# Project 1

BMI Survey

Bashar Khaled Bdewi-S183356 10/23/20 Introduction to statistics

# Introduction

In this project, we look at overweight in Denmark. Overweight, measured by the so-called Body Mass Index (BMI), poses an increasing problem in the Western world and also in Denmark. Overweight is associated with a number of health issues, e.g., hypertension, heart disease, diabetes, etc. This has great impact on the Danish economy, where the cost of health care is increasing. Therefore, it is necessary to put the problem of overweight on the agenda and, in this context, there are many stakeholders. In addition to the public sector, there is almost an entire industry which focuses on lifestyle and the problem of overweight: from Fitness to TV programs, books on how to lead a healthy lifestyle, and on health food to producers of food, fast food, etc.

# The first part ................................ Descriptive analysis

The purpose of the first part of the project is to carry out a descriptive analysis of the data. It can be done using summary statistics and suitable figures.

a) **Write a short description of the data. Which variables are included in the dataset? Are the variables quantitative and/or categorized? (Categorized variables are only introduced in Chapter 8, but they are simply variables which divide the observations into categories/groups - e.g. three categories: low, medium, and high). How many observations are there? Are there any missing values?**

This project aims to investigate the Body Mass Index (BMI) by analyzing a given sample, which has been prepared among the population of Denmark.
Upon closer examination, it can be seen that the respective data material consists of the following 5 variables:

- Height
- Weight
- Gender
- Urbanity
- Fast food

Gender, Urbanity and fast food are categorized variables because they have finite discrete values and divide the observation into groups.

Gender could be only 2 values (male (1) or female (0)).

Urbanity could be in 5 classes (numbers from 1 to 5)

Fast food could be categorized between numbers 1 and 8

Height and weight are quantitative Variables because they could have any number of different values.

height and weight could be any value greater than or equal zero.

Number of observations are the number of rows in the size (we get from the command Dim(D))

⇨ Number of observations = 145

Number of missing values we can get using the command sum(is.na(D)). Number of missing values is zero.

⇨ There are not missing values.

**b) Make a density histogram of the BMI scores. Use this histogram to describe the empirical distribution of the BMI scores. Is the empirical density symmetrical or skewed? Can a BMI score be negative? Is there much variation to be seen in the observations?**

To describe the empirical density, a density histogram for bmi for men and women is compiled together.
Each column describes the relative frequency of the respective group of bmi observations.
To examine whether the empirical density is symmetrical or skewed, the easiest method is to compare the median and mean.

if $\bar{X} < m$ the distribution is left skewed.

if $\bar{X} > m$ the distribution is right skewed.

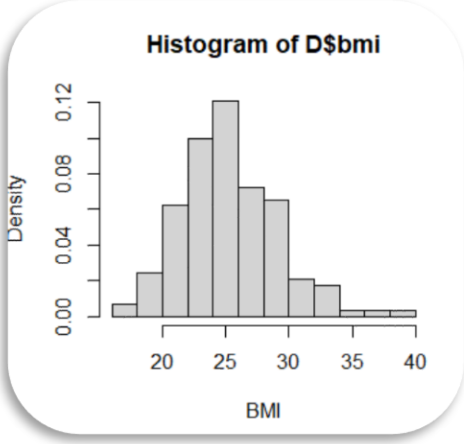
When calculating these parameters, it is seen that the median is equal to 24.96 and the mean value is equal to 25.25. Thus, the mean value is greater than the median, which means that the distribution is right skewed.

The reason why there are no negative outcomes when calculating BMI is to be found in the formula itself. The formula for Bmi is namely:

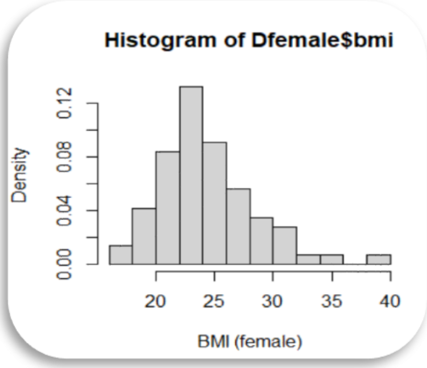
$$BMI = \text{weight } (kg) \ / \ \text{height } (m)^2$$

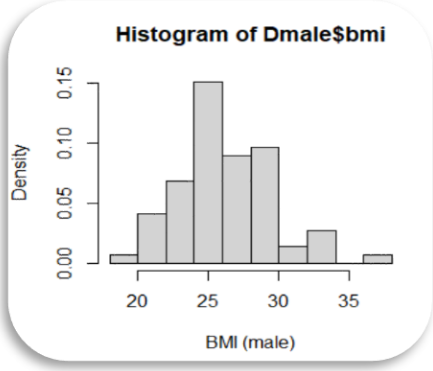
Since neither weight nor height can be negative numbers, the outcome of this can never be negative.

