02402
Introduction to Statistics

Project 1
BMI survey

Danmarks
Tekniske
Universitet

Zoltán György Varga

Contents

Descriptive analysis

a, Question

In the database we have 5 variables. The length of the database is 145 rows, it does not have any missing value. The height, weight and fastfood variables are continuous and the gender, urbanity variables are categorized variables. As Table 1. Shows us every variable is numerous variables, this means we do not have to make any dummy variable to make statistical analysis. Table one shows us the main descriptive statistical values such as the mean of the columns. Table 2 shows us type of the data in every column.

Table 1

	height	weight	gender	urbanity	fastfood
Min.	154.0	50.00	0.0000	1.000	0.000
1st Qu	166.0	65.00	0.0000	3.000	6.000
Median	173.0	75.00	1.0000	4.000	6.000
Mean	173.9	76.74	0.5034	3.669	21.040
3rd Qu.	182.0	87.00	1.0000	5.000	24.000
Max.	196.0	130.00	1.0000	5.000	365.000

Table 2

data.frame': 145 obs. of 5 variables:											
height	int	180	185	180	168	173	161	168	166	181	172
weight	int	80	98	80	60	83	78	82	58	90	52
gender	int	1	1	1	0	1	0	0	0	1	0
urbanity	int	5	1	5	4	5	3	2	2	5	1
fastfood	num	24	6	6	24	24	6	6	24	6	6

After a short descriptive analysis using the following R code, I created the BMI score, which was the basic value of the further investigation.

```
D$bmi <- D$weight/(D$height/100)^2
```

b, Question

```
hist(D$bmi, xlab="BMI",main = "Histogram of:
BMI",col=7,prob=TRUE,xlim=c(mini,maxi))
```

With the previous R code I created a histogram out of the calculated BMI scores. The histogram is Figure 1. As we can see on Figure 1 the distribution of the data is very likely normal distribution. How we can assume that. The highest density of the data is around the mean of the data and around it we have decreasing values. As we can see on Figure 1 the BMI data cannot be negative.

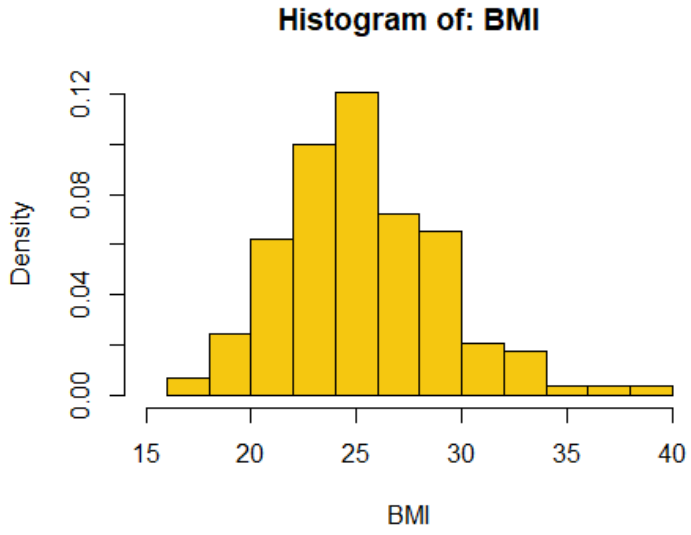

Figure 1

After I created the histogram, I have created two subset out of the original data with the following R code. The two subset is the data split to two group according to the gender of the person.

```
Dfemale <- subset(D, gender == 0)
Dmale <- subset(D, gender == 1)
```

c, Question

```
hist(Dfemale$bmi, xlab="BMI (female)", prob=TRUE,col=2,main ="Histogram of: BMI Female", xlim=c(mini,maxi))
```

```
hist(Dmale$bmi, xlab="BMI (male)", prob=TRUE,col=4,main ="Histogram of: BMI Male", xlim=c(mini,maxi))
```

With the previous R code I created the histograms out of the calculated Female BMI scores. The histogram is Figure 2. As we can see on Figure 2 the distribution of the data is very likely normal distribution. How we can assume that. The highest density of the data is around the mean of the data and around it we have decreasing values. As we can see on Figure 2 the BMI data can not be negative. It is interesting that there is a hiatus in the data around 37 the reason of it must be the division we use on the data. Figure 3 shows us the histogram which belong to the males. I could say the same statements about Figure 3 data. But the hiatus in this case is around 35.

