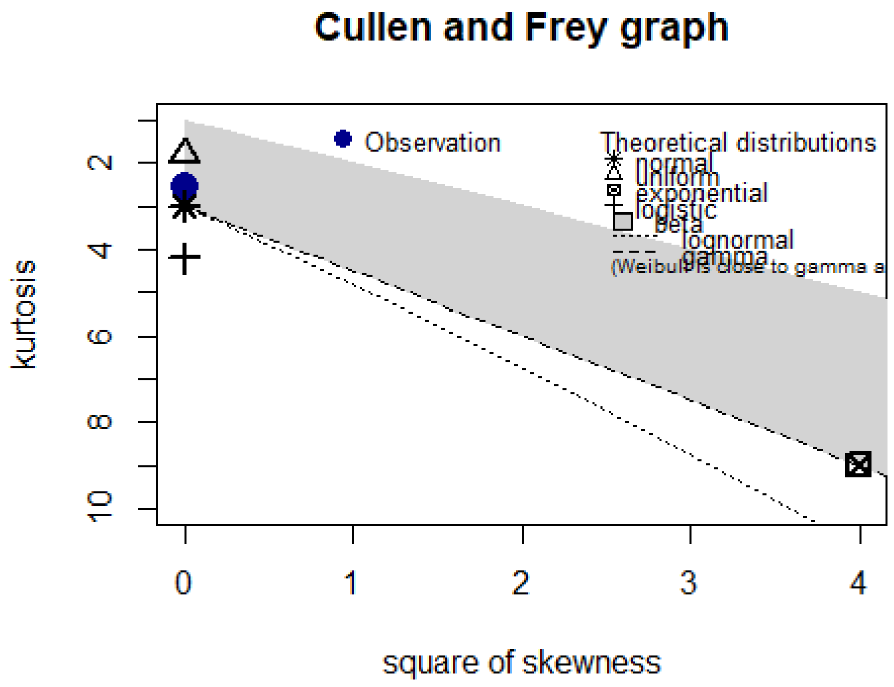
Normality testing

From the Q-Q probability plots for non-diabetic & diabetic PIMA India Women, it can be inferred that the observed values againt normally distributed data(represented by the line). Normally distributed data fall along the line.

```
#Goodness of fit
descdist(pop0$SkinThickness)
```



```
## summary statistics
## ------
## min:  7.39463   max:  59.32231
## median:  28.85484
## mean:  29.40744
## estimated sd:  8.758883
## estimated skewness:  0.1726502
## estimated kurtosis:  3.033787

descdist(pop1$SkinThickness)
```



```
## summary statistics
## ------
## min:  7.173243   max:  54.2981
## median:  29.18616
## mean:  29.51298
## estimated sd:  9.112524
## estimated skewness:  0.03074751
## estimated kurtosis:  2.549341
```

As can be observed from the plot, normal distribution as well as lognormal distribution are the closest to the distribution for SkinThickness for both the cases.

```
#hypothesis testing
z_test(sample0$SkinThickness, sample1$SkinThickness, var(pop0$SkinThickness),
var(pop1$SkinThickness))

## [1] -0.6550233
```

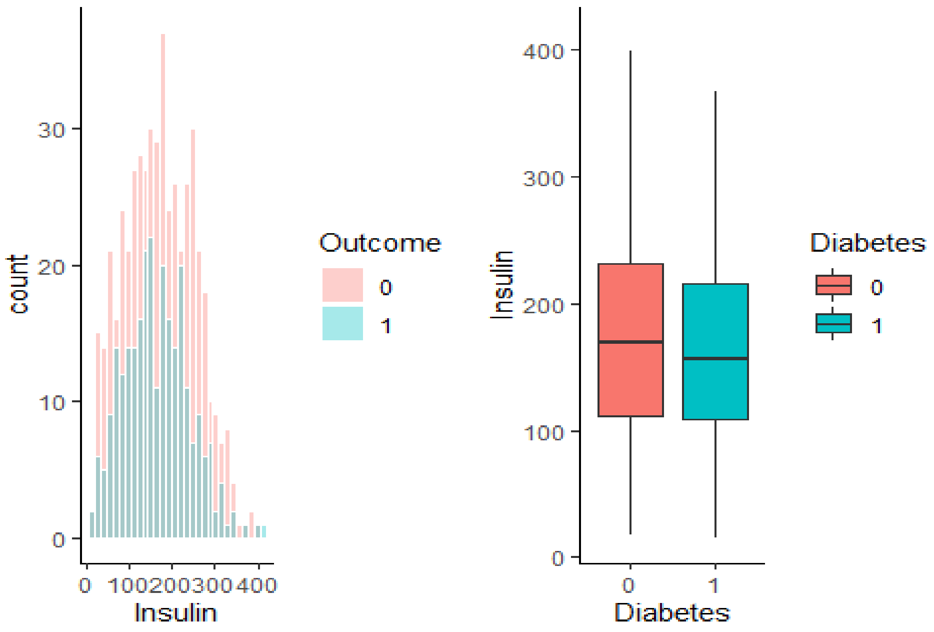
Skin thickness hypothesis

We will check whether outcome is dependent on skin thickness.

H0: µ1 - µ2 = 0
H1: µ1 - µ2! = 0
z-value = 0.4995874

Thus, for a significance level of α = 0.05, we fail to reject the null hypothesis since the z-value lies within the range [−1.96, 1.96] and conclude that there is no significant difference between the mean of skin thickness of two population. So, we can say that skin thickness has no effect on diabetes.

