Project 1: BMI survey

Formalities, structure and expectations for the first mandatory project

The assignment consists of two parts. The first part focuses on descriptive analysis of the data. The second part is primarily about confidence intervals and hypothesis tests.

The assignment is formulated in such a way that it can be solved in small "easy" steps. In practice, the assignment must be solved using the statistical software R. Some R code is provided in order to make it easy to get started with the project. However, the code is not complete, and you are encouraged to explore new features in R while working on the project. For example, you could add suitable titles to the plots, or use R's built-in functions for computing confidence intervals and testing hypotheses.

The results of the analysis must be documented in the report using tables, figures, mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in the appendix. Present the results of your analysis as you would when explaining them to one of your peers.

Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. R code should not be included in the report itself but must be handed in as an appendix (a .R-file). The report and appendix must be handed in under Assignments on Learn at: `Projekt 1: BMI undersøgelse`

The report text should not exceed 6 pages (excluding figures, tables, and the appendix). A normal page contains 2400 characters.

Figures and tables cannot stand alone - it is important that you describe and explain the R output in words.

Figures and tables are not included in the assessment of the length of the report. However, it is not in itself an advantage to include many figures, if they are not relevant!

You may work together in groups, but the report must be written individually. Questions about the project can be addressed to the teaching assistants, see the guidelines on the `Projects` page of the course website.

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.

For one, it is necessary to get an overview of the BMI of the Danish population. However, a study of the factors, that might influence a person's BMI score, e.g. gender, age and education, is needed as well. Other questions arise too: For example, is there an association between a person's BMI score and how often they eat fast food? Is it possible to make the population more aware of the role that fast food plays in the problem of overweight? Should fast food restaurants be required to disclose calorie counts on their menus?

A person's *BMI* (Body Mass Index) score is a measure of the person's overweight, defined by

$$BMI = \frac{weight}{height^2}, \tag{2-1}$$

where *weight* is the person's weight in kg and *height* is the person's height in metres (m). BMI is assessed according to the following table:

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)

The above table shows the WHO classification of BMI scores.[1]

Reading the data into R

Make a folder for the project on your computer. Download the project material from Learn and unzip it to the folder that you just made.

Then, open the data file `bmi1_data.csv` (e.g., in RStudio, File → Open File) in order to see the contents of the file. Note that the first row (referred to as a *header*) contains variable names, and that the subsequent rows contain the actual observations. Variable names and observations of the individual variables are separated by a ';' (therefore `.csv`: "comma separated values", though here it is a semi-colon).

The data file contains the following columns/variables:

Variable	**Explanation**
gender	The respondent's gender
height	The respondent's height in cm
weight	The respondent's weight in kg
urbanity	The size of the city in which the respondent lives
fastfood	Number of days per year on which the respondent eats fast food

Some of these variables are categorized. Further information on the variables in the dataset and how they are coded may be found in Appendix 1 (p. 11).

Open the file `bmi1_english.R`, which contains some R code that may be used in the analysis. First, the "working directory" must be set to the directory on the computer, which contains the files for the project:

[1] `http://www.si-folkesundhed.dk/upload/kap_21_overv%C3%A6gt_og_fedme.pdf`. Here you can also read about obesity as a risk factor for cardiovascular disease and diabetes.

```
## In RStudio the working directory is easily set via the menu
## "Session -> Set Working Directory -> To Source File Location"
## Note: In R only "/" is used for separating in paths
## (i.e. no backslash).
setwd("Replace with path to directory containing project files.")
```

Now the data may be read into R using the following code:

```
## Read data from bmi1_data.csv
D <- read.table("bmi1_data.csv", header=TRUE, sep=";", as.is=TRUE)
```

`D` becomes a "data.frame" (a kind of table), which contains the data that was read into R (see the introduction to R in Section 1.5 of the book).

Descriptive analysis

The purpose of the first part of the project is to carry out a descriptive analysis of the data. In a report it is important to present the data and describe it to the reader. For example, this can be done using summary statistics and suitable figures.