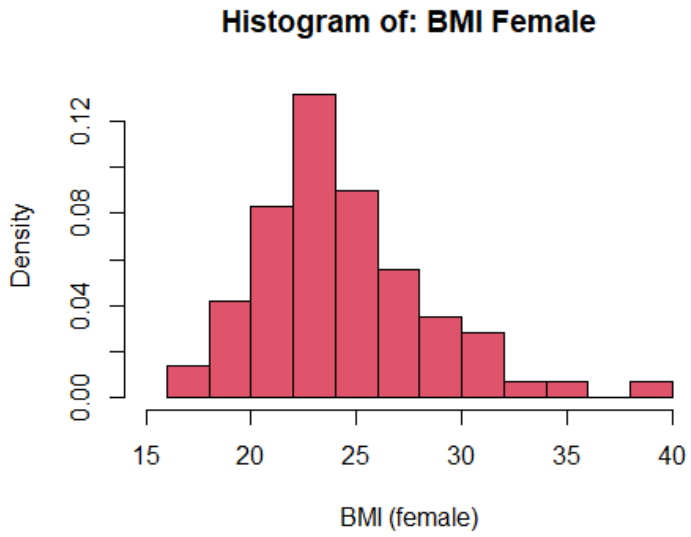

Figure 2

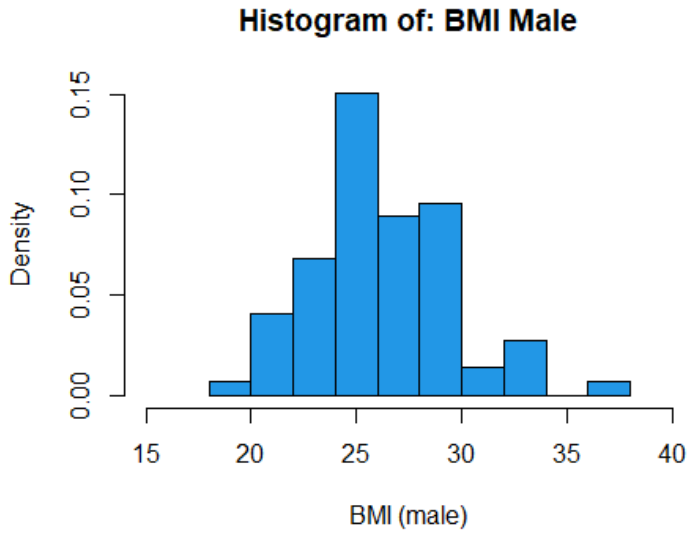

Figure 3

d, Queston

```
boxplot(Dfemale$bmi, Dmale$bmi,col = c(2,4), names=c("Female", "Male"),
xlab="Gender", ylab="BMI")
```

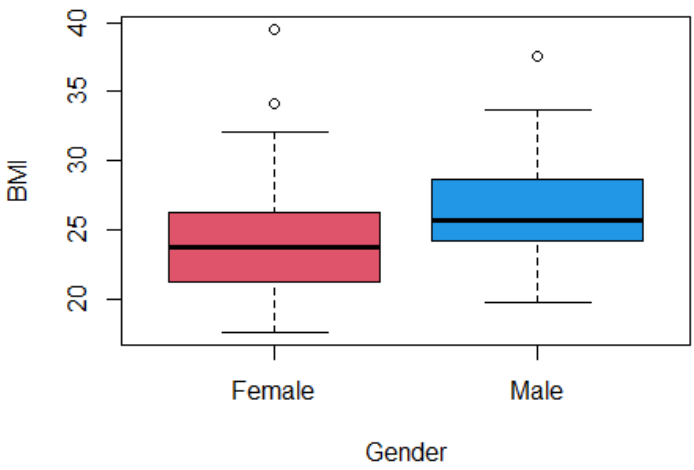

Figure 4

According to the Figure 4 the distribution of the Female is normal distribution, the Male distribution is also look like a normal distribution, but the number of observations do not distribute equally around the median. So the Famale is distributed symmetrical and the Male distributed skewed. As we can see both of the data contain extreme observations. The Female has more extreme observation than the Male.

e, Question

Table 3

	Number of obs.	Sample mean	Sample variance	Sample std. dev.	Lower quartile	Median	Upper Quartile
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
Everyone	145.000	25.248	14.686	3.832	22.589	24.691	27.636
Women	72.000	24.216	16.418	4.052	21.259	23.689	26.291
Men	73.000	26.265	11.069	3.327	24.152	25.726	28.634

Table 3 shows us the number of observations in Men is more than the number of observations in Woman. Table 3 shows us the values in Men has lower sample variance and bigger mean as the Women. This is a plus information compared to the boxplot.

