Danmarks Tekniske Universitet

DTU

Project 1 - Introduction to Statistics 02402

Author

Jacopo Ceccuti s215158

October 11, 2022

1 Descriptive analysis

1.1 a)

The variables included in the data set are: height, weight, gender, urbanity and fast food eaten in a year. Some of them are strictly quantitative like height and weight and some of them are categorized in the following way:

- Gender : can be either 1 (male) or 0 (female)
- Urbanity: represents the size of the city in which the person lives in (has a range of values form 1 to 5).
- Fast food: represents the fast food eaten in a year (has preset values depending on how many days a week the person gets fast food).

The total observations are 145, with 5 different variables. Please note: the observations 125 and 142 have an incorrect value for the variable fast food, from the appendix it should be 182 but in the .csv file is 182.5.

1.2 b)

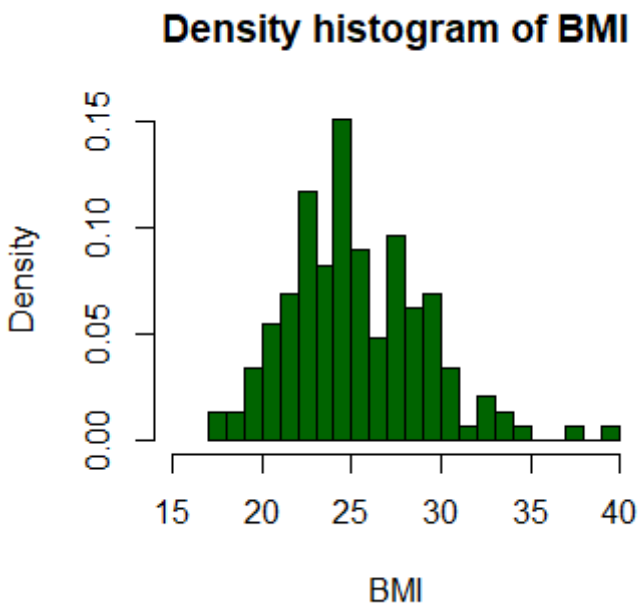

Figure 1: Density histogram of BMI

The empirical density is represented in the histogram above 1. From here it is easy to tell that the empirical distribution is not symmetrical and tends to be right-skewed. The

majority of people have a BMI that lies in between 20 and 26, while there is a big range of variation since some values are close to 17 (severely underweight) and others to 40 (stage 2 obesity). Obviously the BMI can not be negative because that will imply either a negative weight or height.

1.3 c)

1.3.1 Males

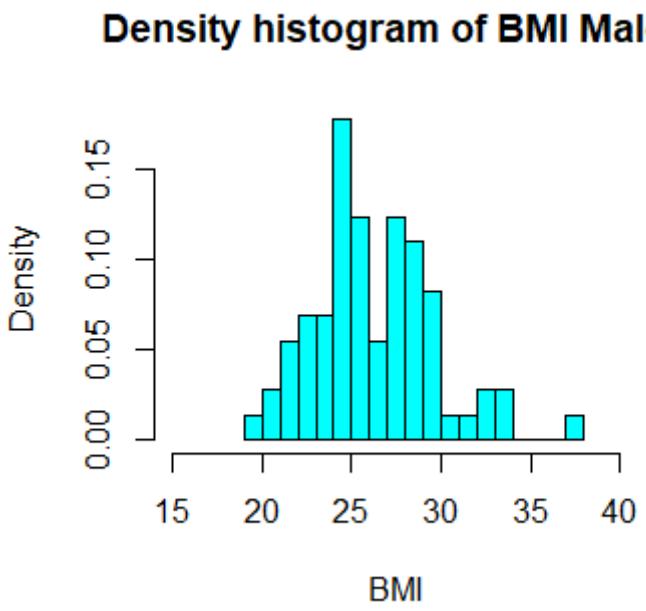

Figure 2: Density histogram of BMI Males

The empirical density for males, represented in the histogram above 2, is more symmetrical than the one found for the whole data set 1. The most amount of people lie in between the BMI values of 23 and 28 and there are not so many discrepancy (it spreads from 19 to 33 BMI) except for an outlier that has 38 of BMI. There still are a lot of big variations (with some cases of obesity) but slightly reduced compared to the gender mixed one 1.