By calculating the standard deviation (3.83), from the result there is not seen very extremes.



c) **Make separate density histograms for the BMI scores of women and men, respectively. Describe the empirical distributions of the BMI scores for men and women using these histograms, like in the previous question. Does there seem to be a gender difference in the distribution of the BMI scores (if so, describe the difference)?**

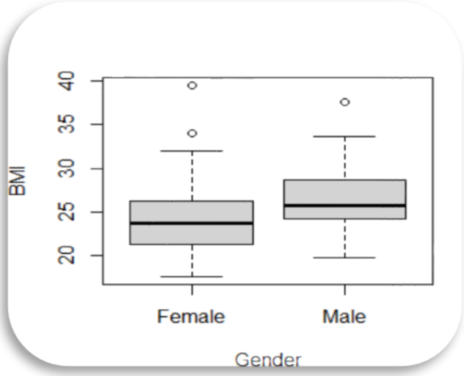




Above are two histograms for resp. women and men. A clear difference in the data material is depicted here if it is separated according to gender. First and foremost, men have an average BMI of 26.27 and women average a BMI of 24.22. However, it is not clearly seen in the two histograms as the longest tails do not represent the mean, which means that they are not normally distributed. If we look again at the comparison between mean and median, it can be seen that the distribution on both histograms is right skewed.
However, the women's histogram is more symmetrical around the median, however, it extends over a slightly larger range. This means that the spread would be a bit larger.

d) **Make a box plot of the BMI scores by gender. Use this plot to describe the empirical distribution of the BMI scores for women and men. Are the distributions symmetrical or skewed? Does there seem to be a difference between the distributions (if so, describe the difference)? Are there extreme observations/outliers?**

When comparing different populations, using a boxplot can be extremely beneficial. When reading the box plot, it is first seen clearly where the median for resp. women and men lie. Here it is also possible to see the fractals for each gender category. The upper quartile of women (Q3) and the lower quartile (Q1) are lower than in men, however, there is a greater spread in women, which is partly due to the fact that there are more extremes in the upward direction. It is also seen that there is a skewed distribution in the men, as the median is not located in the middle of the upper and lower quartile. In contrast, there is a more symmetrical distribution in women between the upper and lower quartiles, which is supported by the location of the median.



e) **Fill in the empty cells in the table above by computing the relevant summary statistics for BMI, first for the full sample (both genders combined), then separately for women and men. Which additional information may be gained from the table, compared to the box plot?**

For further detailing of the empirical distribution of BMI, the primary quantities and key figures are summarized here.

| **Variable**<br>**(BMI)** | **N.Obs**<br>**(n)** | **Mean**<br>**($\bar{x}$)** | **Variance**<br>**($s^2$ )** | **std. dev.**<br>**(s)** | **Q1** | **Median**<br>**(Q2)** | **Q3** |
|---|---|---|---|---|---|---|---|
| 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 |

In relation to the mentioned plots, we can see here exact values which were previously difficult to read. For example. it can be seen here what the exact standard deviation and variance is, where these values before in the box plot, was to be interpreted. The calculated values support that the women on average have a lower bmi in all quartiles. However, there is significantly greater variance in women, which means that the spread is greater. This is due to the more extreme extremes seen in women.

# The second part ................ Statistical analysis

The purpose of the second part of the project is to perform a simple statistical analysis of the BMI for men and women. This includes specifying statistical models for BMI, estimating the parameters of these models, performing hypothesis tests, and computing confidence intervals.