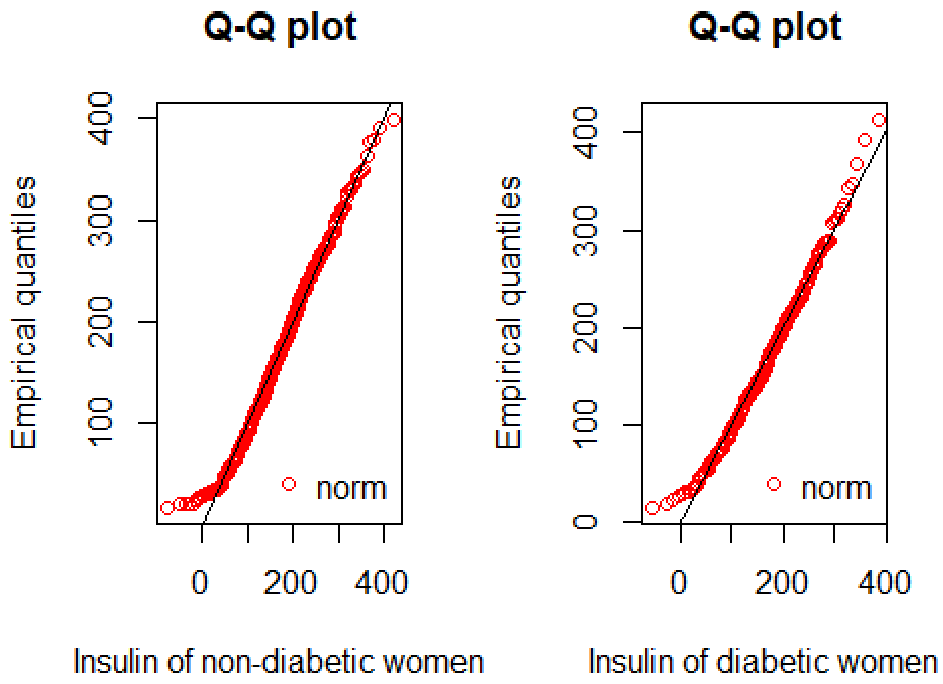
```
#5 Insulin
#Normality Test
kurtosis(df$Insulin) #Calculate Kurtosis

## [1] -0.461546

skewness(df$Insulin) #Calculate Skewness

## [1] 0.2382819

shapiro.test(df$Insulin) #Significance Testing "Shapiro-Wilk's test"

## 
##  Shapiro-Wilk normality test
## 
## data:  df$Insulin
## W = 0.9875, p-value = 3.928e-06

#Plots
p1 <- ggplot(df, aes(x=df$Insulin,fill=df$Outcome))+
  geom_histogram(alpha=0.35, position="identity",colour="white")+
  xlab("Insulin")+labs(fill="Outcome")+theme_classic()
p2 <- ggplot(df, aes(y=df$Insulin,x=df$Outcome,fill=df$Outcome))+
  geom_boxplot(outlier.shape=NA)+
  ylab("Insulin")+xlab("Diabetes")+labs(fill="Diabetes")+theme_classic()
grid.arrange(p1, p2, ncol = 2,nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Inference on Insulin

- Non-diabetic women have slightly lower values of insulin as compared to diabetic women.
- Median of insulin serum is higher in case of diabetic women.
- The distribution is somewhat normal. It has skewness of 0.24 and kurtosis of -0.46

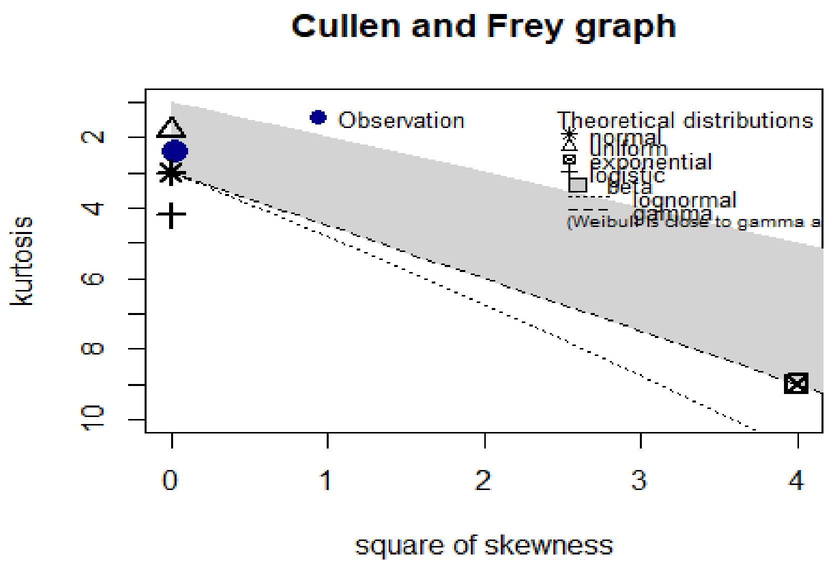
```
fit_insulin0 <- fitdist(pop0$Insulin, "norm")
fit_insulin1 <- fitdist(pop1$Insulin, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp(list(fit_insulin0), legendtext = plot.legend, xlab = 'Insulin of non-d
iabetic women', xlegend = 'bottomright')
qqcomp(list(fit_insulin1), legendtext = plot.legend, xlab = 'Insulin of diabe
tic women', xlegend = 'bottomright')
```



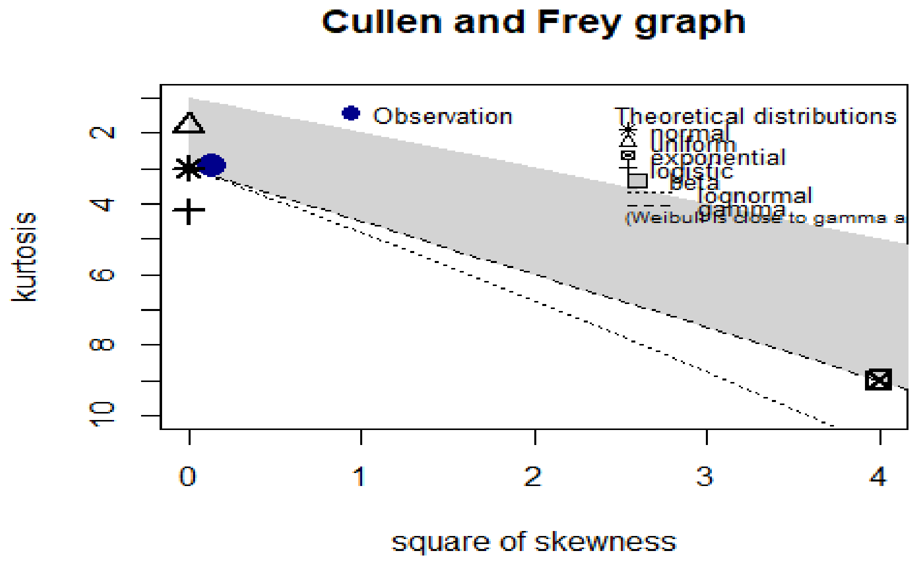
Normality testing

From the Q-Q probability plots for non-diabetic & diabetic PIMA India Women, it can be inferred that the observed values againt normally distributed data(represented by the line). Normally distributed data fall along the line.