1.3.2 Females

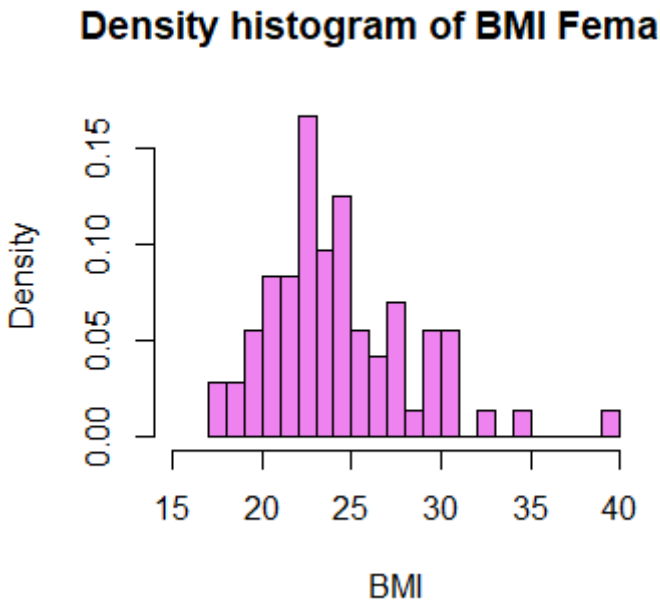

Figure 3: Density histogram of BMI Females

The empirical density for females is not symmetric at all, it is easy to notice that most people lie in the zone that goes from 20 to 25 BMI, which is a bit lower than males. The general discrepancy however is more amplified with BMI values that go form 17 to 40 (this last value is an outlier of type 2 obesity). In general it is possible to say that the female plot represents better the general density histogram 1, this is caused by the wider variation that can be observed in the females data.

1.4 d)

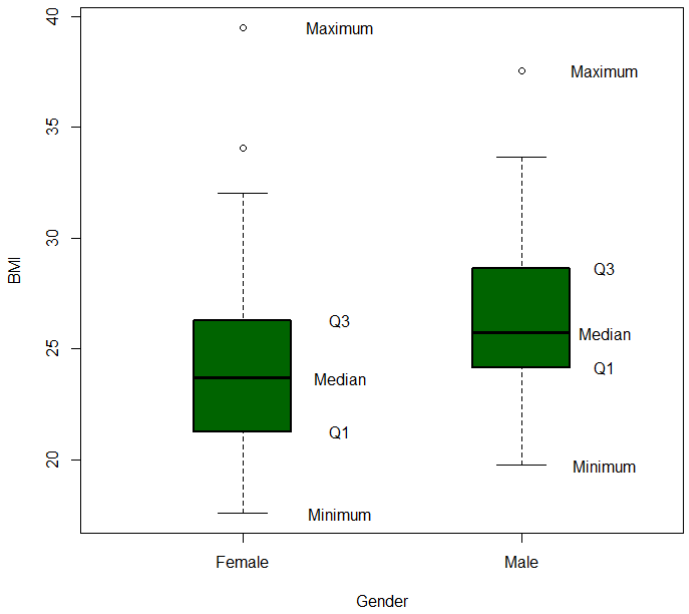

Figure 4: Box-plot of BMI scores by gender

The box plot 4 helps us to point out where the quartiles, the median and the outliers are located in the distributions. It is not possible to say that the distributions are symmetrical (almost skewed); but they can become slightly symmetrical not considering the outliers. The difference between the two distribution is that the male one 2 is more compact with less variation respect to the female one 3. Talking about outliers it is possible to identify two of them for the females and one of them for the males (mentioned above).

1.5 e)

Compared to the box plot 4 using the table 1 it is possible the see the standard deviation and the variance for the general distribution and for each subset.

