# 02402 - Introduction to Statistics

## Project 1: BMI survey analysis

*Author: Kornel Kowalczyk (s202401)*

*Overview:*

The following report presents a descriptive and statistical analysis of BMI survey data collected among Danish population. First part describes results of a survey paper, giving explanation of used attributes and providing basic summary statistics of the dataset. Second part of the report focuses on statistical analysis including statistical model assumption, hypothesis testing both without gender distinction and gender-wise as well as pairwise correlation analysis of dataset attributes.

## *Table of contents*

# 1. Descriptive analysis

The content of the dataset is a result of a BMI survey carried out for Danish population on a sample size of 145 people. Respondents were to answer five questions regarding their physiological parameters, place of residence and diet. Therefore from the following survey 5 attributes were extracted which are used for visualization and analysis of overweight in Denmark.

| Variable | Variable type | Description |
|---|---|---|
| Height | Quantitative | The respondent's height in centimeters |
| Weight | Quantitative | The respondent's weight in kilograms |
| Gender | Categorized | The respondent's gender:<br>**0** - Female<br>**1** - Male |
| Urbanity | Categorized | The size of respondent's place of residence:<br>**1** - Outside urban areas<br>**2** - City with less than 10,000 inhabitants<br>**3** - City with 10,000 to 49,999 inhabitants<br>**4** - City with 50,000 to 99,999 inhabitants<br>**5** - City with over 100,000 inhabitants |
| Fast food | Quantitative | The frequency of respondent eating fast food:<br>**0** - Never<br>**1.0** - Less than 1 time per year<br>**6.0** - 1-11 times per year<br>**24.0** - 1-3 times per month<br>**78.2** - 1-2 times per week<br>**182** - 3-4 times per week<br>**286.7** - 5-6 times per week<br>**365** - Every day |

***Fig. 1. Dataset variables overview.***

The attributes (*fig. 1*) gender and urbanity are categorized while height and weight are quantitative. Fast food parameter is considered quantitative despite the answers being categorized into 8 possible outcomes, as this is the result of recoding the variable into days per year. Dataset is complete as there are neither missing nor corrupted values.

## 1.1. BMI score density histogram

Using collected data we can calculate BMI score for each respondent using the following formula:

$$BMI = \frac{weight}{height^2}$$

BMI score distribution is therefore shown on the histogram (fig. 2).

