Solutions

1. **Descriptive analysis**

a) A list of puls rates is: $70, 64, 80, 74, 92$. What is the median for this list?
□ 72
☒ 74
□ 77
□ 80

b) If the mean of 10 blood pressure changes is negative, then also the standard deviation of these 10 values is negative.
□ True
☒ False
□ Cannot be told.

The standard deviation can never be negative.

c) Which of the following data can be well visualized by a histogram?

1. The blood pressure of 50 patients.
☒ True
□ False

2. The gender of 40 patients.
□ True
☒ False

3. The weight of 50 patients.
☒ True
□ False

4. The number of times each of 430 patients visited a doctor.
☒ True
□ False

A histogram only makes sense for continuous outcomes and enough observations.

d) The distribution of the blood-concentration of a certain doping drug in 120 randomly controlled athletes looks right skewed - which kind of data transformation can change the shape of a distri-bution?
|x| A square-root transformation
□ Standardization
□ An appropriate linear transformation
☒ A log transformation

Square root and log transformation are possible. For left skewed data we could use an exponential function.

e) Which of the following would indicate that a dataset is not bell-shaped?
□ there are no outliers
□ the mean is much larger than the standard deviation
☒ the mean is much smaller than the median
□ the standard deviation is larger than 6

f) Which one of these statistics is least affected by outliers?
□ Mean
□ Interquartile range

2. **Testing**

a) There is statistical evidence on a significance level of 1% that there is no difference in the mean reaction time of young and old men.

□ True ☒ False □ Cannot be told

"Absence of evidence is not evidence of absence": No evidence for the difference doesn't mean that there is evidence for no difference!

b) There is statistical evidence on a significance level of 1% that there is a significant difference in the mean reaction time of young and old men.

□ True ☒ False □ Cannot be told

p-value is >0.01 (siginificance level 1%!), i.e. there is no evidence for a difference .

c) The 99% confidence interval for the mean difference of reaction times does cover the zero.
☒ True □ False □ Cannot be told

d) If the test would have been conducted on a 5%-significance level then the test would have resulted a significant difference in the mean reaction time of young and old men.
☒ True □ False □ Cannot be told

e) If the sample sizes are increased then we have better chances to get a significant result.
☒ True □ False □ Cannot be told

f) It would have been also valid to use the unpaired Wilcoxon-Test.
☒ True □ False □ Cannot be told

g) It would have been better, if the student would have used a paired t-test
□ True ☒ False □ Cannot be told

h) ANOVA would have yielded the same results as the t-test.
☒ True □ False □ Cannot be told

i) It is not possible to get the same results with a linear regression
□ True ☒ False □ Cannot be told

3. **Study design and the role of the different variables**

a) Which of the following is the primary explanatory variable in this study?
□ Exercise
□ Lung capacity
□ Smoking (Yes/No)
☒ Occupation

b) Which of the following is the response variable in this study?
□ Exercise
☒ Lung capacity
□ Smoking (Yes/No)
□ Occupation

c) Which of the statistical methods are appropriate to compare the lung capacity of coal miners and farmers in the study?
☒ Regression
□ Barplot
□ Binomial Test
□ Chi-Square Test

d) What is the study type of this study?
☒ Observational Study
□ Non randomized experimental study
□ Randomized experimental study

4. **Correlation**

a) If the Pearson correlation between blood pressure and body weight of guinea pigs is zero, then we can conclude that body weight has no influence on the blood pressure in these animals.
| | True |x| False
There might be a non-linear relationship, which can't be caputed by the Pearson correlation coefficient

b) If the Pearson correlation is an appropriate measure and yields a positive number then also the Spearman rank correlation would lead a positive number.
□ True ☒ False
It is for example possible that the pearson is 0 (no linear relationship) but the spearman positive (s. lecture). In extrem cases one might be positive and one negative.

c) A scatter plot of the number of medical doctors and the number of people who suffer from diabetes for cities in Switzerland reveals a positive association. What is the most likely explanation for this positive association?

□ The presence of medical doctors encourage people to have an un-healthy life style.

□ Rich cities tend to have more medical doctors and more obese people.

☒ Larger cities tend to have both – more medical doctors and more sick people.

□ Cities with many people suffering from diabetes attract a lot of medical doctors.

5. **Statistical models and their interpretation**

a) If we look only at the variable `amount of coffee`, in which model do we have an effect on the `time to complete a task`?

□ A, B, C, D, E and F.

☒ A, D and F.

□ A, B, C, E and F.

□ only in C and D.

Calculate the mean across the points for morning and evening for one respectively 3 cups. If the points are on a horizontal line, then we have no effect.

b) If we look only at the variable `daytime`, in which model do we have an effect on the `time to complete a task`?

□ C, D, E and F.

☒ A, B, E and F.

□ only in B.

Calculate the average of coffee (2 cups). In the models C and D, the average is the same for morning and evening, i.e. there is no effect.

c) In which model is an interaction present between `daytime` and `amount of coffee` in their effect on the `time to complete a task`?

□ A and F.

☒ C, D, E and F.

□ A, B, E and F.

□ only in B.

If an interaction is present, we allow the slopes to differ between groups.

d) Which variables are assumed as factor variables?

1. `amount of coffee`

   ☒ True

   □ False

   □ Cannot be told.

2. `time to complete a task`

   □ True

   ☒ False

   □ Cannot be told.

3. `daytime`

   ☒ True

   □ False

   □ Cannot be told.

   Factor variables are categorical variables, i.e. variables with multiple categories.

6. **Linear regression**

a) The soil ph has a statistically significant effect on the tree height

☒ True

□ False

□ Cannot be told.

b) There is a significant negative correlation between ph and height
☒ True
□ False
□ Cannot be told.

The estimate for ph is negative and the p-value < 0.05

c) How many trees were included in the study?
□ 28
□ 100
□ 121
□ 122
☒ 123

degress of freedom + number of parameters = 121+2

d) A farmer claims that the height of a tree decreases on average by 0.3 meter when the ph increases by 0.1. Does the result of the regression contradict this statement?
□ Yes
☒ No

Interpretation of the regression parameter: If x (ph) increases by one unit, y (height) decreases by 3 units is the same as if x increases by 0.1 units, y decreases by 0.3 units.

e) Which mean height would you predict for trees grown on a soil with ph=8?
□ 1.8 m
□ 3 m
☒ 4.7 m
□ 5.2 m

f) According to this study we would expect an average height of 29m for 50 trees on a soil with ph=0.
☒ True
□ False
□ Cannot be told.

g) It is possible that the estimated coefficient of the variable ph becomes positive, if an additional explanatory variable is added to the model.
☒ True
□ False
□ Cannot be told.

h) If we want to account for the effect of the mean daily rain volume we should:
□ work only with observations from trees which received the same mean rain volume.
☒ include the mean rain volume into the linear regression model
□ fit a second model which uses only the mean rain volume as explanatory variable.
□ use ANOVA instead of linear regression,
□ use logistic regression instead of linear regression.