Variable *BMI*	**N. of obs.**	**Sample mean**	**Sample var.**	**Sample std. dev.**	**Lower quart.**	**Median**	**Upper quart**
	n	$(\overline{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
Everyone	145	25.25	14.69	3.83	22.59	24.69	27.64
Women	72	24.22	16.42	4.05	21.26	23.69	26.29
Men	73	26.27	11.07	3.33	24.15	25.73	28.63

Table 1: Table with the required values for the data set

2 Statistical analysis

2.1 Confidence intervals and hypothesis tests

2.1.1 f)

The sample was provided, data were took randomly form the population and they have been logged. It is possible to prove that they are identically normal distributed and we will do that at the end of this subsection. In this first part the data make no gender distinction. The model is described by:

$$X_i \sim N(\mu, \sigma^2) \text{ where } i = 1, 2, .., n$$

Hence we have 145 independent variables.
Our goal is to learn about the mean, the variance and the standard deviation of the population and find their precision. Using the already given values and taking the log of them we get: In order to perform model validation we have to look at different criteria such has:

Variable *BMI*	**N. of obs.**	**Sample mean**	**Sample var.**	**Sample std. dev.**
	n	$(\overline{x})$	(s^2)	(s)
Everyone	145	3.218	0.1489	0.0222

Table 2: Table with parameters of the model

histograms and their shape, empirical c.d.f., Q-Q plot and wally plot. In this case we will use Q-Q plots and the wally plot.
Starting with two Q-Q plots in order: one with the normal values of BMI and the other with

the logged values. It is already possible to see a bigger tendency to be normally distributed in the one where the logged values have been used 6.
Also the wally plot is used.

