Assignment 3

The exercises below concern topics that were covered in Lectures 9 and 10. Unless otherwise specified, the significance level is $\alpha = 0.05$.

Theoretical exercises

Hints concerning theoretical exercises

- Random variable $X \sim \text{Bin}(n,p)$ (binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0,1]$) if $X \in \{0,1,2,\ldots,n\}$ and $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$. Interpretation: X is the number of successes in n trials with success probability p in each trial. For example, $X \sim \text{Bin}(1,p)$ takes just two values $\{0,1\}$ with probability $P(X=1) = p = 1 - P(X=0)$. If $X \sim \text{Bin}(n,p)$, $EX = np$, $Var(X) = np(1-p)$.
- If $X_1 \sim \text{Bin}(n_1,p)$, $X_2 \sim \text{Bin}(n_2,p)$ and X_1, X_2 are independent, then $X_1 + X_2 \sim \text{Bin}(n_1+n_2,p)$. In particular, if $X_1,\ldots,X_n \overset{\text{ind}}{\sim} \text{Bin}(1,p)$, then $X = X_1 + \ldots X_n \sim \text{Bin}(n,p)$.
- If $X \sim \text{Bin}(n,p)$ (with unknown p) is observed, the success probability p is estimated by $\hat{p} = X/n$.
- If the goodness-of-fit procedure is applied to a distribution from a family of distributions, (i.e., the parameters of the distribution to be fitted are unknown; e.g., $N(\mu,\sigma^2)$ with unknown μ,σ), then one should use the goodness-of-fit test with estimated parameters and χ^2_{m-p-1}-quantile, where m is the number of cells and p is the number of estimated parameters of the distribution to be fitted.

Exercise 1. The author is a member of of the board of directors of non-profit organization that held a fund-raising auction. He recorded the opening bids suggested by the auctioneer and the final winning bids for several items. The amounts are listed below.

Opening bid	1500	500	500	400	300
Winning bid	650	175	125	275	125

a) Is there sufficient evidence to conclude that there is a linear correlation between the opening bids suggested by the auctioneer and the final winning bids?

b) Compute the intercept and slope of the linear regression line with *Opening bid* as predictor and *Winning bid* as response variable.

c) Use the regression line from b) to predict winning bid if the opening bid were 1000. Derive the residual values for *Opening bids* equal to 300 and 1500. Comment.

Exercise 2. Each test fiber was produced 5 times under similar conditions and its strength was tested on 5 identical devices. Assuming that each fiber has the same strength, the number of breaks of a given fiber should be binomially distributed with five trials and an unknown probability p of fiber break. If all fibers are of the same strength, p should be the same for all of them, if they are of different strength, p should vary from fiber to fiber. The following table summarizes the outcome of the experiment:

Breaks	0	1	2	3	4	5
Frequency	157	69	35	17	1	1

a) Assuming the same strength of all the test fibers, estimate the break probability p.

b) Pooling the last three cells, test the agreement of the observed frequency distribution with the binomial distribution (hint: use the estimated probability of fiber break obtained in a) for the binomial distribution to be fitted) by using Pearson's chi-squared test. You can use R for computing quantiles and p-values.

R-exercises

Hints concerning R

- The R-function `cor()` computes the sample linear correlation coefficient, `cor.test()` performs a test about the correlation coefficient and can be used to compute a confidence interval for the correlation coefficient.
- Command `mod=lm(y~x)` fits a simple linear regression of explanatory variable x on response y and stores the output in `mod`. The output can be studied using `summary(mod)`, `mod$coef` gives the estimated intercept and slope, `mod$res` provides the residuals. To visualize the fitted line, use `abline(mod$coef)`, `abline(mod)`.
- For the analysis of contingency tables the R-function `chisq.test()` can be used, `chisq.test(table)$exp` provides the expected frequency count of the data in the contingency table `table` under the null hypothesis. To see the individual (directed) contributions to the χ^2-statistics, use `chisq.test(table)$res`.

Exercise 3. It has been argued that many cases of infant mortality are caused by teenage mothers who, for various reasons, do not receive proper prenatal care. In the file `mortality.txt` data from the Statistical Abstract of the United States (1995) are listed on teenage birth rate per 1000 (column `teen`) and the infant mortality rate per 1000 live births (column `mort`) for the 48 contiguous states (column `state`).

a) Make a scatterplot of teenage birth rate against the infant mortality rate. Test whether these two variables are correlated or not. Derive the best straight line treating `mort` as y-variable and `teen` as x-variable. Add the best fitted line to the corresponding scatterplot.

b) Perform a simple linear regression analysis by testing about possible linear relationship between the two variables `teen` and `mort`, treating `mort` as response (dependent) variable and `teen` as predictor (independent) variable. Relate the obtained results to the results from part a). Construct a 98%-confidence interval for the slope coefficient from the fitted linear model. Derive also the determination coefficient and the estimate of the error variance.

c) Use the model found in b) to predict the infant mortality rate for the teenage birth rate `teen`=10.

d) Use the following graphical diagnostic tools, the scatterplot of predictor against the residuals and the QQ-plot of the residuals, to verify the assumptions of the resulting model obtained in b). Comment.

Exercise 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). We wish to investigate whether Andy's friends are equally strong opponents.

	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) Create a matrix containing the data and use it to perform a suitable test to answer the question of interest. Take significance level $\alpha = 0.05$.

b) Investigate which entries of the table contributed the most to the test statistics from a).

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? Comment on your findings.