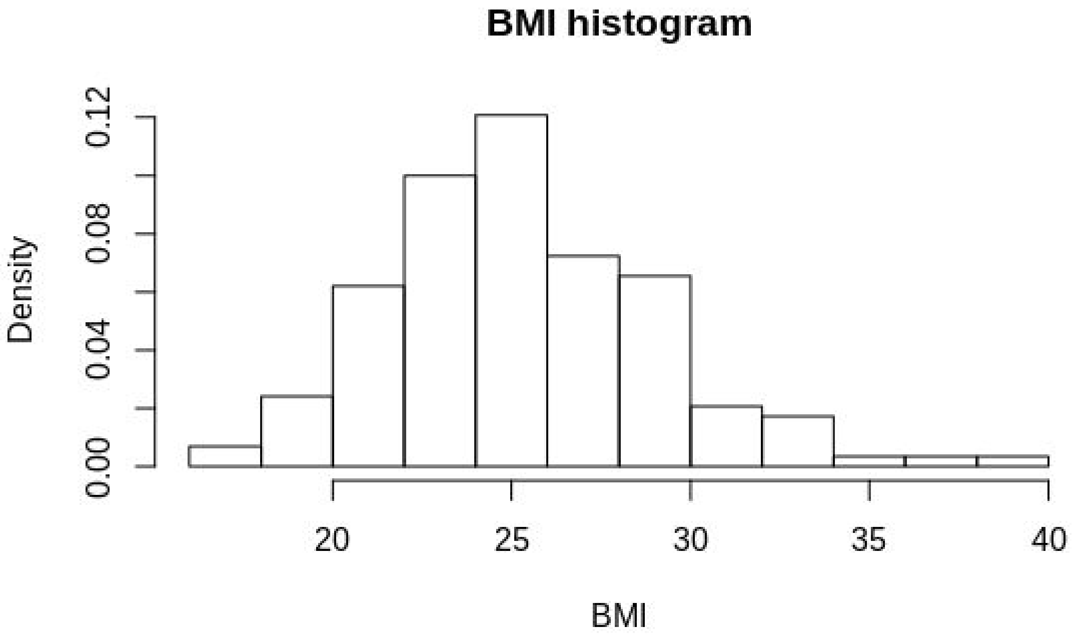

*Fig. 2. BMI survey results histogram.*

As can be seen the distribution is slightly skewed towards right with scores variating from less than 18 up to between 38 and 40. Comparing the results with the WHO classification of BMI scores (fig. 3) most of the respondents are not reaching severe overweight with highest density between 24 and 26 BMI score which is around the upper bound of person's normal weight.

| **BMI score** | **Assessment** |
|---|---|
| Less than 18.5 | The person is underweight |
| Between 18.5 and 25 | The person's weight is normal |
| Between 25 and 30 | The person is moderately overweight |
| Between 30 and 35 | The person is severely overweight (Obesity Class I) |
| Between 35 and 40 | The person is severely overweight (Severe Obesity Class II) |
| Above 40 | The person is severely overweight (Extremely severe obesity Class III) |

*Fig. 3. The WHO classification of BMI scores.*

### 1.2. Gender-wise BMI score density histograms

By dividing dataset for men and women respectively there can be observed that the BMI score distribution for females (fig. 4) is more flat and shifted to the left with highest density between 22 and

24 comparing to distribution for males which is more raised and where highest density occurs between 24 and 26 BMI score.

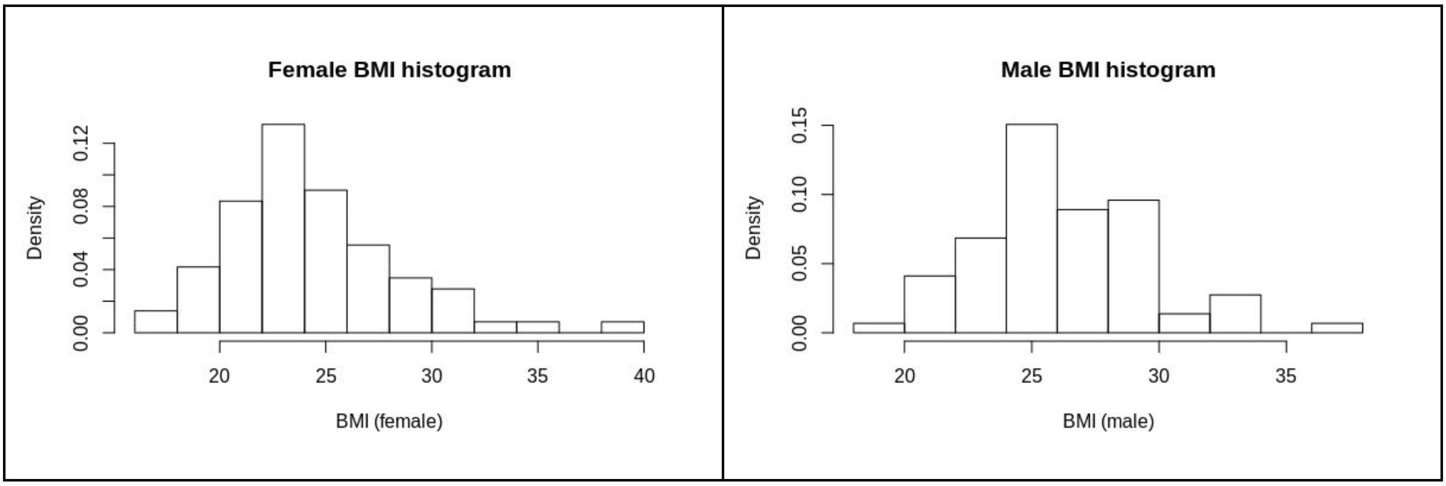

*Fig. 4. BMI survey results histograms by gender.*

## 1.3. Gender-wise BMI score box plot

Boxplots comparison of BMI scores by gender (fig. 5) shows that for female subset distribution is symmetrical while for male subset is right skewed. It also, as well as previous histograms (fig. 4), illustrates that distribution for males is more shifted towards higher BMI score values than one for females. There are also three extreme observations with the highest female BMI score of 39.5 and highest male BMI score of 37.6.

