Home assignment 3

Exercise 3.1

Hand calculation: In exercise 2.4 we used heat treated milk data to see if the number of coliforms is significantly lower than 4.7. When going through the data once again the researcher realizes that there actually are measurements made for each sample before and after heat treatment. Conduct an appropriate test for this situation and determine if the mean number of coliforms is significantly lower after treatment. *In home assignment 2 we assumed normality for the log transformed milk data, this assumption should be used here as well.*

	Before treatment	After treatment
Sample 1	3.9	3.4
Sample 2	5.3	4.1
Sample 3	6.1	4.7
Sample 4	4.9	4.2
Sample 5	9.1	7.6
Sample 6	2.8	1.5
Sample 7	3.5	2.3
Sample 8	3.2	2.6
Sample 9	2.6	2.0
Sample10	5.9	5.2

Exercise 3.2

When statistical tests and confidence intervals are used there are always some assumptions that need to be fulfilled. For t-tests and confidence intervals for means the assumptions are:

- The observations must be independent.
- The observations must be normally distributed or the sample size must be large enough so that the mean is normally distributed according to the Central Limit Theorem.

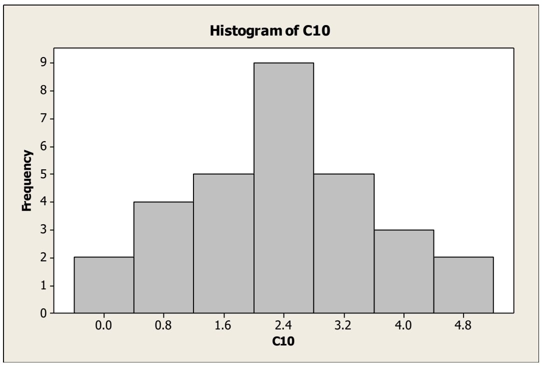
Determine for the situation below if the observations are independent or not. Why? Why not?

i) In a study on coliforms in milk cows at 15 farms are examined. From each farm 3 calves are selected with leads to a total of 45 observations are used in the analysis. Are the 45 observations independent?

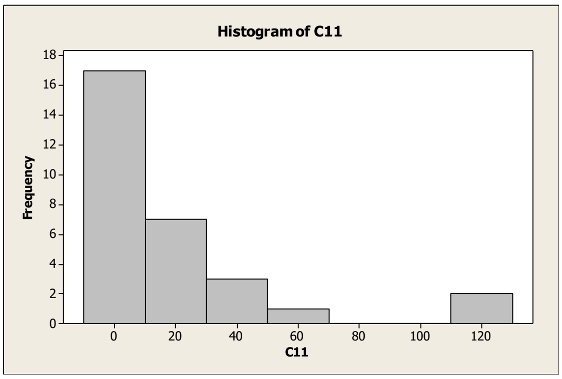
ii) In a study on coliforms in milk cows at 15 farms are examined. From each farm 3 calves are selected and a mean value is computed per farms for these three calves. The dataset consists of 15 mean values, one per farm. Are the 15 mean values independent observations?

Determine for the situation below if the observations are approximately normally distributed. Motivate.

iii)



iv)



v) If observations are not normally distributed: How many observations do we at least need to make to be able to use the central limit theorem. i.e. how large must a sample be so that I always can use the t-test to test the mean or compute the corresponding confidence interval, even if the data is not normally distributed.

Exercise 3.3

Computer Exercise:

For the cordblood data we want to conduct hypothesis tests to see whether or not male and female babies have the same level of **measles** antibodies. We have earlier determined that the distribution of this data is rather skewed, i.e. not normal. On the other hand we have quite many observations and the distribution of the mean should be approximately normal. We have here (at least) 3 different options:

- run a two sample t-test assuming that the mean value is normally distributed due to the large sample size
- log-transform the data and run a two-sample t-test. Observations will after log-transformation be close to a normal distribution
- use the non-parametric Wilcoxon rank sum test

Provide R code and output for the following and answer the questions. Provide everything in only one file.

a) Run three tests in R to compare the level of measles antibodies for male and female babies:

   a.1 a two sample t-test assuming normal distribution

   a.2. a two sample t-test on data that is log-transformed

   a.3 a nonparametric Wilcoxon rank sum test

   Compare the results from the three tests. Are they similar? Are there any significant differences between Measles antibody levels for males and females?

b) Check if the distribution of the log-transformed number of antibodies for Measles for **male** babies is closer to a normal distribution than the original data was (use histogram, boxplots or qq-plots). Comment.

c) Create age classes like for the cats data, making 5 years intervals starting from 15 until 45. (Do this preferably in R, but if you cannot make it work you can also make the changes in the txt file or import the file to Excel and make the changes before you again save the file as txt-file and read it to R)

d) Test if there is a significant difference in Measles antibody levels for babies from mothers aged 20-25 compared to mothers aged 35-40. (Choose either log-transformed data or non-parametric tests as you like). Draw conclusions from the test.