# 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 the statistical software R. Some R 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 in R while working on the project. For example, you could add suitable titles to the plots, or use R's 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. R code should not be included in the report itself but must be handed in as an appendix (a .R-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 R 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 R

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 RStudio, 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 semi-colon).

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 file `skivefjord1_english.R`, which contains some R code that can be used for the analysis. First, the "working directory" must be set to the directory on the computer, which contains the files for the project:

```
## In RStudio the working directory is easily set via the menu
## "Session -> Set Working Directory -> To Source File Location"
## Note: In R only "/" is used for separating in paths
## (i.e. no backslash).
setwd("Replace with path to directory containing project files.")
```

Now the data may be read into R using the following code:

```
## Read data from skivefjord1_data.csv
D <- read.table("skivefjord1_data.csv", header=TRUE, sep=";",
                as.is=TRUE)
```

`D` becomes a "data.frame" (a kind of table), which contains the data that was read into R (see the introduction to R in Section 1.5 of the book).

# 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 following commands to get a simple overview of the data:

```
## Dimensions of D (number of rows and columns)
dim(D)
##  Column/variable names
names(D)
## The first rows/observations
head(D)
## The last rows/observations
tail(D)
## Selected summary statistics
summary(D)
## Another type of summary of the dataset
str(D)
```

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?

The following code may be used to generate a "density histogram" describing the empirical density of the annual nitrate emissions (see Section 1.6.1):

```
## Histogram describing the empirical density of the annual nitrate
## loads (histogram of annual emissions normalized to have an area of 1)
hist(D$Nload, xlab="Nitrate load", prob=TRUE)
```

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 R using, e.g., the `subset` function. See the remark on p. 10 as well.

Use the following R code to separate the data into four parts according to the applicable VMP:

```r
## Subset with VMP0 observations (before the first VMP was implemented)
VMP0 <- subset(D, vmp == 0)
## Check that it is a data.frame with the observations from VMP 0
VMP0
## Subset with VMP1 observations
VMP1 <- subset(D, vmp == 1)
## Subset with VMP2 observations
VMP2 <- subset(D, vmp == 2)
## Subset with VMP3 observations
VMP3 <- subset(D, vmp == 3)
```

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 following R code:

```r
## Plot of nitrate load over time (coloured according to applicable VMP)
plot(D$year, D$Nload, type="b", xlab="Year", ylab="Nitrate load")
lines(VMP0$year, VMP0$Nload, type="b", col=2)
lines(VMP1$year, VMP1$Nload, type="b", col=3)
lines(VMP2$year, VMP2$Nload, type="b", col=4)
lines(VMP3$year, VMP3$Nload, type="b", col=5)
## Add a legend
legend("topright", paste0("VMP", 0:3), lty=1, col=2:5)
```

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?

The following R code makes a box plot of the annual nitrate emission by applicable VMP:

```
## Box plot of nitrate emission by VMP
boxplot(VMP0$Nload, VMP1$Nload, VMP2$Nload, VMP3$Nload,
        names=c("VMP0", "VMP1", "VMP2", "VMP3"),
        xlab="VMP", ylab="Nitrate load")
```

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 | | | | | | | |

R code like the following may be used to fill in the empty cells in the table (see also the remark on p. 11):

```
## Total number of observations during VMP0
## (doesn't include missing values if there are any)
sum(!is.na(VMP0$Nload))
## Sample mean of annual nitrate emissions during VMP0
mean(VMP0$Nload, na.rm=TRUE)
## Sample variance of annual nitrate emissions during VMP0
var(VMP0$Nload, na.rm=TRUE)
## etc.
```

```
##
## The argument 'na.rm=TRUE' ensures that the statistic is
## computed even in cases where there are missing values.
```

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

The following R code may be used 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:

```
## qq-plot for annual nitrate emission during VMP0
qqnorm(VMP0$Nload)
qqline(VMP0$Nload)
```

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 following R code:

```
## CI for the mean annual nitrate emission during VMP0
t.test(VMP0$Nload, conf.level=0.95)$conf.int
```

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 $\alpha$, 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 following R-code:

```
##  Testing hypothesis mu=2000 for annual nitrate emission during VMP0
t.test(VMP0$Nload, mu=2000)
```

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 $\alpha$, 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 following R code:

```r
## Comparison of annual nitrate emission during VMP0 and VMP3
t.test(VMP0$Nload, VMP3$Nload)
```

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 following R code:

```r
## Computing the correlation between nitrate and phosphorus emissions
cor(D[, c("Nload","Pload")], use="pairwise.complete.obs")
```

### |||| Remark 2.1 Extra R tips

This is an optional extra remark about different ways to take subsets in R (useful but not necessary for solving the project):

```
## Optional extra remark about taking subsets in R
##
## A logical vector with a TRUE or FALSE for each row in D,
## indicating the observations from VMP0 as TRUE
D$vmp == 0
## Can be used to take a subset of the data with VMP0 observations
D[D$vmp == 0, ]
## Yields the same result as
subset(D, vmp == 0)
## May be used in a 'for'-loop to plot data from each VMP separately
plot(D$year, D$Nload, type="n")
for(i in 0:3){
  lines(D$year[D$vmp == i], D$Nload[D$vmp == i], type="b", col=i+2)
}
## More complex logical expressions can be made, e.g.:
## Find all observations recorded during VMP0, but after 1984
D[D$year > 1984 & D$vmp == 0, ]
```

### Remark 2.2 Extra R tips

Optional remark with some extra R tips. The table can also be generated more effectively using a ′for′-loop:

```r
## Use a 'for'-loop to calculate the summary statistics for each VMP
## and assign the result to a new data.frame
Tbl <- data.frame()
for(i in 0:3){
  Tbl[i+1,"mean"] <- mean(D$Nload[D$vmp == i])
  Tbl[i+1,"var"] <- var(D$Nload[D$vmp == i])
}
## See what Tbl contains
Tbl

## In R there are even more condensed ways to perform such
## calculations, e.g.:
aggregate(D$Nload, by=list(D$vmp), function(x){ c(mean(x), var(x)) })
## See more useful functions with: ?apply, ?aggregate and ?lapply
## For extremely efficient data handling see, e.g., the packages:
## dplyr, tidyr, reshape2 and ggplot2

## LaTeX tips:
##
## The R package "xtable" can generate LaTeX tables written to a file
## and thereby they can automatically be included in a .tex document.
##
## The R package "knitr" can be used very elegantly to generate .tex
## documents with R code written directly in the document. This
## document and the book were generated using knitr.
```