Statistical analysis

f, Question

We assume that the log-transform BMI has log-normal distribution because the log and it means that the model can be interpreted as a normal distributed model. The formula for this model is $LN(\alpha, \beta^2)$ where $\alpha = the\ mean\ of\ the\ log\ values$ and $\beta = the\ standerd\ deviation\ of\ the \log values$. Because of this I calculated the $\alpha\ and\ \beta$ values using the following equation.

$$\alpha = \frac{\sum_{i=1}^{n} x_{logbmi}}{n} = 3.218$$

$$\beta = \sqrt{\frac{\sum_{i=1}^{n} (x_{logbmi} - \bar{x})^2}{n}} = 0.148$$

$$\mu = e^{\alpha + \beta^2/2} = e^{3.218 + {0.148^2}/{2}} = 25.247$$

$$\sigma = e^{2*\alpha+\beta^2} * \left(e^{\beta^2} - 1\right) = e^{2*3.218+0.148^2} * \left(e^{0.148^2} - 1\right) = 14.286$$

And the model for the log-normal distributed logarithmic BMI scores is the following $L(3.218, 0.148^2)$

Using the following R code I created Figure 5 which show us that the assumption were right about the normal distribution the image of the plot is almost a straight line.

```
D$logbmi <- log(D$bmi)
qqnorm(D$logbmi)
qqline(D$logbmi)
```

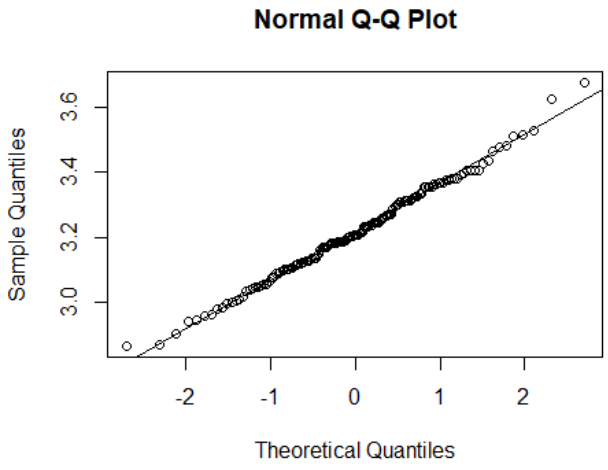

Figure 5

Using the following R code I created Figure 6 which show us the ECDF plot of the log BMI. The plot give us more ground to say that the distribution is normal distribution.

```
plot(ecdf(D$logbmi), verticals = TRUE,main='ecdf:BMI')
xseq <- seq(0.9*min(D$logbmi), 1.1*max(D$logbmi), length.out = 100)
lines(xseq, pnorm(xseq, mean(D$logbmi), sd(D$logbmi)),col=2)
```

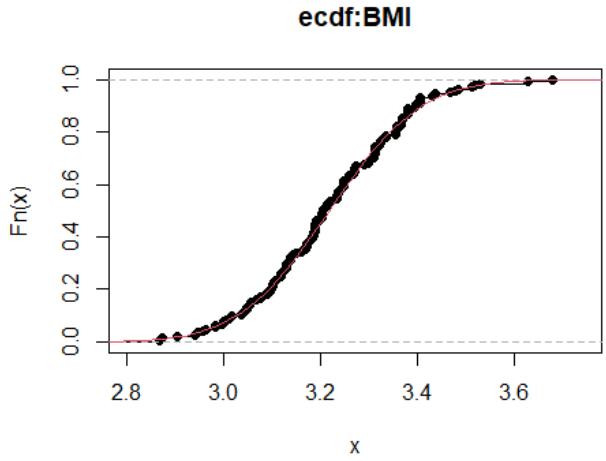

Figure 6

Lastly execute normal Q-Q plot to 9 random but similar cases. The Figure 7 shows us the results. As we can see the result is very similar to our original case- This mean our assumption about the normal distribution were right.

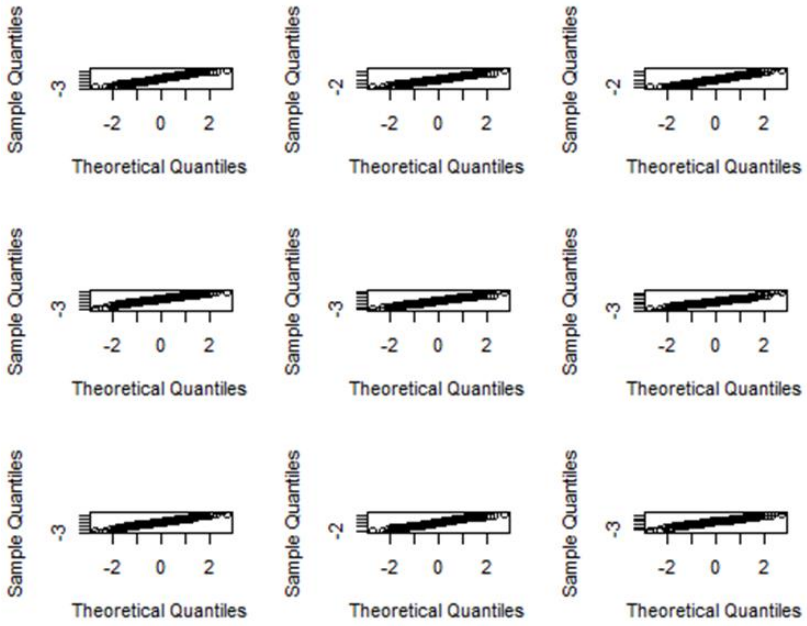

Figure 7

The Central Limit Theorem can be used when we assume that the distribution is differ from normal distribution. The Central Limit Theorem assumes that if the n number of observation is big enough the mean of the X observations is independent from the distribution of X population.