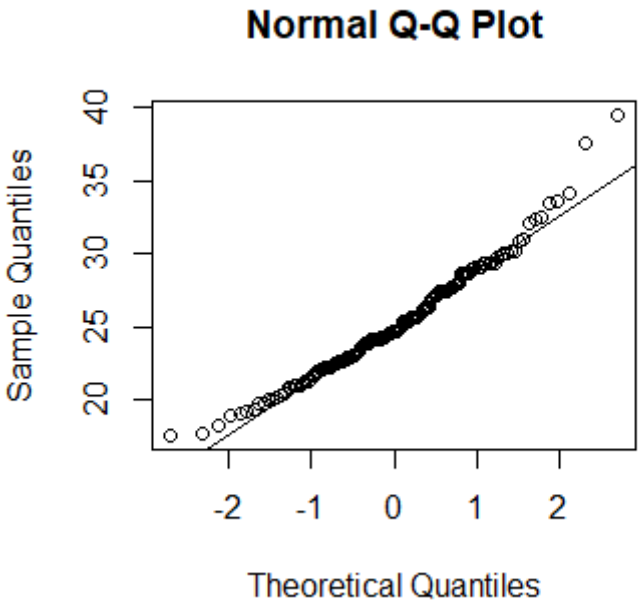

Figure 5: Q-Q plot of BMI scores

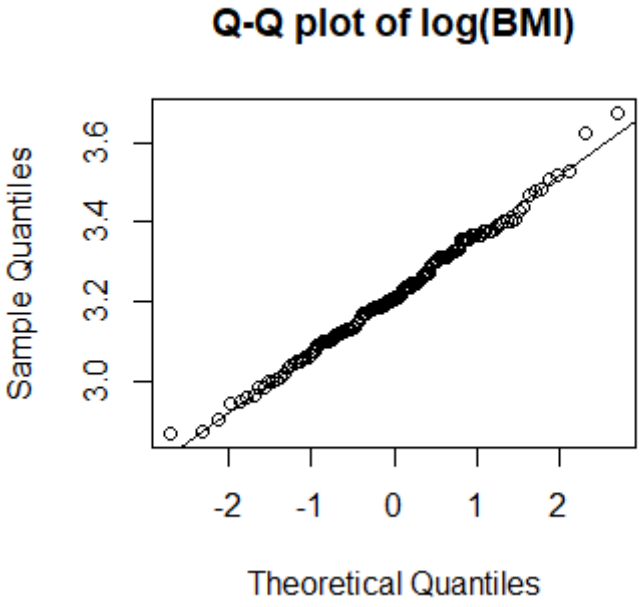

Figure 6: Q-Q plot of logged BMI scores

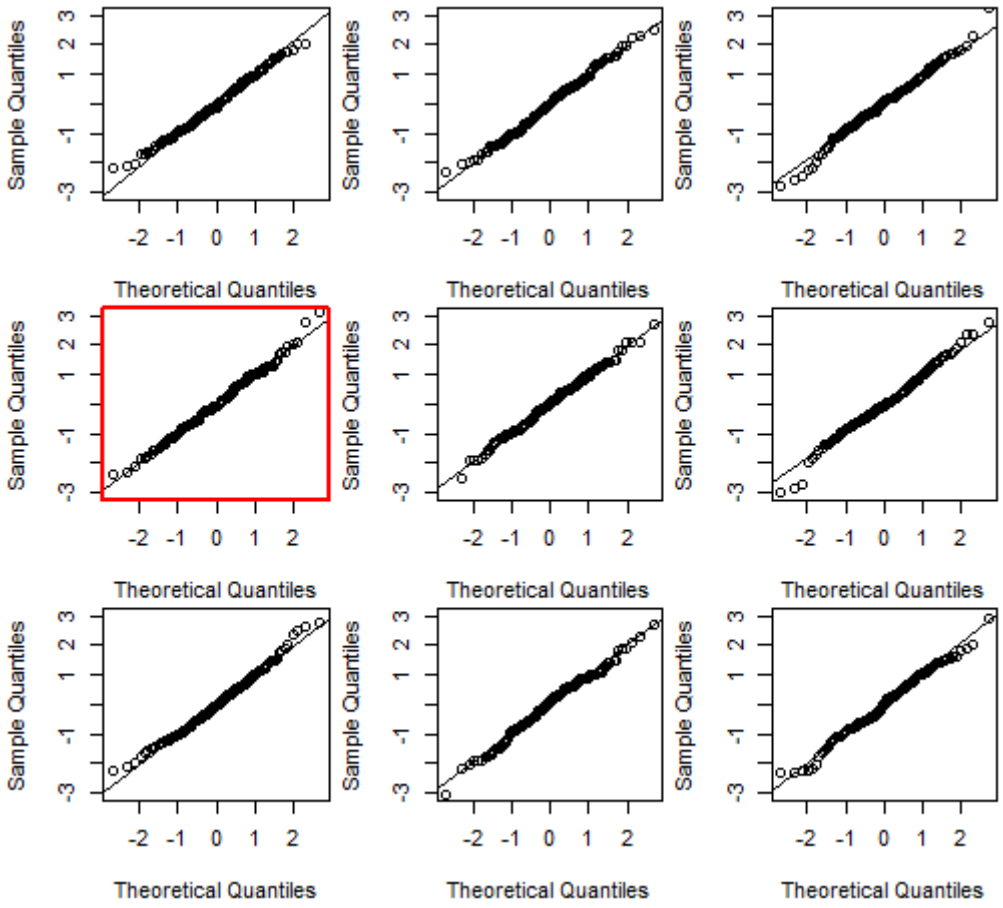

Figure 7: Wally plot of gender mixed data

The power of the wally plot is that if you cannot find your distribution in the 9 plots (generated normal distributions) then your data must be normally distributed. Considering that my personal guess was the bottom left one and they all look pretty similar is possible to say that they are identically normal distributed.

2.1.2 g)

The formula for the 95% confidence interval (CI) for the mean of the log-transformed BMI score is:

$$\overline{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$$3.218 \pm t_{0.975} \cdot \frac{0.0222}{\sqrt{145}}$$

Thus the confidence intervals for the mean are:

$$[2.689, 3.745]$$

To determine the 95% CI of the median the function t.test has been used following section 3.1.9 page 167 in the textbook:

$$[24.366, 25.587]$$

2.1.3 h)

Using the 95% CI we can say that $\alpha = 0.05$, this represents the probability of the assumption being wrong (our mean is out of the CI assuming $\mu = log(25)$), also known as type two error.
Formula for the test statistics:

$$t_{obs} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Distribution of the test statistics:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Where $t(n-1)$ is the t dist. with n-1 deg. of freedom (144 in our case).
The p-values is computed using:

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

The observed p-value is 0.9206, since it is greater than α and large we accept H_0 (not found significance against H_0). It is also possible to state that more than half of the population is, at least, moderately overweight.

