Projekt 1: Vandmiljø in Skive fjord

Formaliteter, opgavestruktur og forventninger til 1. obligatoriske opgave

Opgaven består af to dele. I første del skal der laves en deskriptiv analyse af data. Anden del handler primært om konfidensintervaller og hypotesetests.

Der er lagt op til, at man skal arbejde med opgaven i små "lette" trin. Opgaven skal i praksis løses ved hjælp af programmet R. Der er udarbejdet R-kode, som gør det nemt at komme i gang med projektet. Koden er dog ikke fuldstændig, og I opfordres til at udforske R samtidig med at I laver projektet. F.eks. kan I arbejde med at lave "pæne" titler til graferne eller benytte R's indbyggede funktioner til beregning af konfidensintervaller og test af hypoteser.

Besvarelsen skal dokumentere den gennemførte analyse ved tabeller, grafer, matematisk notation, samt tekst der beskriver analysens resultater. Relevante grafer og tabeller skal indgå i sammenhæng med teksten - ikke som bilag. Præsenter resultaterne fra jeres analyser på samme måde, som I ville videreformidle dem til andre fagfæller.

Inddel besvarelsen i et underafsnit for hver af de stillede spørgsmål.

Besvarelsen skal afleveres som pdf-fil. R-kode bør ikke indgå i besvarelsen, men vedlægges som bilag (i form af en .R-fil). Besvarelsen samt bilag afleveres under Opgaver på Learn ved: `Projekt 1:Vandmiljø in Skive fjord`

En samlet besvarelse bør ikke overstige 6 sider (ekskl. plots, tabeller og bilag). En normal side udgør 2400 anslag.

Grafer og tabeller kan IKKE stå alene - det er altså vigtigt, at I beskriver og fortolker

outputtet fra R med ord.

Grafer og tabeller indgår ikke i opgørelsen af besvarelsens længde. Det er dog IKKE i sig selv en fordel at medtage mange plots, hvis de ikke er relevante!

I må gerne arbejde sammen i grupper, men besvarelsen af opgaven skal skrives individuelt. Spørgsmål omkring projektet kan rettes til hjælpelæren, se retningslinjerne på siden `Projects` på kursets hjemmeside.

Problemstilling

I dette projekt skal målinger taget i Skive fjord i perioden fra 1982 til 2006 analyseres. I denne periode blev forskellige lovgivninger indført for at nedbringe den skadelige udledning af nitrat og fosfor til vandmiljøet. De indførte love hedder Vandmiljøplaner (VMP'er)[1]. Formålet med opgaven er at undersøge, om VMP'erne har ført til en signifikant reduktion i udledningen af nitrat.

Det kan altid diskuteres, hvad et godt vandmiljø er. De fleste kan dog blive enige om, at det ikke er særlig godt hvis vandet er grumset og lugter grimt. Hovedårsagen til at vand i åer, søer og fjorde bliver grønligt og ulækkert er for høje niveauer af phytoplankton, som også kan føre til iltsvind. Dette kan have fatale følger for dyrelivet i vandet. Mængden af phytoplankton afhænger positivt af niveauet af nitrat (mere nitrat giver flere phytoplankton), så det vil være godt for vandmiljøet at nedbringe niveauet af nitrat. Fosfor er et andet nødvendigt næringsstof for phytoplankton.

Den største nitratudledning i Danmark kommer fra gødning brugt i landbruget. Denne udledning forsøges reguleret med lovgivningen. Der er implementeret tre VMP'er i Danmark:

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

I hver VMP blev grænserne for det tilladte niveau af gødning reduceret.

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

Indlæsning af data

Lav en mappe til projektet på din computer. Download materialet til projektet fra Learn og udpak ("unzip") det til den mappe, du lige har lavet.

Åbn derefter data-filen `skivefjord1_data.csv` (f.eks. i RStudio, File → Open File) for at se filens indhold. Bemærk at første linje indeholder variabelnavne (kaldes en *header*), og at de efterfølgende linjer indeholder de egentlige observationer. Observationerne af de enkelte variable er adskilt af et ';' (deraf `.csv`: "comma separated values", her dog semikolon).