**f) Specify a statistical model for log-transformed BMI, making no distinction between men and women (see Remark 3.2). Estimate the parameters of the model (mean and standard deviation). Perform model validation (see Chapter 3 and Section 3.1.8). Since, in this case, confidence intervals and hypothesis tests involve the distribution of an average, it might also be useful to include the central limit theorem (Theorem 3.14) in the discussion.**

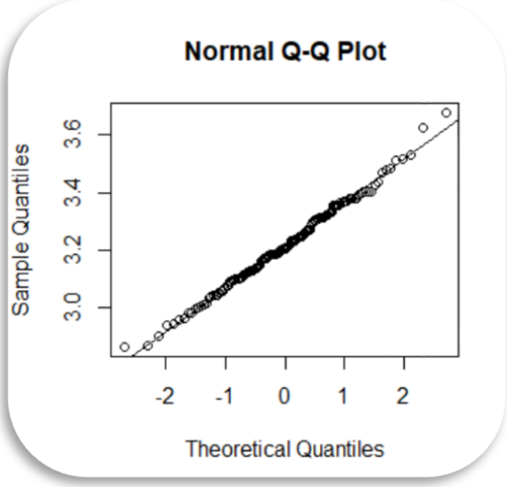
As previously mentioned, the given sample is not normally distributed, so it may be advantageous to perform a transformation of data, thereby improving normality.

A Q-Q plot is therefore prepared to act as model control. Based on this Q-Q plot, it can be assumed that the natural logarithm of the BMI results is normally distributed.

This is because almost all observations are on the line. Thus, it is prescribed that the closer the observations are to the line, the more normally distributed are the respective observations.

The static model can thus be set up as follows:

$Xi$ (3.22, 0.1492 ) where $i$ = 1 … 145



**g) State the formula for a 95% confidence interval (CI) for the mean log-transformed BMI score of the population (see Section 3.1.2). Insert values and calculate the interval. Then, determine a 95% CI for the median BMI score of the population (see Section 3.1.9).**

The confidence interval formula for the mean:

$\overline{x} \pm t_{1-\alpha/2} * s$/ sqrt(n)

Since it is desired to write down the formula for 95% confidence interval, the respective t-value will be:

$\overline{x} \pm t_{0,975} * s$/ sqrt(n)

The T-value is inserted in R and calculated by the following R command:

```
qt(0.975.length(D$logbmi)tf1) = 1.977
```

All the values are now inserted into the formula and thus the 95% confidence interval for the mean value is derived

3.22 ± (1.977 $*$ 0.149 / sqrt(145))=[3.196,3.244]

To derive the 95% confidence interval for the median, the exponential value of the mean value of the logarithmically transformed BMI is used, as this is the same method usually used to find the median:
exp(3.196), exp(24.43) → $R\ kommado$(exp ($c$(3.193,3.244)) = [24.43,25.64]

**h) Perform a hypothesis test in order to investigate whether the mean log-transformed BMI score is different from log(25). This can be done by testing the following hypothesis, and corresponds to investigating whether the median BMI score is different from 25:**

**H0 : µlogBMI = log(25),**
**H1 : µlogBMI != log(25).**

It is examined here whether the mean value of the logarithm of BMI is different from log (25) by hypothesis testing of:

H0 : µlogBMI = log(25), H1 : µlogBMI != log(25).

As a general rule, the significance level is determined to be $\alpha$ = 5% ≈ 0.05. Thus, the null hypothesis will be rejected at this level of significance, ie. that it is rejected if there is less than a 5% chance of making a mistake. The test size is given by the following formula:

$t_{obs}$ = ($\overline{x}$ − $\mu_0$ ) / ($s$/sqrt($n$))

The values are entered and the following test size is calculated:

$t_{obs}$ = (3.22 − log [25]) / (0.149 /sqrt(145)) = 0.091

It is then desired to find the p-value. This is done using the following formula:

