Statistics Project 1 - Water Environment in Skive Fjord

Axel Månson Lokrantz

2023-09-05

A) Description of the dataset

The dataset includes 25 annual observations for the variables 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)). The time period reaches from 1982 to 2006, there are missing Pload values for the years 1982, 1983 and 2004 up until 2006. The categorized variables are year and VMP which divide the observations into categories or groups. The quantitative variables are Nload and Pload which includes decimal values.

B) Density histogram

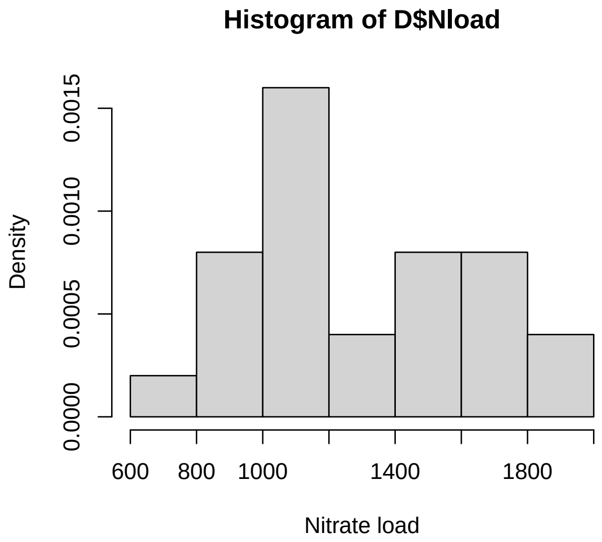

Figure 1: Histogram describing the empirical density of the annual nitrate.

The spread of the histogram refers to how widely the values are distributed. A histogram with a wide spread of values has higher variation. Conversely, a narrow histogram indicates lower variation as the data points concentrate within a smaller range. Through the mean of Nload which is 1272, we can conclude that the histogram is slightly skewed to the left since the probability mass is not symmetrically distributed around the median. The dataset contains no negative values of Nload, therefor there can not be any negative emissions.

C) Annual nitrate over time

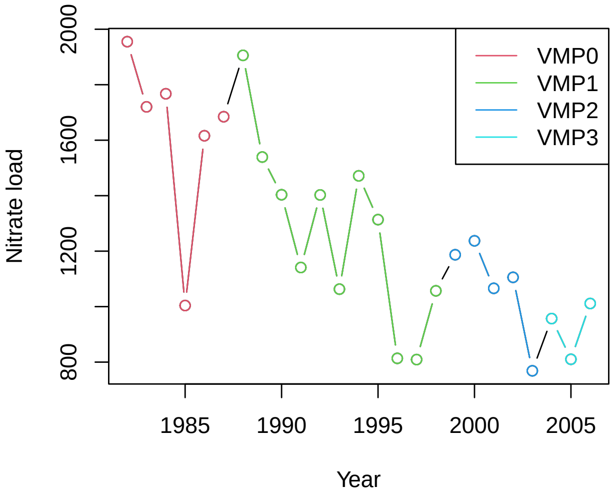

Figure 2: Nitrate load over time (coloured according to applicable VMP).

According to Figure 2, the trend in nitrate emissions has been decreasing over time since the introduction of VMP 1-3. For VMP1, which initially had a nitrate load of approximately 1900, it eventually decreased to around 1000. The same pattern is observed for VMP2. However, in the case of VMP3, the initial nitrate load is slightly lower than the last measurement. There are at least two instances where nitrate emissions differ significantly. In the case of VMP0, there is a substantial decrease in nitrate load between the years 1984 and 1985. Similarly, there is a noticeable dip in nitrate load around the year 1995 to 1996.

D) VMP nitrate load

The median in a box plot is represented by the black line inside the box. If the median is close to the center of the box the distribution is approximately symmetrical. For VMP0, the median is slightly skewed to the right. VMP0 also has an outlier, which is indicated by the white dot, this could be seen in figure 2 where there was a significant decrease in nitrate load 1985. VMP1 has the widest box which indicates that it has the highest spread of the middle 50% of the data. The right whisker is also slightly longer than the left one which indicates that the data is skewed to the right. VMP2 has an outlier, this observation can be seen between the years 1995 and 1996 in figure 2. VMP3 has the most narrow box which suggests a smaller spread of the data.