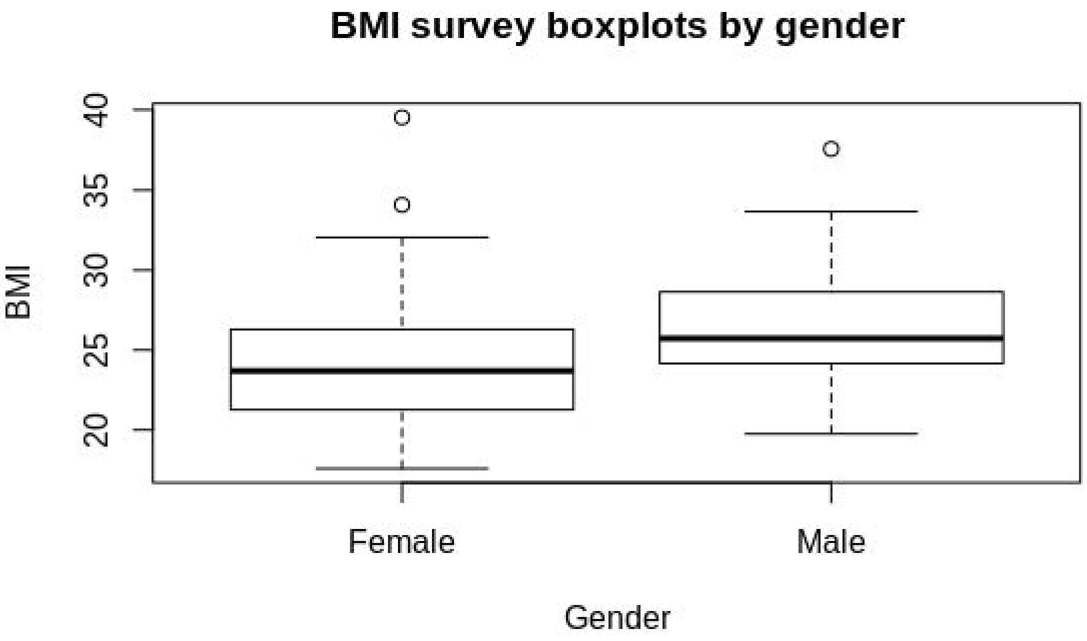

*Fig. 5. BMI survey boxplots by gender*

### 1.4. Summary statistics

The following table (fig. 6) represents summary statistics for the whole dataset as well as for gender wise subsets. It can be seen that the mean BMI of males as well as median is higher than corresponding values for females, while standard deviation being higher for females rather than males.

| ***Variable:*** *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 | 11.06 | 3.33 | 24.15 | 25.73 | 28.63 |

***Fig. 6. Summary statistics of BMI scores.***

## 2. Statistical analysis

Through the descriptive analysis it was found that BMI distribution is right skewed, to better satisfy normality assumptions of a model, the BMI scores are logarithmically transformed which leads to better fitting into the Q-Q plot (fig. 7). It needs to be noted that after the log-transformation the results on the original scale should be interpreted not as a mean but as a median.