Observationerne i datasættet består af årlige værdier for nitrat- og fosforudledningen til Skive fjord. I data-filen er der følgende søjler/variable:

Variabel	**Forklaring**
`year`	Observationsåret
`vmp`	Den gældende vandmiljøplan (0, 1, 2 el. 3)
`Nload`	Årets udledning af nitrat i tons (t)
`Pload`	Årets udledning af fosfor i tons (t)

Åbn filen `skivefjord1_dansk.R`, som indeholder R-kode der kan bruges til analysen. Først skal "working directory" sættes til den mappe på computeren, hvor filerne til projektet er gemt:

```
## I RStudio kan man nemt sætte working directory med menuen
## "Session -> Set Working Directory -> To Source File Location"
## Bemærk: i R bruges kun "/" til separering i stier
## (altså ingen backslash).
setwd("Erstat her med stien til den mappe, hvor projektfilerne er gemt.")
```

Nu kan datasættet indlæses i R med følgende kode:

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

Bemærk, at der i R-koden bruges engelsk. Det er generelt ikke en god ide at bruge æøå osv. ved programmering. `D` bliver en "data.frame" (en slags tabel), som indeholder den indlæste data (se R-introen i kapitel 1.5 i bogen).

Beskrivende analyse (descriptive analysis)

Første del af projektet går ud på at lave en beskrivende analyse af data. I en rapport er det vigtigt at præsentere og beskrive data for læseren. Dette kan f.eks. gøres ved hjælp af opsummerende størrelser/nøgletal ("summary statistics") og passende figurer.

En simpel opsummering af det indlæste datasæt fås ved at køre følgende kode:

```
## Dimensionen af D (antallet af rækker og søjler)
dim(D)
## Søjle-/variabelnavne
names(D)
## De første rækker/observationer
head(D)
## De sidste rækker/observationer
tail(D)
## Udvalgte opsummerende størrelser
summary(D)
## En anden type opsummering af datasættet
str(D)
```

a) Lav en kort beskrivelse af datamaterialet: Hvilke variable indgår i datasættet? Er der tale om *kvantitative* og/eller *kategoriserede* variable? (Kategoriserede variable dukker først op i kapitel 8, men det er bare variable, som inddeler observationerne i kategorier - f.eks. tre kategorier: lav, mellem og høj). Hvor mange observationer er der? Hvilken periode dækker observationerne over (hvornår er første hhv. sidste observation foretaget)? Er der manglende værdier for nogen af variablene?

Et "density histogram" der beskriver den empiriske tæthed af de årlige målinger af nitratudledningen (se kapitel 1.6.1) kan laves ved hjælp af følgende kode:

```
## Histogram der beskriver den empiriske tæthed for de
## årlige målinger af nitratudledningen (histogram for
## de årlige målinger normaliseret så arealet er lig 1)
hist(D$Nload, xlab="Nitratudledning", prob=TRUE)
```

b) Lav et density histogram for de årlige observationer af nitratudledningen. Beskriv observationernes fordelingen ud fra dette histogram. Er den empiriske tæthed symmetrisk eller skæv? Kan nitratudledningen være negativ? Er der stor spredning i observationerne?

Bemærk: I en *skæv* fordeling er sandsynlighedsmassen ikke symmetrisk fordelt omkring medianen. For en venstreskæv fordeling gælder der, at den længste hale ligger til venstre for midten (almindeligvis vil gennemsnittet også ligge til venstre for medianen). Tilsvarende gælder der, at for en højreskæv fordeling ligger den længste hale til højre for midten (almindeligvis med gennemsnit til højre for medianen).

En meget anvendelig operation i dataanalyse er at opdele data i delmængder. Funktionen `subset` kan bruges til at udtage en delmængde, se også bemærkningen på side 10.

Benyt den følgende R-kode til at lave fire deldatasæt, hvor data opdeles efter den gældende VMP:

```
## Delmængde for VMP0 (dvs. før første VMP)
VMP0 <- subset(D, vmp == 0)
## Check at det er en data.frame med alle værdier fra VMP 0
VMP0
## En delmængde for VMP1
VMP1 <- subset(D, vmp == 1)
## En delmængde for VMP2
VMP2 <- subset(D, vmp == 2)
## En delmængde for VMP3
VMP3 <- subset(D, vmp == 3)
```

Ved observationer foretaget regelmæssigt over tid omtales data ofte som en *tidsrække*. Data for nitratudledningen udgør således en tidsrække. For tidsrækker er det ofte relevant at lave grafer, der viser udviklingen over tid. Et plot der viser udviklingen i nitratudledning over tid (farvet efter den gældende VMP) kan laves med følgende R-kode:

```
## Plot af nitratudledning over tid (farvet efter gældende VMP)
plot(D$year, D$Nload, type="b", xlab="År", ylab="Nitratudledning")
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)
## Tilføj en legend
legend("topright", paste0("VMP", 0:3), lty=1, col=2:5)
```

c) Lav et plot der illustrerer nitratudledning over tid (farvet efter den gældende VMP). Beskriv derefter udviklingen i nitratudledningen henover årene i ord. Ser

det ud til at udledningen stiger eller falder? Er udviklingen forskellig under de forskellige VMP'er? Er der nogen år, hvor nitratudledningen er bemærkelsesværdigt anderledes?

