Exercises for the Course
“Methods of Scientific Research”

Institut für Medizinische Informatik und Statistik
Medizinische Fakultät der Christian-Albrechts-Universität Kiel
Universitäts-Klinikum Schleswig-Holstein Campus Kiel
Brunswiker Straße 10
24105 Kiel

Exercise 1 (Descriptive Statistics)

Load the library MASS.

1. (*measures of location and dispersion*)

a. Derive the minimum and the maximum of the birth weight of the babies of the data set birthwt (column bwt). Calculate the mean and standard deviation.

b. Order the birth weight of the babies in increasing order (use the function *sort*).
Which elements of the sorted vector correspond to median and quartiles?

c. Now derive the median of only the first ten babies of the unsorted vector.

d. Now calculate the median and quartile values of the original vector birthwt$bwt by the *summary* and *quantile* functions of R. Additionally, calculate the 0.05 and 0.95 quantiles.

e. Compare mean and median.

2. (*histogram, box-whisker plot*)

a. Draw a histogram of the birth weight of the babies.
What kind of histogram is it?
Mark the mean and median (e.g. by hand).

b. Draw a box-whisker plot of the birth weight of the babies.
What conclusions can you draw from this plot?

Exercise 2 (Probability Theory)

1. (*discrete random variable, probability function, expected value*)

Approximately 5% of Germans are so-called “non-responders” to hepatitis B vaccination, i.e. they do not gain immunization from the usual three-step treatment (injections at 0, 1 and 6 months).

a. Calculate the values of the probability function of the random variable “number of non-responders among 5 treated individuals” (“X” for short) either by hand or by using the *choose* function of R (do not use the function *dbinom*).
Make a plot using the *plot* function with the possible values of X on the x axis and the corresponding values of the probability function on the y axis. Add meaningful labels to the x and y axes.

b. Repeat exercise a. but now with π=0.10 and by using the function *dbinom* instead of the function *choose*. For the plot use red crosses as plot symbols.

c. Calculate the expected value of X from the probability function (π=0.05).

d. Calculate the probability of finding at most one non-responder among 5 treated individuals. Calculate the probability of finding at least one non-responder among 5 treated individuals.

2. (*continuous random variable, density function, normal distribution*)

The body height of male Kiel students can be assumed to follow a normal distribution with an expected value of 183 cm and a variance of 36 cm^2.

a. Draw the density function of the random variable “body height of a male Kiel student” (“Y” for short) by the *plot* function.

b. Calculate the probability that a randomly chosen male Kiel student is a most 189 cm tall.

c. Calculate the probability that a randomly chosen male Kiel student is between 184.5 cm and 192 cm tall.

Exercise 3 (Parameter Estimation)

1. (*binomial distribution, point estimation*)

From the dataset birthwt in library MASS we want to estimate the probability π that a given newborn is underweight (column birthwt$low).

a. What estimator $\hat{\pi}$ can you use and why is it sensible? Check the three important properties of an estimator mentioned in the lecture.

b. What estimate do you derive from the data? Attention, you have to transform birthwt$low into a factor first.

2. (*normal distribution, point estimation, confidence interval*)

Let us assume that we know that the body height of male Kiel students follows a normal distribution with an expected value of 183 cm and a variance of 36 cm^2.

a. Sample 100 body heights of students using the function *rnorm*. From this data estimate the expected value by the mean, the theoretical standard deviation σ by the empirical standard deviation s
and calculate the 95% confidence interval (for unknown σ).

b. Repeat this procedure nine times such that you have 10 samples with estimated means and confidence intervals.
Note for how many times the true expected value (i.e. 183 cm) is included in the confidence intervals.

c. Draw also one sample of size 10 and one of size 1000 and again calculate 95% confidence intervals.
Compare the values between the samples of different size (n=10, 100 and 1000).

d. For one of your samples with n=100 calculate also the 90% and 99% confidence intervals.
Compare the three confidence intervals with confidences 90%, 95% and 99%.

Exercise 4 (Epidemiology)

1. (*cohort study, case-control study*)

a. Compare the two generic epidemiological studies, i.e. cohort study and case-control study, theoretically with regard to their design and evidential power, using the criteria in Table A.1.