```
#Goodness of fit
descdist(pop0$Insulin)
```



```
## summary statistics
## ------
## min:  17.65466   max:  398.8461
## median:  168.9909
## mean:  171.0786
## estimated sd:  81.05305
## estimated skewness:  0.1699785
## estimated kurtosis:  2.40272

descdist(pop1$Insulin)
```



```
## summary statistics
## ------
## min:  15.2536   max:  413.6139
## median:  155.5973
## mean:  164.1655
## estimated sd:  76.14117
## estimated skewness:  0.3742918
## estimated kurtosis:  2.928103
```

As can be observed from the plot, normal distribution as well as lognormal distribution are the closest to the distribution for Insulin for both the cases.

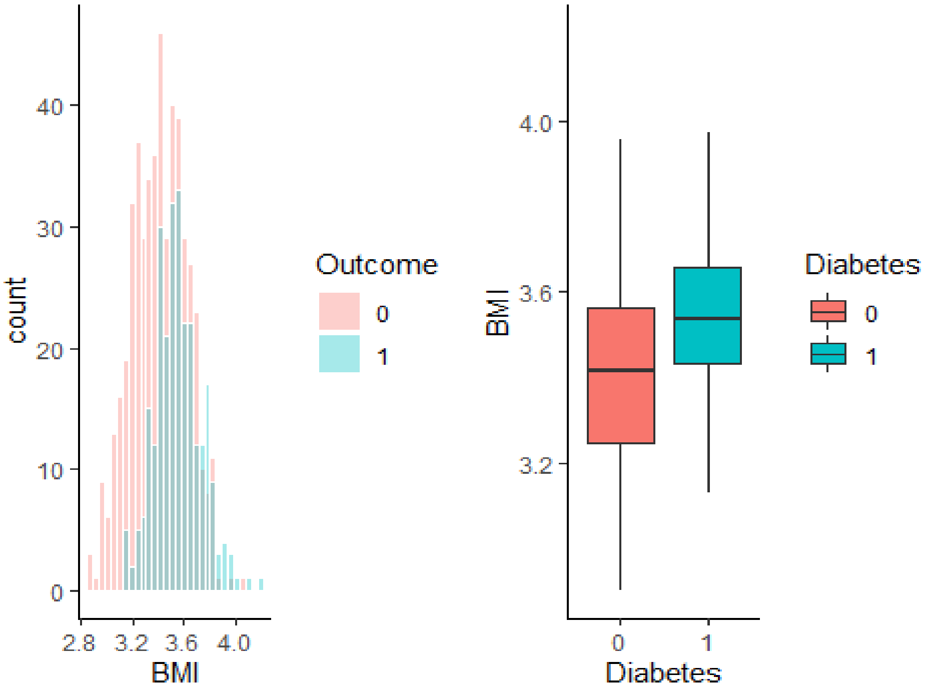
```
#hypothesis testing
z_test(sample0$Insulin, sample1$Insulin, var(pop0$Insulin), var(pop1$Insulin)
)

## [1] 1.338075
```

Insulin hypothesis

We will check whether outcome is dependent on insulin.
H0: μ1 - μ2 = 0
H1: μ1 - μ2 != 0
z-value = 1.145054
Thus, for a significance level of α = 0.05, we fail to reject the null hypothesis since the z-value lies within the range [−1.96, 1.96] and conclude that there is no significant difference between the mean of insulin of two samples. So, we can say that insulin has no effect on diabetes.

```
#6 BMI
#Normality Test
kurtosis(df$BMI) #Calculate Kurtosis

## [1] -0.1085503

skewness(df$BMI) #Calculate Skewness

## [1] -0.05172912

shapiro.test(df$BMI) #Significance Testing "Shapiro-Wilk's test"

## 
##  Shapiro-Wilk normality test
## 
## data:  df$BMI
## W = 0.99714, p-value = 0.1958

#Plots
p1 <- ggplot(df, aes(x=df$BMI,fill=df$Outcome))+
  geom_histogram(alpha=0.35, position="identity",colour="white")+
  xlab("BMI")+labs(fill="Outcome")+theme_classic()
```

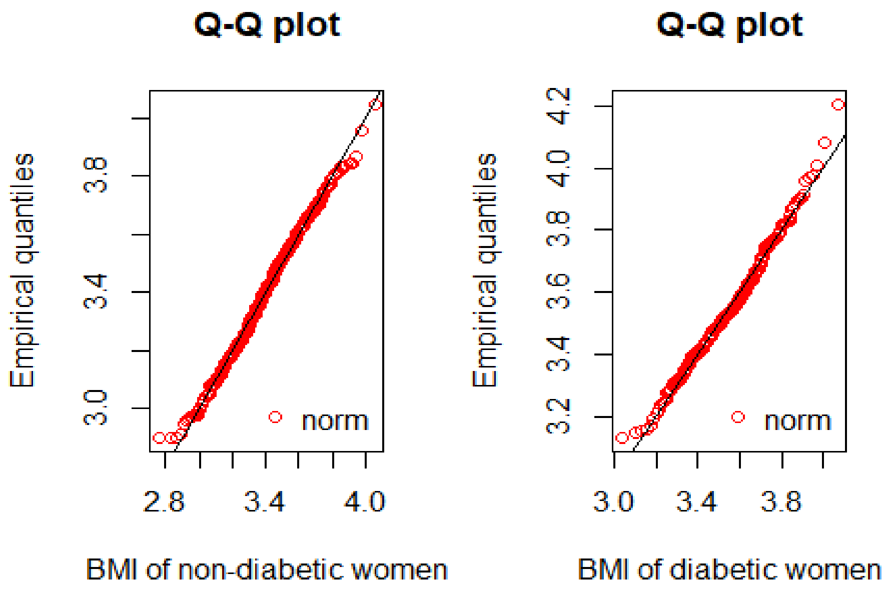
```
p2 <- ggplot(df, aes(y=df$BMI,x=df$Outcome,fill=df$Outcome))+
  geom_boxplot(outlier.shape=NA)+
  ylab("BMI")+xlab("Diabetes")+labs(fill="Diabetes")+theme_classic()
grid.arrange(p1, p2, ncol = 2,nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Inference on BMI

- The BMI of diabetic women is higher than non-diabetic women.
- Even the median of BMI is higher in case of diabetic women. So, it can be said that women with diabetes are obese. Obesity is one of the factors for diabetes and obese women have more likelihood of having diabetes.
- The distribution is normal and can be confirmed from qq plot. Skewness of BMI distribution is -0.05 and kurtosis is -0.1

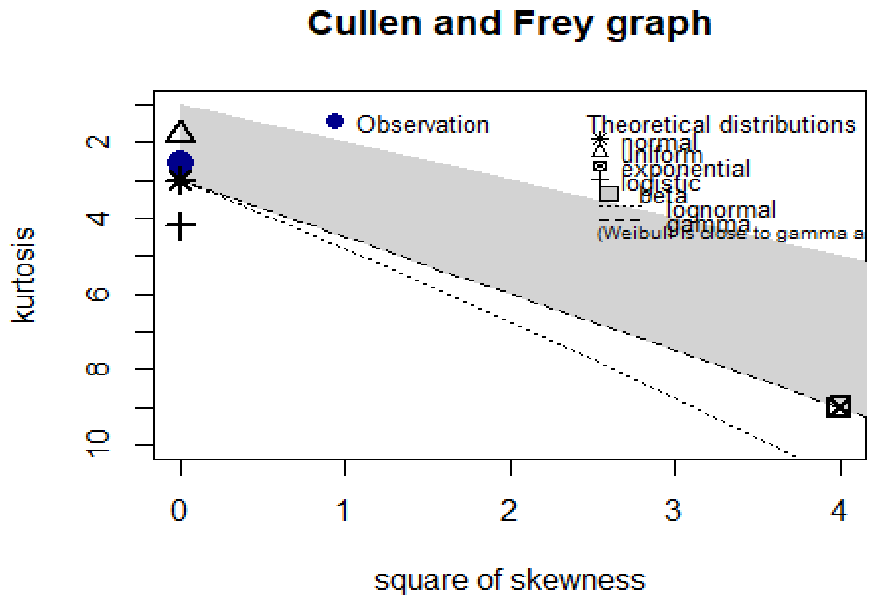
```
fit_bmi0 <- fitdist(pop0$BMI, "norm")
fit_bmi1 <- fitdist(pop1$BMI, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp(list(fit_bmi0), legendtext = plot.legend, xlab = 'BMI of non-diabetic
women', xlegend = 'bottomright')
qqcomp(list(fit_bmi1), legendtext = plot.legend, xlab = 'BMI of diabetic wome
n', xlegend = 'bottomright')
```



