SDA 2022 — Assignment 6

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 can use the *R*-functions `fisher.test` and `phyper` for the cumulative distribution function of a hypergeometric distribution – see `help(fisher.test)` and `help(phyper)` – and the functions `bootstrapcat` and `maxcontributionscat`.[1] When performing statistical tests, *clearly* state the null hypothesis and test statistic together with its distribution under the null hypothesis.

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 6.1 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).

In this exercise (from part b. to part d.) we will only consider the subset of states with a population of at least 4,000,000.

a. **(Class)** 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. **(Class)** Make a plot of the expend *rate*, i.e. state expenditures in the state per (1000) citizen, versus crime rate. Based on this plot, how do you judge the correlation between `crime` and expend rate?

c. Test whether the variables `crime` and expend rate are dependent by using Kendall's and also Spearman's rank correlation tests. Take the significance level $\alpha = 5\%$.

d. Read Section 6.4 and Example 6.7 in the syllabus. Perform a permutation test for testing dependence between `crime` and expend rate, as explained in Section 6.4.3, based on Spearman's rank correlation coefficient.[2] Take the significance level $\alpha = 5\%$.

e. Use simulations to find an approximate value for the asymptotic relative efficiency of Kendall's rank correlation test with respect to Spearman's rank correlation test, when the data pairs are bivariate t-distributed with scale matrix $\Sigma = \left(\begin{smallmatrix} 1 & 0.3 \\ 0.3 & 1 \end{smallmatrix}\right)$ and six degrees of freedom.[3] Use the sample sizes $n = 48, 50, 52$ to find out which of the three values $0.96, 1, 1.04$ seems to be closest to the true asymptotic relative efficiency.
(Note: $52/50 = 1.04 \approx 50/48$ and $50/52 \approx 0.96 = 48/50$.)
For this, generate at least B =5,000 independent test outcomes per sample size to find an approximation of the true power. Which test seems to be more powerful?

[1] You can find these functions in the file `functions_Ch7.txt`.

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

[3] You can generate such bivariate t-distributed data with the help of the command `rmvt(n, sigma=matrix(c(1,rho,rho,1), 2,2), df=6)` after having installed and loaded the *R* package `mvtnorm`.

You could use a modification of the R code on slides 7 and 8 of the Lecture 8 handout slides to simulate the power of both tests.

Hand in: answers to parts c.–e.

Exercise 6.2 In a study about the consequences of a certain virus infection, scientists wished to use the numbers of cases in a certain town to infer for all of humanity, given sufficient medical equipment, whether there is a relationship between the gender and the fatalities among the infected.

The following table represents the dataset:

infected	deaths	recoveries	total
men	30	1067	1097
women	17	1120	1137
total	47	2187	2234

For each test in this exercise, we use the significance level $\alpha = 5\%$.

a. Describe the null and alternative hypotheses to be tested and perform Fisher's exact test. Use the *R* command `fisher.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* `matrix(c(24,...),nrow=2,ncol=2)`.

b. The scientists also analyzed the numbers of cases in a different city and got the impression that men are more often among the fatalities than women. Investigate this suspicion based on the current dataset; state the null and alternative hypotheses to be tested and perform Fisher's exact test. Use the R command `fisher.test`.

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

Hand in: your answers to all parts.

Exercise 6.3 The file `nausea.txt` contains data about post-operative nausea after medication against nausea. The patients, who complained about post-operative nausea, were randomly assigned to one of the different medicines or to a placebo. One of the medicines, Pentobarbital, was administered in two different doses. The first column in the file contains the **total** number of patients that were given the medicine or placebo, and the second column contains the number of cases of nausea (after medication) in that group. In this exercise, we are going to analyze the relationship between occurrence of nausea and type of drug. For all tests considered in this exercise, we use the significance level $\alpha = 5\%$.

a. Which of the three models II A, II B and II C is most suitable for these data?
Note: always motivate your answers!

b. Conduct a suitable hypothesis test to test the hypotheses that belong to the model you have chosen in a.; also state the hypotheses. Also check the rule of thumb if you are using a chi-square approximation.

c. Use the option `simulate.p.value` of the R command `chisq.test` to check whether the p-value that you found in part b. is reliable.

d. Compute the contributions and the standardized residuals for the test(s) that you performed in part b. Which 'categories' stand out?

For bootstrap procedures for statistics other than the standard chi-square statistic you can use the function `bootstrapcat`. Such a procedure can be rather time consuming. Use at least $B = 1000$ bootstrap iterations.

e. Test the same (undirected) hypotheses as in part b., but now with a bootstrap procedure using the statistic $T =$*'The largest of the absolute values of the contributions'*. To compute the observed value of this statistic the R-function `maxcontributionscat` available on Canvas can be used.
Note that even though the undirected hypotheses are tested, the test is right-tailed, i.e. it rejects for relatively large values of the test statistic.

f. Conduct a one-sided Fisher's exact test to find out whether Chlorpromazine works better than the placebo. Do not forget to state appropriate null and alternative hypotheses in terms of the categorical variables that belong to the first row and the first column of the (2×2)-matrix.

Hand in: your answers to all parts.