Introduction to statistics 02402
Project 1 - BMI survey

Tuesday, 03 13, 2018

Nina Munkholt Jakobsen, Department of Applied Mathematics and Computer Science

Frederik Rander Andersen, s164146

DTU Fotonik
Department of Photonics Engineering

Contents

I
Descriptive analysis

1 a) - Data description and introduction

Body Mass Index or BMI is a formula to describe whether a person is overweight or not. It only considers height and weight, not things such as body fat. The formula is simply:

$$BMI = \frac{weight(kg)}{height(cm)^2}$$

LaTeX

The data we are given is a number of measurements done on 145 persons. The information given is: height, weight, gender, urbanity and fastfood. Height is in cm, weight in kg, gender as 0(female) or 1(male), urbanity ranging from 1 to 5 given as the population of the place where the person lives, 0 being outside urban areas and 5 being in a city with more than 100000 inhabitant. Fastfood is given as how often the person eats fastfood given in days per year. Of the given variables, height, weight and fastfood are quantitative while gender and urbanity are categorized variables. We are given 145 observations and there are no missing values.

2 b) - Density histogram

The figure below shows a density histogram of the BMI observations. The red lines separate each of the BMI categories and the dotted line is a density estimate derived from the given data.

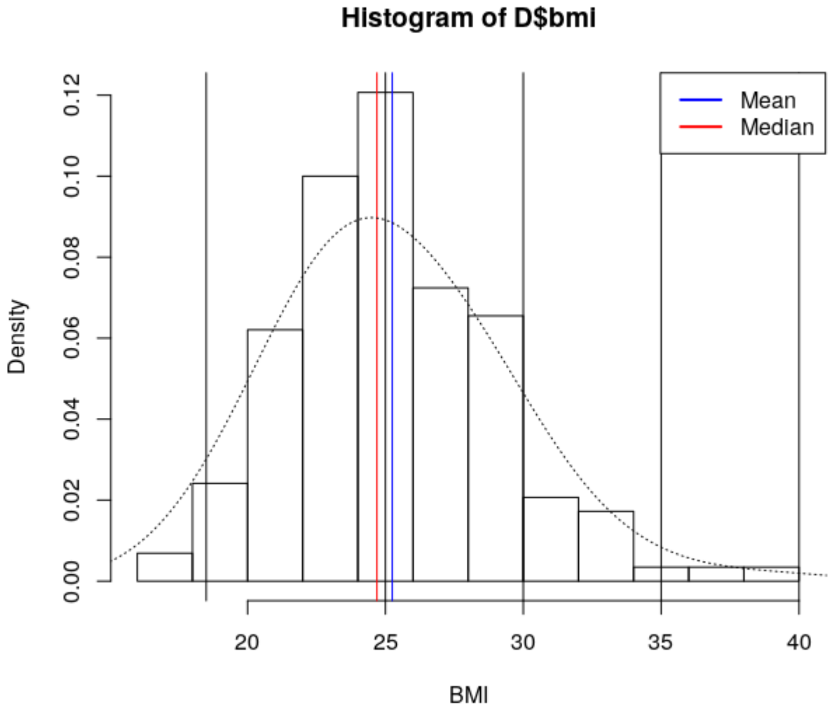

Figure 1.1: Density histogram of BMI

From the histogram we can see that most of the observed BMI's lie within 20-30 $\frac{kg}{m^2}$ which are the two categories: normal weight and moderately overweight. This means that the majority of people are around normal weight or moderately overweight. If we plot the median(24.69) and mean(25.25), we see that the distribution is right skewed since the mean is to the right of the median. BMI scores cannot be negative since what we are getting out of the formula is $\frac{kg}{m^2}$ and a person cannot have negative weight or height. The standard deviation is found to be 3.83, which shows that there is some variation in the observations since we see a lot of different BMI values, with the minimum being 17.6 and the maximum being 39.5.

