Appendix A

Preparation and examination

Content *Page*

Case study 1: Providing insight into consumer populations

In this case study you can 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 respondents were also asked about their highest education and in what sector they are currently employed. The client's management wants to specifically know what category of customers rate their product the highest.

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

- A short introduction about the reason and topic of the report.
- A description of each variable in the **`shaving.csv`** data set, including what it represents and what its measurement level is.
- A table containing the descriptive statistics for the rating in the sample of customers, split by the respondents' gender. Report, only where meaningful, the frequencies or means for each variable. Discuss what you can learn from this table.
- For the highest rating gender, a table containing the descriptive statistics for the rating in the sample of customers, split by education level. Discuss what you can learn from this table.
- For the highest rating gender and education combination, a table containing the descriptive statistics for the rating in the sample of customers, split by job sector. Discuss what you can learn from this table.
- For the highest rating gender, education, and job combination, report the mean, median, and mode of the rating. Using these measures, describe the distribution of the rating in this particular category of customers. Discuss what observations you can make from this distribution with respect to skewness.
- An argumentation about whether you think this sample is reliable enough (or not) to make a good decision about all the customers of the company.
- On the basis of your results, the answer to the question(s) posed by the client's management and an explanation of how you arrived at this answer.

Case study 2: Finding the ideal location for a start-up

In this case study you can 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 sets **`UNresolutions.csv`** [5], **`US.csv`** and **`Russia.csv`** from the online resources. The case you will be working on this week is the following:

You and your team are working for a consultancy agency that advises start-up businesses about financial opportunities. You have been hired by a startup business that has developed a technological innovation with applications in both the agriculture and health sectors. However, your client has an internal conflict between its two co-founders. One wants to introduce their innovative product in the US first and the other wants to introduce the product in Russia first. You have been hired to provide advice about whether the start-up should first go to Russia, or first go to the United States of America, and in what sector it is best to introduce their innovation. Due to the technological nature of their product, it is important for the client that, whatever country they settle in, expresses views that are in line with that of the United Nations.

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

- A short introduction about the reason and topic of the report.
- A description of each variable in the **`resolutions.csv`** data set, including what it represents and what its measurement level is.
- A figure that shows the amount of positive votes of the United States of America and the Russian Federation together, and over time. Discuss what observations you can make from this figure.
- A figure that shows the average income per sector for the United States of America. Discuss what observations you can make from this figure.
- A figure that shows the average income per sector for the Russian federation. Discuss what observations you can make from this figure.
- On the basis of your results, the answer to the question(s) posed by the client's management and an explanation of how you arrived at this answer.

[5] These source of this data set is publicly available at https://doi.org/10.7910/DVN/LEJUQZ.

Case study 3: Estimating the total misstatement in an audit

In this case study you can 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 **`BuildIt.csv`** from the online resources. The case you will be working on this week is the following:

Your team works in an audit of a construction company BuildIt and has to determine to which extent the financial statements of the company are presented in a true and fair way. To make a statement about the total misstatement in the financial statements, your teams must take a sample recorded transactions in the financial statements, audit these transactions, and write down their true value (audit value). Next, you investigate the difference between the book value and the audit value in the sample to estimate the total error in the population. According to industry-specific demands, the maximum allowed misstatement is five percent of the total value of the population. The client is interested in whether they comply with these industry demands.

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

- A short introduction about the reason and topic of the report.
- What the variables in the data set represent on a practical level and what their measurement level is.
- An explanation of the chosen sample size and how you selected the sample.
- A calculation where you estimate the mean difference between the book values and the audit values using a confidence interval.
- A clear explanation of the interpretation of this confidence interval.
- Your opinion about the size of the chosen sample and its effect on the precision of the confidence interval.
- On the basis of your results, the answer to the question(s) posed by the client's management and an explanation of how you arrived at this answer.

Case study 4: Determining predictors of credit rating

In this case study you can 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 **`creditRating.csv`** from the online resources. The case you will be working on this week is the following:

Your team works for a large bank and is in charge of designing an algorithm that automatically assigns people a credit rating, indicating whether they are likely to default on a loan they received from the bank. To predict this credit rating, management has asked of your team to find out what factors should be considered in the prediction.