2. (*odds ratio, relative risk*)

Study 1: A retrospective analysis of 58 hospitalized leukemia cases revealed that some of them used to work in a nearby synthetic rubber factory for over a year. Therefore, a comparable large control group of patients with a leisure accident was questioned with respect to working in this factory.
Study 2: Because of the results of study 1, two large groups from all over Germany were prospectively monitored for 5 years and the number or leukemia incidents recorded. The first group consisted of synthetic rubber factory workers, whereas the second group was composed of metal workers.

a. You can find the data from study 1 and study 2 on the internet on *http://www.uni-kiel.de/medinfo/lehre/medlife*. Copy the files onto your computer (use the right mouse click). Load the data from study 1 into the workspace and call it “study_1”. Get an overview of the study. The variables are coded in the following way:
leukemia: aff: diseased with leukemia; non-aff: not diseased with leukemia
rubber: rubber: working in a rubber factory; no rubber: not working in a rubber factory
Generate a two-times-two table with affected/non affected as columns (here leukemia) and exposed/non exposed as rows (here working in a rubber factory).

b. What kind of study is study 1?
Estimate the respective effect measure and calculate a 95% confidence interval for it.
For the confidence interval you will need the function *qnorm* or get the value from the lecture.
or

c. Now load the data from study 2 into the workspace
and determine the type of the study.
As in a. and b., generate a two-times-two table, estimate the respective effect measure and calculate a 95% confidence interval for it.

d. From the relative risk, calculate the attributable risk (AR) and the population attributable risk (PAR) assuming that 0.05% of the German population work in a synthetic rubber factory. This is a straight forward calculation, which can be done either manually with the help of a calculator or with R.

Exercise 5 (Diagnostic Testing)

1. (*conditional probability, Bayes theorem; use of R not necessary*)

HIV infections can be probed by means of the ELISA test. This procedure, which targets antibodies against HIV in the proband's blood, but not the virus itself, has a sensitivity and specificity of 99.5%, respectively. According to the Robert Koch Institute, the prevalence of HIV infection equals 0.01% in the "low risk" part of the German population (i.e., heterosexual, no drug abuser), compared to 15% among intravenous drug abusers.

a. Calculate the positive and negative predictive value of the ELISA test for both subpopulations.

b. What would be the positive predictive value of a second ELISA test carried out for a low risk proband with a positive result in the first test?

2. (*ROC curve, sensitivity, specificity*)

Gestational diabetes (GD) can be diagnosed by two alternative procedures that are less cumbersome than the usual oral 100g glucose tolerance test (the current "gold standard" for GD diagnostics). This includes the 50g glucose challenge test (GCT) and the measurement of the fasting plasma glucose concentration (FPGC). In the late 1990s, both the evidential value and the reliability were compared between the two procedures in a prospective study by Perucchini *et al*. (BMJ 319:812, 1999). Of a total of 520 randomly selected pregnant women, 53 were diagnosed with gestational diabetes using the oral 100g glucose tolerance test. Prior to this assessment, the same women were also subjected to GCT and FPGC measurement.

a. Install the package pROC and load it. You can find the data from the study at *http://www.uni-kiel.de/medinfo/lehre/medlife.* Load it into your workspace (use option dec=","). 

The variables are coded in the following way:
conc_fpgc: fasting plasma glucose concentration in mmol/l measured by the FPGC
conc_gct: plasma glucose concentration in mmol/l measured by the GCT
gold: gold standard (100g glucose tolerance test); 1: gestational diabetes present, 0: gestational diabetes absent

b. For each diagnostic procedure (i.e. FPGC and GCT) plot an ROC curve. 

c. For each diagnostic procedure determine the Youden threshold and calculate the corresponding sensitivity and specificity. 

d. For each diagnostic procedure and the Youden threshold calculate false/true positives and negatives  and make a two-times-two table with disease (present/absent, here GD) as rows and test result (positive/negative, here FPGC or GCT) as columns.
Now calculate sensitivity and specificity manually from the table and compare with the results of c.

e. Plot the two ROC curves in one diagram (option "add=TRUE" in the function plot(roc_obj)) in different colours. 

f. From a medical professional perspective, which procedure would be more suitable for routine GD screening? Why?

