Problem 1: Statistical Tests

In order to compare the milk production of two farms a random sample of 10 dairy cows of farm A and a random sample of 15 cows of farm B are taken and the weekly yield of milk is recorded (in kg, dataset `farms_milk.asc`):

Farm A: 141.6, 128.2, 153.4, 161.4, 156.5, 129.9, 110.3, 148.8, 168.6, 120.9

Farm B: 139.8, 163.3, 135.5, 191.5, 183.7, 172.7, 220.2, 150.5, 163.0, 160.5, 200.0, 153.2, 140.7, 155.5, 165.2

Check whether there is a significant difference between the two herds at the 5% and 1% level of significance and comment on the results with respect to the level of significance.

Submission of your report

Submit either a Word or a pdf file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 2: Effects of Biofilms

The effects of four biofilm types on the recruitment of serpulid larvae has been tested by leaving the substrates for several days in shallow marine waters.

At the end of the experiment the number of newly recruited serpulid worms have been counted:

sterile unfilmed subtrate (SL):	63, 113, 123, 70, 120, 83, 95
unnetted lab films (UL):	143, 81, 101, 155, 151, 193, 163
netted lab films (NL):	191, 159, 139, 161, 179, 97, 157
netted field biofilms (NF):	128, 194, 108, 116, 140, 160, 87

Check whether there is a significant effect of various biofilms on the recruitment of serpulid worms (data set: `biofilm_2022.asc`). Perform your checks on both a 1% and a 5% level of significance.

Submission of your report

Submit either a Word or a pdf file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 3: Linear Regression

Ruminants (cattle, sheep and goats) produce meat by digesting complex carbohydrates. As shown in the literature the meat production can be increased by supplementary protein meals. The dataset `ruminant_meat_2022.asc` contains the data obtained from such experiments, i.e. the daily protein intake (normalized to live weight), and the increase of live weight per day.

Establish a linear model relating the weight increase per day to the daily protein intake. Try to use three different functions for the model (straight line, parabolic and logarithmic) and provide an answer to the following questions:

1) Which of the proposed models would you select as the best one?

2) If you look at the parabolic model: what are the arguments against it or in favor of it?

3) Using the selected "best" model: assuming that the cattle's daily intake of additional protein is 2.5 g/kg, what is the expected weight increase per day? Do not forget to specify the confidence interval of your estimate (95%).

Submission of your report

Submit either a Word or a pdf file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 4: Multiple Regression

The data set `bodyfat_2022.asc` contains the percentage of body fat, age, weight, height, and ten body circumference measurements for 174 men. Try to establish a model for estimating the body fat from the given variables.

1) Calculate an MLR model using all variables (model 1) and check which of the variables contribute significantly to the model (at the 5% level of significance).

2) Discard all non-significant variables and recalculate the model (model2). Compare this model with a model obtained by stepwise regression (model 3).

3) Explain the differences between model 2 and model 3. Which of the two models is better? Explain why.

4) Check for multicollinearities in model 3.

5) Do you think that person "R42" who exhibits an unusual low body height has any influence on model 3?

Submission of your report

Submit either a Word or a pdf file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 5: PCA of Breast Cancer Data

The dataset `breast_cancer.idt` describes 9 characteristics of 683 samples of breast tumor mass (obtained by fine needle aspirate). The data come together with a classification regarding the malignance of the tumor.

Try to get an overview of the data by means of principal component analysis. Create score/score-plots for the most important principal components covering at least 85% of the total variance and try to find an answer to the following questions:

- How many principal components are sufficient to describe the dataset?
- What is the reason that there is not much difference between mean centered and standardized data?
- How could we decide whether a particular tissue is malignant? Suppose that you know the parameters of two samples:

Parameter	**Sample 1**	**Sample 2**
ClumpTh	6	2
UnifCellSz	10	3
UnifCellShape	9	1
MargAdhesion	10	2
SEpiCellSz	4	5
BareNuc	10	1
BlandChrom	8	2
NormNuc	10	1
Mitoses	1	1

  Which of the two samples is most probably malignant?
- Multiply the variable 3 (UnifCellSz) by a factor of 10 and repeat the PCA (both for mean centered and standardized data). Which differences you see? Explain your findings.

