Bachelor Informatik:
Exam Statistics WS 2019/20
Online Version
Prüfer: Falkenberg, Schrader

Name:

Matriculation Number: Klicken oder tippen Sie hier, um Text einzugeben.

I hereby confirm that I am physically fit to take the examination and that I have revised the exam alone without the help of third parties.	☒

Some hints:

- Please mention, not only the solution but the derivation of the solution has to be given.
- Insert your solutions to the individual questions at the appropriate places in the document and upload the document to Moodle at the end of the exam.
- If you want to add a hand written solution, please copy the sheet with hand written solution by your smartphone and copy the corresponding file in the document.
- To copy a by R generated diagram click on the plots window "Export" and then click "Copy to Clipboard". Copy the diagram by pressing the "Copy Plot" button in the "Copy Plot to Clipboard" windows. With ctrl-v you can paste the diagram in the document.

Descriptive Statistics

The csv file class.data.csv contains data from a sample of 100 employees of a large company. The file includes information on the first and last name, sex, age, health insurance number, canteen valuation, body weight, and height of the employees.

Answer the following questions. Write your answer directly into this document below the corresponding question. Copy your R-code and a copy of your generated diagrams into this document, too.

1. Import the csv file into a tibble class.data. Use the R function read_cvs2() to do this.

Write your code here!

2. Specify the type and scale for all variables in the dataset.

Please write your answer here!

3. The weight is in pounds and the height is in inches. Change the values of the variables so that the weight is given in kg and the height is given in m.
   Hint: 1 inch = 2.54 cm, 1 lb = 0.45 kg

Write your R-Code here!

4. Add a new variable bmi to the tibble class.data, which contains the “Body Mass Index”. The Body Mass index is defined as “weight in kg / (height in m)^2”

Write your R-Code here!

5. Add a new variable bmi.category to the tibble class.data, which contains the values of the Body Mass Index categories “underweight” (bmi < 18.5), “normalweight” (18.5 <= bmi < 24.5), “overweight” (24.5 <= bmi < 30” and “obesity” (bmi > 30).

Write your R-Code here!

6. Calculate the quartiles, mean, standard deviation, min, max for the variable height grouped by gender.

Write your R-Code here!

Copy your R-Output here!

7. Create a side by side boxplot of the variable height with respect to the variable gender.

Write your R-Code here!

Copy your diagram here!

8. How do the values of the variables body height differ with respect to gender in terms of location and variability? Use the side by side boxplot to answer the question!

Write your answer here!

9. Create a histogram of the variable bmi where the classes given by the categories of bmi.

Write your R-Code here!

Copy your diagram here!

10. Determine the parameters of the linear regression line weight = a + b*height.

Write your R-Code here!

Coefficients:

11. Interpret the parameters a and b of the regression line.

Write your answer here!

12. Plot a scatterplot of all pairs (height, weight) including the regression line weight = a + b*height.

Write your R-Code here!

Copy your diagram here!

13. Determine a contingency table of the variables gender and bmi.category.

Write your R-Code here!

Copy your R-output here!

Probability

1. In a bag are 8 fair dice and 2 dice without a six. One die is randomly picked out of the bag and rolled 2 times.
   a) Calculate the probabilities for getting no sixes for all possible values of the numbers of sixes in rolling the picked die 2 times.

Write your answer here!

   b) In rolling the picked die twice, calculate the probabilities for getting the different possible values of sixes.

Write your answer here!

   c) How many times n should the randomly picked die be rolled at least that the probability of a fair die if there are no sixes in rolling the die n times is at most 0.5?

Write your answer here!

2. Every German consumes on average 123 l of drinking water per day. Assume that the water consumption of Germans per day can be described by independent and identically normally distributed random variables with an expected value of 123 l and a standard deviation of 15 l.

   a) Find a symmetric interval around 123 l which contains 95% of the daily water comsumption of a German.

Write your answer here!

   b) What is the probability that a 4 person household consumes more than 550 l of water per day?

Write your answer here!

   c) In one street live 100 people. How much water must be provided to these people per day so that this amount is sufficient in 95% of all cases?

Write your answer here!

   d) The municipal utilities can provide 50000 l of water per day in a settlement to be planned. What is the maximum number of people per day that can be supplied with water in this settlement area, so that this is sufficient in 99% of all cases?

Write your answer here!

Inferential Statistics

1. A brewery regularly checks the fill level of its 0.5 l beer bottles. The following values are from a random sample:

   0.487 0.522 0.513 0.512 0.514 0.500 0.527 0.419 0.477 0.432

   Assume that the fill level is normally distributed.

   a) Calculate a 95% lower confidence bound for the mean of the fill level.

Write your answer here!

   b) Calculate a 99% confidence interval for the standard deviation of the fill level.

Write your answer here!

   c) Assume that the standard deviation of the fill level is known to be 0.05. How many beer bottles should be at least sampled that the length of the 95% confidence interval for the mean of the fill level is less than 0.05?

Write your answer here!

2. A producer of chocolate bars hypothesizes that his production does not adhere to the weight standard of 100 g. As a measure of quality control, he weighs 15 bars and obtains the following results in gram.

   96.4 97.64 98.48 97.67 100.11 95.29 99.8 98.8 100.53 99.41 97.64 101.11 93.43 96.99 97.92

   It is assumed that the weight of the bars follows a normal distribution and that the production process is standardized in the sense that the variation is controlled by σ= 2.

   a) The producer wants to show that the expected weight is smaller than 100 g. What are the appropiate hypotheses are to use?

Write your answer here!

   b) Conduct the test for the hypothesis in a) with alpha = 0.05 and calculate the p-value.

Write your answer here!

   c) Calculate the probability of a type II error if the true expected value is 99.

Write your answer here!