

Chapter 1

Introduction, descriptive statistics,
Python and data visualization

Exercises

Contents

1 Introduction, descriptive statistics, Python and data visualization

Exercises	1
1.1 Infant birth weight	3
1.2 Course grades	4
1.3 Cholesterol	5
1.4 Project start	7
1.5 Descriptive statistics	8
1.6 Reading speed	9

1.1 Infant birth weight

|||| Exercise 1.1 Infant birth weight

In a study of different occupational groups the infant birth weight was recorded for randomly selected babies born by hairdressers, who had their first child. The following table shows the weight in grams (observations specified in sorted order) for 10 female births and 10 male births:

Females (x)	2474	2547	2830	3219	3429	3448	3677	3872	4001	4116
Males (y)	2844	2863	2963	3239	3379	3449	3582	3926	4151	4356

Solve at least the following questions a)–c) first manually and then by Python. It is OK to use Python as an alternative to your pocket calculator for the manual part, but avoid built-in functions that produce the result without forcing you to think about how to compute it during the manual part.

- What is the sample mean, variance and standard deviation of the female births? Express in your own words the story told by these numbers.
- Compute the same summary statistics for the male births, and compare the mean and standard deviation to the results for the female births. Based on these ten observations alone, would you say the difference is large?
- Find the five quartiles for each sample and draw the two box plots with pen and paper (i.e. not using Python).
- Are there any “extreme” observations in the two samples (use the *modified box plot* definition of extremeness)?
- What are the coefficients of variation in the two groups?

1.2 Course grades

||| Exercise 1.2 Course grades

To compare the difficulty of two different courses at a university the following grade distributions (given as number of pupils who achieved each grade) were registered:

	Course 1	Course 2	Total
Grade 12	20	14	34
Grade 10	14	14	28
Grade 7	16	27	43
Grade 4	20	22	42
Grade 2	12	27	39
Grade 0	16	17	33
Grade -3	10	22	32
Total	108	143	251

- a) What is the median of the 251 achieved grades?
- b) What are the quartiles and the IQR (Inter Quartile Range)?

1.3 Cholesterol

|||| Exercise 1.3 Cholesterol

In a clinical trial of a cholesterol-lowering agent, 15 patients' cholesterol (in mmol L^{-1}) was measured before treatment and 3 weeks after starting treatment. Data are listed below:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	9.1	8.0	7.7	10.0	9.6	7.9	9.0	7.1	8.3	9.6	8.2	9.2	7.3	8.5	9.5
After	8.2	6.4	6.6	8.5	8.0	5.8	7.8	7.2	6.7	9.8	7.1	7.7	6.0	6.6	8.4

- What is the median of the cholesterol measurements for the patients before treatment, and similarly after treatment?
- Find the standard deviations of the cholesterol measurements of the patients before and after treatment.
- Find the sample covariance between the cholesterol measurements before and after treatment.
- Find the sample correlation between the cholesterol measurements before and after treatment.
- Compute the 15 differences ($\text{Dif} = \text{Before} - \text{After}$) and do various summary statistics: sample mean, sample variance, sample standard deviation. Also draw a boxplot.

- f) Considering such data, the key question is whether the average decrease in cholesterol can be “shown statistically”. Which summary statistics and/or plots would you examine to get an initial idea of the answer? (The formal answer is given in Chapter 3.)

1.4 Project start

||| Exercise 1.4 Project start

- a) Go to DTU Learn and read the first project description. Follow the steps to import the data into Python and get started with the exploratory data analysis.

1.5 Descriptive statistics

|||| Exercise 1.5 Descriptive statistics

The table below shows a sample of 10 measurements. The data is reported in sorted order.

Measurement (x)	256	258	266	296	314	318	326	353	380	391
---------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

- a) What is the mean of the sample?
- b) Find the minimum, maximum and the quartiles (Q_1 , Q_2 and Q_3) of the sample.
- c) What is the sample range and the IQR (Inter Quartile Range)?
- d) Compute the 35th percentile of the sample.
- e) Are there any “extreme” values in the sample? (Use the *modified box plot* definition of extremeness.)
- f) Draw a boxplot of the sample (using pen and paper).

The table below shows another sample of only 5 measurements (not reported in sorted order).

Measurement (y)	2	5	3	4	3
---------------------	---	---	---	---	---

- g) Draw a cumulative distribution plot of the sample (using pen and paper).

1.6 Reading speed

|||| Exercise 1.6 Reading speed

In an elementary school, 5 pupils participated in a short training class to improve reading speed. The number of pages each pupil could read within a certain amount of time was measured before and after the training:

Pupil	1	2	3	4	5
Before training	3	2.5	4	3	3.5
After training	3.5	3	5	2.5	4

- Draw a scatter plot of the data (using pen and paper). Express in your own words the story told by this plot.
- What is the mean number of pages read, both before and after the training? By how much did the mean increase/decrease?
- What is the standard deviation of the number of pages read, both before and after the training? How does the spread compare?
- What is the CV (coefficient of variation) of the number of pages read, both before and after the training? What does the comparison tell you?
- Express in your own words the story told by the numbers calculated above.
- Find the sample covariance between observations before and after training. What does the sign of this number tell you about the relationship between pupils' reading speed before and after training?

- g) Find the sample correlation between observations before and after training. Express in your own words how you would interpret this result. Does the result fit with the plot you made in question (a)?
- h) Calculate the 5 differences ($\text{Dif} = \text{After} - \text{Before}$) to obtain the achieved “effect” for each pupil. Calculate the mean of these differences. Express in your own words how you could use these differences (instead of the raw data) to analyse the effect of the training class on reading speed.
- i) Observing such data, the big question is whether an average increase in reading speed can be “shown statistically”. How to formally answer this question is presented later in the course, but consider now how you would answer such a question.