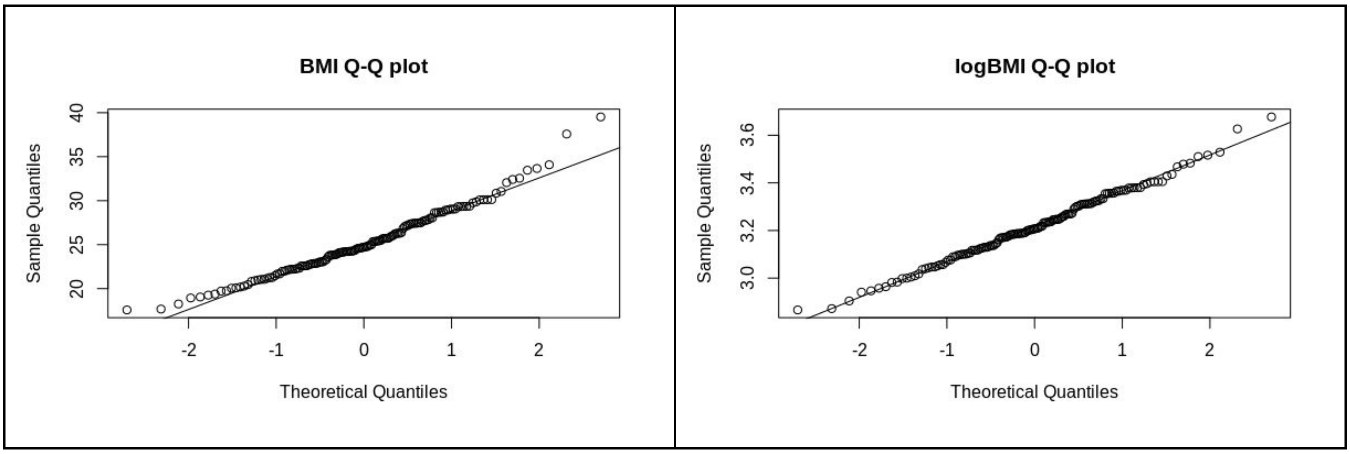

***Fig. 7. BMI and logBMI Q-Q plots comparison.***

As well as for BMI the summary statistics of logBMI scores is also provided (fig. 8).

| ***Variable:***<br>*logBMI* | ***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 | 3.22 | 0.02 | 0.15 | 3.12 | 3.21 | 3.32 |
| Female | 72 | 3.17 | 0.03 | 0.16 | 3.06 | 3.17 | 3.27 |
| Male | 73 | 3.26 | 0.02 | 0.12 | 3.18 | 3.25 | 3.36 |

***Fig. 8. Summary statistics of logBMI scores.***

## 2.1. Statistical model

It is assumed that the statistical model for logBMI follows a normal distribution. Mean and standard deviation are estimated from the sample and are put into formula.

$$X_{everyone} \sim N\left( 3.22 \,,\; 0.15^2 \right)$$

Empirical cumulative density function (fig. 9.) is calculated in order to provide model validation. As can be seen on the figure data points are fitting closely to the modelled function, therefore as a result it is considered that the proposed model is valid.