3 c) - Gender density histograms

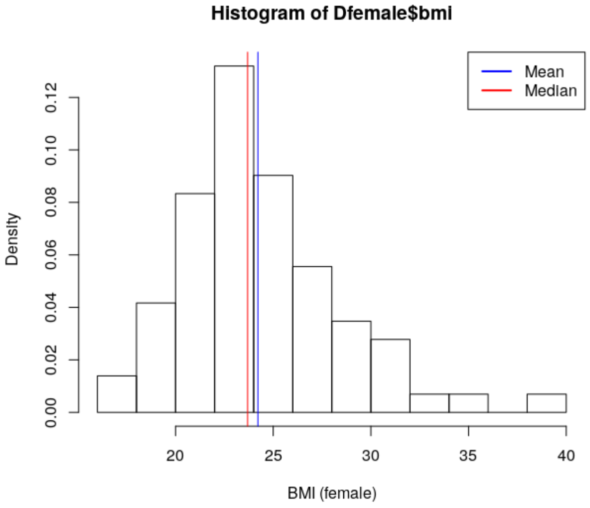

Figure 1.2: Histogram of female BMI scores

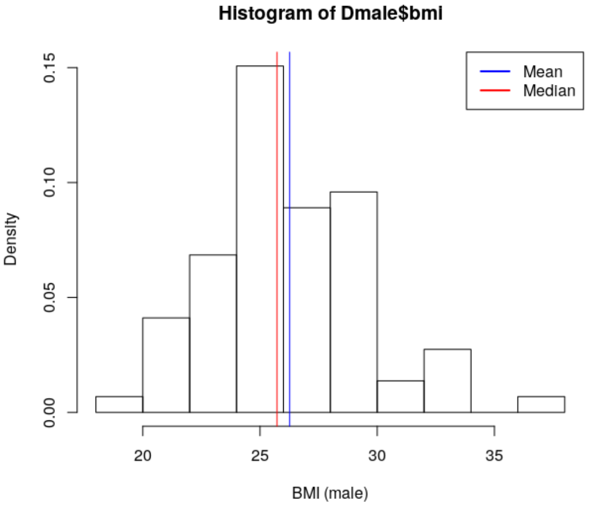

Figure 1.3: Histogram of male BMI scores

We can see from the figures above that both histograms are right-skewed since the mean(blue) is on the right side of the median(red) in both density histograms. For the female BMI observations, it is seen that they are primarily distributed around 20-26 which means that most of the female weights are within the normal weight category. If we look at the male density histogram, we see that most of the male weights are within 23-30, which shows that the men are generally a bit higher in their BMI than females. Male BMIs are generally in the normal to moderately weight categories.

The standard deviation for the female observations is 4.1 so there is a quite a bit of variation in the values with the minimum being 17.6 and the maximum being 39.5 with the mean 24.2 and median 23.7. The standard deviation for the male observations is 3.3 so there is still some variation but not as much as in the female observations. The minimum is 19.8 and the maximum is 37.6 with the mean 26.3 and the median 25.7.

Generally the male BMI scores are higher than the female BMI scores and not as spread out.