Normality test

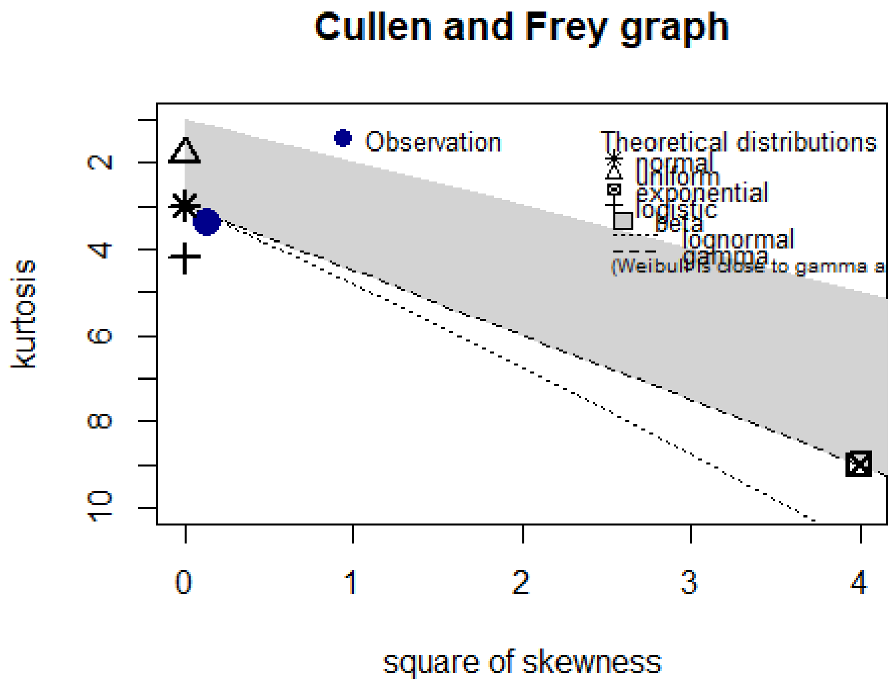
From the Q-Q probability plots for non-diabetic & diabetic PIMA India Women, it can be inferred that the observed values againt normally distributed data(represented by the line). Normally distributed data fall along the line.

```
#Goodness of fit
descdist(pop0$BMI)
```



```
## summary statistics
## ------
## min:  2.901422   max:  4.048301
## median:  3.414443
## mean:  3.407562
## estimated sd:  0.2102338
## estimated skewness:  -0.03121963
## estimated kurtosis:  2.53469
```

```
descdist(pop1$BMI)
```



```
## summary statistics
## ------
## min:  3.131137   max:  4.206184
## median:  3.535145
## mean:  3.550628
## estimated sd:  0.1782751
## estimated skewness:  0.365818
## estimated kurtosis:  3.399488
```

As can be observed from the plot, normal distribution as well as lognormal distribution are the closest to the distribution for BMI for both the cases.

```
#hypothesis testing
z_test(sample0$BMI, sample1$BMI, var(pop0$BMI), var(pop1$BMI))

## [1] -9.142559
```

BMI hypothesis

We will check whether outcome is dependent on BMI.
H0: µ1 - µ2 = 0
H1: µ1 - µ2 != 0
z-value = -9.726171
Thus, for a significance level of α = 0.05, we reject the null hypothesis since the z-value lies outside the range [−1.96, 1.96] and conclude that there is significant difference between the mean of BMI of two population. So, we can say that BMI has effect on diabetes.

```
#Confidence interval of mean for BMI of diabetic women
mean_sample<-exp(mean(sample1$BMI))
sd_pop<-exp(sd(pop1$BMI))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 34.26131

upper_limit

## [1] 35.19831
```

Confidence interval of mean for BMI of diabetic women

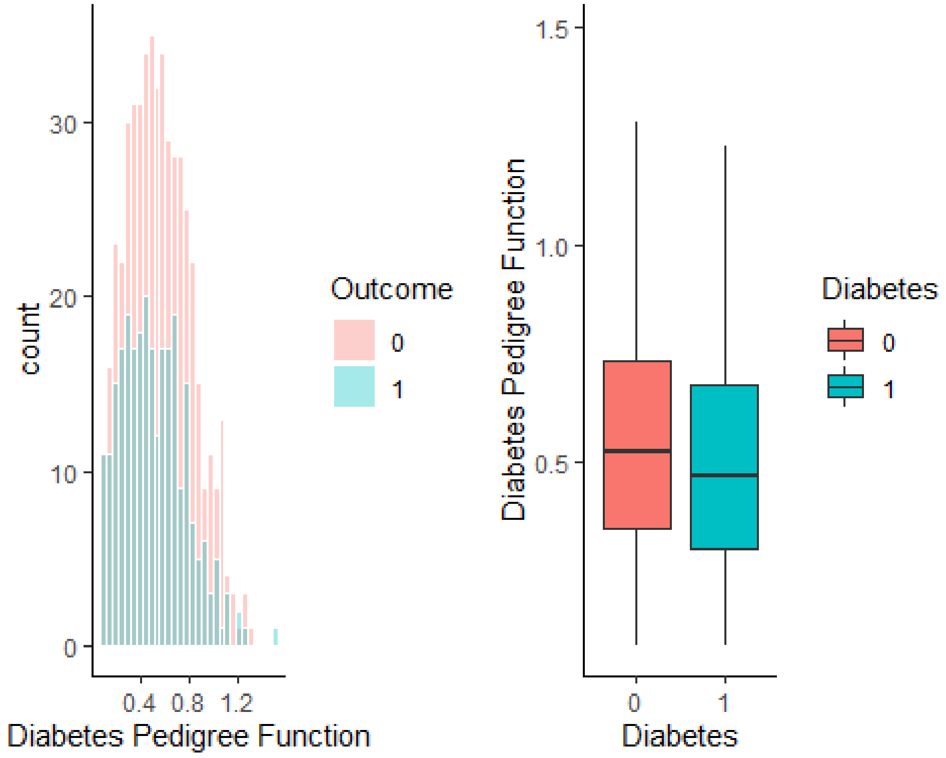
Sample mean of diabetic women = 34.96717 kg/m^2
Standard deviation of population of diabetic women = 1.195154 kg/m^2
Sample size = 25
α = 0.05
After calculation we find that the BMI of diabetic women lies between 34.49 and 35.43. So, if a woman is diabetic, we can say with 95% confidence that her BMI lies in this range.

```
#Confidence interval of mean for BMI of non-diabetic women
mean_sample<-exp(mean(sample0$BMI))
sd_pop<-exp(sd(pop0$BMI))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 29.79226

upper_limit

## [1] 30.75969
```

Confidence interval of mean for BMI of non-diabetic women

Sample mean of non-diabetic women = 30.217 kg/m^2
Standard deviation of population of non-diabetic women = 1.233967 kg/m^2
Sample size = 25
α = 0.05
After calculation we find that the BMI of non-diabetic women lies between 29.73328 and 30.70071. So, if a woman is non-diabetic, we can say with 95% confidence that her BMI lies in this range.