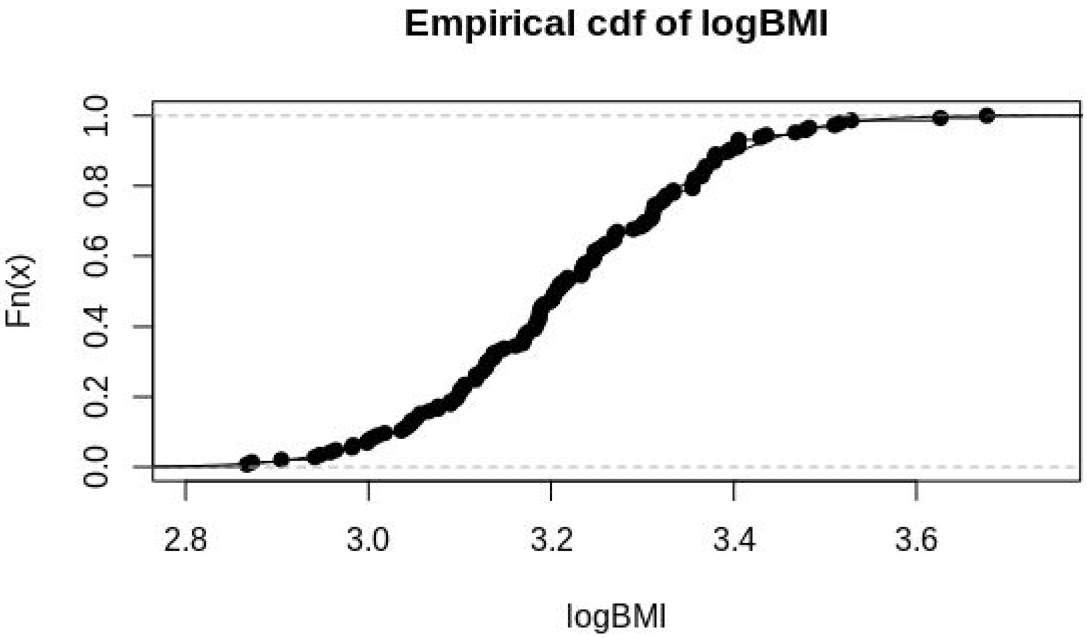

***Fig. 9. Empirical cumulative density function of log-transformed BMI***

## 2.2. Confidence interval

To estimate range of plausible values of model median, 95% confidence interval for log-transformed BMI mean is calculated using following formula:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}} = \bar{x} \pm 1.98 \cdot \frac{0.15}{\sqrt{145}} = \bar{x} \pm 0.02$$

$$3.22 \pm 0.02 = [3.19,\ 3.24].$$

Therefore after back-transform 95% confidence interval for estimated model median is as follows:

$$24.69 \pm 0.6 = [24.37,\ 25.59].$$

## 2.3. Hypothesis test

To investigate if over half of Danish population is overweight, there need to be inspected if median BMI is 25 as it is an upper limit for normal weight according to WHO classification (fig. 3). The hypothesis is that mean of log-transformed BMI is equal to log(25).

$$H_0\colon \mu_{\log BMI} = \log(25)\ ,$$

$$H_1\colon \mu_{\log BMI} \neq \log(25)\ .$$

Significance level $\alpha = 0.05$ is chosen and test statistic is calculated using following formula:

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.22 - \log(25)}{0.15/\sqrt{145}} = -0.10$$

next, using $t_{obs}$ the *p-value* is calculated with degrees of freedom equal to n-1 which is 144:

$$p - value = 2 \cdot P\left(T > |t_{obs}|\right) = 0.92.$$

As calculated *p-value* is higher than significance level $\alpha = 0.05$ The null hypothesis, as BMI median for Danish population is equal to 25, cannot be rejected. According to this and calculated before the confidence interval it is statistically possible that the median of BMI score is equal to 25, which gives the conclusion that over half of the Danish population is overweight.

## 2.4. Gender-wise statistical models

For gender-wise logBMI models normal distribution is assumed as well. Mean and standard deviation are calculated from the sample for female and male respectively and distributions for both are as follows:

$$X_{female} \sim N(3.17\ ,\ 0.16^2)$$

$$X_{male} \sim N(3.26\ ,\ 0.12^2)\ .$$

As can be seen on the Q-Q and ecdf plots (fig. 10) data points are fitting well to the proposed function, thus it can be said that models for, both female and male, logarithmically transformed BMI distributions are valid.