Følgende R-kode laver et boxplot for nitratudledningen opdelt efter VMP:

```
## Boxplot af nitratudledning opdelt efter VMP
boxplot(VMP0$Nload, VMP1$Nload, VMP2$Nload, VMP3$Nload,
        names=c("VMP0", "VMP1", "VMP2", "VMP3"),
        xlab="VMP", ylab="Nitratudledning")
```

d) Lav et boxplot for nitratudledningen opdelt efter VMP. Benyt derefter plottet til at beskrive den observerede fordeling af nitratudledningen under hver af de fire VMP'er. Er fordelingerne symmetriske eller skæve? Ser det umiddelbart ud til, at der er forskelle mellem fordelingerne (hvis ja, hvilke)? Er der ekstreme observationer/outliers?

Man kan også beskrive den empiriske fordeling af den årlige nitratudledning under hver VMP ved hjælp af opsummerende størrelser/nøgletal som i følgende tabel:

VMP	**Antal obs.**	**Stikprøve-gennem-snit**	**Stikprøve-varians**	**Stikprøve standard-afvigelse**	**Nedre kvartil**	**Median**	**Øvre kvartil**
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
VMP0							
VMP1							
VMP2							
VMP3							

For at udfylde de tomme celler i tabellen kan man f.eks. benytte R-kode som følgende (se også bemærkning på side 11 for tricks til udregningerne):

```
## Antal observationer af nitratudledningen under VMP0
## (medregner ej eventuelle manglende værdier)
sum(!is.na(VMP0$Nload))
## Stikprøvegennemsnit for nitratudledningen under VMP0
mean(VMP0$Nload, na.rm=TRUE)
## Stikprøvevarians for nitratudledningen under VMP0
var(VMP0$Nload, na.rm=TRUE)
## osv.
## Argumentet 'na.rm=TRUE' sørger for at størrelsen
## udregnes selvom der eventuelt er manglende værdier
```

e) Udfyld tabellen ovenfor med de opsummerende størrelser for de fire VMP'er. Beskriv hvilken ekstra information kan udledes fra tabellen sammenlignet med boxplottet?

Statistisk analyse

Andel del af projektet går ud på at lave en simpel statistisk analyse vedrørende den årlige nitratudledning og effekten af VMP'erne. Der skal opstilles statistiske modeller for nitratudledningen. Modellernes parametre skal estimeres, og der skal udføres hypotesetests og beregnes konfidensintervaller.

Konfidensintervaller og hypotesetests

Følgende R-kode kan benyttes til at lave et qq-plot med henblik på at vurdere, om den årlige udledning af nitrat under VMP0 (altså før man lavede vandmiljøplaner) kan antages at være normalfordelt:

```
## qq-plot for nitratudledning under VMP0
qqnorm(VMP0$Nload)
qqline(VMP0$Nload)
```

f) Opskriv separate statistiske modeller for nitratudledningen under hver af de fire VMP'er (se bemærkning 3.2). Estimer parametrene i de fire modeller (middelværdi og standardafvigelse). Foretag modelkontrol af de antagede forudsætninger (se kapitel 3 samt afsnit 3.1.8 i bogen). Idet konfidensintervaller og hypotesetests her involverer fordelingen af gennemsnit kan det være nyttigt også at inddrage den centrale grænseværdisætning (sætning 3.14) i argumentationen.

I praksis vil der opstå situationer, hvor man på baggrund af f.eks. modelkontrollen *ikke* kan tillade sig at antage, at en models forudsætninger er opfyldte. Da vil man ofte overveje, om det kunne hjælpe at foretage en transformation af data (se kapitel 3.1.9 i bogen). Bemærk at efter en transformation ændres fortolkningen af resultaterne på den oprindelige skala. Det er *ikke* meningen, at I skal lave en transformation af data i dette projekt.

g) Angiv formlen for et 95% konfidensinterval (KI) for middelværdien af den årlige nitratudledning under VMP0 (se sektion 3.1.2 i bogen). Indsæt tal og beregn intervallet. Beregn tilsvarende konfidensintervaller for middelværdien af nitratudledningen under de tre andre VMP'er og udfyld tabellen nedenfor.

