SDA 2019 — Assignment 10

For these exercises you can use the *R*-functions `cor` and `cor.test` for the (test on) different correlation, see `help(cor)` and `help(cor.test)`.
For the categorical data you should use the *R*-functions `fisher.test` and `phyper` for the cumulative distribution function of a hypergeometric distribution. See `help(fisher.test)` and `help(phyper)`.

Make a concise report of *all* your answers in *one single PDF file*, with only *relevant R* code *in an appendix*. It is important to make clear in your answers how you have solved the questions. Graphs should look neat (label the axes, give titles, use correct dimensions etc.). Multiple graphs can be put into one figure using the command `par(mfrow=c(k,l))`, see `help(par)`. Sometimes there might be additional information on what exactly has to be handed in. **Read the file AssignmentFormat.pdf on Canvas carefully**.

Exercise 10.1 The data in the file `expensescrime.txt` were obtained to determine factors related to state expenditures on fighting criminality (courts, police, etc.). The variables are: `state` (indicating the state in the USA), `expend` (state expenditures on fighting criminality in $1000), `bad` (number of persons under judicial supervision), `crime` (crime rate per 100000), `lawyers` (number of lawyers in the state), `employ` (number of persons gainfully employed by and performing services for a government) and `pop` (population of the state in 1000).

a. Make plots for every pair of variables to judge their relationship. (Don't include the variable `state`.) Use the *R*-function `pairs`. (Don't hand in these plots.)

b. Make a plot of the employment *rate* versus crime rate. Based on these plots, how do you judge the correlation between `crime` and employment rate?
Hint: first think about how to obtain the employment rate.

c. Perform Kendall's rank correlation test for the variables `crime` and employment rate using the *R*-function `cor.test`.

d. Read Section 6.4 and Example 6.7 in the syllabus. Perform a permutation test for testing dependence between `crime` and employment rate, as explained in Section 6.4.3, based on Kendall's rank correlation coefficient.[1]

e. What is your conclusion about the correlation between the two variables `crime` and employment rate, based on the outcomes in parts c and d?

Hand in: your plot (in b.) and answers to parts b.–e.

[1] Because computing all 51! possible permutations is not possible, you should generate a large number (e.g. 1000) of permutations using the function `sample` and approximate the p-value based on these values.

Exercise 10.2 In a study on the relation between movie genre preference and gender, 120 randomly selected cinema visitors were asked about whether they like sci-fi movies or not. Moreover, their gender was scored. In the table below you find the categorical counts for gender and preference for sci-fi movies.

gender / sci-fi	like	don't like	total
women	43	27	70
men	38	12	50
total	81	39	120

In this exercise, we are using the significance level $\alpha = 10\%$.

a. Describe the null and alternative hypotheses to be tested and perform Fisher's exact test.
Hint: In R you should store a contingency table as an object of type `matrix`. *For these data you can for example use the command* `mytable=matrix(c(43,38,27,12),nrow=2,ncol=2)`.

b. In the R command `fisher.test` it is also possible to choose the option `alternative="less"`. Which alternative hypothesis is tested with this option? Formulate this alternative hypothesis in two ways: first, with reference to the probabilities

$p_{11} = P(\text{gender} = \text{female, sci-fi preference} = \text{TRUE})$,
$p_{1\cdot} = P(\text{gender} = \text{female})$,
$p_{\cdot 1} = P(\text{sci-fi preference} = \text{TRUE})$,

and second, in your own words.

c. Perform the test of part b.

d. Find the p-value from part c. with the help of a suitable application of the command `phyper`.

Hand in: your answers to all parts.