Final exam

Exam available: February 16, 2022

Deadline to submit: March 16, 2022 at 11:59 pm

Submission: Upload to Moodle, in case of problems by email to: stefan.konigorski@hpi.de

To be submitted: 2 files: (i) a Word/pdf/html document containing only the requested analyses and results (i.e. results, tables and graphs) and their requested description/interpretation, and (ii) a file with the R code for calculating the results (R Markdown, with comments which R code belongs to which question). Clearly write to which question the output and the R code belong. Any extensive unnecessary and irrelevant computations can yield point deductions. Results can be given with 2 or 3 decimal places. To assess statistical significance in hypothesis testing, the significance level α=0.05 should be used.

Points:

Question 1:	5 points
Question 2:	5 points
Question 3:	13 points
Question 4:	15 points
Question 5:	12 points
Question 6:	15 points
Total:	**65 points**

Background to the questions:

In the questions of this exam, different data analysis steps of an epidemiological study will be performed and R Markdown will be used for documentation and reporting of results. The main aim is to investigate blood pressure, in the form of systolic blood pressure (*sys12*) and high blood pressure.

Variables that will be investigated for their association with blood pressure/hypertension are waist-to-hip-ratio *whr* (which is an anthropometric measure which is informative of body fat distribution and has been associated with disease outcomes in cohort studies), *bmiB* (body mass index), *sex* (sex) and *age2* (age).

Question 1 - R Markdown [5 points]

As described on page 1, two files should be submitted: a Word/pdf/html document with explained results, and an Rmd file with the R code for the calculation of the results.

Create an R Markdown file containing all relevant R code (in R chunks) that was used to calculate the results. Also include text in this R Markdown script to answer all questions so that all the requested results of the analyses (i.e. results, tables and graphs) are included and described/interpreted. Then knit the R Markdown script to a Word/pdf/html document and submit these two files. [5 points]

Alternatively (if you have problems with knitting), a manually generated Word/pdf/html file with the explained results, and an Rmd file with the R code can be submitted. This means that no points can be obtained for question 1, but all other questions are unaffected.

Question 2 - Import, extract and save data [5 points]

a) Download the SPSS data file KiGGS03_06.sav from moodle and import it into R. [2 points]

b) Create a new dataframe in R named *kiggs*, which contains all variables (and only these) for the analysis (*sys12*, *whr*, *bmiB*, *sex*, *age2)*. [2 points]

c) Run the formatting steps in the provided Rmd file data_formatting.Rmd. Save this formatted dataframe on your computer as a RData file, e.g. on your desktop. [1 points]

Question 3 - Descriptive statistics and graphs [13 points]

a) Describe the variables *sys12*, *whr*, *bmiB* with regard to the following criteria:

- What is the scale (measurement level) of each of these 3 variables (nominal, ordinal, metric)? [1.5 points]
- For each variable, decide which descriptive statistic is best suited to describe it, and explain why. Available for selection: (i) Frequencies, (ii) mean/standard deviation. [1.5 points]
- Calculate the descriptive statistics that you have chosen and report them in text or in a table [3 points].
- Also indicate how many missing values each variable has, and how many observations have complete data for all 3 variables (i.e. no missing values for any of the variables). [2 points]

b) For each variable, select whether a barplot or a histogram is more suitable for displaying their distribution. [2 points]

Then create these 3 diagrams using functions in the ggplot package. [3 points]. You can also create the diagrams using functions in base R, but then only maximally get 1.5 of the 3 points.

Note: If necessary, for this question and all further questions, transform factor variables to numeric variables.

Question 4 - Correlation [15 points]

Here, investigate the association of WHR and BMI with blood pressure.

a) Calculate estimates of the Pearson correlation coefficient and the Spearman correlation coefficient of *sys12* with *whr* and of *sys12* with *bmiB*. Perform 2-sided significance tests at the significance level of 0.05 to test whether the correlations are equal to 0. Give the two estimated correlation coefficients, their respective estimated 95% confidence intervals and corresponding p-values [9 points].

b) Interpret the result of the significance tests for the two correlation coefficients [4 points]:

- Is *sys12* associated with *whr* if the Pearson correlation coefficient is used as measure of association?
- Is *sys12* associated with *bmiB* if the Pearson correlation coefficient is used as measure of association?
- Is sys12 associated with whr if the Spearman correlation coefficient is used as measure of association?
- Is sys12 associated with bmiB if the Spearman correlation coefficient is used as measure of association?

c) The Pearson correlation coefficient is well-suited when the data is normally-distributed, the Spearman correlation coefficient does not make any assumptions on the distribution of the data and computes an association based on ranks of the values only. Which of the two correlation coefficients is a more appropriate measure here, and why? [2 point]

Note: If necessary, transform factor variables to numeric variables.

Question 5 - Logistic Regression [12 points]

Now, examine if there is an association of *whr* with *sys12*, accounting for differences in the age and sex of children.

a) To do this, create a binary variable from *sys12* (*sys12* lower than 120 vs. *sys12* higher or equal to 120) as outcome for the analysis. [2 points]

b) Now calculate a logistic regression with this outcome and the predictors *whr*, *sex*, *age2*. [2 points]

c) To answer the question of whether the WHR of children is associated with high blood pressure adjusting for possible influencing factors, consider the significance test of the regression coefficient of *whr* in this regression. Report the regression coefficient of *whr*, its 95% confidence interval, and the p-value of the significance test. Interpret the exponentiated regression coefficient, the 95% confidence interval, and the results of the hypothesis test. [6 points]

d) Do the results make sense or is there anything that let's you doubt the validity of the modelling or the results? [2 points]

Question 6 – Sample size calculation [15 points]

Now, let's switch to a different study. The aim is to perform a sample size calculation for a new study, whose aim is to investigate the effect of walking daily 5000 steps more (compared to usual) on systolic blood pressure.

a) Look at the literature or think for yourself based on expert knowledge what effect size you would expect. State the effect size that you are assuming and explain why. [4 points]

b) Choose an appropriate statistical model for the sample size calculation and explain why. [4 points]

c) Now compute the minimum necessary sample size for a power of 80% and a significance threshold of alpha = 0.05, for example by using a function in the R package *pwr*. What is the sample size? [6 points]

d) Do you think this is a good study, or do you see any major weaknesses in the study design? [1 point]