Submission of your report

Submit a PDF file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 6: Estimating the moisture content of corn

One of the major parameters to be measured before storing corn is its moisture content. A high water content will possibly result in fungi growing on the stored corn. One (very quick) way to measure the moisture is to acquire an infrared spectrum of ground corn and use a mathematical model to estimate its water content from the spectrum.

The dataset `corn_2022.asc` contains reduced infrared spectra of 80 corn samples, and their moisture contents.

Try to establish a multilinear model for the estimation of the moisture using PLS based on standardized data. Give an answer to the following questions:

- Find the optimum number of factors
- Compare the "actual vs. estimated" plot for 2,4,6 and 8 factors. What do you observe when you increase the number of factors?
- Compare the regression coefficients of the original variables to the absorption spectrum of water (see for example https://omlc.org/spectra/water/data/palmer74.txt for details). Perform this comparison using 3, 6, 10 and 35 factors.
- Why is it a bad idea to use MLR instead of PLS for modelling the moisture content?
- Is there something special about a model using 35 factors?

Submission of your report

Submit a PDF file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 7: Recognizing Dry Beans

One common task in food processing is to automatically recognize (=classify) certain cultivars. The dataset `drybeans_2022.asc` contains geometric measures of 2000 dry beans of seven different cultivars automatically obtained by an imaging device.

Try to create three classifiers which are able to distinguish the cultivars SEKER, BOMBAY and SIRA from the other six cultivars, respectively. The three classifiers should be based on PLS/DA and RF (random forests), so that you end up all together with 6 different classifiers. Do not forget to optimize the number of factors in the case of PLS/DA and find the optimum number of trees of the random forest.

Compare the six optimized classifiers by comparing the MCC (Matthews Correlation Coefficient) obtained from 10-fold cross-validation and discuss your findings.

Find an answer to the following questions:

- Why is the RF based classifier for classes 6 and 1 (cultivars SIRA and SEKER) considerably better than the corresponding PLS/DA classifier
- Why performs PLS/DA better than RF for class 3 (cultivar BOMBAY)?
- What are the drawback of using too many or too few trees in an RF classifier?

Submission of your report

Submit a PDF file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 8: Clustering of Red Wines

Adulterated wines are always a problem for market authorities. One way to uncover the adulteration of wines is to perform a chemical and physical analysis of the constituents of wine. The dataset `wines_biostat_2022` contains the results of an analysis of 55 Italian red wines (Barolo, Grignolino, Barbera).

Exercise:

Perform a cluster analysis of this dataset using all variables and Ward's method.

1) Do you find any misclassification (i.e. a wine which belongs to the "wrong" subtree of the dendrogram)?
2) Compare the dendrograms obtained from standardized data and unscaled data. Why are there so many misclassifications in the case of unscaled data?
3) What do you think are the most influential variables when using unscaled data?
4) Try to use PCA to get an idea how severe the misclassification is. What are your findings?

Submission of your report

Submit either a Word or a pdf file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).
The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).
Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)

Problem 9: Metal inhibition of oxygen uptake

In order to find out the effect of certain metals on the inhibition of oxygen uptake of activated sludge a factorial experimental design has been set up. The results obtained from a two-level, four-factor experiment allow to study the effect of zinc (Zn), cobalt (Co), antimony (Sb) and calcium (Ca). The design used the following metal concentrations:

- Zn: 0 and 10 mg/L
- Co: 0 and 1 mg/L
- Sb: 0 and 1 mg/L
- Ca: 100 and 600 mg/L

The cumulative oxygen uptake (in mg/L) has been measured in 20 hours of reaction time. The entire dataset is contained in file `sludge_2022.asc`

Calculate and interpret the main and interaction effects of the four metals. Which of the four metals increase the oxygen uptake? What would you recommend the manager of the sewage plant? Restrict the interactions to second order interactions only.

Submission of your report

Submit either a Word or a pdf file containing the description of your solution. **Please do not forget to put your name at the beginning of the report**. Please describe what you have done and why you have done that. List your results and discuss the results (if you feel that a discussion is required).

The report file has to be (re)named according to the following template:

`biostat_XX_lastname_firstname.pdf`

(with XX being the problem number).

Please **do not** submit several reports in a single file!! (Submitting several reports in a single file makes it harder for me to manage and correct your reports.)