Appendix A

Preparation and examination

Content *Page*

Case study 1: Insight into consumer populations

In this case study you may work in groups (max. 4 students) to perform a statistical analysis in R, write a small report, and present the results to your client. For this case study you will need the data set **`shaving.csv`** from the online resources. The case you will be working on this week is the following:

Your team is working in the advisory department of a Big 4 Accounting firm that has been hired by a client to gain insight about their consumer population. The client in question is a company that produces shaving products, and they have already asked a random sample of 300 customers to provide a rating of their flagship product. Other than giving a rating of the product, the participants were also asked about their highest education and in what sector they work. The client's management wants to specifically know what type of customers rate their product the highest.

Write a small report to the client's management where you give your recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

1. A short introduction about the reason and topic of the report.
2. What the variables in the data set represent on a practical level and what their measurement level is.
3. A table containing the descriptive statistics for the rating in the sample of customers, split by male or female. Report, only where meaningful, the frequencies or means for each variable. Discuss what observations you can make using this table.
4. For the highest scoring gender, a table containing the descriptive statistics for the rating in the sample of customers, split by education level.
5. For the highest scoring education and gender, a table containing the descriptive statistics for the rating in the sample of customers, split by job sector. Discuss what observations you can make using this table.
6. For the highest scoring education and gender and job sector, report the mean, median, and mode of the rating. Using these measures, describe the distribution of the rating in this group. Discuss what observations you can make from this distribution.
7. An argumentation about whether you think this sample is reliable enough (or not) to make statements like this.
8. Your answer to the question posed by the client's management.

Case study 2: ...

In this case study you may work in groups (max. 4 students) to perform a statistical analysis in R, write a small report, and present the results to your client. For this case study you will need the data set **x.csv** from the online resources. The case you will be working on this week is the following:

Case text

Write a small report to the client's management where you give your recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

1. A short introduction about the reason and topic of the report.
2. What the variables in the data set represent on a practical level and what their measurement level is.
3. Your answer to the question posed by the client's management.

Case study 3: ...

In this case study you may work in groups (max. 4 students) to perform a statistical analysis in R, write a small report, and present the results to your client. For this case study you will need the data set **`x.csv`** from the online resources. The case you will be working on this week is the following:

Case text

Write a small report to the client's management where you give your recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

1. A short introduction about the reason and topic of the report.
2. What the variables in the data set represent on a practical level and what their measurement level is.
3. Your answer to the question posed by the client's management.

Case study 4: ...

In this case study you may work in groups (max. 4 students) to perform a statistical analysis in R, write a small report, and present the results to your client. For this case study you will need the data set **`x.csv`** from the online resources. The case you will be working on this week is the following:

Case text

Write a small report to the client's management where you give your recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

1. A short introduction about the reason and topic of the report.
2. What the variables in the data set represent on a practical level and what their measurement level is.
3. Your answer to the question posed by the client's management.

Case study 5: ...

In this case study you may work in groups (max. 4 students) to perform a statistical analysis in R, write a small report, and present the results to your client. For this case study you will need the data set **`x.csv`** from the online resources. The case you will be working on this week is the following:

Case text

Write a small report to the client's management where you give your recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

1. A short introduction about the reason and topic of the report.
2. What the variables in the data set represent on a practical level and what their measurement level is.
3. Your answer to the question posed by the client's management.

Case study 6: ...

In this case study you may work in groups (max. 4 students) to perform a statistical analysis in R, write a small report, and present the results to your client. For this case study you will need the data set **x.csv** from the online resources. The case you will be working on this week is the following:

Case text

Write a small report to the client's management where you give your recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

1. A short introduction about the reason and topic of the report.
2. What the variables in the data set represent on a practical level and what their measurement level is.
3. Your answer to the question posed by the client's management.

Case study 7: Assessing financial fraud using Benford's law

In this case study you may work in groups (max. 4 students) to perform a statistical analysis in R, write a small report, and present the results to your client. For this case study you will need the data set **`sinoForest.csv`** from the online resources. The case you will be working on this week is the following:

You and your team are part of the audit division at a Big 4 accounting firm that is hired to perform an audit of the Sino-Forest corporation, (which was) the leading commercial forest plantation operator in China[5]*. To determine whether it is likely that the data have been tampered with, your team decides to apply the principles of Benford's law to the data.*

Write a small report to the client's management where you give your recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

1. A short introduction about the reason and topic of the report.
2. What the variables in the data set represent on a practical level and what their measurement level is.
3. Your answer to the question posed by the client's management.

[5]The Sino-Forest corporation was accused of fraud in 2011, and on July 13, 2017, it was decided that four individuals at the corporation had indeed committed fraud.