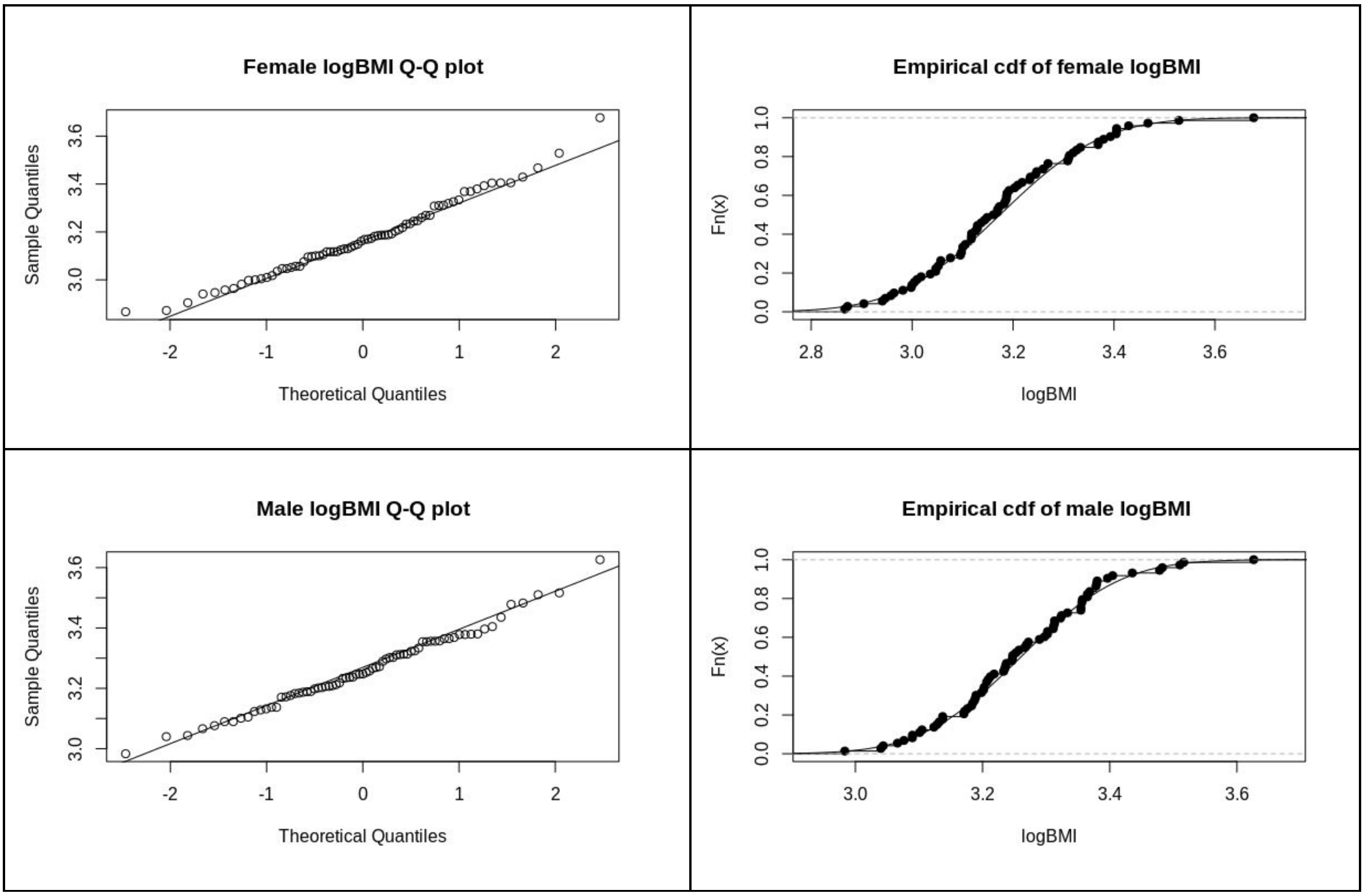

***Fig. 10. Q-Q plots and ecdf's of log-transformed BMI's for men and women respectively.***

## 2.5. Gender-wise confidence intervals

To find accuracy of estimated gender-wise BMI score medians, similarly like before confidence intervals are calculated and bounds for women and men are presented in the table below (fig. 11).

|  | Lower bound of CI | Upper bound of CI |
|---|---|---|
| **Women** | 23.02 | 24.82 |
| **Men** | 25.32 | 26.83 |

***Fig. 11. Median BMI scores confidence intervals for women and men respectively.***

## 2.6. Gender-wise hypothesis test

Hypothesis test is now performed to investigate if there is a difference between the BMI of women and men. Null hypothesis is as follows:

$$H_0 \colon \delta_{\log BMI} = \log(0) \, ,$$
$$H_1 \colon \delta_{\log BMI} \neq \log(0) \, .$$

Significance level $\alpha = 0.05$ is chosen and the Welch two-sample t-test statistic is calculated:

$$t_{obs} = \frac{\left(\bar{x}_{female} - \bar{x}_{male}\right) - \delta_{\log BMI}}{\sqrt{\frac{s^2_{female}}{n_{female}} + \frac{s^2_{male}}{n_{male}}}} = \frac{(3.17 - 3.26) - 0}{\sqrt{\frac{0.16^2}{72} + \frac{0.12^2}{73}}} = -3.64$$

as well as degrees of freedom:

