Der er følgende tre grundelementer i opgaverne nedenfor:

- at lære at bruge, eller repetere brugen af R til statistiske analyser herunder håndtering af data (opg. 1.1–1.3, 1.9)
- at repetere ensidet variansanalyse og regressionsanalyse, (opg. 1.3–1.5)
- at indøve og overveje begreber om forsøg og faktorer (opg. 1.6–1.8)

Selvom forusætningerne for faget omfatter de første to punkter, kan der være brug for en repetition, og nogle kan pga. overgangsordninger og studieskift komme med andre forudsætninger. Det er derfor op til dig selv at vurdere, hvor dit største behov er for at sætte ind i denne uge, og jo mere du mangler at samle op, jo større indsats må du yde. Du bør sættte ind på at komme til at føle dig tryg ved at bruge R, men sørg også for at følge med i de nye begreber om faktorer.

Følgende anbefalinger kan være til hjælp:

- Hvis du føler dig rimelig sikker i brugen af R (f.eks. fra SD1), anbefales det at du løser opgaverne 1.3 og 1.5.
- Hvis du er ny (eller rusten) R-bruger, så kan du starte med opgaverne opg. 1.1, 1.9 og 1.2. For at komme helt med på vognen, bør du nok hjemme supplere med at kigge på opgaverne 1.3 og 1.5 inden torsdag.

Hvad angår R skal du (som minimum) kunne besvare følgende spørgsmål efter denne uge:

- Hvordan benytter jeg R som regnemaskine ?
  - sædvanlige regningsarter
  - logaritme, eksponentialfunktion, potensopløftning, kvadratrod
- Hvordan indtaster jeg selv et datasæt i R ?
- Hvordan laver jeg deskriptiv statistik på et datasæt (en vektor) ?
  - beregning af middelværdi, median, varians og spredning
  - tegning af histogram og boxplot
- Hvordan laver jeg en ensidet variansanalyse og linear regression?
  - udskrivning af parameterestimater
  - udskrivning af residualer og tegning et residualplot
  - test af hypotesen om at der ikke er forskel på grupper

Selv om det er en fordel at arbejde sammen i mindre grupper er det *vigtigt*, at du selv bruger PCen til R-programmering. Det er derfor bedst hvis I sidder med hver jeres PC i gruppen.

Bemærk endelig at selv om denne uges øvelser handler mest om brug af R, så er den statistiske analysemetode og tolkningen af resultaterne det vigtige i kurset. Allerede fra på torsdag vil dette være i fokus for alle opgaverne, mens R blot skal opfattes somet naturligt hjælpemiddel.

Opgave 1.1: Basale beregninger i R. Variabel.

Start R-programmet og udfør nedenstående. (jf. R-vejl.: afsn. 1 og s. 11)

a) Foretag nogle simple beregninger med addition, multiplikation, osv. Fx $10 + 5 * 3$. Vær opmærksom på brug af parenteser. Prøv fx at sammenligne det sjuskede $8/2 * 2$ med de mere driftssikre udtryk $8/(2 * 2)$ og $(8/2) * 2$.

b) Prøv potensopløftning 0.5^6. Hvad sker der, hvis du har negativ eksponent på et positivt hhv. negativt tal.

c) Beregn $(2 \cdot 4) - \frac{2}{7}$ og $2 \cdot (4 - \frac{2}{7})$.

d) Efter mekanisk ukrudtsbekæmpelse på en 5 kvm parcel har man registreret en biomasse af ukrudt på 41.9. Gem dette resultat i en variabel med et passende navn. Tilsvarende fås resultatet 211.5 i en ubehandlet parcel. Gem tilsvarende dette resultat.

e) Beregn logaritmen til kvotienten mellem de to resultater (bekæmpet i forhold til kontrol). Brug naturlig logaritme.

f) Beregn dernæst logaritmerne til hver af de to biomasser og gem differencen mellem dem (bekæmpet minus kontrol) som en variabel.

g) Brug R til at vise, at de to resultater fra (e) og (f) er de samme. Hvorfor er de det? Prøv det samme med titalslogaritmen. Gælder det stadig, at de to resultater er ens?

h) R har en række indbyggede funktioner som benyttes ved at skrive `fkt(..argumenter..)`, fx `y= log(x)`. Du kan få hjælp til brugen af funktionen ved at skrive `?log`. Prøv det og se i hjælpen, hvordan man beregner titals-logaritme og naturlig logaritme. Find to måder at skrive udtrykket for titals-logaritmen til x.

