EXAM FOR APPLIED BIOSTATISTICS I

Name: _____

Matriculation number: _____

INSTRUCTIONS

The exam starts at 08:00 am and you have two hours to complete it. Please send the answers to

stephan.peischl@bioinformatics.unibe.ch

when you are done. Ideally you submit your answers as a pdf file. Alternative formats that I accept are: text files, scans of handwritten answers in sufficiently high quality (handwriting and scan resolution!), word documents. When sending attachments, make sure that they are not too big. If you have trouble sending a file, just send me an email and I will get in touch with you to find a solution. I will also be online in my Zoom room in case there are questions or other problems:

Join Zoom Meeting:

https://unibe-ch.zoom.us/j/5606529364?pwd=Q2Z1K3B6b2xpNnZueGh2Vy9SelVyUT09

Remember that this is a new situation for all of us and if something doesn't work out the way it was planned or if there are some technical problems, keep calm and carry on - we will find a solution!

RULES

This is an open book exam. You can use any resource you want: lecture notes, books, even searching the internet. However, it is not allowed to communicate about the exam questions during the exam, you cannot copy and paste from any book, webpage, etc. Therefore, it will be critical that you try to be concise in answers to the questions and not just write down everything that comes to your mind. I want to see what you think is relevant - not just how well you can copy a text book chapter or wikipedia entry using your own words.

Please remember to save your progress regularly during the exam!

Try to be concise! Give short answers!

GRADING SCHEME

Grade = 0.2 * grade from exercises + 0.8 * grade from exam.

The grade from the exam is determined as:

points / (maximum possible points) * 5+1

QUESTION 1 (8 POINTS)

A) Pick *two* of the following three concepts from the course and write one paragraph about it (maximum 1/2 A4 page). Try to summarize the most important features you have learned about the concept you have chosen. Be concise and don't just list everything that comes to your mind! I am interested in what you think is important, not everything you know.

Choose *two* topics:

- Difference between parameters and estimates
- Standard error of the sample mean and its relationship to uncertainty of estimates
- Relationship between hypothesis tests and confidence intervals in the context of a one-sample t-test

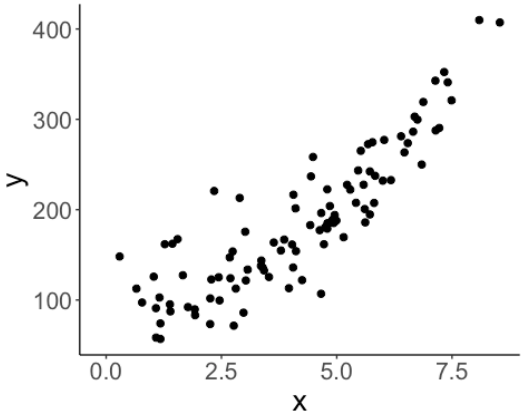
B) Discuss the idea of choosing a null hypothesis in the context of hypothesis testing. Why do we assume a null hypothesis? What is the idea behind this? Is this useful or misleading? Some people say we should abandon classical hypothesis testing and p-values completely – what is your opinion?

QUESTION 2 (OPTIONAL QUESTION, 2 BONUSPOINTS):

Write down one question about statistics that you want to ask me. Something that is not clear to you or something that you are not sure you understood correctly. There are no stupid questions and you cannot “lose” points here (bonuspoints can only improve your grade but not lower it).

QUESTION 3 (4 POINTS):

Below you see an example for a data set. The data describes points scored on an exam (y) as a function of time invested in learning (x). The lecturer performed a linear regression on the data.




```
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.478 -25.605  -6.363  25.297  97.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.441      9.047   4.691  8.9e-06 ***
## x             34.544      1.959  17.631  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.41 on 97 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7622, Adjusted R-squared:  0.7597
## F-statistic: 310.9 on 1 and 97 DF,  p-value: < 2.2e-16
```

A student asks how much they would have needed to learn to get a score of 500 points (the maximum score that is achievable). The lecturer respond: *"… the model predicts a value of* $\hat{y} = 42.4 + 34.5\ x$ *and hence* $\hat{y} \approx 500$ *for* $x = 13.2$*. Therefore you should have learned at least 13 hrs or more to obtain 500 points."*