4 d) - Box plot

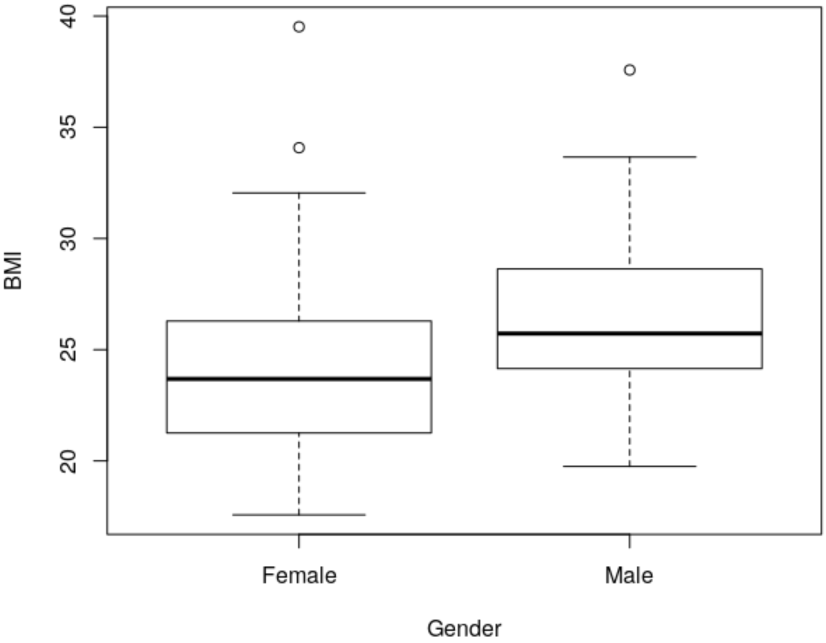

Figure 1.4: Box plot of BMI scores for each gender

From the female BMI box plot we can conclude that the distribution is right-skewed, however not by much. The male BMI box plot is much more clearly right-skewed. There are some differences which are the same differences as observed before; the male BMI scores are generally higher and less spread out. It is perhaps a bit more easily observed on the box plot. We can easily distinguish extreme outliers on the box plot as they are plotted as single points. There are two for the female observations and one for the male observations.

5 e) - BMI summary statistics

Table 1.4: BMI score summary statistics table

BMI	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	25.25	14.69	3.83	22.59	24.69	27.64
Female	72	24.22	16.42	4.05	21.26	23.69	26.29
Male	73	26.27	11.07	3.33	24.15	25.73	28.63

Compared to the box plot, we can see a bit more information from this table. Number of observations, variance and standard deviation are not included in the box plot. The standard deviation in particular can provide some more information about the spread of the observations.

II
Statistical Analysis

1 f) - Statistical model for BMI

We try to plot the logarithmic-transformed BMI values as a QQ plot. The QQ plot can be used to check whether the logarithmic-transformed BMI values follow a normal distribution.

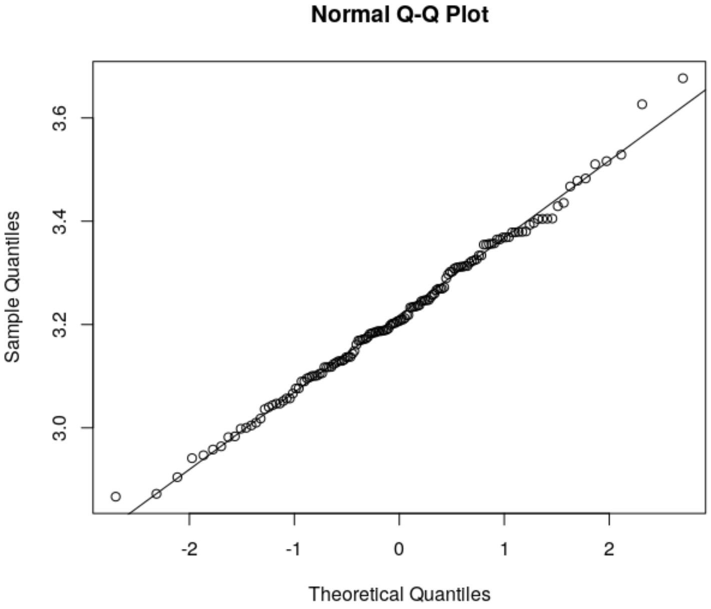

Figure 2.1: QQ-plot normal distribution

From the above figure we can see that since the values nicely follows a straight line, the BMI follow a normal distribution. The logarithmic mean and logarithmic standard deviation are estimated to be 3.22 and 0.15 respectively. When looking at the QQ plot we can see that these values seem to be quite accurate as the values follows a straight line (for a normal distribution) quite nicely. The mean can be seen where the theoretical quantiles = 0 and the standard deviation is how much the values deviate from the straight line. Since the log values follow the normal distribution line, we have validated that the log-transformed observations follow a normal distribution. Furthermore it is validated since all individual observations are independent. If we consider the Central Limit Theorem, we can conclude that if we had more observations the observations would follow a normal distribution better, and if we had less, it would not be as close to a normal distribution as it is.