In our case where the distribution is normal distribution. It is not important to use the Central Limit Theorem

g, Question

The formula to calculate the confidence interval for the mean is the following.

$CI_{95\%} = \bar{x} \pm t_{1-\alpha/2} * \frac{s}{\sqrt{n}}$ where $t_{1-\alpha/2}$ is the (1 − α/2) quantile from the t-distribution with n – 1 degrees of freedom

$$3.218 \pm 1.976 * \frac{\sqrt{14.286}}{\sqrt{145}} = [24.627,25.868]$$

The following R code shows us the result of the t-test which execute the same calculation.

```
t.test(D$logbmi)

##
## One Sample t-test
##
## data: D$logbmi
## t = 260.25, df = 144, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 3.193203 3.242078
## sample estimates:
## mean of x
## 3.217641
```

h, Question

First I defined the μ_0 value it is log (25) as was stated in the question. I calculated the t_{obs}. The formula of t_{obs}.is the following:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.217 - \log(25)}{0.149/\sqrt{145}} = -0.0999$$

With the value of t_{obs} I can calculate the p value, which show us the null hypothesis can be rejected or not.

I calculated the probability cumulative density of the Student t-distribution on the t_{obs} with 144 degrees of freedom the value is 0.539.

$$p - value = 2 * P(T > |t_{obs}|) = 2 * (1 - 0.539) = 0.9205$$

The p value is almost 1 which means we can not reject the null hypothesis. The following R code show us the same result which means the calculation were correct.

```
t.test(D$logbmi,mu=log(25))

##
## One Sample t-test
##
## data: D$logbmi
## t = -0.099913, df = 144, p-value = 0.9206
```

```
## alternative hypothesis: true mean is not equal to 3.218876
## 95 percent confidence interval:
## 3.193203 3.242078
## sample estimates:
## mean of x
## 3.217641
```

i, Question

We assume that the log-transform BMI Male has log-normal distribution because the log and it means that the model can be interpreted as a normal distributed model. The formula for this model is $LN(\alpha, \beta^2)$ where $\alpha = the\ mean\ of\ the\ log\ values$ and $\beta = the\ standerd\ deviation\ of\ the\ \log values$. Because of this I calculated the $\alpha\ and\ \beta$ values using the following equation.

$$\alpha = \frac{\sum_{i=1}^{n} x_{logbmiMale}}{n} = 3.261$$

$$\beta = \sqrt{\frac{\sum_{i=1}^{n}(x_{logbmiMale} - \bar{x})^2}{n}} = 0.124$$

$$\mu = e^{\alpha+\beta^2/2} = e^{3.261+{0.124^2}/{2}} = 26.266$$

$$\sigma = e^{2*\alpha+\beta^2} * \left(e^{\beta^2} - 1\right) = e^{2*3.261+0.124^2} * \left(e^{0.124^2} - 1\right) = 10.674$$

And the model for the log-normal distributed logarithmic BMI scores is the following $L(3.261, 0.124^2)$

Using the following R code I created Figure 7 which show us that the assumption were right about the normal distribution the image of the plot is almost a straight line.

```
Dmale$logbmi <- log(Dmale$bmi)
qqnorm(Dmale$logbmi)
qqline(Dmale$logbmi)
```

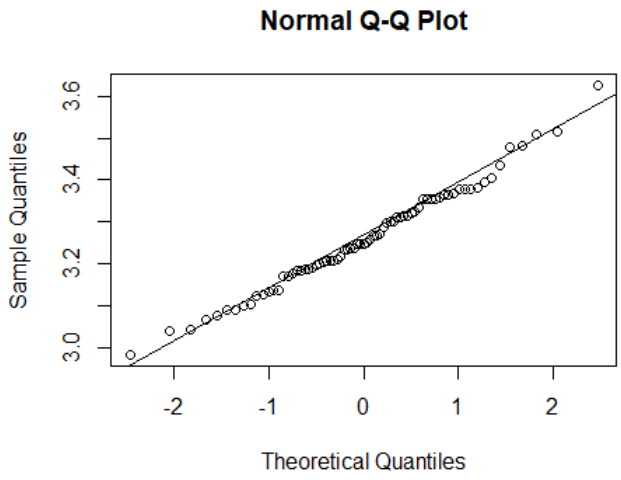

Figure 8

Using the following R code I created Figure 8 which show us the ECDF plot of the log BMI Male. The plot give us more ground to say that the distribution is normal distribution.

```
plot(ecdf(Dmale$logbmi), verticals = TRUE,main='ecdf:BMI Male')
xseq <- seq(0.9*min(Dmale$logbmi), 1.1*max(Dmale$logbmi), length.out = 100)
lines(xseq, pnorm(xseq, mean(Dmale$logbmi), sd(Dmale$logbmi)),col=2)
```