Exercise 6 (Statistical Testing I)

1. (*hypotheses, significance level, one-sample t-test, p value*)

The mean birth weight of a baby in the US population is around 3.200 grams. The cohort birthwt in library MASS was recruited to study the question whether smoking of the mother during pregnancy reduces the birth weight of the baby.

a. Formulate the scientific question outlined above as a statistical decision problem (one-sided): Characterize the null hypothesis (H_0) and the alternative hypothesis (H_A) and determine a sensible significance level.

b. Which statistical test can be used to decide between H_0 and H_A assuming that the birth weight for smoking mothers follows a normal distribution?

c. What is the critical value of the test statistic identified in (b) for the significance level of (a)?

d. Calculate mean and standard deviation of the birth weight of babies with mothers who smoked.
From these values derive manually the test statistic identified in (b)
and interpret the result.

e. Now perform the statistical test of (b) with R.
Compare the test statistic with that derived in (d).
With R you also get the respective p value. What information can you obtain from it?

Exercise 7 (Statistical Testing II)

1. (*sample size calculation, two-sample t-test, Wilcoxon rank sum test*)

The question whether smoking of the mother influences the birth weight of the baby shall be investigated in a study containing smoking and non-smoking mothers. The mean birth weight and the standard deviation is around 3200 grams and 700 grams, respectively, for non-smoking mothers. For smoking mothers, a birth weight around 2850 grams and the same standard deviation is assumed.

a. Assume that the birth weight follows a normal distribution for smoking and non-smoking mothers. Which test can be used in the study?
How many mothers in each group (smokers and non-smokers) must be at least included in the study to verify the observed effect with 80% power at the 5% significance level (use a two-sided alternative hypothesis)?

b. Now load the dataset birthwt in library MASS. Is the sample size sufficient?
Perform the test from (a) using the appropriate columns.

c. For smoking and non-smoking mothers inspect the boxplot of the weight of the mothers (column birthwt$lwt)
and decide whether it follows a normal distribution.
Now choose and perform an appropriate test to compare the weight of the smoking and non-smoking mothers.

2. (χ^2 *test, multiple testing*)

A prospective, double blinded, placebo-controlled study was carried out to determine the efficacy of the drug Bulliforton for the treatment of postprandial digestion problems. The primary endpoint (EP) of the study was a reduction in UADS ("upper abdominal discomfort severity") scores after 4 weeks ($UADS_4$) by at least 200 point relative to the baseline ($UADS_0$). The two secondary EPs were (i) a reduction in UADS score by at least 200 points after 2 weeks and (ii) a UADS score of less than 150 points after 4 weeks.

You can find the data from the study at http://www.uni-kiel.de/medinfo/lehre/medlife. Load it into your workspace.

The variables are coded in the following way:

- treatment: group the patient was randomized into (placebo/verum)
- primary EP: result of the primary EP (1: success, 0: primary EP negative)
- sec EP (i), sec EP (ii): accordingly

a. Formulate the scientific question of the Bulliforton study regarding the primary EP as a statistical decision problem by specifying null and alternative hypothesis and the corresponding statistical test.

b. Transform the primary EP data of the study into a two-times-two table with treatment (verum/placebo) as rows and primary EP (1/0) as columns and perform the statistical test identified in (a). What is the result of the study with respect to the primary EP?

c. Do the same for the secondary EPs. How would you interpret these results?

Exercise 8 (Correlation and linear Regression)

1. (*linear regression, Pearson and Spearman correlation coefficient*)

Using a laser polarimeter, the mean thickness (in μm) of the retinal nerve fibre layer (RNFL) of 10 volunteers was measured and related to the mean visual field sensitivity (VFS; in dB) of the probands. This study was aimed at modelling the functional relationship between the two entities. Such a mathematical model would be highly relevant in glaucoma diagnostics because it appears as if a measurable destruction of the nerve fibre layer antedates the functionality loss of the retina by a long way (a phenomenon known as “functional reserve”). You can find the data from the study at http://www.uni-kiel.de/medinfo/lehre/medlife. Load it into your workspace.

a. From the data of the study, calculate the intercept and slope of the least square regression line, the Pearson correlation coefficient r_{XY}, and the coefficient of determination R^2.

b. Is the Pearson correlation coefficient r_{XY} calculated in (a) significantly different from zero?

c. Perform a scatter plot of the data. Add the least square regression line using the function *abline*.
How would you interpret the result of the linear regression analysis?

d. Calculate the Spearman correlation coefficient ρ_{XY} for the data and compare it to the Pearson correlation coefficient r_{XY} from (a). How can the difference between ρ_{XY} and r_{XY} be explained?