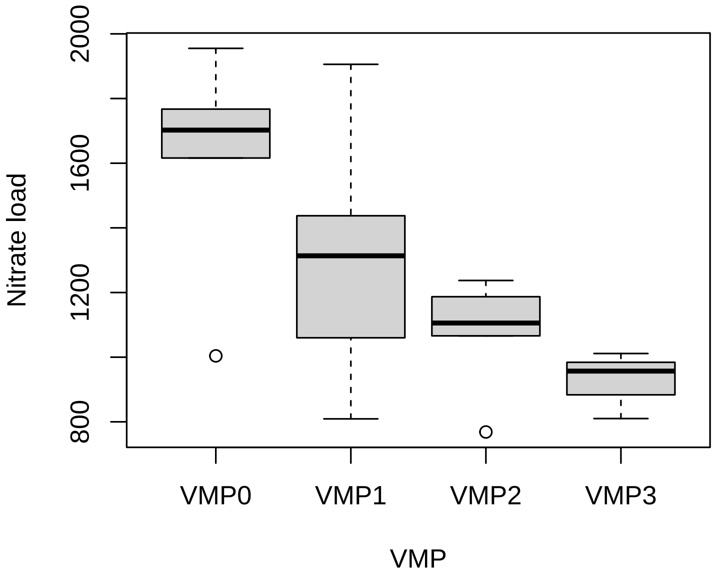

Figure 3: Nitrate load over time (coloured according to applicable VMP).

E) Summary statistics of VMP

VMP	Number of obs.	Sample Mean	Sample Variance
	n	$\hat{x}$	s^2
0	6	1624.55	105522.2
1	11	1265.44	107970.2
2	5	1072.9	33428.26
3	3	926.16	10828.93

Table 1: Table 1

VMP	Std. dev.	Lower quartile	Median	Upper quartiles
	s	Q1	Q2	Q3
0	324.84	1633.15	1702.37	1755.55
1	328.59	1059.99	1313.57	1437.41
2	182.83	1066	1105.62	1186.98
3	104.06	883.56	956.94	984.15

Table 2: Table 2

Unlike the box plot, the two tables provide information about the number of observations in each category. Furthermore, the tables offer more precise values for the standard deviation, variance, and the sample mean, whereas the box plot only displays the median.

F) VMP QQ-plots

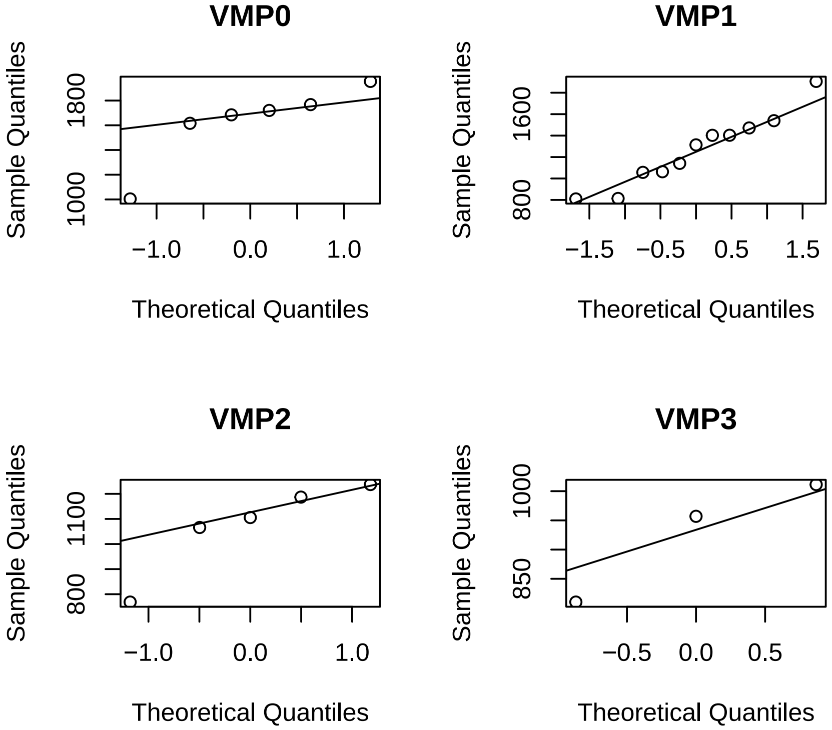

Figure 4: QQ-plots of VMP 0-4.