	Nedre grænse af KI	**Øvre grænse af KI**
VMP0		
VMP1		
VMP2		
VMP3		

Sammenlign det beregnede konfidensinterval for VMP0 med resultatet af følgende R-kode:

```
## Konfidensinterval for middelværdi af årlig nitratudledning under VMP0
t.test(VMP0$Nload, conf.level=0.95)$conf.int
```

I forarbejdet til VMP'erne blev det vurderet, at nitratudledningen til Skive fjord var i omegnen af 2000 ton/år.

h) Udfør et hypotesetest med henblik på at undersøge, om middelværdien af nitratudledningen til Skive fjord før VMP'erne (dvs. under VMP0) afviger signifikant fra 2000 ton/år. Dette kan gøres ved at teste følgende hypotese:

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

Angiv signifikansniveauet α, formlen for teststørrelsen samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord. Kommenter på om det var nødvendigt at udføre det statistiske test, eller om samme konklusion kunne opnås ved konfidensintervallet alene.

Sammenlign resultaterne for test af hypotesen med resultaterne af følgende R-kode:

```
## Test af hypotesen mu=2000 for årlig nitratudledning under VMP0
t.test(VMP0$Nload, mu=2000)
```

Vi ønsker nu også at undersøge, om VMP'erne har haft en signifikant effekt på nitratudledningen i Skive fjord.

i) Undersøg ved et hypotesetest, om der kan påvises en forskel på middelværdien af den årlige nitratudledning under VMP0 og VMP3. Opskriv hypotesen og angiv signifikansniveauet α, formlen for teststørrelsen samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord. Kan der konstateres en signifikant forskel i nitratudledning mellem VMP0 og VMP3? Hvis ja, hvor er udledningen mindst? Har VMP3 virket efter hensigten?

Sammenlign resultaterne for test af hypotesen med resultaterne af følgende R-kode:

```
## Sammenligning af årlig nitratudledning under VMP0 og VMP3
t.test(VMP0$Nload, VMP3$Nload)
```

j) Kommenter om det var nødvendigt at udføre hypotesetestet i forrige spørgsmål, eller om samme konklusion kunne opnås ud fra konfidensintervallerne alene? (Se bemærkning 3.59 i bogen).

Korrelation

I forbindelse med efterfølgende opbygning af modeller til beskrivelse af mængden af phytoplankton, f.eks. i forbindelse med projekt 2, vil vi sætte fokus på yderligere sammenhænge mellem udvalgte variable.

k) Angiv formlen for beregning af korrelationen mellem nitratudledningen (Nload) og fosforudledningen (Pload). Indsæt tal og beregn korrelationen (indsæt kun i korrelationsformlen, dvs. sæt kun tre tal ind!). Lav desuden et scatterplot der illustrerer sammenhængen mellem de to variable. Vurder om sammenhængen mellem plottet og korrelationen er som forventet.

Sammenlign den beregnede korrelation med resultatet fra følgende R-kode:

```
## Beregning af korrelation mellem Nload og Pload
cor(D[, c("Nload","Pload")], use="pairwise.complete.obs")
```

Remark 2.1 Ekstra tips til R

Dette er en valgfri ekstra bemærkning om R-kodning (ikke nødvendig for at løse opgaven). Der er mange måder hvorpå man kan udtage en delmængde i R.

```
## Ekstra bemærkning om måder at udtage delmænger i R
##
## En logisk (logical) vektor med sandt (TRUE) eller falsk (FALSE)
## som for hver række i D undersøger om observationen er fra VMP0
D$vmp == 0
## Vektoren kan bruges til at udvælge alle observationer fra VMP0
D[D$vmp == 0, ]
## En vektor af denne type kunne også bruges til at plotte data
## fra hver VMP i en for-løkke
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)
}
## Mere komplekse logiske udtryk kan laves, f.eks.:
## Find alle observationer fra efter 1984 og fra VMP 0
D[D$year > 1984 & D$vmp == 0, ]
```

Remark 2.2 Ekstra tips til R

Endnu en bemærkning med ekstra R-tips for de interesserede. Man kan f.eks. lave tabellen mere effektivt med en for-løkke.

```
## Lav en for-løkke med beregning af et par opsummerende størrelser
## og gem resultatet i en ny 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])
}
## Se hvad der er i Tbl
Tbl

## I R er der endnu mere kortfattede måder sådanne udregninger kan
## udføres. For eksempel
aggregate(D$Nload, by=list(D$vmp), function(x){ c(mean(x), var(x)) })
## Se flere smarte funktioner med: ?apply, ?aggregate og ?lapply
## og for ekstremt effektiv databehandling se f.eks. pakkerne: dplyr,
## tidyr, reshape2 og ggplot2.

## LaTeX tips:
##
## R-pakken "xtable" kan generere LaTeX tabeller og skrive dem direkte
## ind i en fil, som derefter kan inkluderes i et .tex dokument.
##
## R-pakken "knitr" kan anvendes meget elegant til at lave et .tex
## dokument der inkluderer R koden direkte i dokumentet. Dette
## dokument og bogen er lavet med knitr.
```