2 g) - Confidence interval formula

We insert the numbers into the formula for confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

$$3.2176 \pm 1.9765 \cdot \frac{0.1489}{\sqrt{145}}$$

From this we get the confidence interval to be: [3.193 ; 3.242], which is correct if we check with the t.test function in r. Now we find the confidence interval for the median BMI scores by using the exponential function on the already calculated confidence interval.

$$e^{3.193}; e^{3.242}$$

Which gives us a median BMI confidence interval [24.366 ; 25.587]

3 h) - hypothesis test

The significance level α is set to 5%, the formula for the test statistics is:

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

$$\frac{3.218 - log(25)}{0.1489/\sqrt{145}} = -0.0999 = t_{obs}$$

The test statistics follow a normal distribution and we have 144 degrees of freedom. We have already calculated the test statistics and the p-value is found to be 0.92. This is done by using the R command: P(T>x)=2(1-pt(D$logbmi,144)).

We hereby accept the null-hypothesis as the p-value is larger than α. Now we can conclude that since the null-hypothesis is accepted, we cannot deny that the average BMI score is 25, which is right on the tipping-point between the normal weight category and the overweight category.

4 i) - statistical models for female and male BMI

We will assume that both male and female BMI scores are normally distributed. So we will create a QQ plot for both in order to check this assumption and validate the model.

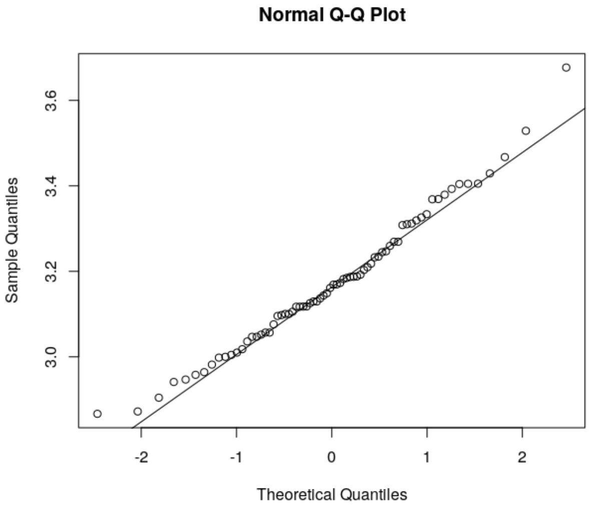

Figure 2.2: QQ plot of female log-transformed BMI

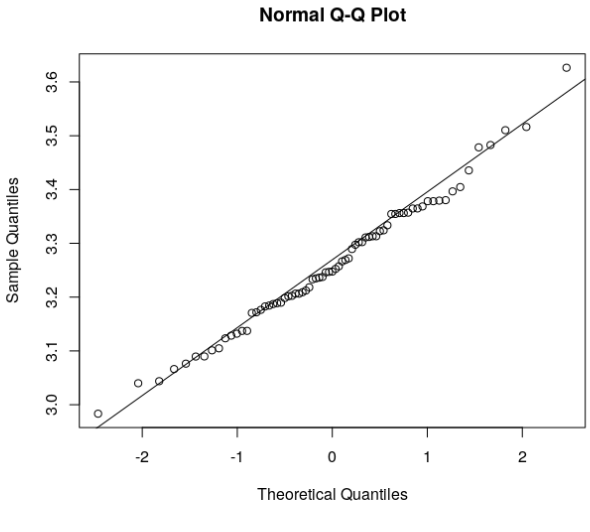

Figure 2.3: QQ plot of male log-transformed BMI

We can see from the QQ plots that both male and female log transformed BMI scores follow a normal distribution nicely. The mean can be estimated to 3.17 for females and 3.26 for men. The standard deviation is 0.16 for females and 0.12 for males. Overall we see that the male values do not vary as much as the females and the mean is a bit higher, meaning their BMI will generally be higher than the females. At least according to the data.

