# Statistical Methods Fall 2021

## Assignment 3: Estimation and testing

**Deadline: see Canvas**

*Topics of this assignment*

The exercises below concern topics that were covered in Lectures 5, 6, 7 and 8: estimation of population proportion and mean, tests for population proportion or mean based on one sample, tests for differences of two means or proportions based on two samples (see the respective sections in Chapters 6, 7, and 8 of the book and the slides of Lectures 5–8). Before solving 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;
– the observed value of the test statistic (the observed score);
– the $P$-value or the critical values;
– 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 three theoretical exercises below use Tables 2 and 3 from the Appendix in the book (also given on Canvas) to find 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 3.1** You have been employed by an airline and given the task of estimating the percentage of flights that arrive on time, i.e. no later than 20 minutes after the scheduled arrival time. How many flights should you survey in order to be 95% confident that your estimate is within two percentage points of the true population percentage? Solve this problem under each of the following assumptions:

a) nothing is known about the percentage of on-time flights of your airline;

b) for a previous year, 90% of all flights of your airline were on time (based on an official statistic from the government).

*Also state the formulas that you are using.*

**Exercise 3.2** Consider a study about the brain volumes (in $cm^3$) of 30 pairs of twins. The aim of the study was to quantify the difference in the population mean volumes of first- and second-born twins. The average volume of the 30 first-born twins was $\bar{x}_1 = 1124.3$, and that of the 30 second-borns was $\bar{x}_2 = 1118.1$. Some more statistics that you may or may not use are: $s_1 = 130.5, s_2 = 124.7, s_d = 57.8$. Construct a 90% confidence interval for the difference of the population means.
*Also state the formulas that you are using.*

**Exercise 3.3** Consider a study about the brain volumes (in $cm^3$) of 25 first-borns and 20 (unrelated) second-borns. The aim of the study was to test whether there is a difference in the population mean volumes of first- and second-born babies. The average volume of the 25 first-borns was $\bar{x}_1 = 1131.3$, and that of the 20 second-borns was $\bar{x}_2 = 1123.8$. Some more statistics that you may or may not use are: $s_1 = 129.0, s_2 = 127.2, s_p = 128.2$. Conduct the statistical test of interest and use $\alpha = 5\%$. Do we need to make specific assumptions in order to be allowed to use the test?
*(See the first page of the assignment for detailed instructions about testing)*

## R-exercises

*Hints concerning R:*

- The *R*-functions `pnorm` and `qnorm` can be used for computing probabilities and quantiles of normally distributed random variables. Similarly, `pt(...,df=k)` and `qt(...,df=k)` can be used to compute probabilities and quantiles of a $t$-distributed random variable with $k$ degrees of freedom.
- For computing a confidence interval based on a $t$-distribution and for performing a $t$-test the *R*-function `t.test` can also be used. For example, to perform a $t$-test with significance level 5% for testing the null hypothesis $\mu = 10$ against the alternative hypothesis $\mu > 10$, using the data set `example` use the command `t.test(example,mu=10,alt="greater")`. The arguments of `t.test` can be adjusted to test with other significance levels and other null and alternative hypotheses as well: study `help(t.test)`.
- Note that the function `t.test` reports also the $P$-value. Use `t.test(...)$p.value` to access it without printing the whole output of `t.test()`.
- In order to get a proportion of elements of the vector `x` that are, for instance, less than some value `M` you can use `mean(x<M)`.
- The *R*-function `t.test` can also be used for *two*-sample problems: put the values of the two samples into two vectors, `x` and `y`, say; the command `t.test(x,y, ...)` performs a two-sample test (based on the $t$-distribution) and computes a confidence interval as well, with default significance level $\alpha = 5\%$. Values of other arguments, like `"paired = TRUE"` to perform a paired $t$-test, can be specified by inserting them at the position of the dots.

**Exercise 3.4** Alice and Bob work evening shifts in a supermarket. Alice has complained to the manager that she works, on average, much more than Bob. The manager claims that on average they both work the same amount of time, i.e. the competing claim is that the average working hours are different. After a short discussion between the manager and Alice, the manager randomly selected 50 evenings when Alice and Bob both worked. The datasets `Alice` and `Bob` in the file `Assign3.RData` contain the number of hours they have worked per evening.

a) Give an estimate and also a 90% confidence interval for the difference of mean working time per evening of Alice and Bob.
*Also state the formulas that you are using.*
(Remark: sometimes estimates, i.e. statistics which are single numbers, are also called *point estimates*, whereas confidence intervals are sometimes called *interval estimates*.)

b) Investigate the manager's claim with a suitable test. Take significance level 10%. Motivate your choice.
*(See the first page of the assignment for detailed instructions about testing)*

c) Now investigate Alice's claim with a suitable test. Take significance level 1%. Motivate your choice.
*(See the first page of the assignment for detailed instructions about testing)*

**Exercise 3.5** Alice from the previous exercise has another concern. By contract the employees are supposed to work 3.8 hours per evening. Alice claims that the proportion of evenings on which she worked more than 3.8 hours is larger than the proportion of evenings during which Bob worked more than 3.8 hours.
*In contrast to Exercise 3.4, we now suppose that the data given in `Assign3.RData` were all collected on different evenings for Suppose again that the data were collected on 100 different evenings. We will reuse that dataset in this exercise.*

a) Based on the data find an estimate for the difference in proportion of evenings that Alice and Bob have worked more than 3.8 hours.

b) Investigate Alice's claim with a suitable test. Take significance level 1%. Motivate your choice.
*(See the first page of the assignment for detailed instructions about testing)*