Final exam

Overview: 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.

Exam available: February 6, 2025

Deadline to submit: March 6, 2025 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 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 $\alpha=0.05$ should be used.

Important note: Use of generative AI (e.g. ChatGPT) is not allowed for creating the solutions to this exam. This will be checked and its use can lead to failure of the exam.

Points:

Question 1:	5 points
Question 2:	6 points
Question 3:	12 points
Question 4:	12 points
Question 5:	20 points
Question 6:	7 points
Total:	**62 points**

Question 1 - R Markdown [5 points]

As described on page 1, two files should be submitted: (1) a Word or pdf or html document with explained results, and (2) 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_xyz* (where you should replace *xyz* with your name) which contains the following variables (and only these): sex (*sex*), age (*age2*), the children's birthweight in grams (*e017a.k*), and whether the children have ever had pertussis (*e02001*), a measles infection (e02002), a scarlet fever (Streptococcus infection; *e02006*), a salmonella infection (*e02009*). [3 points]

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

Note: depending on how you imported your dataframe, the variable name might be *e017a.k* or *e017a_k.*

Question 3 - Data transformations and data checks [12 points]

a) The variables *e02001, e02002, e02006, e02009* contain different diseases children might experience. As an alternative to computing all analyses with these 4 variables, we will combine them in one variable called *burden.* This new variable shall contain the number of diseases experienced by the children and is to be used in all further questions. Carry out the following steps:
   - Check that the variables *e02001, e02002, e02006, e02009* are all factors. If they are not, transform them into factors. [2 points]
   - Set the value "don't know" of all four variables to NA for all children. [1 point]
   - Delete this now empty factor level from the variables. [1 point]
   - Check whether these two steps worked as intended. [1 point]
   - Now calculate the new variable *burden* as the number of diseases the children had (this is a number between 0 and 4). [4 points]

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

c) What could be an alternative way to summarize the four variables into a summary variable? [1 point]

Question 4 - Descriptive statistics [12 points]

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

a) Create one table with absolute frequencies of all three variables using the summary_table() function or a similar function. [8 points]

b) Also indicate how many missing values each of these three variables has, and how many observations have complete data for all three variables. This can be computed outside of the table from task (a) above. [4 points]

Question 5 – Regression [20 points]

Here, the aim is to use a regression model or multiple regression models to investigate whether the birthweight of a child has an effect on getting child infections.

a) Choose whether you want to use the *burden* variable or the single variables *e02001, e02002, e02006, e02009*. State the reason for your choice. [1 point]

b) Choose an appropriate regression model (linear regression, logistic regression,…). [2 points]

c) Compute this regression model/these regression models. [3 points]

d) Report the estimated regression coefficients or exponentiated regression coefficients (choose which one is more appropriate based on your model) for the birthweight variable, as well as the p-value(s) from significance tests [4 points].

   Also give estimates of the 95% confidence interval of the coefficients/odds ratios for birthweight [3 points].

   Interpret the results (statistically significant relationship yes/no) for birthweight and also interpret the coefficients/odds ratios. [6 points].

e) In this data set, do you think this analysis is appropriate? Justify why yes/no. [1 point]

Question 6 – Sample size calculation [7 points]

Now, the aim is to perform a sample size calculation for a new study to investigate the effect of birthweight of a child as continuous variable on getting child infections (more than 5 within the first 5 years of life vs less than 5 within the first 5 years of life). We will consider using a t-test for the analysis.

a) Think about what an effect size measure can be. 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. [2 points]

b) Discuss potential disadvantages of using a t-test for the analysis and for the sample size calculation. [2 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? [2 points]

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