Project 1: Water environment in Skive fjord

Formalities, structure and expectations for the first mandatory project

The assignment consists of two parts. The first part focuses on descriptive analysis of the data. The second part is primarily about confidence intervals and hypothesis tests.

The assignment is formulated in such a way that it can be solved in small "easy" steps. In practice, the assignment must be solved using Python. Some Python code is provided in order to make it easy to get started with the project. However, the code is not complete, and you are encouraged to explore new features while working on your code for the project. For example, you could add suitable titles to the plots, or built-in functions for computing confidence intervals and testing hypotheses.

The results of the analysis must be documented in the report using tables, figures, mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in the appendix. Present the results of your analysis as you would when explaining them to one of your peers.

Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. Code should not be included in the report itself but must be handed in as an appendix (a .py or .ipynb file). The report and appendix must be handed in under Assignments on Learn at: `Projekt 1: Vandmiljø in Skive fjord`

The report text should not exceed 6 pages (excluding figures, tables, and the appendix).

A normal page contains 2400 characters.

Figures and tables cannot stand alone - it is important that you describe and explain the Python output in words.

Figures and tables are not included in the assessment of the length of the report. However, it is not in itself an advantage to include many figures, if they are not relevant!

You may work together in groups, but the report must be written individually. Questions about the project can be addressed to the teaching assistants, see the guidelines on the `Projects` page of the course website.

Introduction

This project focuses on data collected from Skive fjord between the years 1982 and 2006. During this time period different legislations were introduced with the purpose of reducing the harmful emission of nitrate and phosphorus to the water environment. These legislations are called "Vandmiljøplaner" (VMPs).[1] The objective of this assignment is to investigate whether the VMPs have led to a significant reduction in the emission of nitrate.

What defines a good water environment is always up for discussion. However, most people would agree that unclear, smelly water is not very appealing. The main cause of greenish and unappealing water in streams, lakes and fjords is too high levels of phytoplankton, which can also lead to deoxygenation. This can have fatal consequences for the animal life in the water. The level of phytoplankton depends positively on, among other things, the level of nitrate in the water. Thus, reducing nitrate emissions would likely have a positive effect on the water environment. Phosphorus is another necessary nutrient for phytoplankton.

The main source of nitrate in inland waters is agriculture fertilization, which is sought to be regulated by the legislations. In Denmark, three VMPs have been implemented:

- `VMP1` in 1987
- `VMP2` in 1998
- `VMP3` in 2003

[1] `http://mst.dk/erhverv/landbrug/saerligt-for-borgere-om-landbrug/baeredygtighed-i-landbruget/vandmiljoeplanerne-et-historisk-overblik/`

In each VMP different initiatives were taken to decrease the level of nitrate emissions.

Reading the data into Python

Make a folder for the project on your computer. Download the project material from Learn and unzip it to the folder that you just made.

Then, open the data file `skivefjord1_data.csv` (e.g., in VS Code (or other editor), File → Open File) in order to see the contents of the file. Note that the first row (referred to as a *header*) contains variable names, and that the subsequent rows contain the actual observations. Variable names and observations of the individual variables are separated by a ';' (therefore `.csv`: "comma separated values", though here it is a semicolon).

The dataset consists of annual observations of the nitrate and phosphorus emissions to Skive fjord. The data file contains the following columns/variables:

Variable	**Explanation**
`year`	Year of observation
`vmp`	The applicable VMP (0, 1, 2 or 3)
`Nload`	The nitrate emission to Skive fjord in tonnes (t)
`Pload`	The phosphorus emission to Skive fjord in tonnes (t)

Open the Jupyter notebook file `skivefjord1_english.ipynb`, which contains some code that may be used in the analysis. First, the "Path" must be set to the directory on the computer, which contains the files for the project. After that, the data may be read into Python using the code in the Jupyter notebook. See "Read Data" section in the notebook.
`D` becomes a "data.frame" (a kind of table), which contains the data that was read into Python.

Descriptive analysis

The purpose of the first part of the project is to carry out a descriptive analysis of the data. In a report it is important to present the data and describe it to the reader. For example, this can be done using summary statistics and suitable figures.

Start by running the commands in section "Simple summary of data" in the notebook, to get an overview of the data.

a) Write a short description of the data. Which variables are included in the dataset? Are the variables *quantitative* and/or *categorized*? (Categorized variables are only introduced in Chapter 8, but they are simply variables which divide the observations into categories/groups - e.g. three categories: low, medium, and high). How many observations are there? Which time period is covered by the observations (year of first and last observations)? Are there any missing values?

In the section "Histogram (empirical density)" in the notebook, you can find code to generate the empirical density of the annual nitrate emissions. (See Section 1.6.1).

b) Make a density histogram of the annual nitrate loads. Use this histogram to describe the empirical distribution of the loads. Is the empirical density symmetrical or skewed? Can the nitrate emission be negative? Is there much variation to be seen in the observations?

Note: In a *skewed* distribution, the probability mass is not symmetrically distributed around the median. In a left-skewed distribution, the left tail is longer than the right tail (and, typically, the mean will lie to the left of the median). Similarly, in a right-skewed distribution, the right tail is the longer of the two (usually, with the mean to the right of the median).

When doing data analysis, it is often useful to be able to divide the data into subsets. This can be done in Python using, e.g., Logical expressions. See also additional tips in the bottom of the notebook.