$pværdi$ = 2 ∗ $P(T$ > $|tobs|$) →

$R\ kommando$ → (2 ∗ (1 − pt(0.091, df = length(D$logbmi) − 1)) = 0.928

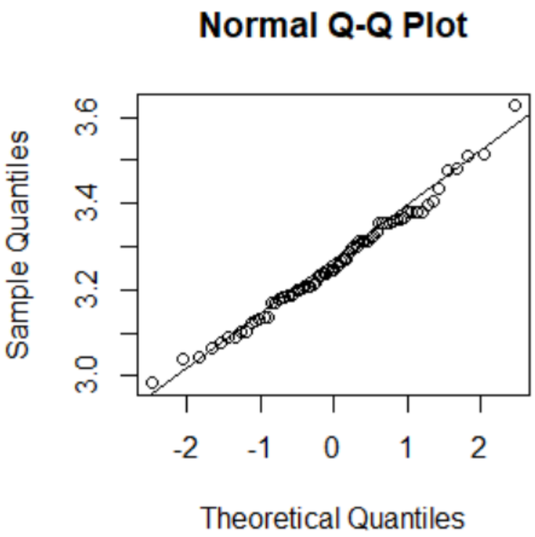
Since the p-value> significance level / 0.928> 0.05, the null hypothesis is accepted. A p-value of this magnitude means that there is little or no evidence against $H0$, so the chance that the null hypothesis is true must be very large. Furthermore, the median is not different from 25 and it can therefore be concluded that half of the population must be overweight

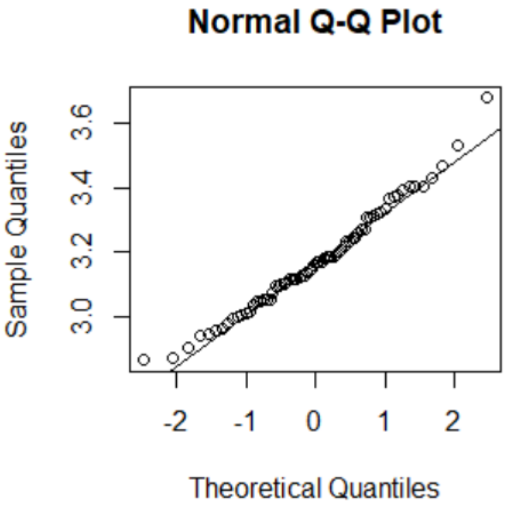
i) **Specify separate statistical models for log-transformed BMI for men and women. Perform model validation for both models. Estimate the parameters of the models (mean and standard deviation for men and women, respectively).**

As before, the same type of model check is performed for the static models of the logarithm for BMI. It is separated here in resp. women and men. A Q-Q plot is set up for both women and men, in order to be able to visually see whether there are normal distributions. Both plots show that observations generally follow the straight line, which is why it can again be assumed that both distributions are normally distributed. As mentioned earlier, it is also seen here that there are some extreme values at the end of the women's Q-Q plot. Again, static models can thus be written up. These are as follows:

Women: $Xi$ (3.17, 0.162 )where $i$ = 1 … 72

men: $Xi$ (3.26, 0.1242 )where $i$ = 1 … 73





j) Calculate 95% confidence intervals for the mean log-transformed BMI score for women and men, respectively (se Section 3.1.2). Use these to determine 95% confidence intervals for the median BMI score of women and men, respectively. Fill in the table below with the confidence intervals for the two medians.

Again, the following formula is used to first find the 95% confidence interval of the logarithm of BMI for women:

$$\overline{x} \pm t_{0.975} * (s \ / \ \text{sqrt}(n))$$

$$t_{0.975} = 1.994$$

The values are inserted:

$$3.17 \pm (1.994 * 0.16/ \text{sqrt}(72)) = [3.132,3.208]$$

The exponential value of the mean of the logarithmically transformed BMI is found:

$$\exp(3.132) = 22.9 \ \& \ \exp(3.208) = 24.73$$

The same procedure is used for men:

$$t_{0.975} = 1.993$$

$$3.26 \pm (1.993 * 0.124 /\text{sqrt}(73)) = [3.231,3.289]$$

$$\exp(3.132) = 25.3 \ \& \ \exp(3.289) = 26.82$$

| | Lower limit of KI | Upper limit of KI |
|---|---|---|
| women | 22.9 | 24.73 |
| men | 25.3 | 26.82 |

k)Perform a hypothesis test in order to investigate whether there is a difference between the BMI of women and men. Specify the hypothesis as well as the significance level α, the formula for the test statistic, and the distribution of the test statistic (remember the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words.