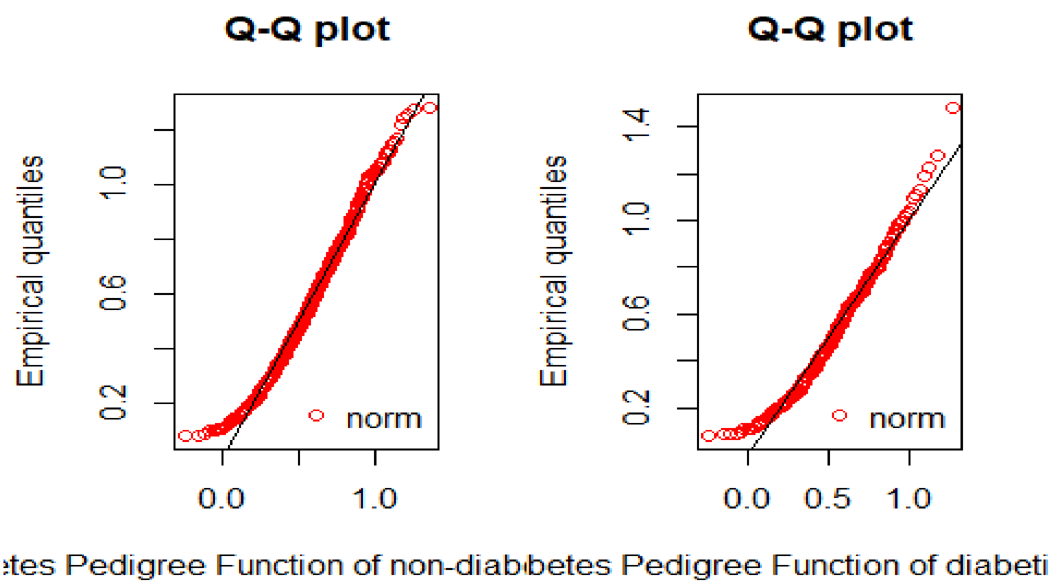
```
#7 Diabetes Pedigree Function
#Normality Test
kurtosis(df$DiabetesPedigreeFunction) #Calculate Kurtosis

## [1] -0.2946569

skewness(df$DiabetesPedigreeFunction) #Calculate Skewness

## [1] 0.452508

shapiro.test(df$DiabetesPedigreeFunction) #Significance Testing "Shapiro-Wilk
's test"

##
##  Shapiro-Wilk normality test
##
## data:  df$DiabetesPedigreeFunction
## W = 0.97697, p-value = 1.215e-09

#Plots
p1 <- ggplot(df, aes(x=df$DiabetesPedigreeFunction,fill=df$Outcome))+
  geom_histogram(alpha=0.35, position="identity",colour="white")+
  xlab("Diabetes Pedigree Function")+labs(fill="Outcome")+theme_classic()
p2 <- ggplot(df, aes(y=df$DiabetesPedigreeFunction,x=df$Outcome,fill=df$Outco
me))+
  geom_boxplot(outlier.shape=NA)+
  ylab("Diabetes Pedigree Function")+xlab("Diabetes")+labs(fill="Diabetes")+t
heme_classic()
grid.arrange(p1, p2, ncol = 2, nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Inference on Diabetes pedigree function

- The plot for diabetes pedigree function is right skewed.
- As we can see from the second plot the value of diabetes pedigree function is almost same for diabetic as well as non-diabetic women. It is difficult to say that this parameter has any effect on diabetes and needs to be statistically validated.
- There are many outliers in both the cases.

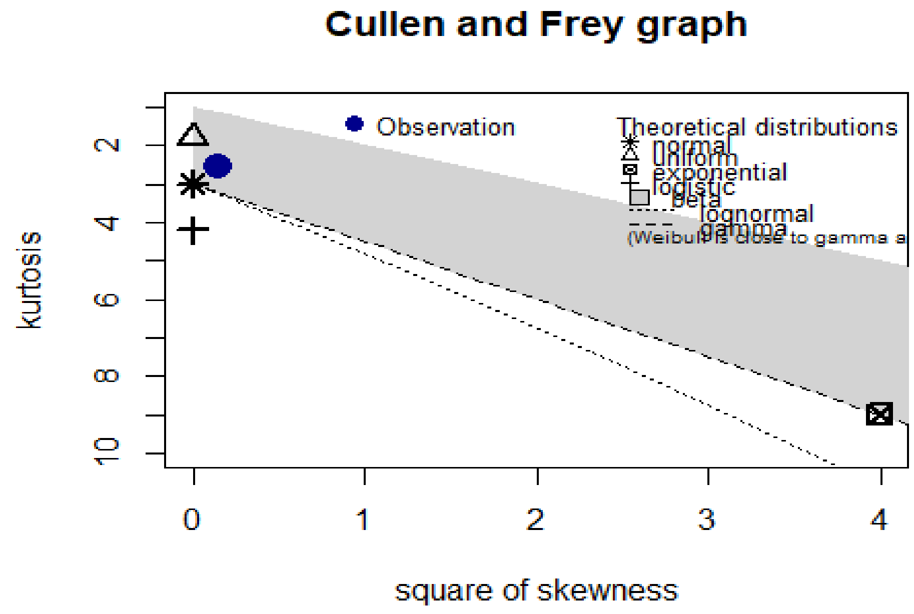
```
fit_dpf0 <- fitdist(pop0$DiabetesPedigreeFunction, "norm")
fit_dpf1 <- fitdist(pop1$DiabetesPedigreeFunction, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp(list(fit_dpf0), legendtext = plot.legend, xlab = 'Diabetes Pedigree Fu
nction of non-diabetic women', xlegend = 'bottomright')
qqcomp(list(fit_dpf1), legendtext = plot.legend, xlab = 'Diabetes Pedigree Fu
nction of diabetic women', xlegend = 'bottomright')
```



Normality testing

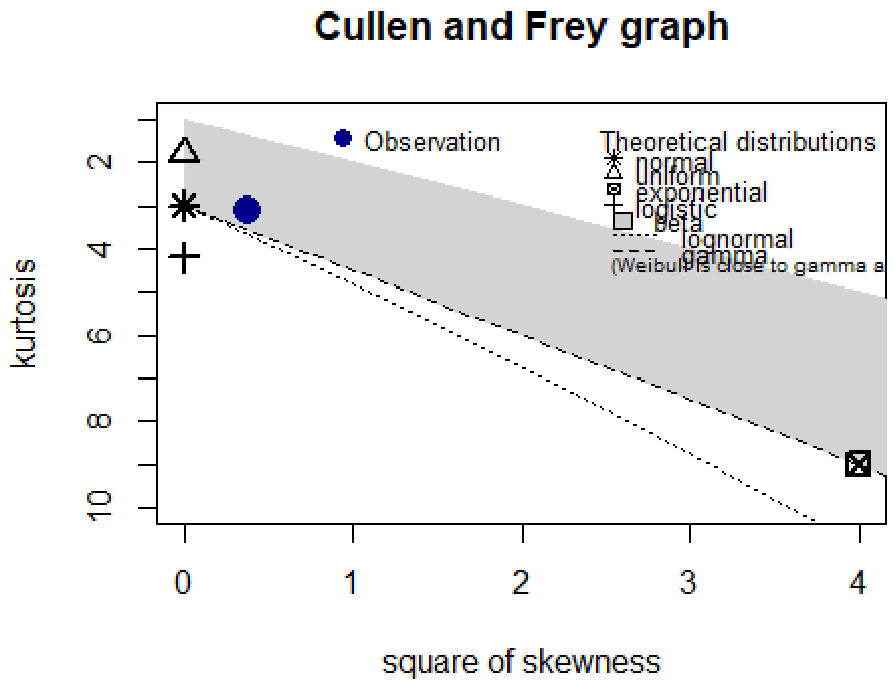
From the Q-Q probability plots for non-diabetic & diabetic PIMA India Women, it can be inferred that the observed values againt normally distributed data(represented by the line). Normally distributed data fall along the line.