$$v = \frac{\left(\frac{s_{female}{}^2}{n_{female}} + \frac{s_{male}{}^2}{n_{male}}\right)^2}{\frac{\left(s_{female}{}^2 / n_{female}\right)^2}{n_{female} - 1} + \frac{\left(s_{male}{}^2 / n_{male}\right)^2}{n_{male} - 1}} = \frac{\left(\frac{0.16^2}{72} + \frac{0.12^2}{73}\right)^2}{\frac{\left(0.16^2/72\right)^2}{72 - 1} + \frac{\left(0.12^2/73\right)^2}{73 - 1}} = 133.75$$

finally the *p-value* is found:

$$p - value = 2 \cdot P\left(T > \left|t_{obs}\right|\right) = 0.0004 \, .$$

Calculated *p-value* is more than 100 times lower than chosen significance level $\alpha = 0.05$ so the null hypothesis is rejected as there are very strong evidence against $H_0$. Result of this hypothesis test is that there is a difference between BMI of men and women.

The same conclusion can be obtained just after calculating the confidence interval of median BMI for both women and men respectively as it can be seen that CI’s don't overlap which means that two groups are significantly different, thus hypothesis testing was not necessary in this case.

## 2.7. Pairwise correlation

The final analysis is performed in order to find correlations between variables. First, the correlation coefficient between BMI and weight is calculated using following formula:

$$r_{BMI,\ Weight}=\frac{s_{BMI,\ Weight}}{s_{BMI}\cdot s_{Weight}}=\frac{48.27}{3.83\cdot 15.21}=0.83$$

As can be seen there exists a strong correlation between BMI score and weight which is, as can be expected, due to linear dependence of BMI score from weight in its formula. Other pairwise correlations are also shown on the table below (fig. 12)

| | **Weight** | **Fast food** | **BMI** |
|---|---|---|---|
| **Weight** | 1.00 | 0.28 | 0.83 |
| **Fast food** | 0.28 | 1.00 | 0.15 |
| **BMI** | 0.83 | 0.15 | 1.00 |

***Fig. 12. Correlation coefficients between weight, fast food, and BMI.***

To better visualize the relations between variables, scatter plots below (fig.13) are showing that except BMI and weight correlation there is none between BMI and fast food consumption as well as between weight and fast food consumption.

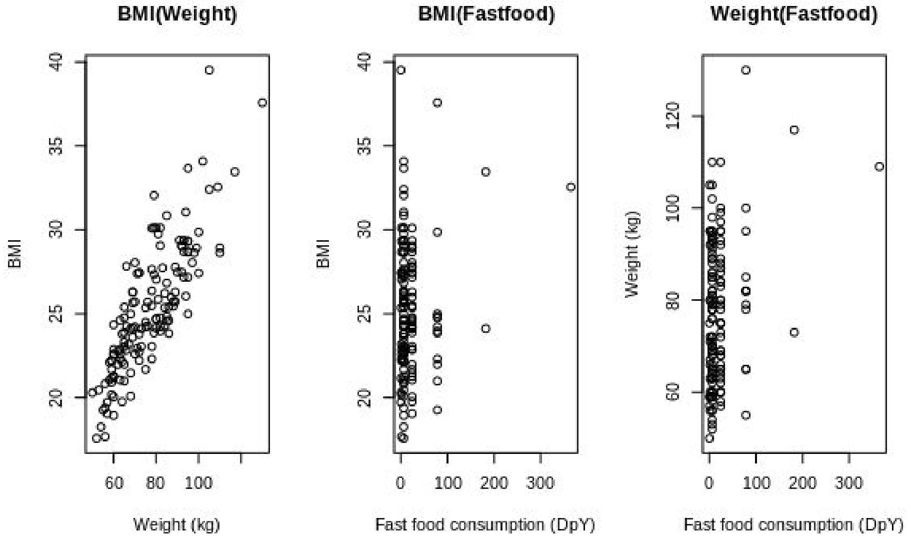

***Fig. 13. Scatterplots of BMI, Weight, Fast food.***

It is not as one may expect but it cannot be concluded from this dataset that there exists statistically significant influence of fast food consumption on the BMI score in the Danish population. It is also hard to reject the potential influence of fast food consumption on BMI scores because of the data points being mostly distributed on the left side of the fast food consumption axis, with little information about BMI scores of people which eat in fast food restaurants more often.