Practical assignment

In this assignment you are asked to choose a data set from the online resources[6], formulate two research questions about the variables in this data set, and explore the data using the knowledge and techniques acquired throughout this workbook. For this, you will need to explore your data set using descriptive statistics and visualization, formulate two hypotheses, perform the two relevant statistical tests and draw conclusions for your research questions.

Your first test should contain a comparison of means or proportions performed using the formulas on page 140, while the second test should contain a linear regression performed using R.

1. **(3 points)** Shortly describe the data set you have chosen and formulate two research questions that you will explore. Remember, the first research question should involve a comparison of two means or two proportions, while the second research question should involve a linear relationship between two variables. Write down the business reasoning behind your research questions.
2. **(2 points)** Discuss the variables in your research questions using the appropriate descriptive statistics. Create relevant figures for your research questions and discuss what you can learn from these figures.
3. **(6 points)** Perform a test on your data set using a comparison of means or proportions.
    a. **(2 points)** Discuss which test you are going to perform and how this relates to the parameters in the first research question. Indicate what parameter(s) you will need to investigate to answer this research question. Next, formulate the appropriate (statistical) null hypothesis H_0 and alternative hypothesis H_1. Make sure that your hypotheses are formulated correctly.
    b. **(2 points)** Calculate the corresponding test statistic (e.g., t, z) from your data using the formulas on page 140. Write down the formulas and explain your calculations.
    c. **(2 points)** Define the critical area for your test statistic using Table 2 (z-values) on page 143 or Table 3 (t-values) on page 144. Based on the critical value, draw the conclusion about your hypotheses. Explain how you got to this answer.
4. **(6 points)** Perform a test for a linear relationship using linear regression in R.
    a. **(2 points)** Discuss which test you are going to perform and how this relates to the second research question. Write down the regression equation and indicate what parameter(s) you will need to investigate to answer this research question. Formulate the appropriate (statistical) null and alternative hypothesis. Make sure that your hypotheses are formulated correctly.
    b. **(2 points)** Perform the linear regression using R. Provide the relevant R output in your paper and incorporate the relevant statistics in your answer.
    c. **(2 points)** Based on your results in R, what do you conclude about your hypotheses? Explain how you got to this answer.
5. **(1 point)** Discuss what your results imply for your population of interest and how they relate to your research questions.
6. **(2 points)** Discuss how representative your sample is and which risk applies to your conclusions: *type-I* or *type-II*, and what this implies for your business scenario.

[6]The data sets that you can choose from can be found in the folder *Practical assignment*.

Final exam 2019 - 2020

Question 1 (10 points)

1a) (2 points) Give the correct measurement levels *(nominal, ordinal, interval, ratio)* for the following variables:

- Rating (low, medium, high)
- Distance (in meters)
- Hair color (blonde, brown, black, red)
- Temperature (in $^{\circ}C$)

Use the following numbers to answer the next questions.

You are given the following sample:

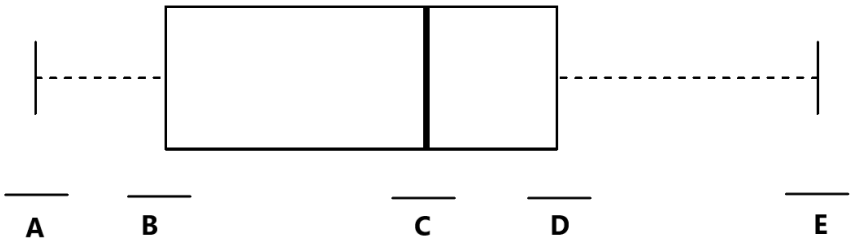
4 2 5 6 2 1 2 5 3 3 7 4 5 1 7 5 3

1b) (3 points) Calculate the mean, median, and mode of this sample.

1c) (2 points) Is the distribution of these values *positively skewed* or *negatively skewed*? Explain your answer using the relation between the mean, the median, and the mode.

1d) (2 points) Find the minimum, the lower quartile, the upper quartile, and the maximum of this sample.

1e) (1 point) Give the values that are supposed to be on the dots *A, B, C, D* and *E* on the basis of your calculations in the previous questions.



Question 2 (10 points)

2a) (2 points) Write down the Central Limit Theorem.

One of the four assumptions for parametric tests is normally distributed data.

2b) (2 points) Explain why parametric tests require normally distributed data.

2c) (1 point) What test do you use to objectively test the normality of a distribution?

2d) (1 point) What are the other three assumptions for parametric tests based on the normal distribution?

Use Table 2 on page 143 to answer the following two questions.

