EDDA: Assignment 1

Throughout this assignment tests should be performed using a level of 0.05, unless otherwise specified.

Exercise 1. Birthweights

The data set `birthweight.txt` contains the birthweights of 188 newborn babies. We are interested in finding the underlying (population) mean μ of birthweights.

a) Check normality of the data. Compute a point estimate for μ. Derive, assuming normality (irrespective of your conclusion about normality od the data), a bounded 90% confidence interval for μ.
b) An expert claims that the mean birthweight is bigger than 2800, verify this claim by using a t-test. What is the outcome of the test if you take $\alpha = 0.1$? And other values of α?
c) In the R-output of the test from b), also a confidence interval is given, but why is it different from the confidence interval found in a) and why is it one-sided?

Exercise 2. Power function of the t-test

We study the *power function* of the two-sample t-test (see Section 1.9 of Assignment 0). For `n=m=30`, `mu=180`, `nu=175` and `sd=5`, generate 1000 samples `x=rnorm(n,mu,sd)` and `y=rnorm(m,nu,sd)`, and record the 1000 p-values for testing H_0: `mu=nu`. You can evaluate the power (at point `nu=175`) of this t-test as fraction of p-values that are smaller than 0.05.

a) Set `n=m=30`, `mu=180` and `sd=5`. Calculate now the power of the t-test for every value of `nu` in the grid `seq(175,185,by=0.25)`. Plot the power as a function of `nu`.
b) Set `n=m=100`, `mu=180` and `sd=5`. Repeat the preceding exercise. Add the plot to the preceding plot.
c) Set `n=m=30`, `mu=180` and `sd=15`. Repeat the preceding exercise.
d) Explain your findings.

Exercise 3. Telecommunication company

A telecommunication company has entered the market for mobile phones in a new country. The company's marketing manager conducts a survey of 200 new subscribers for mobile phones. The results of the survey are in the data set `telephone.txt`, which contains the first month bills $X_1, \ldots X_{200}$, in euros.

a) Make an appropriate plot of this data set. What marketing advice(s) would you give to the marketing manager? Are there any inconsistencies in the data? If so, try to fix these.
b) By using a bootstrap test with the test statistic $T = \text{median}(X_1, \ldots, X_{200})$, test whether the data `telephone.txt` stems from the exponential distribution $\text{Exp}(\lambda)$ with some λ from $[0.01, 0.1]$.
c) Construct a 95% bootstrap confidence interval for the population median of the sample.
d) Assuming $X_1, \ldots X_n \sim \text{Exp}(\lambda)$ and using the central limit theorem for the sample mean, estimate λ and construct again a 95% confidence interval for the population median. Comment on your findings.
e) Using an appropriate test, test the null hypothesis that the median bill is bigger or equal to 40 euro against the alternative that the median bill is smaller than 40 euro. Next, design and perform a test to check whether the fraction of the bills less than 10 euro is less than 25%.

Exercise 4. Energy drink

To study the effect of energy drink a sample of 24 high school pupils were randomized to drinking either a softdrink or an energy drink after running for 60 meters. After half an hour they were asked to run again. For both sprints they were asked to sprint as fast they could, and the sprinting time was measured. The data is given in the file `run.txt`. [Courtesy class 5E, Stedelijk Gymnasium Leiden, 2010.]

a) Disregarding the type of drink, test whether the run times before drink and after are correlated.
b) Test separately, for both the softdrink and the energy drink conditions, whether there is a difference in speed in the two running tasks.
c) For each pupil compute the time difference between the two running tasks. Test whether these time differences are effected by the type of drink.
d) Can you think of a plausible objection to the design of the experiment in b) if the main aim was to test whether drinking the energy drink speeds up the running? Is there a similar objection to the design of the experiment in c)? Comment on all your findings in this exercise.

Exercise 5. Chick weights

The dataset `chickwts` is a data frame included in the standard `R` installation, to view it, type `chickwts` at the `R` prompt. This data frame contains 71 observations on newly-hatched chicks which were randomly allocated among six groups. Each group was given a different feed supplement for six weeks, after which their weight (in grams) was measured. The data frame consists of a numeric column giving the weights, and a factor column giving the name of the feed supplement.

a) Test whether the distributions of the chicken weights for `meatmeal` and `sunflower` groups are different by performing three tests: the two samples t-test (argue whether the data are paired or not), the Mann-Whitney test and the Kolmogorov-Smirnov test. Comment on your findings.

b) Conduct a one-way ANOVA to determine whether the type of feed supplement has an effect on the weight of the chicks. Give the estimated chick weights for each of the six feed supplements. What is the best feed supplement?

c) Check the ANOVA model assumptions by using relevant diagnostic tools.

d) Does the Kruskal-Wallis test arrive at the same conclusion about the effect of feed supplement as the test in b)? Explain possible differences between conclusions of the Kruskal-Wallis and ANOVA tests.