Exercise 9 (Statistical Modeling)

1. (*multiple linear regression*)

A representative sample of 100 workers from a cadmium-processing chemical plant was assessed for a possible relationship between vital lung capacity (vc, in liters) and job tenure (jt, in years). The age of workers (in years) was also recorded. You can find the data from the study at http://www.uni-kiel.de/medinfo/lehre/medlife. Load it into your workspace (read.table("vitcap.dat", header=TRUE)).

a. Perform two simple linear regressions with either job tenure or age as single explanatory variable.
Then, perform a multiple linear regression with both job tenure and age as explanatory variables in the model.
Write down the resulting three model equations.
Produce scatter plots with the regression line added for the two simple linear regressions.

b. How can these results be interpreted?

2. (*logistic regression, odds ratio*)

The cohort birthwt in library MASS was recruited to study the question whether smoking of the mother during pregnancy reduces the birth weight of the baby. In addition to birth weight and smoking status of the mother several other potential explanatory variables were recorded.

a. Perform a logistic regression with the low birth weight status (column low) as response variable and smoking status, age of the mother, weight of the mother, hypertension and physician visits in first trimester as explanatory variables (columns smoke, age, lwt, ht and ftv).
Which model equation results after backward selection?

b. How can the results of the regression analysis be interpreted? To answer this question, transform the regression coefficients of the two dichotomous influential variables smoke and hypertension into odds ratios. or

c. Calculate the probability of giving birth to an underweight baby for a non-smoking woman without hypertension weighing 160 pounds. Compare the result to the risk of a smoking woman with hypertension and a weight of 100 pounds.

Preparation for next tutorial: Install the package “survival” onto your laptop.

Exercise 10 (Survival Analysis)

1. (*Kaplan-Meier estimate, log-rank test*)

In a clinical study, 10 newly diagnosed cancer patients were randomly assigned to one of two different chemotherapies, C1 or C2. The question was whether therapy C2 may significantly extend post-therapeutic survival compared to C1. The results are given in Table 10.1 and at http://www.uni-kiel.de/medinfo/lehre/medlife.

Table 10.1: Survival (in days) of cancer patients after chemotherapy

C1	4	18(+)	55	66(+)	90	101	148	207(+)	283	441(+)
C2	26(+)	70	93	105(+)	193	229(+)	242	455(+)	518	595

(+): right-censored observation

a. In the following table, fill in all information necessary to estimate the Kaplan-Meier survival function for therapy C2.

t_i	n_i	d_i	$\hat{P}(T > t_{i-1})$	$\hat{P}(T > t_i \mid T > t_{i-1}) = (n_i - d_i)/ n_i$	$\hat{P}(T > t_i) = \hat{P}(T > t_{i-1}) \cdot (n_i - d_i)/ n_i$

b. Plot Kaplan-Meier estimates of the survival function (i.e., “Kaplan-Meier curves”) for therapies C1 and C2. What is the qualitative difference between these two curves?

c. Calculate the log-rank statistic and interpret the result.

Solution:

1a.

t_i	n_i	d_i	$\hat{P}(T > t_{i-1})$	$\hat{P}(T > t_i \mid T > t_{i-1}) = (n_i - d_i)/ n_i$	$\hat{P}(T > t_i) = \hat{P}(T > t_{i-1}) \cdot (n_i - d_i)/ n_i$
70	9	1	1.000	8/9=0.889	0.889
93	8	1	0.889	7/8=0.875	0.778
193	6	1	0.778	5/6=0.833	0.648
242	4	1	0.648	3/4=0.750	0.486
518	2	1	0.486	1/2=0.500	0.243
595	1	1	0.243	0/1=0	0

Table A.1: Criteria for the comparison of cohort and case-control studies

data generation (retrospective vs. prospective)
data quality
costs
group formation (explanatory vs. response variable)
occurrence of response (before or after beginning of study)
availability of incidence information
scientific credibility
effect measure