2e) (2 points) Calculate the 95% upper confidence bound of the population mean μ for a random sample with sample size $n = 16$, a sample mean $\bar{x} = 38$ and a sample standard deviation $s = 4$.

2f) (2 points) Calculate the 95% two-sided confidence interval of the population mean μ for a random sample with sample size $n = 81$, a sample mean $\bar{x} = 250$ and sample standard deviation $s = 34$.

Question 3 (15 points)

Recently, there has been a lot of media attention towards the amount of lead in the water in the Netherlands. In the Netherlands, the maximum legal amount of lead per liter of water is 10 micro grams. A water researcher is hired to check the lead levels of a high school in the Netherlands. From a previous grand measurement within the school, it is known that the variance in lead in all tabs of the school is $\sigma^2 = 2.3$. The water researcher takes a sample of water from 50 tabs in the school and measures the amount of lead per liter (in micro grams) for these observations. He finds a sample mean of $\bar{x} = 7.5$.

3a) (2 points) Write down the statistical null hypothesis H_0 and the statistical alternative hypothesis H_1 if the water researcher wants to show that, with 95% confidence, the mean lead level (in micro grams) in the tabs is lower than the legal maximum amount.

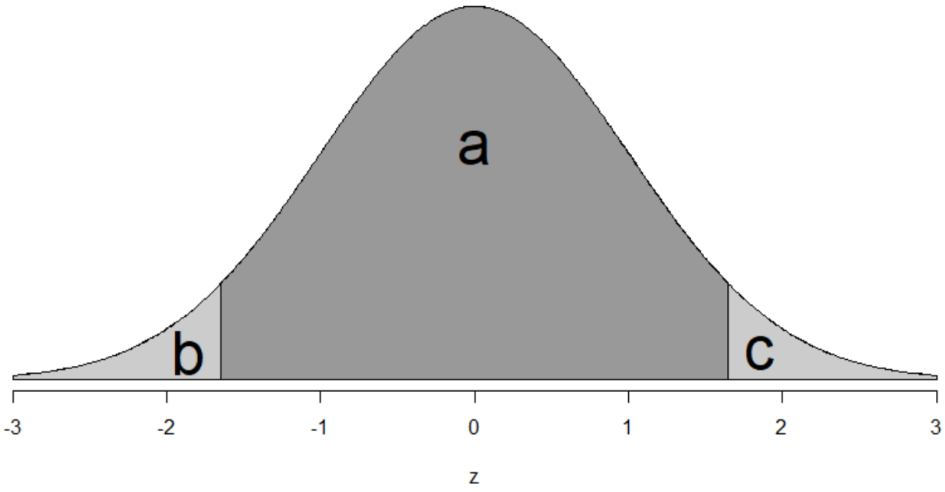
3b) (2 points) Give the critical z-value for this test. You may base this critical z-value on the information in Table 2 on page 143.

3c) (2 points) Calculate the z-value from the sample.

3d) (4 points) Draw a conclusion about on the basis of the sample z-value. Include the following four elements:

- How the calculated z-value relates to the critical z-value.
- Whether this implies that H_0 is rejected or not.
- What this tells you about the mean lead level μ in the school.
- What type of error is relevant (*type-I* or *type-II*).

The figure below depicts the probability density function for a standard normal distribution. The z-score is displayed on the x-axis.



3e) (3 points) Give the letter of the area that represents the p-value of the last test in the figure above. Explain your answer using the definition of the p-value.

3f) (2 points) State whether, for this test, the p-value is lower or higher than 0.05. Argue how you arrived at this answer.

Question 4 (15 points)

Researchers watched the Netflix documentary 'The Game Changers' and want to test the effect saturated fats have on the blood flow in a body. They did the following experiment: Ten healthy male subjects fasted for 10 hours and then got a high-fat meal. On another occasion they also fasted for 10 hours and then got a low-fat meal. Their endothelial function (how well the blood flows through small arteries to the various tissues) was measured 4 hours after this meal.

The mean endothelial function was 8.2% with a standard deviation of 3.7% after eating the high-fat meal and 13.7% with a standard deviation of 3.3% after the same men ate the low-fat meal. The average difference was 4.9% with a standard deviation of the difference of 6.1%. The researchers want to show that, with 95% confidence, the endothelial function is significantly lower after eating a high-fat meal than after eating a low-fat meal. They have already found the critical t-value to be 2.262 for this case.

4a) (11 points) Perform a two-sample t-test for this case; include the following steps:

- Explain whether this an independent or a dependent sample t-test,
- Write down the statistical hypotheses,
- Calculate the sample t-score,
- Draw the conclusion; make sure you include the following four elements:
  - How the calculated t-value relates to the critical t-value,
  - Whether this implies that H_0 is rejected or not.
  - What this tells you about the the endothelial function.
  - What type of error is relevant (*type-I* or *type-II*).