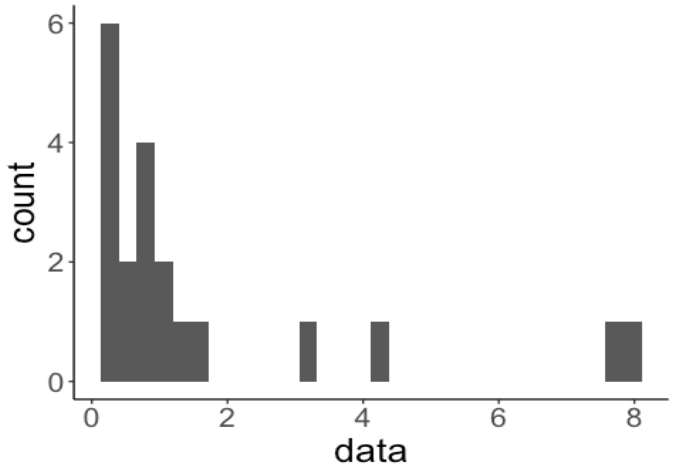
Answer the following questions / perform the following tasks:

- Is the method used by the researchers appropriate. If yes: why? If no: why not?
- Is there a problem with the conclusions drawn from the analysis? If yes, what is the problem and how could you fix it?

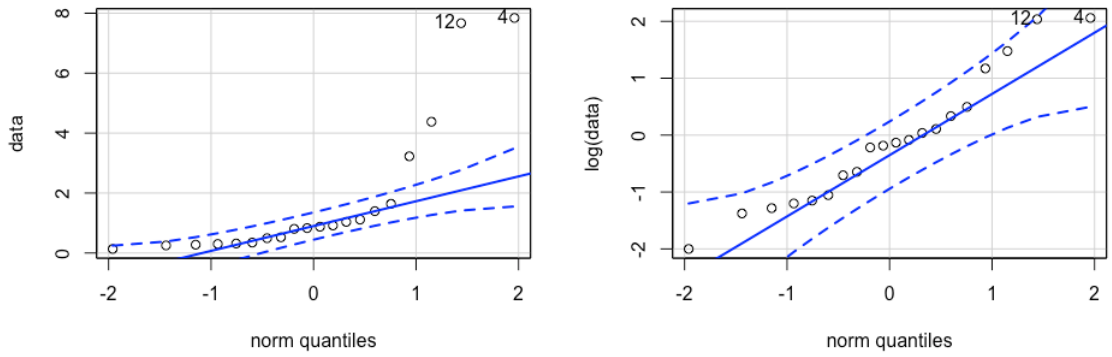
- How could you improve / extend the analysis?
- Add the regression line to the plot (or make a sketch of how the plot and the regression line would roughly look).
- How do you think would the data look like for students that learned 10 hrs or more? Make a sketch and explain.

QUESTION 4 (3 POINTS):

The following histogram shows some raw data with sample size $n = 20$:



and the corresponding qq-Plots of the raw data and the log-transformed data:



You are supposed to test if the mean of the ***original raw data (= non-transformed)*** is significantly different from 1. Write down how you would do this.

- Which test would you use? Why?
- Write down the null hypothesis as precisely as possible!

(You don't need to write the R code or do the analysis, just write down what the steps would be.)

QUESTION 4 (3 POINTS):

Consider data that describes the size of cancer cell colonies grown on organoids. The control group received the best currently available treatment and the treatment group received a new drug. The goal of the study is to test if the new drug works better than the old one. The researcher in your department shows you the output of a ***t-test*** they performed. The data is normalized such that mean and standard deviation in the control group is (approximately) 0 and 1, respectively. In other words, we measure the increase in tumor size in units of standard deviations of the control group. A value of 0 in the treatment group indicates that tumors are the same size as in the control, and a value of -1 would indicate a decrease by 1 standard deviation. Assume that the analysis was performed correctly and that the data is indeed normally distributed. Here is the output of the t.test:

```
##
##  Two Sample t-test
##
## data:  control and treatment
## t = 2.4429, df = 4e+05, p-value = 0.01457
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.001528236 0.013932442
## sample estimates:
##    mean of x    mean of y
## -0.002748602 -0.010478941
```

- What is the sample size (approximately)?
- Is there a significant difference between the mean of the two groups?
- Is it justified to say the new drug work better? If yes, why? If, no, why not?