Statistical Methods Fall 2021

Assignment 4: Correlation, regression and contingency tables

Deadline: see Canvas

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.

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;
- also check the assumptions required for retrieving the 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
A political scientist wishes to analyze the relationship between the 7-day-incidence rates (SARS-CoV-2; data from Dec. 6, 2021) and the recent governmental election results of a certain (extreme) political party. He used 16 data points that correspond to the 16 federal states of that country. He computed a sample correlation of 0.926. Check with the help of a hypothesis test whether there is sufficient evidence to support a claim of a linear correlation of the incidence numbers and the election results. Take $\alpha = 1\%$.
Also argue whether a strong vote for that party seems to go along with a rather high or low incidence rate.
Follow the detailed instructions about testing presented above.

Exercise 4.2
A sport scientist claims that more baseball players have birthdays in the months immediately following July 31, because that was the cutoff date for nonschool baseball leagues. Here is a sample of frequency counts of months of birthdates of American-born major league baseball players (starting with January):

$$64, 55, 61, 58, 56, 52, 52, 84, 70, 72, 66, 62.$$

Check with the help of a hypothesis test whether there is sufficient evidence to warrant rejection of the claim that American-born major league baseball players are born in different months with the same frequency. Take $\alpha = 10\%$.
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.
- You can load an `.RData` file into the workspace using the command `load(...)`.

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 (without conducting a hypothesis test), 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 1%.
Follow the detailed instructions about testing presented in the first page of this assignment.

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 and argue whether the requirements are indeed met.

Exercise 4.4 Andy uses a mobile app to play games of trivia with his friends. On different evenings, he made appointments with either of Bob, Cecilia, David, Emma, or Freddy and played one-on-one with the chosen friend for the whole evening. They have previously agreed on playing, respectively, 283, 149, 83, 69, and 160 rounds of the game. Both players have to answer a number of randomly selected questions, and the player who correctly answers more questions wins. The table below contains results of 744 games Andy played with his friends (e.g., Andy won 179 games against Bob).

	Won	Lost	Draw	Total
Bob	179	47	57	283
Cecilia	96	17	36	149
David	52	13	18	83
Emma	39	15	15	69
Freddy	84	37	39	160
Total	450	129	165	744

a) In order to investigate whether Andy's friends are equally strong opponents, should you use a test of independence or a test of homogeneity? Motivate your answer and formulate the null and alternative hypothesis.

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

c) How many games against Freddy would Andy be expected to win, if the null hypothesis were true and he played 160 games against Freddy?

d) Use a suitable test to test whether the probability to win against Freddy are smaller than the probability to win against Bob. Use $\alpha = 10\%$.
Follow the detailed instructions about testing presented in the first page of this assignment.