Final exam

Exam available: February 8, 2023

Deadline to submit: March 8, 2023 at 11:59 pm CET

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 unrelevant 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:	6 points
Question 3:	10 points
Question 4:	12 points
Question 5:	12 points
Question 6:	10 points
Total:	**55 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 question is whether parents' smoking behaviour has an influence on their children's health.

Smoking was assessed by the four variables *E070M* (mother smoking), *E070V* (father smoking), *E072* (mother smoking in pregnancy), *E074* (mother smoking while breastfeeding).

Children's health was measured by the variable *arztZ01* (number of paediatrician visits).

As possible influencing factors, *sex* (sex), *age2* (age), *schichtz* (social class), and *e065z* (total sleep per day) will be included.

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 [6 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 (*E070M*, *E070V*, *E072*, *E074*, *arztZ01*, *sex*, *age2*, *schichtz*, *e065z)*. [3 points]

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

Question 3 - Data transformations and data checks [10 points]

a) In order to avoid measuring smoking with 4 variables and to calculate all analyses with these 4 variables, we will combine them in one variable called *burdenS.* This new variable shall contain the total smoke exposure to which the children were exposed and is to be used in all further questions to measure parents' smoking. Carry out the following steps:

- Check that the variables *E070M*, *E070V*, *E072*, *E074* are all factors. If they are not, transform them into factors. [2 points]
- Set the value "has not breastfed" of variable *E074* to NA for all children. [1 point]
- Delete this now empty factor level from the variable. [1 point]
- Check whether these two steps worked as intended. [1 point]
- Now calculate the new variable *burdenS* as the sum of the ranks of the four variables *E070M*, *E070V*, *E072*, *E074* for each person (i.e. sum of the numerical factor levels). [2 points]

What is the meaning of this new variable, does a high value mean that the children were exposed to high levels of smoking, or that they were exposed to low levels of smoking? [1 point]

b) Add this variabes *burdenS* to the dataset kiggs, and save it in its updated form as an RData file (overwrite the previous file). [2 points]

Question 4 - Descriptive statistics [12 points]

Consider the variables *age2, sex, arztZ01, burdenS* and describe them with regard to the following criteria:

- Calculate absolute frequency tables for *age2*, *sex*, *burdenS*, and mean + standard deviation for *arztZ01*. Display them in a table or describe them in continuous text. [8 points]
- Also indicate how many missing values each of these 4 variables has, and how many observations have complete data for these 4 variables. [4 points]

Question 5 - Linear Regression [12 points]

The final step is to examine whether there is an association between child health and smoking exposure, taking into account the possible influencing factors sex, age, social class and sleep.

a) Calculate a linear regression, with *arztZ01* as outcome and the predictors *burdenS*, *sex*, *age2*, *schichtz* and *e065z*. [2 points]

   Check for each predictor how you take it into the regression model (factor, ordinal or metric) and justify for each variable why you did it that way (e.g. because the variable has the measurement level xyz) [3 points].

b) To answer the question of whether the smoking behavior of parents has an influence on the health of children, adjusting for possible influencing factors, consider the significance test of the regression coefficient of *burdenS* in this regression. Report the regression coefficient of *burdenS*, interpret the coefficient, report its 95% confidence interval, and report its p-value of the significance test [4 points].

   What is your conclusion: Is there an association or not? In which direction? [2 Points]

c) Since there is evidence that individuals drawn from the same area are correlated with each other, but we are not interested in the effect of the area on the health ... what would be a suitable strategy for accounting for this correlation? [1 point]

Question 6 – Sample size calculation [10 points]

Now, let's switch to a different study. The aim is to perform a sample size calculation for a new observational study, whose aim is to investigate the effect of pregnant women smoking 10 cigarettes per day (compared to not smoking at all) on the birth weight of their babies.

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. [3 points]

b) Choose an appropriate statistical model for the sample size calculation and explain why. [3 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? [3 points]

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