```
#Goodness of fit
descdist(pop0$DiabetesPedigreeFunction)
```



```
## summary statistics
## ------
## min:  0.08082815   max:  1.282164
## median:  0.5276332
## mean:  0.5532705
## estimated sd:  0.2611969
## estimated skewness:  0.3815929
## estimated kurtosis:  2.55703
```

```
descdist(pop1$DiabetesPedigreeFunction)
```



```
## summary statistics
## ------
## min:  0.08206073   max:  1.481349
## median:  0.4681086
## mean:  0.5074465
## estimated sd:  0.2624959
## estimated skewness:  0.6085985
## estimated kurtosis:  3.128991
```

As can be observed from the plot, normal distribution as well as lognormal distribution are the closest to the distribution for DiabetesPedigreeFunction for both the cases.

```
#hypothesis testing
z_test(sample0$DiabetesPedigreeFunction, sample1$DiabetesPedigreeFunction, va
r(pop0$DiabetesPedigreeFunction), var(pop1$DiabetesPedigreeFunction))

## [1] 1.825984
```

Diabetes Pedigree Function hypothesis

We will check whether outcome is dependent on Diabetes Pedigree Function.
H0: μ1 - μ2 = 0
H1: μ1 - μ2 != 0
z-value = 2.219293
Thus, for a significance level of α = 0.05, we reject the null hypothesis since the z-value lies outside the range [−1.96, 1.96] and conclude that there is significant difference between the mean of Diabetes Pedigree Function of two population. So, we can say that Diabetes Pedigree Function has effect on diabetes.

```
#Confidence interval of mean for dpf of diabetic women
mean_sample<-exp(mean(sample1$DiabetesPedigreeFunction))
sd_pop<-exp(sd(pop1$DiabetesPedigreeFunction))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 1.161285

upper_limit

## [1] 2.180619
```

Confidence interval of mean for Diabetes Pedigree Function of diabetic women

Sample mean of diabetic women = 1.662086
Standard deviation of population of diabetic women = 1.300171
Sample size = 25
α = 0.05
After calculation we find that the Diabetes Pedigree Function of diabetic women lies between 1.15 and 2.17. So, if a woman is diabetic, we can say with 95% confidence that her Diabetes Pedigree Function lies in this range.

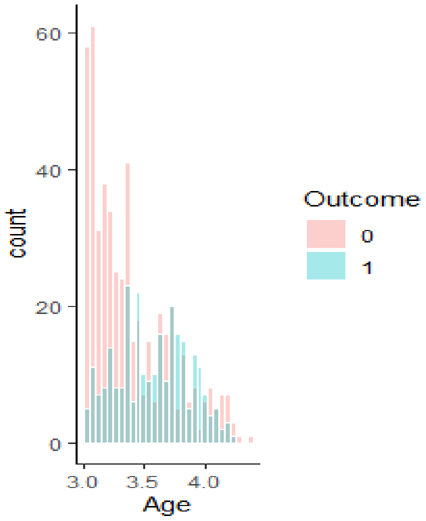
```
#Confidence interval of mean for dpf of non-diabetic women
mean_sample<-exp(mean(sample0$DiabetesPedigreeFunction))
sd_pop<-exp(sd(pop0$DiabetesPedigreeFunction))
size_sample<-25
z_alpha<--(round(qnorm(0.05/2),2))
error<-z_alpha*sd_pop/sqrt(size_sample)
lower_limit<-mean_sample-error
upper_limit<-mean_sample+error
lower_limit

## [1] 1.226216

upper_limit
```

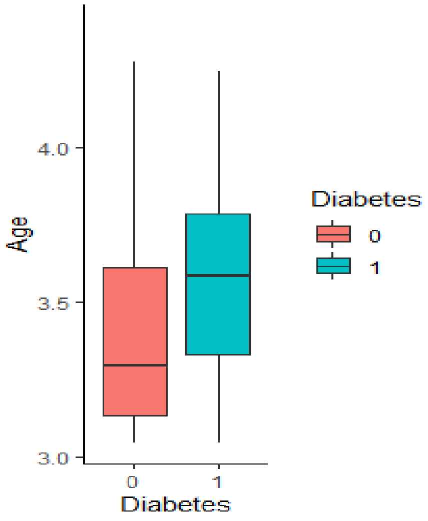
```
## [1] 2.244227
```

Confidence interval of mean for Diabetes Pedigree Function of non-diabetic women

Sample mean of non-diabetic women = 1.740104
Standard deviation of population of non-diabetic women = 1.298483
Sample size = 25
α = 0.05
After calculation we find that the Diabetes Pedigree Function of non-diabetic women lies between 1.231099 and 2.24911. So, if a woman is non-diabetic, we can say with 95% confidence that her Diabetes Pedigree Function lies in this range.