Use the Python code in the section "Data subsets" in the notebook to make subsets of the data - one subset for each VMP.
When observations are recorded regularly over time, the data is often referred to as a *time series*. Thus, the annual nitrate emissions constitute a time series. For time series, it is often relevant to make figures illustrating the data over time. A plot illustrating the annual nitrate loads over time (coloured by the applicable VMP) can be made using the Python code in section "Plot development of nitrate emissions" in the notebook.

c) Make a plot illustrating the annual nitrate emission over time (coloured according to the applicable VMP). Describe the development of the nitrate emission over time in words. Does the level of emission seem to increase or decrease? Does the development over time seem to depend on the applicable VMP? Are there any years where the nitrate emission is notably different?

In the section "Box plot by VMP" in the notebook, you can find code to make a box plot of the annual nitrate emission by applicable VMP.

d) Make a box plot of the annual nitrate loads by VMP. Use this plot to describe the empirical distribution of the annual emission during each of the four VMPs. Are the distributions symmetrical or skewed? Does there seem to be a difference between the distributions (if so, describe the difference)? Are there extreme observations/outliers?

The empirical distribution of the annual nitrate emission during each of the four VMPs may also be quantified using summary statistics as in the following table:

VMP	**Number of obs.**	**Sample mean**	**Sample variance**	**Std. dev.**	**Lower quartile**	**Median**	**Upper quartile**
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
VMP0							
VMP1							
VMP2							
VMP3							

In the section "Key summary statistics for VMP's" in the notebook, you can find code to fill in the empty cells in the table. Se also more tips in the bottom of the notebook.

e) Fill in the empty cells in the table above by computing the relevant summary statistics for each of the four VMPs. Which additional information may be gained from the table, compared to the box plot?

Statistical analysis

The purpose of the second part of the project is to perform a simple statistical analysis of the annual nitrate emission and the effect of the VMPs. This includes specifying statistical models for annual emission, estimating the parameters of these models, performing hypothesis tests, and computing confidence intervals.

Confidence intervals and hypothesis tests

In the section "QQ-plot for model validation" in the notebook, you can find code to make a qq-plot. This plot can be used to investigate whether the annual nitrate loads during VMP0 may be assumed to be normal distributed.

f) Specify separate statistical models describing the annual nitrate emission during each of the four VMPs (see Remark 3.2). Estimate the parameters of the models (mean and standard deviation). Perform model validation (see Chapter 3 and Section 3.1.8). Since, in this case, confidence intervals and hypothesis tests involve the distribution of an average, it might also be useful to include the central limit theorem (Theorem 3.14) in the discussion.

In practice, situations will arise where it is *not* appropriate to assume that the assumptions of a model are satisfied. In these cases, one often considers whether a transformation of the data might improve the situation (see Section 3.1.9). Note that after a transformation, the interpretation of the results on the original scale changes. In this specific project, however, the intention is *not* for you to transform the data.

g) State the formula for a 95% confidence interval (CI) for the mean annual nitrate emission during VMP0 (see Section 3.1.2). Insert values and calculate the interval. Compute corresponding intervals for the three other VMPs and fill in the table below.

	Lower limit of CI	**Upper limit of CI**
VMP0		
VMP1		
VMP2		
VMP3		

Compare the CI for VMP0 computed above with the result of the Python code in the section "One-sample t-test".
Prior to the implementation of the VMPs the annual nitrate emission to Skive fjord was estimated to be approximately 2000 tonnes/year.

h) Perform a hypothesis test with the purpose of investigating whether the mean annual nitrate emission during VMP0 is significantly different than 2000 tonnes/year. This can be done by testing the following hypothesis:

$$H_0 : \mu_{\text{VMP0}} = 2000,$$
$$H_1 : \mu_{\text{VMP0}} \neq 2000.$$

Specify the significance level α, the formula for the test statistic, as well as the distribution of the test statistic (remember to include the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words. In particular, comment on whether it was necessary to perform the hypothesis test or whether the same conclusion could have been reached using the confidence interval for VMP0.

Compare the results of the test with the results of the Python code in the section "One-sample t-test".
Now we would like to investigate whether the VMPs have had a significant effect on the nitrate emission to Skive fjord.

i) Perform a hypothesis test in order to investigate whether the mean annual nitrate emission differs between VMP0 and VMP3. Specify the hypothesis as well as the significance level α, the formula for the test statistic, and the distribution of the test statistic (remember the degrees of freedom). Insert relevant values and compute the test statistic and p-value. Write a conclusion in words. Do the annual nitrate emissions differ significantly between the two VMPs? If so, in which case are the emissions lower? Has VMP3 worked according to plan?

Compare the results from the hypothesis test with the results of the Python code in the section "Welch t-test".

j) Comment on whether it was necessary to carry out the statistical test in the previous question, or if the same conclusion could have been drawn from the confidence intervals alone? (See Remark 3.59).

Correlation

In relation to the subsequent development of models for the description of the level of phytoplankton, e.g. in Project 2, we will focus on the correlation between two variables of interest.

k) State the formula for computing the correlation between the nitrate (*Nload*) and phosphorus (*Pload*) loads. Insert values and determine the correlation (note, insert only numbers in the correlation formula, i.e. three numbers). Make a scatter plot of one variable against the other. Assess whether the relation between the plot and the correlation is as you would expect.

Compare the correlation computed above to the result of the Python code in the section "Correlation".