Statistical data analysis

Name: _____

Signature: _____

Labs	
Exam (written + oral)	≥ 25
Total	≥ 50
Grade	

Instructions: the solution time is 120 minutes, clearly answer as many questions as possible, work with the terms used in the course, employ math (notation, expressions, equations) as often as possible, you can use calculators.

Statistical minimum. *(10 b)* Answer the following questions:

(a) *(6 b)* Define a likelihood function (often denoted simply the likelihood). Explain the purpose of maximum likelihood estimation. How would you employ the likelihood in statistical hypothesis testing?

(b) *(4 b)* Explain the difference between point and interval estimate of statistical parameters.

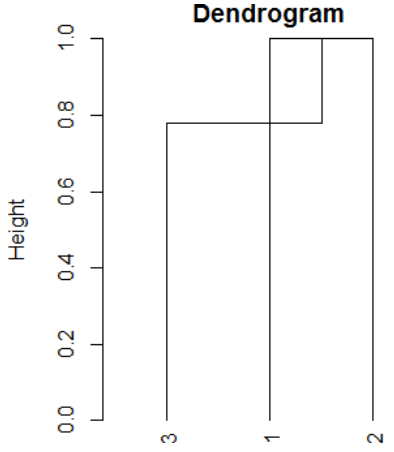
Hierarchical clustering. *(10 b)* Discuss hierarchical clustering, answer the questions and issues below.

(a) *(3 b)* Formally define hierarchical agglomerative clustering (input, output, algorithm and its parameters).

(b) *(2 b)* What is a taxonomy and dendrogram? Thoroughly interpret their purpose in hierarchical clustering (why is a taxonomy more informative than a mere partition of the input set, where can we find similarity between a pair of objects and/or their clusters in a dendrogram, etc.).

(c) *(3 b)* Estimate time complexity of hierarchical agglomerative clustering. The result is important, however, the reasoning behind it is equally important.

(d) *(2 b)* In the dendrogram below, there is so-called inversion, i.e. a scenario under which the similarity during agglomerative clustering does not decrease but increases. Does this scenario really occur? Under which conditions?



Multivariate regression. *(10 b)* You are a mechanical locksmith and you are trying to find out how the shaft machining error is related to the machine tool parameter setting. You have compiled a multivariate linear model. The model expresses the relationship between the production error (the difference between the ideal shaft diameter and the actual shaft diameter, ProdError) and the setting of ten different continuous machine parameters (P1-P10). Below is the output you received:

```
summary(lm(ProdError ~ P1+P2+P3+P4+P5+P6+P7+P8+P9+P10))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2638     0.2556   1.032    0.329
P1            0.2471     0.2564   0.964    0.360
P2           -0.6112     0.1979  -3.089    0.013 *
P3            0.2728     0.2341   1.165    0.274
P4            0.1093     0.2061   0.530    0.609
P5           -0.3165     0.4674  -0.677    0.515
P6           -0.4419     0.2660  -1.661    0.131
P7            0.1244     0.3152   0.395    0.702
P8            0.2452     0.2657   0.923    0.380
P9            0.1287     0.3093   0.416    0.687
P10           0.3544     0.2956   1.199    0.261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7549 on 9 degrees of freedom
Multiple R-squared:  0.7237,        Adjusted R-squared:  0.4168
F-statistic: 2.358 on 10 and 9 DF,  p-value: 0.1062
```

(a) *(2 b)* Decide whether at least one of the machine parameters (independent variables) is useful for estimating a manufacturing error (ProdError). In other words, formally decide whether you can decline $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{10} = 0$. Justify correctly.

(b) *(2 b)* Can the model description be used to find out the number of samples of the different shafts the model has been assembled from? If so, calculate the number.

(c) *(2 b)* How do we consider this number when evaluating the utility of the model? How do the type I error and type II error change with increasing number of samples if we maintain a constant level of significance α?

(d) *(4 b)* Describe in detail the way in which you would verify your model over the samples that you currently have. You can create additional auxiliary models. Describe the validation method, define the error function, and specify with which baseline you will compare the calculated error.

Logistic regression. *(10 b)* Suppose you have data on the last year's SAN course student group. For each student, we know the number of hours he/she was preparing for the exam ($hours$), his/her bachelor study grade point average (avg), and whether or not he/she came from the bachelor OI program (OI). The target binary variable Y denotes whether the given student earned an A SAN grade. You learned a logistic model and got the coefficients $\beta_0 = -1$, $\beta_{hours} = 0.05$, $\beta_{avg} = -1$ and $\beta_{OI} = 1$.

(a) *(3 b)* What is the interpretation of the coefficients in the logistic model? Compare with the linear regression, where the coefficients express the average change of the output with the unit change of the given independent variable while keeping the values of the other independent variables constant. Explain step by step for β_0, β_{hours} and β_{OI}.

(b) *(2 b)* Calculate how the student who came from the OI bachelor program, who was preparing for 20 hours and his avg was 2 will be classified according to your model?

(c) *(2 b)* How long would the aforementioned student have to prepare for the exam to have just 50% probability of getting an A grade?

(d) *(3 b)* On this model/example, illustrate the term confounding variable. Define the term, show a simple relationship between the variables. You can customize the model anyhow.

Robust statistics. *(10 b)* Robustly estimate the location from the sample below. Use three different methods. {-1.84, 1.18, 0.0499, -0.751, -0.00707, -2.05, -1.47, -0.0520, -0.991, -0.945}.

(a) *(2 b)* Method 1:

(b) *(2 b)* Method 2:

(c) *(2 b)* Method 3:

(d) *(2 b)* Describe the criteria that determine the quality of a robust location estimate.

(e) *(2 b)* Discuss the advantages and disadvantages of chosen methods according to the criteria described in the previous subtask.