Statistical Methods Fall 2017

Assignment 4: Correlation, regression and contingency tables

Deadline: December 13, 23.59h

Topics of this assignment

The exercises below concern topics that were covered in Lectures 9 and 10: correlation, regression and contingency tables (see Sections 9.2 (incl. Part 2), 9.3 (incl. Part 3), 9.4, 10.2 and 10.3 (incl. Part 2) of the book and the slides of Lectures 9 and 10). Before making the assignment, study these topics. Numbers of exercises in the book refer to the twelfth edition (Pearson New International Edition).

How to make the exercises? See Assignment 1.

If you are asked to perform a test, do not only give the conclusion of your test, but report:
- the hypotheses in terms of the population parameter of interest;
- the significance level;
- the test statistic and its distribution under the null hypothesis;
- the observed value of the test statistic (the observed score);
- the P-value or the critical region;
- whether or not the null hypothesis is rejected and why.

If applicable, also phrase your conclusion in terms of the context of the problem.

Theoretical exercises

For the two theoretical exercises below use Tables 3 and 4 from the Appendix in the book to find probabilities and/or critical values. Do not use R. If you need to use a t-distribution with the number of degrees of freedom not included in Table 3, report the number of degrees of freedom, and use the critical value based on a t-distribution with the next lower number of degrees of freedom found in the table.

Exercise 4.1
Exercise 4 from Section 9.2 (Weight Loss and Correlation). Take significance level $\alpha = 1\%$.
(Follow the detailed instructions about testing presented above).

Exercise 4.2
Exercise 10 from Section 10.2 (Baseball Player Births; without the last question). Take significance level $\alpha = 5\%$.
(Follow the detailed instructions about testing presented above).

R-exercises

Do not use tables from the Appendix in the book. Use R to find probabilities and/or critical values.

Hints concerning R:

- The *R*-function `cor()` computes the sample linear correlation coefficient. The *R*-function `cor.test()` can be used to compute a confidence interval for the population correlation coefficient, and to perform a test concerning the population correlation coefficient. At the same time it also gives the sample correlation coefficient; here it is called 'sample estimate' (for the population linear correlation coefficient).
- For analysis of the linear regression model the R-function `lm()` can be used. Let the measurements of the explanatory variable be in the vector `x` and the measurements of the outcome variable be in `y`. Then `lmsim=lm(y~x)` fits a simple linear regression model and stores the output in `lmsim`. The output is a list, which can be studied using `summary(lmsim)`. To obtain the estimated coefficients for the intercept and slope the command `lmsim$coef` can be used. Similarly, `lmsim$res` provides the residuals. The standard errors of the estimated coefficients can be obtained (apart from inspection of `summary(lmsim)`) with the command `summary(lmsim)$coef[,2]`. To visualise the regression equation, the command `abline(lmsim$coef)` can be used. See also the slides of Lecture 9.
- For the analysis of contingency tables the function `chisq.test()` can be used. The command `chisq.test(table)$exp` provides the expected frequency count of the data in the fictitious contingency table `table` under the null hypothesis. See also the slides of Lecture 10.
- Recall: for computing probabilities and quantiles of normally, t-, chisquare, etc. distributed random variables the *R*-functions `pnorm, pt, pchisq, ...`, and `qnorm, qt, qchisq, ...` can be used. For the t- and chisquare distributions the number of degrees of freedom needs to be specified.

Exercise 4.3 There is considerable variation among individuals in their perception of crime, and in particular of which specific acts constitute a crime. A study was made to investigate which variables, like age, level of education, parental income, etc., may influence this perception. The file `crimemale.txt` contains part of the results of this study: data are given for 18 male college students who were asked how many of the following 25 acts they perceive as being a crime: *aggravated assault, armed robbery, arson, atheism, auto theft, burglary, civil disobedience, communism, drug addiction, embezzlement, forcible rape, gambling, homosexuality, land fraud, nazism, payola, price fixing, prostitution, sexual abuse of child, sex discrimination, shoplifting, striking, strip mining, treason, vandalism.* The column `crimes` contains the number of acts the students perceive as crimes, `age` the ages of the students, and `income` the incomes of the parents (in $1000).

a) Make for the data of the male students a scatterplot in which the x variable is `age` and the y variable is `crimes`. Compute also the sample linear correlation coefficient. Based on the plot and the sample linear correlation coefficient, do you think there is linear correlation between the two variables?

b) Repeat part a with `income` as the x variable.

c) Perform a linear regression analysis—i.e. formulate the regression model and compute the estimates of the unknown parameter values—with the variable `income` as the explanatory variable and the variable `crimes` as the response variable. Report the estimated values of the

intercept and slope that determine the 'best' line and draw this 'best' line in the corresponding scatterplot.

d) Using the results of the regression analysis of part c, test the claim that there is no linear relationship between the two variables `income` and `crimes`. Take significance level 5%. (See the first page of the assignment for detailed instructions about testing).

e) In order to perform the test of part d, certain requirements have to be met. What are these requirements? Provide a suitable plot (or plots) and report whether the requirements are indeed met.

Exercise 4.4 During the previous Christmas party, Dennis was asked to give an estimate of the general chances that Germany would win (or lose) against Italy in matches of the national men's soccer teams. He always thought that he knew a lot about soccer, so he was confident that the following guess was appropriate (from a German point of view):

	Wins	Draws	Defeats
men	40%	30%	30%

The following table shows the actual results (of both German men's and women's national soccer results against Italy):

	Wins	Draws	Defeats
men	8	12	15
women	15	8	4

a) Compute the expected frequency of German men's wins against Italy under the assumption that Dennis' guess was true.

b) Use the significance level $\alpha = 10\%$ to test Dennis' guess with a goodness-of-fit test.
(See the first page of the assignment for detailed instructions about testing).

Furthermore, Dennis claimed that the German men's and women's national soccer teams have the same chances of winning (and of losing) against the respective Italian teams.

c) Should you use a test of independence or a test of homogeneity to test Dennis' second claim? Motivate your answer and formulate the null and alternative hypothesis.

d) Create a matrix `results` containing the data and use it to perform the test of part c). Take significance level $\alpha = 5\%$.
(See the first page of the assignment for detailed instructions about testing).

e) How many games would the German men have won against Italy if the second claim was true?

f) Now combine Draws and Defeats among each, men and women. This results in a 2×2 contingency table. Use Fisher's exact test and significance level $\alpha = 1\%$ to test the directed claim

H_a : "Men have worse chances to win a soccer match against Italy than women."

(See the first page of the assignment for detailed instructions about testing).