Table 2.3: Calculated median BMI CI

	Lower bound of CI	Upper bound of CI
Women	23.02	24.82
Men	25.32	26.83

5 j) - median confidence interval

We calculate the CI with the same formula as previously used When compared with the results from the R code, we see that the values are the same.

6 k) - hypothesis test

So with H_0 being that there is no difference in male and female BMI scores. $\alpha = 5\%$. The formula for the test statistic:

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

Which is calculated with the values from table 1.4 to be -3.33. We find the degrees of freedom by using the formula:

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

$$v = \frac{(\frac{(4.05)^2}{72} + \frac{(3.33)^2)}{73})^2}{\frac{((4.05)^2/(72))^2}{71} + \frac{((3.33)^2/(73))^2}{72}} = 133.75$$

Now the p-value is found to be 0.000392 which means that we can reject the null-hypothesis, since the p-value is much lower than α. For our data it means that there is a difference between the distributions for male and female BMI scores.

7 l) - Comment on the hypothesis test

If we take a look at the confidence intervals in table 2.3. From these we could have concluded that the two score-groups are very significantly different, since the CI intervals do not overlap.

8 m) - Correlation between BMI and weight

With BMI as x and weight as y, the formula for calculating correlation is then:

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

The correlation is now calculated and we get the following result. We have a strong correlation between weight and BMI, exactly 0.828. Now we calculate the correlation between weight, fastfood and BMI. We get the following correlation:

Table 2.3: Correlation

weight	0.279	0.828
0.279	fastfood	0.153
0.828	0.153	bmi

We plot the correlation:

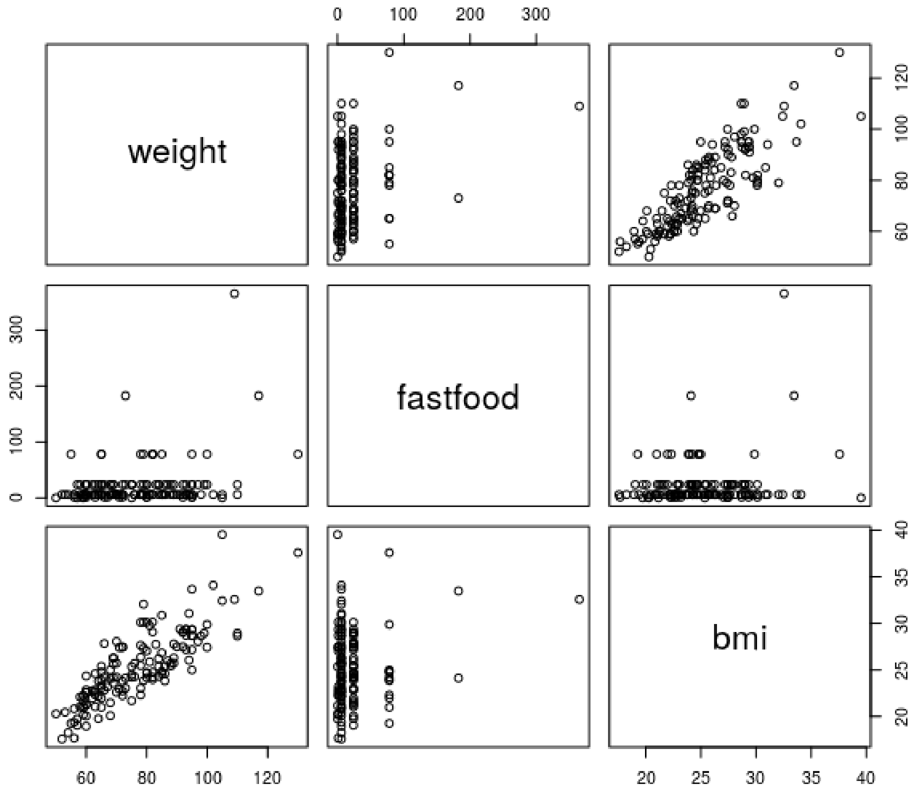

Figure 2.4: Scatter plot of correlation

From the correlation values and the plot, we can see that there is only a significant correlation between weight and BMI. The other values are too low to conclude any correlation.