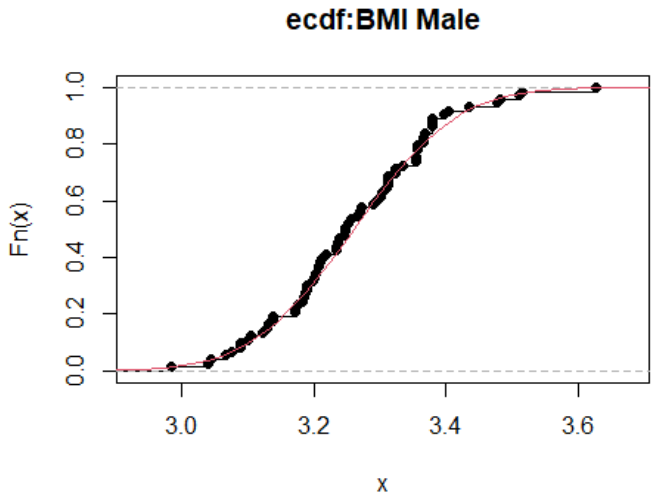

Figure 9

Lastly execute normal Q-Q plot to 9 random but similar cases. The Figure 10 shows us the results. As we can see the result is very similar to our original case- This mean our assumption about the normal distribution were right.

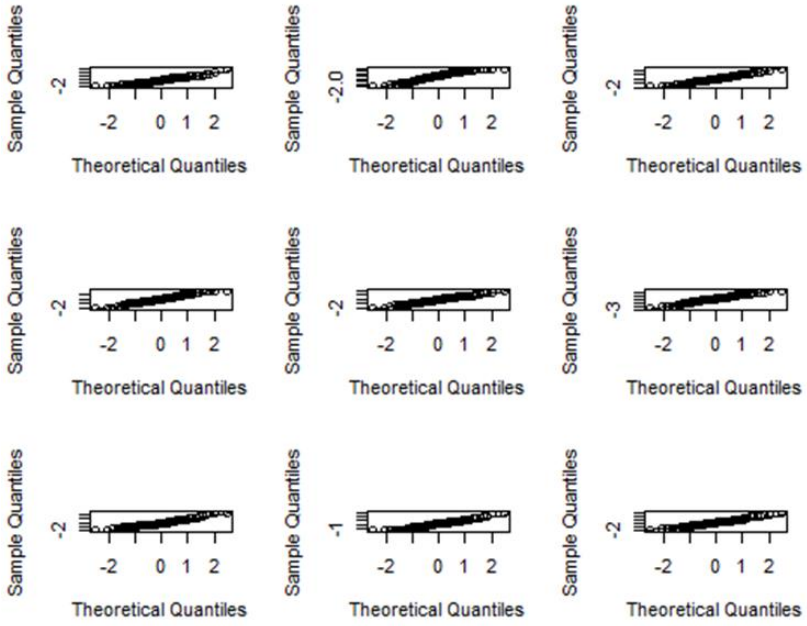

Figure 10

```
Dfemale$logbmi <- log(Dfemale$bmi)
qqnorm(Dfemale$logbmi)
qqline(Dfemale$logbmi)
```

We assume that the log-transform BMI has log-normal distribution because the log and it means that the model can be interpreted as a normal distributed model. The formula for this model is $LN(\alpha, \beta^2)$ where $\alpha = the\ mean\ of\ the\ log\ values$ and $\beta = the\ standerd\ deviation\ of\ the\ \log values$. Because of this I calculated the $\alpha\ and\ \beta$ values using the following equation.

$$\alpha = \frac{\sum_{i=1}^{n} x_{logbmiFemale}}{n} = 3.174$$

$$\beta = \sqrt{\frac{\sum_{i=1}^{n}(x_{logbmiFemale} - \bar{x})^2}{n}} = 0.159$$

$$\mu = e^{\alpha+\beta^2/2} = e^{3.174+0.159^2/2} = 24.213$$

$$\sigma = e^{2*\alpha+\beta^2} * \left(e^{\beta^2} - 1\right) = e^{2*3.174+0.159^2} * \left(e^{0.159^2} - 1\right) = 15.180$$

And the model for the log-normal distributed logarithmic BMI scores is the following $L(3.174, 0.159^2)$

Using the following R code I created Figure 9 which show us that the assumption were right about the normal distribution the image of the plot is almost a straight line.

