Practical: Descriptive statistics

0. Introduction to the dataset *surgery.txt*

This dataset contains data from 343 persons with overweight and obesity (BMI>25), that were treated using one of two possible surgical interventions: either a gastric band or a gastric bypass. Both interventions reduce the active volume of the stomach, but the latter one is much more invasive. BMI had been measured before (variable **BMIpre**) and after (**BMIpost**) the intervention.

Variables include:

id:	identification of the patient
surgery:	type of surgery 0 = gastric band; 1 = gastric bypass
sex:	gender of the patient
age:	age in years
weight:	weight in kilogram before the operation
length:	body length in meter
BMIpre:	BMI (= weight/length^2) before the operation
BMIpost:	BMI (= weight/length^2) after the operation
complic:	did complications occur? 0=no; 1 = yes
comorb:	did comorbidity occur (=presence of multiple pathologies) 0=no; 1=yes
depressi:	did the patient suffer from depression (0=no; 1=yes)
diabetes:	did the patient suffer from diabetes (0=no; 1=yes)
SES:	socio-economic status (1=low; 2=middle; 3=high)

1. Data reading

Read in the dataset and analyze the read-in dataset using `str(myData)`.

The variable **SES** (socio-economic status) is a categorical variable, but was read in as a numeric. Convert this variable to a factor and assign the value labels “low”, ”middle” and “high” to the 3 levels of the variable.

Define the variables **complic, comorb, depressi, diabetes** as factors. For all these factors, a zero means “no” and a 1 means “yes”. Assign these value labels through the function `factor()` and check the change using `str()` or using the RStudio “Environment” quadrant.

2. Descriptive statistics

You can summarize your data using `summary()`.

- Calculate the following: mean, median, minimum, maximum, first and third quartile, using separate commands.

Question
Why do you need to add the argument `na.rm=TRUE` to these commands? What happens if you don't? Check the help for the `mean()` function for the solution.

For a categorical variable a frequency table is more appropriate, using `table(myData$sex)`

- Calculate the mean age for females and males and correlation coefficient between weights and lengths

Questions:

1) Several types of correlation coefficients exist. Go to the help of the `cor()` function and find which correlation coefficient is calculated by default.
2) Add the appropriate argument to the `cor()` function in the previous command so that the Spearman correlation is calculated instead.

Some more summary statistics…

1) What is the maximum value of **weight** in the entire dataset?
2) How many missing values are there for **age**?
3) What is the mean value for **BMIpre** for the two surgical techniques (variable **surgery**)?
4) What is the Pearson correlation between the BMI before the operation (**BMIpre**) and the BMI after the operation (**BMIpost**)?

3. Graphics

- Generate a bar chart of a categorical variable for the gender.
- Generate a bar chart graph with mean age in males and females
- Make a histogram of a continuous variable "age".
- Make a scatterplot of 2 continuous variables *lengths* and *weights* (using formula notation `Y~X`) with different colors for each **SES**-value
- Add the three regression lines for each of the 3 SES groups
- Make a boxplot of age and a separate boxplots per group we use the formula-notation (`Y~X`).

Questions:

1) Make a histogram of **BMIpre**.
2) Make a bar graph of the socio-economic status (**SES**).
3) Make a boxplot of **BMIpre** for the 2 surgical intervention types defined by **surgery**.
4) Make a boxplot of **BMIpost** for the 2 surgical intervention types defined by **surgery**.
5) Make a scatterplot of **BMIpre** (Y-axis) versus **age**.
6) Make a scatterplot of **BMIpre** (Y-axis) versus **age**, using different markers for males and females. Add a line to the plot for each group.

4. Outlier detection

Prior to any statistical analysis, it is good practice to start with making graphs and descriptive statistics. This not only gives you an idea about the properties of your variables, but it also helps to identify errors in your dataset – values that are completely impossible or extremely unlikely. They have to be corrected or removed from the dataset.

- The previous boxplot of **BMIpost** versus **surgery**, shows one striking outlier. Identify this outlier.

In all upcoming exercises, you should work on the dataset with this outlier corrected to 29.41. You can proceed in several ways:

- Export the corrected dataset to a text file. In the next exercises, you can then start by reading in the corrected dataset.
- Start all your following exercises with reading in the original dataset, followed by the command that corrects the outlier.

5. Testing for normality

The choice of a statistical hypothesis test depends on the distribution of your (numeric) variables. Parametric testing, such as a t-test is only allowed if we have sufficient data points (at least 10) in each group, and if, in addition, the dependent variable has an (approximately) normal distribution per group.

- Check the normality using, histograms, QQ plots and Shapiro’s test, for **BMIpre** vs **BMIpost**
- What do you think?