A QQ-plot is a graphic representation used to assess whether a data set follows a particular distribution, such as normal distribution. Figure 4 displays all the VMPs and their data points. The plots compares the quantiles of the data to the quantiles of theoretical normal distribution which is represented as a straight line. As seen in the diagram all of the VMPs closely follow the line, which suggest that they are normally distributed. The central limit theorem state that the distribution of the sample mean of a large number of independent random variables will approach a normal distribution. However, in this case the data set contains a relatively small number of observations for all of the VMPs. Therefor it is difficult to detect anomalies in the data. To reliably ensure that the sample mean follows normal distribution more data points are needed. How many points or observations that are needed depend on the original distribution and what the desired level of confidence is

G) 95% Confidence Interval

A 95% confidence interval means that if you were to take a lot of random samples from the same population and calculate a confidence interval from each sample, you would expect 95% of those intervals to contain the true population parameter. In other words, you can be almost certain that the parameters fall within the calculated interval.

	Lower Limit of CI	Upper Limit of CI
VMP0	1283.65	1965.45
VMP1	1044.69	1486.19
VMP2	845.88	1299.91
VMP3	667.65	1184.66

Table 3: 95% Confidence Interval for the VMPs

As seen in table 3, the interval width decrease for all the VMPs compared to VMP0 which suggest that the estimates are becoming more precise. A smaller numerical value in the interval indicate greater certainty, while a wide confidence interval, indicate a higher level of uncertainty. The formula for a 95% confidence interval is expressed:

$$\text{CI} = \bar{x} \pm Z \left(\frac{s}{\sqrt{n}} \right)$$

where you take the mean plus minus the desired confidence level times the sample standard deviation divided by the square root of the sample size.

H) Hypothesis Test

$$H_0 : \mu VMP0 = 2000,$$
$$H_1 : \mu VMP0 \neq 2000.$$

With a significance level of 5%, and 5 degrees of freedom $(n - 1)$ we get a p-value of 0.03663 which is smaller than our significance level. When consulting a table for p-values, the value falls into the interval of $0.001 < p < 0.01$, suggesting that we have very strong evidence against the H_0 hypothesis. From the previous question, when calculating the critical values of VMP_0, we can observe that a mean value of 2000 is not part of the confidence interval, which further strengthens our evidence against the H_0 hypothesis. Since both tests give us the same answer, it is not necessary to perform them both, but doing so can potentially further strengthen our evidence.

I) Hypothesis Test of VMP0 and VMP3

We performe a Welch t-test to investigate if the annual nitrate emission differs between VMP0 and VMP3.

T-value: The t-value is 4.7969 which represent the test statistic and quantifies how many standard errors the sample mean of VMP0 is away from VMP3. A higher t-value indicate a larger difference.

Degrees of Freedom: The degrees of freedom for the t-test is approximately 6.5711.

P-value: The p-value is 0.002349 and represents the probability of observing a t-value as extreme as the one calculated if the true difference in means between VMP3 and VMP0 were zero. Since 0.002349 is less than the typical significance level of 5% it indicates strong evidence against the H0 hypothesis.

Alternative Hypothesis: The alternative hypothesis which states that the true difference in mean is not equal to zero.

Confidence Interval: The confidence interval for the difference in which the mean is likely to fall within with a 95% confidence (349.5108, 1047.2628). The interval does not include the 0 which is strong evidence against the H0 Welch hypothesis.

Through these observations we can conclude that the annual nitrate emissions differs significantly between VMP0 and VMP3 and that VMP3 has worked according to plan in terms of reducing nitrate emissions.