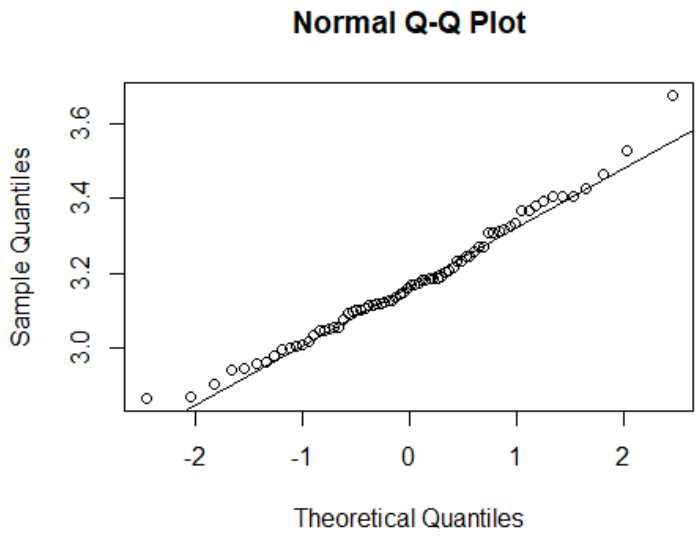

Figure 11

Using the following R code I created Figure 10 which show us the ECDF plot of the log BMI Female. The plot give us more ground to say that the distribution is normal distribution.

```
plot(ecdf(Dfemale$logbmi), verticals = TRUE,main='ecdf:BMI Female')
xseq <- seq(0.9*min(Dfemale$logbmi), 1.1*max(Dfemale$logbmi), length.out =
100)
lines(xseq, pnorm(xseq, mean(Dfemale$logbmi), sd(Dfemale$logbmi)),col=2)
```

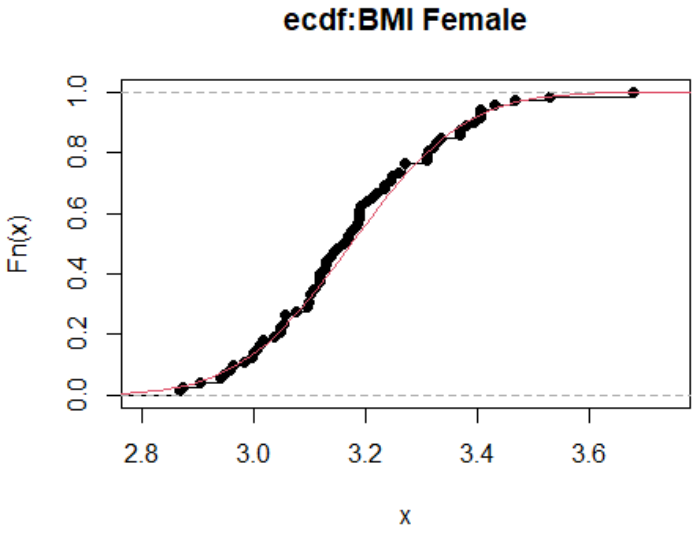

Figure 12

Lastly execute normal Q-Q plot to 9 random but similar cases. The Figure 13 shows us the results. As we can see the result is very similar to our original case- This mean our assumption about the normal distribution were right.