4b) (1 point) When an independent two-sample t-test for the mean does not satisfy the homogeneity of variance assumption, what alternative test can you use?

4c) (3 points) Explain why a dependent two-sample t-test for the mean does not have to satisfy the homogeneity of variance assumption.

Question 5 (15 points)

A baby food factory has a linear regression model to predict the production price of baby food (€) on the basis of its sugar content (mg) and vitamins (mg) in the food. They have fitted this model to a sample ($n = 50$) of the 200 types of baby food they produce. The results indicate that, on average, the base price of baby food without any sugar or vitamins is €0,60. Every mg of vitamins that the factory adds to the baby food raises its production price by €0,10. Every mg of sugar that is added to the baby food raises its production price by €0,15.

5a) (3 points) Write down the regression equation for the linear model for all 200 types of baby food in the population. Use the names 'price' for the cost price (€) of the food, 'sugar' for the amount (mg) of sugar in the food, and 'vitamins' for the vitamins (mg) in the food. Do not fill in the values of the parameters yet.

5b) (1 point) Fill in the values of the parameters in the regression equation using the information provided by the text.

The results of the fitted model indicate that the model sum of squares is 200. The total sum of squares of the model is 250, and the residual sum of squares of the model is 50.

5c) (1 point) Calculate the explained variance (R^2) of this linear model.

5d) (1 point) Interpret the R^2 of this linear model.

5e) (3 points) Explain the problem with using the R^2 to compare the fit of several linear models. What alternative statistic did you learn that you can use to more reliably compare two linear models?

In order to differentiate between baby food for 0-9, 9-12, and 12+ months, the factory's data scientist adds two variables to the data set. The first variable ('9-12') contains a 1 for baby food in the category 9-12 months, and 0 otherwise. The second variable ('12+') contains a 1 for baby food in the category 12+ months, and 0 otherwise.

5f) (1 point) What type of variable are the variables '9-12' and '12+'?

5g) (1 point) What test is this linear model equivalent to?

5h) (4 points) Calculate the F-value of this linear model.

Question 6 (10 points)

The Netherlands is a rainy country and historically most rain falls during the fall and early winter and is evenly spread over these four months. Students at a university in the Netherlands believe that climate change does not only affect the temperature or the amount of rainfall but also when the rain falls. They looked up the historical distribution of total rainfall in their university town for September until December on a weather website and also measured the total rainfall in the town themselves in this period in 2019. The results are shown in the table below.

	September	October	November	December	Total
Historical distribution (%)	23.7	26.3	25.4	24.6	100
2019 measurements (ml)	102	126	113	88	429

6a) (8 points) Perform a Pearson chi-square (X^2) test to find out if the rainfall distribution in the four months September - December in 2019 significantly deviates from the historical distribution at a 90% confidence level. Use a critical chi-square ($df = 3$) of 6.251; include the following steps:

- How the calculated X^2 score relates to the critical X^2 value.
- Whether this implies that H_0 is rejected or not.
- What this tells you about the 2019 rainfall distribution.
- What type of error is relevant (*type-I* or *type-II*).

6b) (2 points) What is the minimum required amount for the expected frequency in all categories for the Chi-square test and what test can be done when this requirement is not met?

Question 7 (15 points)

Your friend works as an employee of an advertising company. The company does the branding for a new startup and your friend is in charge of the advertising budget. The first six weeks of advertisement have just ended, and to assess the effectiveness of the advertising campaign your friend is interested in the relationship between the money that they have spent and the number of products that were sold as a consequence. Therefore, your friend collects data on the amount of money spent (variable x) and the number of products sold (variable y). The data is provided in the table below.

i	x	y
1	100	40
2	120	60
3	125	45
4	110	30
5	170	70
6	150	60

7a) (6 points) Calculate the covariance of the variables x and y.

7b) (5 points) Calculate the correlation between the variables x and y.

Your friend wants you to assess, with 95% confidence, whether the correlation is truly positive in the population.

7c) (1 point) Write down the statistical null hypothesis H_0 and the statistical alternative hypothesis H_1 for a left-tailed one-sided test of the correlation in the population.

7d) (1 point) Calculate the sample t-value for a test of a correlation against zero.

The critical t-value for a left-tailed one-sided test with 5 degrees of freedom and 95% confidence is 2.015.

7e) (2 points) Draw the conclusion about the population correlation for your friend. Include the following elements:

- How the calculated t-value relates to the critical t-value.
- Whether this implies that H_0 is rejected or not.

This page is intentionally left blank.