Start by running the following commands to get a simple overview of the data:

```
## Dimensions of D (number of rows and columns)
dim(D)
##  Column/variable names
names(D)
## The first rows/observations
head(D)
## The last rows/observations
tail(D)
## Selected summary statistics
summary(D)
## Another type of summary of the dataset
str(D)
```

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? Remember to consult the extended description of the variables in Appendix 1 (p. 11).

Before we can proceed with the analysis, we need to compute BMI scores corresponding to the observations of height and weight, and add these to the dataset as a new variable bmi. This may be done in R using the following code:

```
## Calculate BMI scores and add new variable to D
D$bmi <- D$weight/(D$height/100)^2
```

The following code may be used to generate a "density histogram" describing the empirical density of the BMI scores in the dataset (see Section 1.6.1):

```
## Histogram describing the empirical density of the BMI scores
## (histogram of the BMI scores normalized to have an area of 1)
hist(D$bmi, xlab="BMI", prob=TRUE)
```

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?

Note: In a *skewed* distribution, the probability mass is not symmetrically distributed around the median. In a left-skewed distribution, the left tail is longer than the right tail (and, typically, the mean will lie to the left of the median). Similarly, in a right-skewed distribution, the right tail is the longer of the two (usually, with the mean to the right of the median).

When doing data analysis, it is often useful to be able to divide the data into subsets. This can be done in R using, e.g., the subset function. See also the remark on p. 11.

Use the following R code to make two subsets of the data, one with the data for women and another with the data for men:

```
## Divide data into two subsets according to gender
Dfemale <- subset(D, gender == 0)
Dmale <- subset(D, gender == 1)
```

Now, the following code may be used to make separate density histograms of the BMI scores for women and men:

```
## Density histograms describing the empirical density
## of the BMI scores of women and men, respectively.
hist(Dfemale$bmi, xlab="BMI (female)", prob=TRUE)
hist(Dmale$bmi, xlab="BMI (male)", prob=TRUE)
```

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)?

The following R code makes a box plot of the BMI scores by gender:

```
## Box plot of BMI scores by gender
boxplot(Dfemale$bmi, Dmale$bmi, names=c("Female", "Male"),
        xlab="Gender", ylab="BMI")
```

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?

The empirical distribution of the BMI scores (for both genders combined, as well as separately for men and women) may also be quantified using summary statistics as in the following table:

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							
Women							
Men							

R code like the following may be used to fill in the empty cells in the table above (see also the remark on p. 12):

```
## Total number of observations
## (doesn't include missing values if there are any)
sum(!is.na(D$bmi))
## Sample mean (both genders combined)
mean(D$bmi, na.rm=TRUE)
## Sample variance (both genders combined)
var(D$bmi, na.rm=TRUE)
## etc.
##
## The argument 'na.rm=TRUE' ensures that the statistic is
## computed even in cases where there are missing values.
```

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?

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.

Confidence intervals and hypothesis tests

The following R code may be used to make a qq-plot. This plot can be used to investigate whether the natural logarithm of the BMI scores may be assumed to be normal distributed:

```
## New variable 'logbmi' with log-transformed BMI
D$logbmi <- log(D$bmi)
## qq-plot of log-transformed BMI
qqnorm(D$logbmi)
qqline(D$logbmi)
```

In practice, situations arise where it is *not* appropriate to assume that the assumptions of a model are satisfied. In these cases, one often considers whether a transformation of the data might improve the situation. Often, a logarithmic transformation can be helpful. (See Section 3.1.9). Note that after a log-transformation, the interpretation of the results on the original scale changes from the mean to the median, see 3.1.9. In this project, it has already been assessed that the statistical analysis should be performed using log-transformed BMI scores instead of the BMI scores themselves.

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.

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).

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:

$$H_0 : \mu_{\text{logBMI}} = \log(25),$$
$$H_1 : \mu_{\text{logBMI}} \neq \log(25).$$

Specify the significance level α, the formula for the test statistic, as well as the distribution of the test statistic (remember to include the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words. In particular, comment on whether it can be concluded that over half of the population is overweight.