2.1.4 i)

Now our discussion will focus on the two subsets, male and female, after the log-transformation. Fist it will be shown briefly that they also are normally distributed using the same methods as in subsection f.
Starting with males:

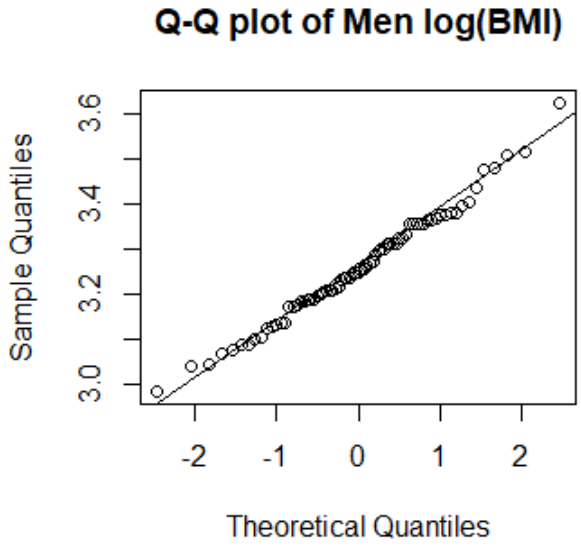

Figure 8: Q-Q plot of logged BMI scores for men

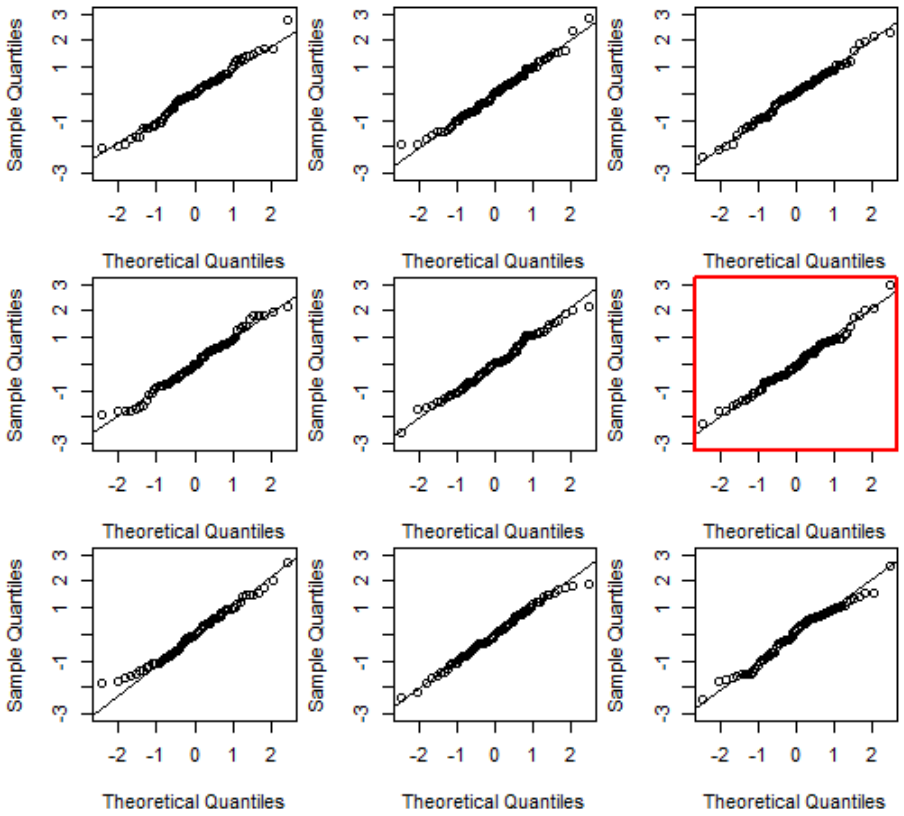

Figure 9: Wally plot of men

As before it is possible to say that the data are normally distributed.
Even for the females it is hard to find back the plot, thus they are normally distributed.