J) Hypothesis test conclusion

While confidence intervals provide a range of plausible values for the population means, it is also possible, as stated in remark 3.59, to determine whether two groups are significantly different. When interpreting two independent samples, if their confidence intervals do not overlap, we can conclude that the two groups are significantly different. If the intervals do overlap, we cannot draw any conclusion. However, referring to table 3, we observe that VMP0 has an interval of (1283.65, 1965.45) and VMP3 has an interval of (667.65, 1184.66). Since these two intervals do not overlap, we can conclude that the two groups are significantly different. While the Welch t-test provides even stronger evidence against the null hypothesis (H0), it was not necessary in this case to determine that the two data sets are significantly different.

K) Correlation between Nload & Pload

The formula used to calculate the correlation coefficients in R using the function "cor" uses Pearson correlation coefficient as default and can be seen below.

$$r = \frac{n\sum(XY) - \sum X \sum Y}{\sqrt{[n\sum(X^2) - (\sum X)^2][n\sum(Y^2) - (\sum Y)^2]}}$$

From the correlation function in R we get the following output:

	Nload	**Pload**
Nload	1.0000000	0.6244952
Pload	0.6244952	1.0000000

Table 4: Correlation Matrix

The correlation coefficient can take values from -1 to 1. A positive value (0.6245 in this case) indicates a positive linnear relationship between the variables. If nitrate emission increases so will phosphorus emissions according to our data.

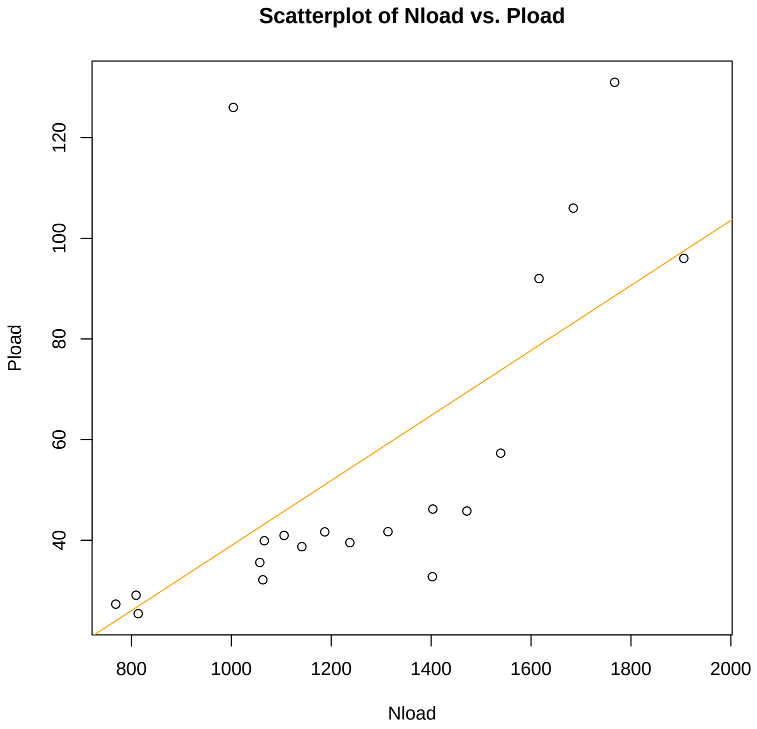

Figure 5: Scatter plot of correlation between nitrate emission and phosphorus emissions

Figure 5 displays a scatter plot of the correlation of the two variables. The upward pointing orange line represent the best fit linear relationship between the two variables. The level of phytoplankton depend largely on the emission of nitrate and phosphorus. High levels of emission will increase the phytoplankton population which in turn has a negative impact on the ecosystem and causes deoxygenation of the water. It is important to understand that a scatter plot does not necessarily imply strong relationship between variables. Sometimes, patterns in data may be due to random chance, a few points of data can give the illusion of correlation when there is no real relationship at all. There could also be a third variable that influence both the variables which we are examining, this can create the appearance of correlation when, in reality, the two variables are not directly related.