Compare the results of the test with the results of the following R code:

```r
## Testing hypothesis mu=log(25) for log-transformed BMI
t.test(D$logbmi, mu=log(25))
```

We will now investigate whether a difference can be detected between the BMI of women and men.

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).

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.

	Lower bound of CI	**Upper bound of CI**
Women		
Men		

Compare the CI's for women with the results of the following R code:

```r
## Consider data for women only
Dfemale <- subset(D, gender == 0)
## Compute CI for mean log-BMI score of a woman
KI <- t.test(Dfemale$logbmi, conf.level=0.95)$conf.int
KI
## "Back-transform" to get a CI for median BMI score of a woman
exp(KI)
```

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.

Compare the results from the hypothesis test with the results of the following R code:

```
## Comparison of mean logBMI for women and men
t.test(D$logbmi[D$gender == 0], D$logbmi[D$gender == 1])
```

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).

Correlation

For the subsequent development of models describing BMI, e.g. in Project 2, 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 (note, insert only numbers in the correlation formula, i.e. three numbers). 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.

Compare the correlations computed above to the results of the following R code:

```
## Computing correlations between selected variables
cor(D[,c("weight","fastfood","bmi")], use="pairwise.complete.obs")
```

Appendix 1 Description of variables in the dataset and their coding

Variable	Meaning	Code
gender	The respondent's gender	0) Female 1) Male
height	The respondent's height in cm	
weight	The respondent's weight in kg	
urbanity	"In your own opinion, do you live in a big city, a city, or outside urban areas?"	1) Outside urban areas 2) City with less than 10,000 inhabitants 3) City with 10,000 to 49,999 inhabitants 4) City with 50,000 to 99,999 inhabitants 5) City with over 100,000 inhabitants
fastfood	"How often do you eat "fast food" e.g. McDonald's, Burger King, KFC, Sunset, Subway, Dominos, food from gas stations, hot dog stands, DSB kiosks, 7-Eleven, pizzerias, etc. " **Note**, that this variable has been "recoded" to "days per year" (DPY) (the value shown in parenthesis for each category).	1) Never (0) 2) Less than 1 time per year (1.0) 3) 1-11 times per year (6.0) 4) 1-3 times per month (24.0) 5) 1-2 times per week (78.2) 6) 3-4 times per week (182) 7) 5-6 times per week (286.7) 8) Every day (365)

Remark .1 Extra R tips

This is an optional extra remark about different ways to take subsets in R (useful but not necessary for solving the project):

```
## Optional extra remark about taking subsets in R
##
## A logical vector with a TRUE or FALSE for each value
## of a column in D, e.g.: Find all women in the data
D$gender == 0
## Can be used to find all the data for women
D[D$gender == 0, ]
## Alternatively, use the 'subset' function
subset(D, gender == 0)
## More complex logical expressions can be made, e.g.:
## Find all women who weigh less than 55 kg
subset(D, gender == 0 & weight < 55)
```

Remark .2 Extra R tips

Optional remark with some extra R tips. The table can also be generated more effectively using a ′for′-loop:

```r
## Use a 'for'-loop to calculate the summary statistics
## and assign the result to a new data.frame
Tbl <- data.frame()
for(i in 0:1){
  Tbl[i+1, "mean"] <- mean(D$bmi[D$gender == i])
  Tbl[i+1, "var"] <- var(D$bmi[D$gender == i])
}
row.names(Tbl) <- c("Women","Men")
## View the contents of Tbl
Tbl

## In R there are also more condensed ways to do such calculations.
## For example,
aggregate(D$bmi, by=list(D$gender), function(x){
  c(mean=mean(x), var=var(x))
})
## See more useful functions with: ?apply, ?aggregate and ?lapply
## For extremely efficient data handling see, e.g., the packages:
## dplyr, tidyr, reshape2 and ggplot2

## LaTeX tips:
##
## The R package "xtable" can generate LaTeX tables written to a file
## and thereby they can automatically be included in a .tex document.
##
## The R package "knitr" can be used very elegantly to generate .tex
## documents with R code written directly in the document. This
## document and the book were generated using knitr.
```