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

- A short introduction about the reason and topic of the report.
- What the variables in the data set represent on a practical level and what their measurement level is.
- A visualization of the linear relationships in the data set, providing an indication of what to look for.
- A linear regression predicting the credit rating of a person using all other sensible variables in the data set.
- A check of the assumptions of your linear regression analysis and a statement of whether these are violated.
- On the basis of your results, the answer to the question(s) posed by the client's management and an explanation of how you arrived at this answer.

Case study 5: Assessing effectiveness of changes via an A/B test

In this case study you can 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 **`abTest.csv`** from the online resources. The case you will be working on this week is the following:

Your team works as a data analytic task force for a social media company. Your company is constantly testing different layouts for the company web page via an A/B test. In your current A/B test, you want to determine whether a green "Like" button on your web page works better than the default blue "Like" button. Management is interested in whether they should put the green button on the web page instead of the blue button.

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

- A short introduction about the reason and topic of the report.
- What the variables in the data set represent on a practical level and what their measurement level is.
- A statistical formulation of the hypotheses that you are investigating.
- What test you will be using to investigate these hypotheses. Also elaborate on whether the assumptions of the parametric variant of this test are violated and if so, what non-parametric test to use.
- A test of the mean return frequency in the blue group against the mean return frequency in the green group.
- On the basis of your results, the answer to the question(s) posed by the client's management and an explanation of how you arrived at this answer.

Case study 6: Performing a multi-group survey

In this case study you can 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 insights and recommendations. The report should describe your analysis and how you arrived at your conclusions. It should at least contain the following points:

- A short introduction about the reason and topic of the report.
- What the variables in the data set represent on a practical level and what their measurement level is.
- On the basis of your results, the answer to the question(s) posed by the client's management and an explanation of how you arrived at this answer.

Case study 7: Assessing potential data tampering using Benford's law

In this case study you can 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`** [6] 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[7]*. You client is interested in knowing whether the data is likely to be tampered with. To determine whether this tampering has likely occurred, your team decides to apply the principles of Benford's law to the data set.*

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

- A short introduction about the reason and topic of the report, including a description of what Benford's law represents and how it can be tested.
- What the variables in the data set represent on a practical level and what their measurement level is.
- A table containing the frequency of the first digits in the data set, alongside a table containing the expected frequency of the first digits under Benford's law.
- A test of the frequencies of the first digits in the data against their expected frequencies under Benford's law.
- On the basis of your results, the answer to the question(s) posed by the client's management and an explanation of how you arrived at this answer.

[6] This data set is featured in the R package `benford.analysis`.

[7] 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 individual assignment you are asked to investigate a data set, formulate one or more research questions, and explore these questions using the knowledge and techniques you acquired in this course. More concretely, you will need to explore these data using descriptive statistics and visualization, investigate two research questions, perform the appropriate statistical tests and draw sound conclusions. The first test should contain a comparison of means or proportions performed using the formulas on page 138, the second test should contain a linear regression performed using R. There are 5 questions for a total of 20 points. Your assignment should be in essay form and should describe your analysis. All included tables should be formatted in *APA* style.

The scenario is as follows: you are a data analyst at a small supermarket chain that has recently been publicly criticised for its high work pressure. Management has ordered an internal investigation and has asked 300 out of the 500 employees at the firm to fill in a questionnaire about their experiences with high work pressure. You are given the task to find out what factors underlie this high level of stress amongst employees.

The data set that you will investigate is unique, and can be found in Canvas under as your student number (e.g., `1020183.csv`).