```
#8 Age
#Normality Test
kurtosis(df$Age) #Calculate Kurtosis

## [1] -0.6767798

skewness(df$Age) #Calculate Skewness

## [1] 0.5993976

shapiro.test(df$Age) #Significance Testing "Shapiro-Wilk's test"

##
##  Shapiro-Wilk normality test
##
## data:  df$Age
## W = 0.92853, p-value < 2.2e-16

#Plots
p1 <- ggplot(df, aes(x=df$Age,fill=df$Outcome))+
  geom_histogram(alpha=0.35, position="identity",colour="white")+
  xlab("Age")+labs(fill="Outcome")+theme_classic()
p2 <- ggplot(df, aes(y=df$Age,x=df$Outcome,fill=df$Outcome))+
  geom_boxplot(outlier.shape=NA)+
  ylab("Age")+xlab("Diabetes")+labs(fill="Diabetes")+theme_classic()
grid.arrange(p1, p2, ncol = 2, nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

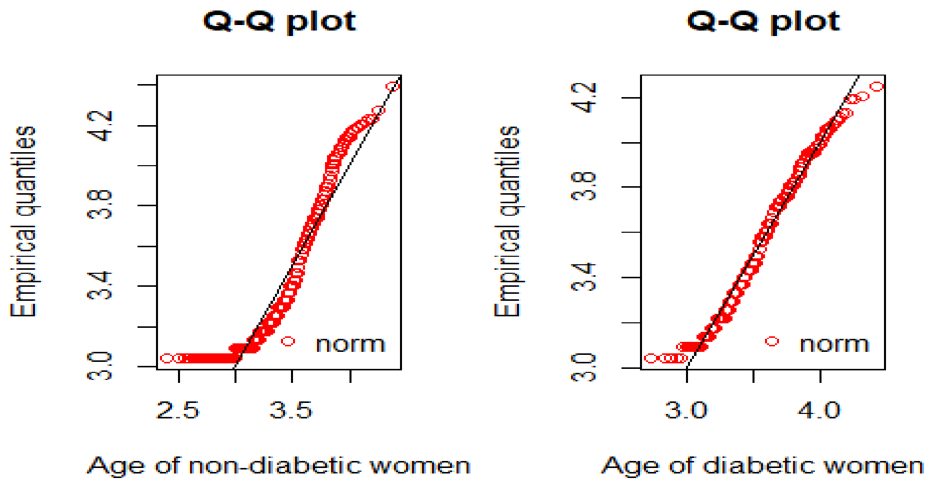




Inference on Age

- The tendency of having diabetes increases as age increases and this can be clearly seen from the segmented histograms.
- But diabetes, itself doesn't seem to have an influence of longevity. Maybe it impacts quality of life which is not measured in this data set. Median of age is higher in case of diabetic women.
- The distribution of age has skewness of 0.599 and kurtosis of -0.6767

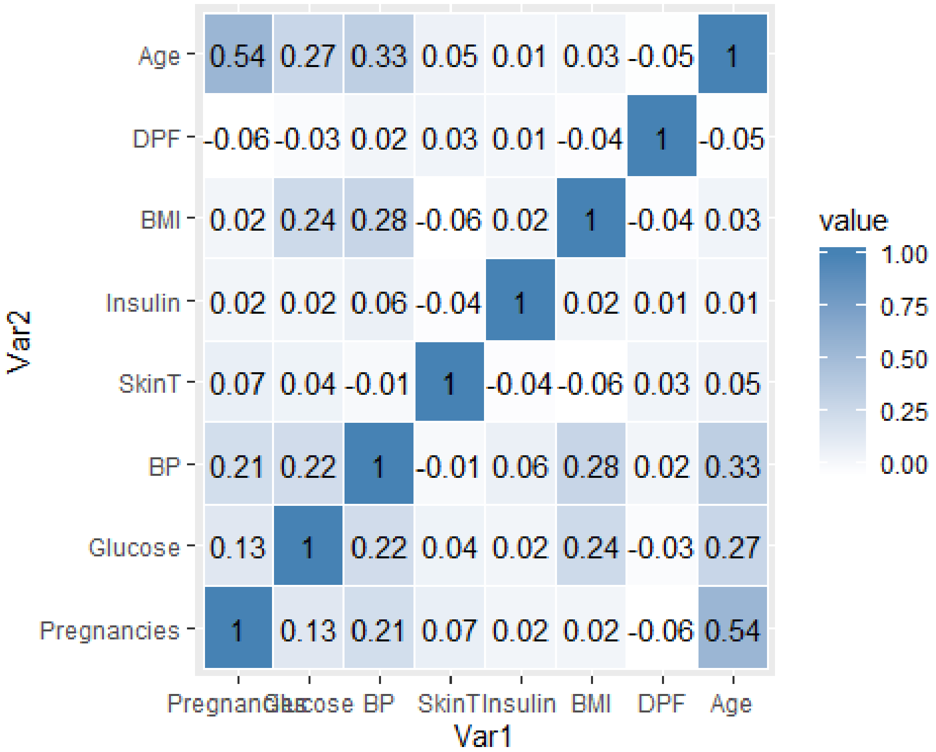
```
fit_age0 <- fitdist(pop0$Age, "norm")
fit_age1 <- fitdist(pop1$Age, "norm")
par(mfrow=c(1,2))
plot.legend <- c("norm")
qqcomp(list(fit_age0), legendtext = plot.legend, xlab = 'Age of non-diabetic
women', xlegend = 'bottomright')
qqcomp(list(fit_age1), legendtext = plot.legend, xlab = 'Age of diabetic wome
n', xlegend = 'bottomright')
```



Normality testing

From the Q-Q probability plots for non-diabetic & diabetic PIMA India Women, it can be inferred that the observed values againt normally distributed data(represented by the line). Normally distributed data fall along the line.

```
#correlation heat map
names(df1)[3] <- "BP"
names(df1)[4] <- "SkinT"
names(df1)[7] <- "DPF"

correlat <- round(cor(df1[, setdiff(names(df1), 'Outcome')]),2)
correlat1 <- melt(correlat)
ggplot(data = correlat1, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile( colour = "white") +
  scale_fill_gradient(low = "white",high = "steelblue")+
  geom_text(aes(label=value), size=4)
```



	Pregnancies	Glucose	BP	SkinT	Insulin	BMI	DPF	Age
Age	0.54	0.27	0.33	0.05	0.01	0.03	-0.05	1
DPF	-0.06	-0.03	0.02	0.03	0.01	-0.04	1	-0.05
BMI	0.02	0.24	0.28	-0.06	0.02	1	-0.04	0.03
Insulin	0.02	0.02	0.06	-0.04	1	0.02	0.01	0.01
SkinT	0.07	0.04	-0.01	1	-0.04	-0.06	0.03	0.05
BP	0.21	0.22	1	-0.01	0.06	0.28	0.02	0.33
Glucose	0.13	1	0.22	0.04	0.02	0.24	-0.03	0.27
Pregnancies	1	0.13	0.21	0.07	0.02	0.02	-0.06	0.54