The following hypothesis is tested to investigate whether a difference between women's and men's BMI can be detected:

$H_0$: $\mu_k = \mu_k$ → (There is no difference between men and women's $BMI$)

$H_1$: $\mu_k \neq \mu_m$ → (There is difference between men and women's $BMI$)

Likewise, a significance level of $\alpha = 5\% \approx 0.05$ is used

The formula for the test size is rewritten into a two-sample so that it can be compared:

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

The values from the static models are inserted:

$$t_{obs} = \frac{(3.17 - 3.26) - 0}{\sqrt{\frac{0.16^2}{72} + \frac{0.124^2}{73}}} = -3.78$$

Next, the number of degrees of freedom is determined following formula:

$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}} = 133.75$$

Then the p-value is derived:

$2 * (1 - pt(3.78, df = 133.75)) \rightarrow P_{\text{value}} = 0.00023$

It can thus be concluded that the null hypothesis is then rejected( 0.00023 <0.05). It can therefore be deduced that there is a lot of evidence against $H_0$, i.e. the chance that the null hypothesis is true, when observing data, is extremely small. Thus, the null hypothesis is rejected, which means that the difference between men's and women's BMI is significant.

l) Comment on whether it was necessary to carry out the hypothesis test in the previous question, or if the same conclusion could have been drawn from the confidence intervals alone? (See Remark 3.59).

Previous calculations from the hypothesis test must be said to be unnecessary. This is because even the same result could have been derived by correctly reading the confidence intervals. If the intervals are examined for resp. women and men [22.9: 24.73] & [25.3: 26.82] are seen not overlapping. When the respective confidence intervals of two sets of observations do not overlap, it means that they are significantly different.

# Correlation

For the subsequent development of models describing BMI, we will also focus on the correlations between selected variables.

m) State the formula for computing the correlation between BMI and weight. Insert values and calculate the correlation. Furthermore, compute the remaining pairwise correlations involving BMI, weight and fast food. Make pairwise scatter plots of these variables. Assess whether the relation between the plots and the correlations is as you would expect.

When two observation variables are available for each unit, it may be interesting to quantify the relationship between the two. Thus, it is advantageous to quantify how the two variables can vary with each other, their covariance and their empirical correlation coefficient. To derive the correlation, the following formula is used:

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

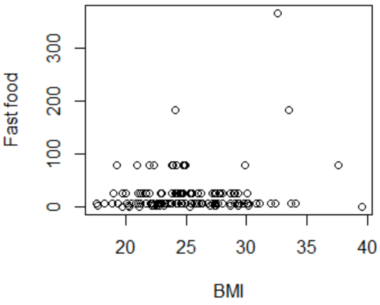
For the sake of clarity, some examples of the results of the different correlations are inserted in a table:

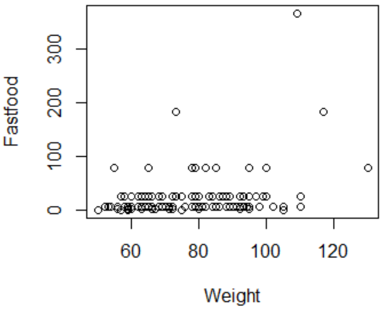
| **correlation** | **R code** | **Result R code** |
|---|---|---|
| BMI: weight | cov(D$bmi,D$wight) | 0.828 |
| BMI: Fastfood | cov(D$bmi,D$fastfood) | 0.153 |
| Weight: Fastfood | cov(D$weight,D$fastfood) | 0.279 |

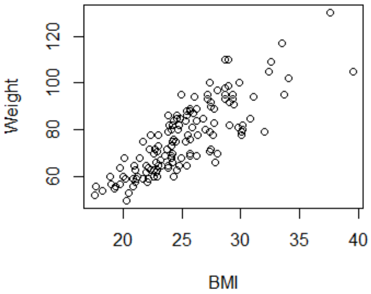
If one dives into the meaning of a correlation, it indicates a relationship between two variables. Thus, a change in one will predict a change in the other. Initially, the correlation between BMI and weight is examined. The calculation shows a strong correlation, which is also clarified by the scatter plot. The reason for this is to be found in the formula for BMI. The unit weight acts count in the formula, and relative to BMI, the tendency is for it to increase approximately linearly as the BMI increases.

Conversely, there is not a large correlation between BMI and fast food. It has a low correlation value of 0.153, which is due to lack of parameters in relation to health analysis. It is not known whether the people are more active, or for example have high muscle mass.

Likewise, there is not a large correlation between weight and fast food. It can therefore be concluded that, in this case, there is no correlation between either BMI


Fast food
BMI


Fastfood
Weight


Weight
BMI