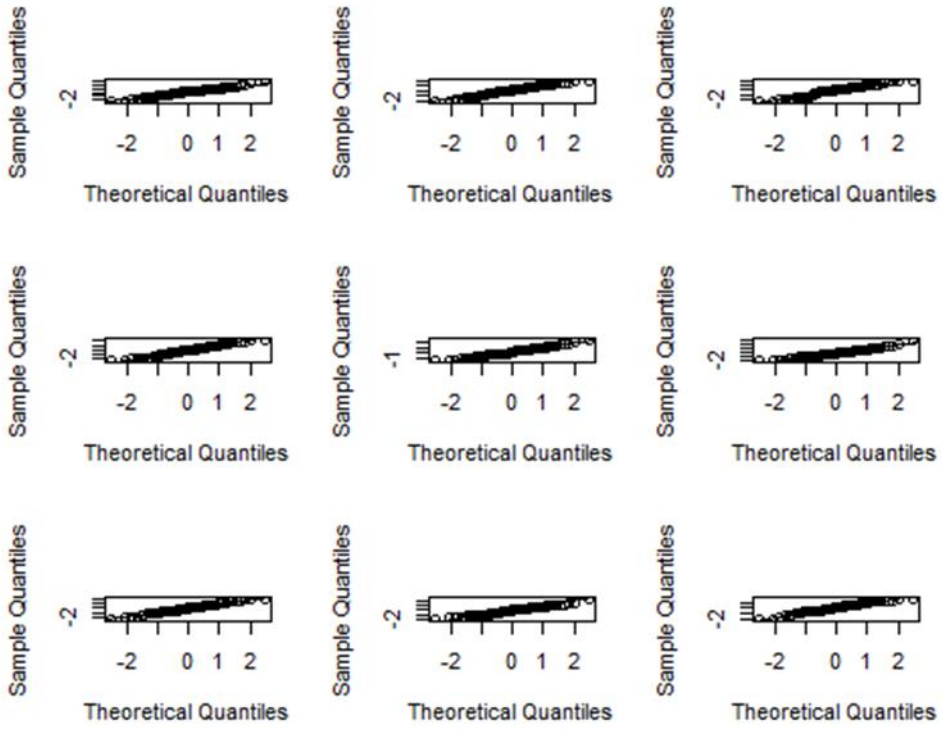

Figure 13

j, Question

Table 4

	Lower bound of CI	Upper bound of CI
Women	23.297	25.503
Men	25.128	27.028

The formula to calculate the confidence interval for the mean is the following. The result can be seen in Table 4.

$CI_{95\%} = \bar{x} \pm t_{1-\alpha/2} * \frac{s}{\sqrt{n}}$ where $t_{1-\alpha/2}$ is the (1 – α/2) quantile from the t-distribution with n – 1 degrees of freedom

$$t_{95\%Woman} = 1.9939\ degrees\ of\ freedom = 71$$

$$t_{95\%Men} = 1.9934\ degrees\ of\ freedom = 72$$

$$24.213 \pm 1.9939 * \frac{\sqrt{15.180}}{\sqrt{72}} = [23.297, 25.503]$$

$$26.266 \pm 1.9934 * \frac{\sqrt{10.674}}{\sqrt{73}} = [25.128, 27.028\ ]$$

The following R code generate the confidence interval for the women. The generated values is the same, therefore the values in Table 4 should be correct.

```
Dfemale <- subset(D, gender == 0)
KI <- t.test(Dfemale$logbmi, conf.level=0.95)$conf.int
KI

## [1] 3.136525 3.211669
## attr(,"conf.level")
## [1] 0.95

exp(KI)

## [1] 23.02372 24.82047
## attr(,"conf.level")
## [1] 0.95
```

k, Question

In this case I had to make a two sample hypothesis test which is different of the previously presented one sample hypothesis test. First I defined the μ_0 value it is 0 because we want to compare two data to each other. I calculated the t_{obs}. The formula of t_{obs}.in this case is the following:

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.174 - 3.261) - 0}{\sqrt{\frac{0.0255^2}{72} + \frac{0.0153^2}{73}}} = 0.00039$$

With the value of t_{obs} I can calculate the p value, which show us the null hypothesis can be rejected or not.

I calculated the probability cumulative density of the Student t-distribution on the t_{obs} with the calculated degree of freedom.

$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\frac{0.0255^2}{72} + \frac{0.0153^2}{73}}{\frac{(0.0255^2/72)^2}{72 - 1} + \frac{0.0153^2/73)^2}{73 - 1}} = 133.750$$

$$p - value = 2 * P(T > |t_{obs}|) = 2 * (1 - 0.9998) = 0.000391$$

The p value is almost 1 which means we can reject the null hypothesis. The following R code show us the same result which means the calculation were correct.

```
t.test(D$logbmi[D$gender == 0], D$logbmi[D$gender == 1])

##
## Welch Two Sample t-test
##
## data: D$logbmi[D$gender == 0] and D$logbmi[D$gender == 1]
## t = -3.6375, df = 133.75, p-value = 0.000392
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.13352005 -0.03946156
## sample estimates:
## mean of x mean of y
## 3.174097 3.260588
```

l, Question

It was necessary to carry out the hypothesis test because the two confidence interval have overlapped. This means we cannot stated that the two value do not have any connection and the connection need more investigation.