Inference from 'r' values and heat map

- No two variables share strong linear relationships
- Age & Pregnancies have moderate positive linear relationship
- Rest of the combination of variables show low to zero linear relationship. Hence, we can say that the variables are independent of each other

Logistic Regression Model Building

A full model will be built with Outcome as the response variable with rest of the 8 variables. Stepwise variable selection method was used to identify the most important variables.

```
#Logistic Regression
set.seed(123)
model.glm = glm(Outcome~.,data = diabetes, family = binomial)
step_model = step(model.glm)
```

```
## Start:  AIC=748.42
## Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
##     Insulin + BMI + DiabetesPedigreeFunction + Age
##
##                            Df Deviance    AIC
## - SkinThickness             1   730.49 746.49
## - DiabetesPedigreeFunction  1   732.39 748.39
## <none>                          730.42 748.42
## - Insulin                   1   732.88 748.88
## - Age                       1   733.39 749.39
## - BloodPressure             1   736.63 752.63
## - Pregnancies               1   744.93 760.93
## - BMI                       1   776.59 792.59
## - Glucose                   1   857.40 873.40
##
## Step:  AIC=746.49
## Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +
##     DiabetesPedigreeFunction + Age
##
##                            Df Deviance    AIC
## - DiabetesPedigreeFunction  1   732.47 746.47
## <none>                          730.49 746.49
## - Insulin                   1   732.94 746.94
## - Age                       1   733.45 747.45
## - BloodPressure             1   736.70 750.70
## - Pregnancies               1   744.93 758.93
## - BMI                       1   776.92 790.92
## - Glucose                   1   857.42 871.42
##
## Step:  AIC=746.47
## Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +
##     Age
##
##                 Df Deviance    AIC
## <none>               732.47 746.47
## - Insulin        1   734.99 746.99
## - Age            1   735.61 747.61
## - BloodPressure  1   739.15 751.15
## - Pregnancies    1   747.56 759.56
## - BMI               1   779.92 791.92
## - Glucose           1   859.43 871.43
```

- The final model chosen with AIC as the criterion for selection generated a logistic regression model with the lowest AIC value of 739.45.52 as shown below.
- The important variables necessary for model building are – Insulin, Age, Blood Pressure, DPF, Pregnancies, BMI and Glucose.

```
# Iteration 1:
lrmodel1 <- glm(Outcome ~Pregnancies + Glucose + BloodPressure + Insulin + BM
I +
                  DiabetesPedigreeFunction + Age

                ,family = "binomial", data =diabetes)
summary(lrmodel1)

##
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +
##     Insulin + BMI + DiabetesPedigreeFunction + Age, family = "binomial",
##     data = diabetes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2038  -0.7233  -0.4228   0.7581   2.9084
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.408239   0.723883 -10.234  < 2e-16 ***
## Pregnancies               0.119181   0.031872   3.739 0.000184 ***
## Glucose                   0.034032   0.003408   9.987  < 2e-16 ***
## BloodPressure            -0.012713   0.005132  -2.477 0.013233 *
## Insulin                  -0.001827   0.001171  -1.561 0.118566
## BMI                       0.090116   0.014244   6.327  2.5e-10 ***
## DiabetesPedigreeFunction -0.499028   0.355992  -1.402 0.160976
## Age                       0.016024   0.009275   1.728 0.084070 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 730.49  on 760  degrees of freedom
## AIC: 746.49
##
## Number of Fisher Scoring iterations: 5
```

- After building another LR model based on these above variables, from the summary we can see the p-value for Insulin and Age is greater than 0.05, thus we remove for the next iteration.

```
lrmodel2 <- glm(Outcome ~ Pregnancies + Glucose + BloodPressure  + BMI +
                  DiabetesPedigreeFunction

                ,family = "binomial", data =diabetes)
summary(lrmodel2)
```

```
## 
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure + 
##     BMI + DiabetesPedigreeFunction, family = "binomial", data = diabetes)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2277  -0.7299  -0.4315   0.7625   2.9866  
## 
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)    
## (Intercept)              -7.375731   0.680657 -10.836  < 2e-16 ***
## Pregnancies               0.145801   0.027635   5.276 1.32e-07 ***
## Glucose                   0.035031   0.003362  10.420  < 2e-16 ***
## BloodPressure            -0.011521   0.005037  -2.287   0.0222 *  
## BMI                       0.087370   0.014086   6.202 5.56e-10 ***
## DiabetesPedigreeFunction -0.524585   0.354885  -1.478   0.1394    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 736.22  on 762  degrees of freedom
## AIC: 748.22
## 
## Number of Fisher Scoring iterations: 5
```

Finally, all the variables in the 2nd iteration show p-value less than 0.05. Therefore, the important factors are Pregnancies, Glucose, Blood Pressure, BMI and DPF. Thus, out of all the variables in the data, the variables achieved at the end of 2nd iteration show that these variables when combined and built a LR model, can produce the outcome together.

Summary for Logistic Regression

- Data set contains many zero values and they have been imputed on the basis of Outcome (0 and 1) and the cleaned data has been used for screening and logistic regression model building
- Approximately, 34% cases are diabetic and 66% are non-diabetic
- Visual screening of boxplot and categorized histogram shows that some of the field seem to affect the outcome (0 or 1)
- The hypothesis based on visual observations were confirmed using the statistical significance test (Hypothesis Testing)

• After plotting the correlation matrix, it is observed that the moderate correlation exists between some fields. Since we are using these variables for further model building this observation is critical, otherwise this will make the model's performance biased • The order of importance based on p-value is Glucose – BMI – Pregnancies – Blood Pressure – Diabetes Pedigree Function.

Conclusion

The PIMA Indian Women's diabetes data set was analyzed and explored in detail. Statistical validation was done using hypothesis testing for each parameter. Parameters that are responsible for diabetes were identified and then the interval of mean of those values were calculated. The interval of these dependent parameters were calculated seperately for diabetic and non-diabetic women. The result from the hypothesis testing showed that the major contributing factors for diabetes of PIMA Indian women are BMI, glucose level, blood pressure and diabetes pedigree function. The analysis also included studying bivariate relationship between variables using correlation and heat map. It was found that there is moderate correlation between pregnancy and BMI and weak correlation between rest of the rest of them. So, all the parameters are independent. The results of statistical analysis were verified using logistic regression. It was found that even with logistic regression the major contributing factors were same as that were found using statistical analysis. Both statistical anaylsis and classification model yielded the same result.

References

https://www.kaggle.com/uciml/pima-indians-diabetes-database

https://rpubs.com/jayarapm/PIMAIndianWomenDiabetes

https://www.collaborat.com/pima-diabetes-data-discovery-predictive-model/

https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16

https://en.wikipedia.org/wiki/Diabetes_mellitus

WHO technical report series (1985). Technical Report 727, WHO Study Group