1. **(3 points)** Discuss to what extent there is a difference or a relation between the variables *Gender*, *Age*, and *Stress* in your data set using the appropriate descriptive statistics.
    a. **(1 point)** Create a table that provides the sample size, the mean, the variance, and standard deviation for the *Age* and *Stress* level of the group of *Males* and the group of *Females* in the data set. Include this table in your report and discuss what you can learn from it.
    b. **(1 point)** Create two visualizations that provide information about the spread within these groups with respect to these two variables. Include these visualizations in your report and discuss what you can learn from it.
    c. **(1 point)** Discuss the representativeness of this sample and the validity of your conclusions when making decisions about whether the high stress levels are caused by an *Age* or *Gender* differences on the basis of this information.
2. **(3 points)** Formulate two research question(s) that you will explore using the two required statistical tests to find out what causes the high stress levels among employees. Use new variables in each research question.
    a. **(1 point)** Create a plot that has all the variables plotted against each other to discover possible relations in the data set, and discuss what you can learn from it.
    b. **(1 point)** Write down the reasoning behind your research question(s). This reasoning should be based on theoretical and visual (from the figure) information.
    c. **(1 point)** Indicate what variable(s) and statistical parameter(s) you will investigate in your research questions.
3. **(6 points)** Test your first research question using a test of means or proportions.
    a. **(2 points)** Discuss the test you are going to perform and how this helps you determine factors that may contribute to the high stress levels. Formulate the appropriate (statistical) null and alternative hypothesis. Make sure that your hypotheses are formulated in terms of the statistical parameters you formulated earlier.
    b. (2 points) Calculate the corresponding test statistic (e.g., t, z, F, X^2) from your data using the formulas learned in this course. Include the full calculation (including filled in formulas) in your report. You may use R to perform the calculations.
    c. **(2 points)** Define the critical area for your test statistic using the formulas learned in this course. Based on your previous calculations, draw a sound four-part conclusion. Argue how you got to this answer.
4. **(6 points)** Test your second research question using linear regression.
    a. **(2 points)** Discuss the test you are going to perform and how this helps you determine factors that may contribute to the high stress levels. Formulate the appropriate (statistical) null and alternative hypothesis. Make sure that your hypotheses are formulated in terms of the statistical parameters you formulated earlier.
    b. **(2 points)** Perform the corresponding linear regression using R. Provide the relevant output in your paper and discuss what you can learn from it.
    c. **(2 points)** Based on the results of your analysis in R, formulate a sound four-part conclusion about your hypotheses? Argue how you got to this answer and discuss the relevant statistics that informed your decision.
5. **(2 points)** Discuss what risks are associated with making decisions on the basis of your performed tests, and what this means for your firm's management business scenario.

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:

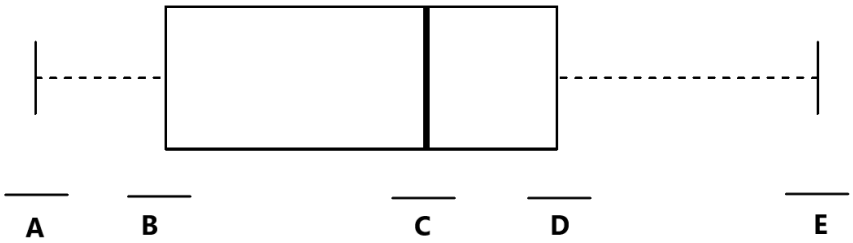
4 2 5 6 2 1 2 5 3 3 7 4 5 1 7 5 3

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 141 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.

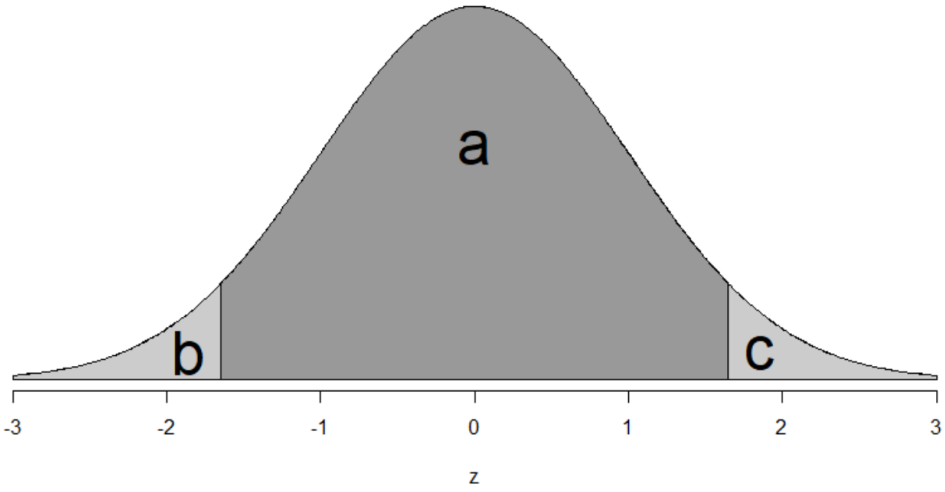
3b) (2 points) Give the critical z-value for this test. You may base this critical z-value on the information in Table 2 on page 141.

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.