Correlation

m, Question

The formula to calculate the correlation between two variable is the following:

$$\rho_{xy} = \frac{\delta_{xy}}{\delta_x * \delta_y} = \frac{\sum_{i=1}^{n}(x - \bar{x}) * (y - \bar{y}) * \frac{1}{n}}{\delta_x * \delta_y} = \frac{47.939}{15.208 * 3.832} = 0.828261$$

The value of the correlation mans that there is a strong positive connection between the two values.

```
plot(x = D$weight,y = D$bmi,main = 'Correalations Weight-BMI',xlab =
'Weight',ylab = 'BMI',col='green')
```

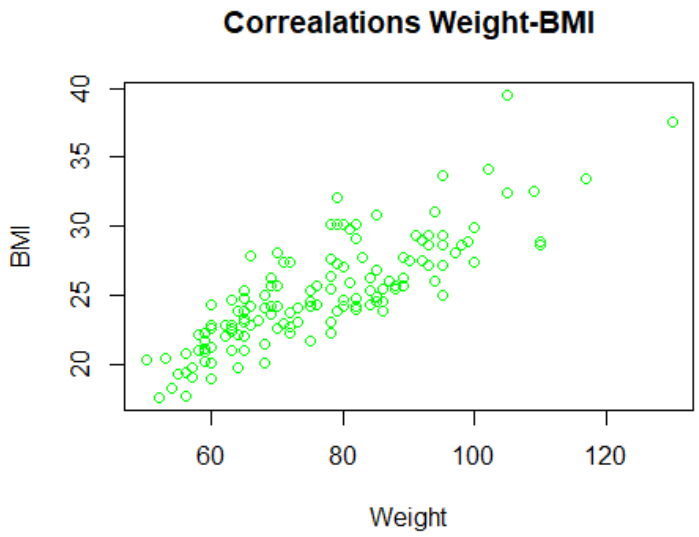

Figure 14

```
plot(x = D$fastfood,y = D$bmi,col='blue',main = 'Correalations Fastfood-
BMI',xlab = 'Fastfood',ylab = 'BMI')
```

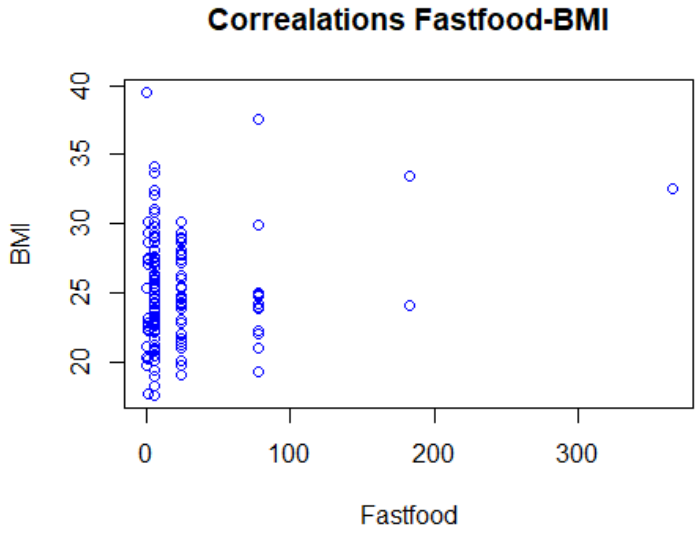

Figure 15

```
plot(y = D$weight,x = D$fastfood,col='red',main = 'Correalations Weight-
Fastfood',xlab = 'Weight',ylab = 'Fastfood')
```

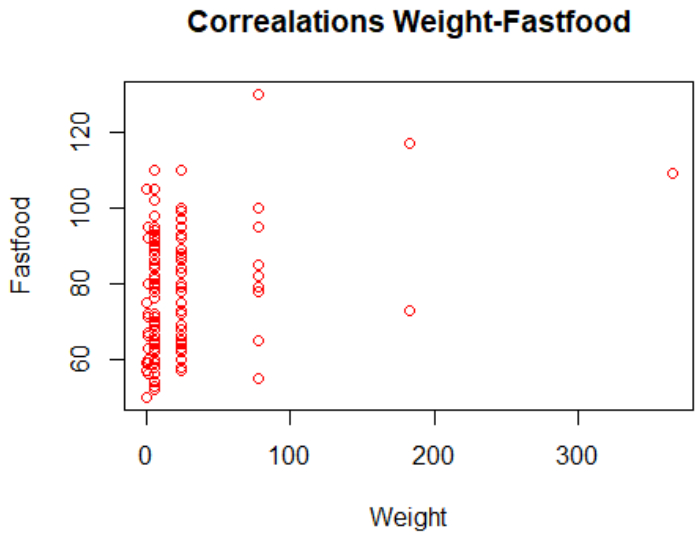

Figure 16

Figure 11, Figure 12, Figure 13 shows the scatter plot of two values in the database. With the plot I would like to find out the expected correlation between the values. According to Figure 11 there is a strong correlation but taking a look on Figure 12 and 13 I expect low correlation. The following R code just show evidence to my assumption.

```
cor(D[,c("weight","fastfood","bmi")], use="pairwise.complete.obs")
```

Table 5

	weight	fastfood	bmi
weight	1.0000000	0.2793223	0.8282610
fastfood	0.2793223	1.0000000	0.1531578
bmi	0.8282610	0.1531578	1.0000000

Table 5 shows us the correlation matrix between the selected values.