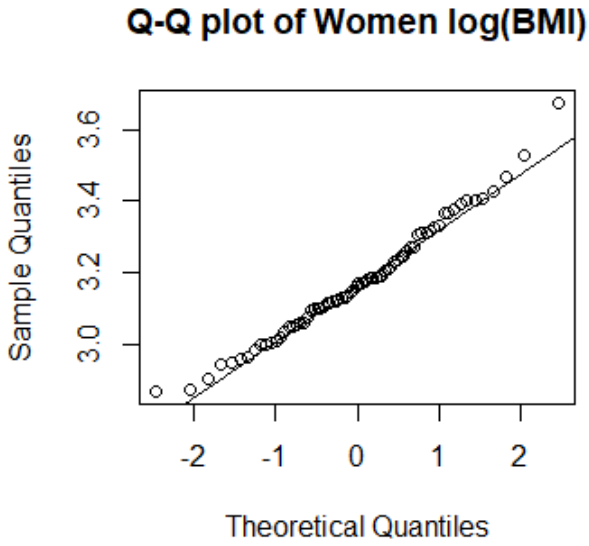

Figure 10: Q-Q plot of logged BMI scores for women

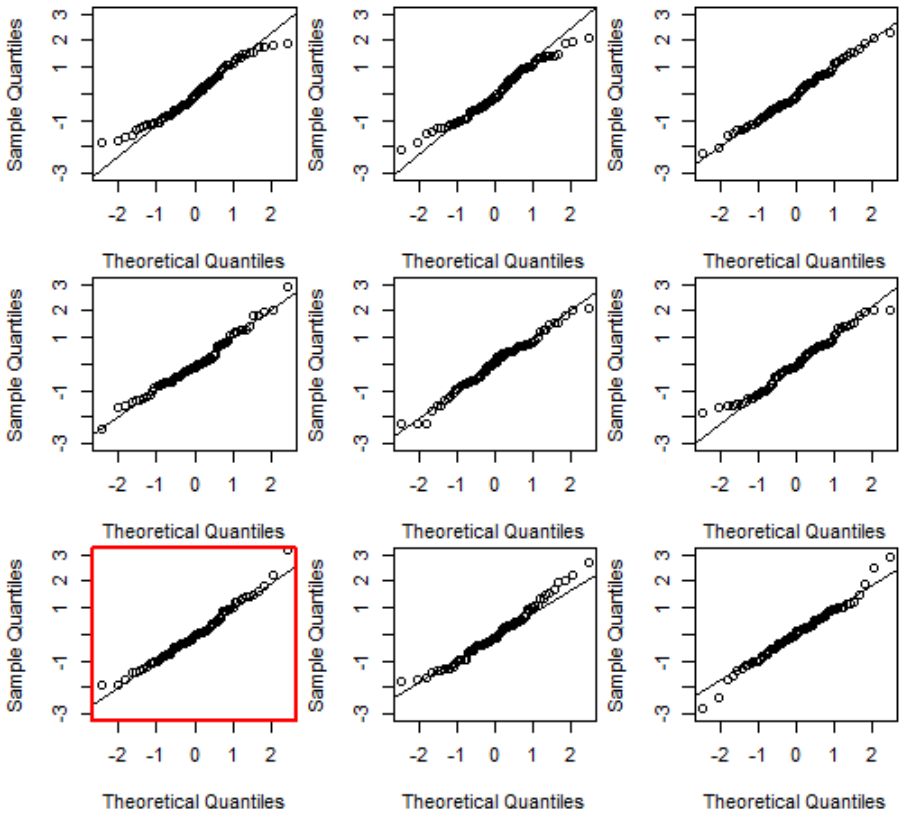

Figure 11: Wally plot of women

The parameters of the models are:

	Sample mean	**Sample std. dev.**
Women	3.1741	0.1599
Men	3.2605	0.1240

Table 3: Estimates of mean and standard deviation

2.1.5 j)

The 95% CI for the mean have been calculated and following what was already done in section g the 95% CI for the median have been found:

	Lower bound of CI	**Upper bound of CI**
Women	23.0237	24.8204
Men	25.3221	26.8294

Table 4: Confidence intervals for the medians

2.1.6 k)

The hypothesis that is going to be tested is that the difference between μ_{log_men} and μ_{log_women} is zero.

$$H_0 : \mu_{log\_men} - \mu_{log\_women} = 0$$

$$H_1 : \mu_{log\_men} - \mu_{log\_women} \neq 0$$

Using the 95% CI we can say that $\alpha = 0.05$, this represents the probability of the assumption being wrong. The formula for the test statistics is:

$$t_{obs} = \frac{(\overline{x_1} - \overline{x_2}) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

The distribution of test statistics is:

$$T = \frac{(\overline{X_1} - \overline{X_2}) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

With v degrees of freedom:

$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

The p-value is found using:

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

The found p-value is: 0.000382
It is possible to state that since the p-value is much smaller than α and it is very small in general we reject the hypothesis where the two genders have the same BMI (high evidence against it so reject H_0), thus there is a difference between the females BMI and males one (H_1). Comparing the results with the given R code we notice that our CIs are positive, but this depends the order in which the samples for the Welch test are chosen.

2.1.7 l)

The same conclusion could have definitely been drown from just the CIs. Using remark 3.59 from the text-book.
"In the case of two independent variables with added CIs we can say that:
when the two CIs do not overlap the two groups are significantly different."
And this is our case, see table 4.

2.2 Correlation

2.2.1 m)

In order to compute the correlation we need to use the covariance, found like this:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

and use it in the correlation formula, that is:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y}$$

In particular the correlation formula between BMI and weight is:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y} = \frac{48.27}{15.21 \cdot 3.83} = 0.83$$

With the correlations between BMI and weight, BMI and fast food, weight and fast food the scatter plots have been made. The results for correlations of BMI and fast food and fast food and weight have a low value of correlations thus they are not included here because the representation is not significant. The correlation between BMI and weight is a clear indication that as people's weight increases also their weight does the same. We can see that the trend is positive and there is a high level of linear relation between these two as we could expect.

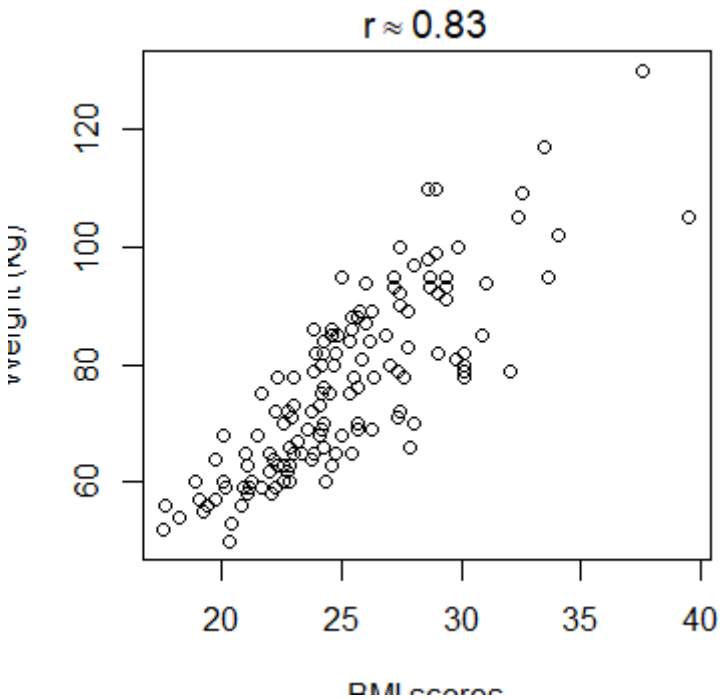

Figure 12: Scatter plot of BMI vs. weight