Opgave 1.2: Indlæsning af datasæt. Vektorer. Ensidet variansanalyse.

Et datasæt kan indlæses i R som beskrevet i R-vejledningens afsnit 2.

a) Se hvad filen `potter.txt` (link fra uge-hjemmesiden) indeholder. Hertil skal ikke bruges R.

b) Indlæs datasættet `potter.txt` fra uge-hjemmesiden til et dataframe ved navn `potter`.

c) Undersøg effekten af kommandoerne `potter`, `names(potter)`, `dim(potter)`, `Nitrogen` og `potter$Nitrogen` . Forklar hvad resultaterne betyder.

d) Benyt `attach()` på datasættet `potter` og udskriv vektorerne Nitrogen og Treat på skærmen (R-vejl. s. 9, eller BMS s. 212).

e) Anvend funktionerne **mean(), var(), median(), sd() og summary(), plot(), hist(), boxplot()** på vektoren Nitrogen og overvej, hvad der sker (R-vejl., afsn. 3).

e) Lav en vektor, Nitrogen2, der indeholder elementerne 1, 6, 11, 16, 21 og 26 fra vektoren Nitrogen. Udskriv Nitrogen2 på skærmen (R-vejl., s. 10).

f) Lav en ny vektor Nitrogen3, som indeholder de elementer fra Nitrogen, som er større end 25 (R-vejl., s. 11). Udskriv Nitrogen3 på skærmen.

g) Anvend funktionen `lm()` til at lave en ensidet variansanalyse til modellering af sammenhængen mellem Nitrogen og Treat (behandling) (R-vejl. afsn. 4.1.1). Gem resultatet som et objekt ved navn `model`.

h) Diskuter udskriften på skærmen, som fremkommer ved kommandoerne `predict(model)`, `coef(model)`, `resid(model)` og `confint(model)`.

i) Beskriv effekten af kommandoen `plot(predict(model), resid(model))`.

j) Lav et objekt, `model2`, ved at køre kommandoen `lm(Nitrogen ~ 1)`. Herved fittes en model, hvor der er samme middelværdi i alle grupper givet ved faktoren `Treat`.

k) Man kan teste hypotesen om at behandlingen ikke påvirker nitrogen-mængden ved at benytte kommandoen `anova(model2, model)`. Undersøg hvad denne kommando udskriver på skærmen.

Opgave 1.3: Udvidet ensidet variansanalyse.

I et forsøg med kartofler og med to års gødskning med fosfor var der i alt 9 forskellige gødsknings-behandlinger (som kombination over de to år). Udbyttet (i hkg pr. ha ved andet års dyrkning) for hver af de 27 parceller fremgår af følgende tabel (hvor de 27 parceller er nummereret inden for hver behandling).

	Fosfor-behandling								
Parcelnr.	1	2	3	4	5	6	7	8	9
1	330	346	409	368	371	409	360	382	398
2	320	350	363	340	373	410	396	407	415
3	355	369	414	366	403	402	406	425	433

a) Indlæs datasættet i to vektorer `beh` og `udb` ved brug af `rep(..)` og `c(..)`, jf. R-vejl. s. 5. For at spare dig for tastearbejdet findes de 27 udbyttetal i filen `fosfor-udbytter.txt` på ugeplanen for uge 1. Brug "copy-and-paste" til at få tallene overført til R. Begge vektorer skal have længde 27 og repræsentere de 27 parceller i samme rækkefølge.

b) Benyt `plot(..)` til at tegne udbyttet op mod behandlingen `beh`.

c) Tegn et boxplot over samtlige udbytter. Hvad viser det? Er det relevant at se på? (R-vejl., afsn. 3).

d) Lav en ny vektor `behfac` som er en *faktor* med de 9 behandlinger. Benyt `plot(..)` til at tegne udbyttet op mod `behfac` (R-vejl., afsn. 4.1.1).

e) Sammenlign resultaterne af delopgave b) og d). Hvad er mest velegnet her? Hvorfor?

f) Benyt `lm(..)` til at tilpasse modellen for ensidet variansanalyse af udbyttet som funktion af behandlingen. Skal du her bruge `beh` eller `behfac`?

g) Benyt både `summary(..)` og `predict(..)` til finde parameterestimaterne svarende til de ni forskellige behandlinger.

h) Benyt `anova(..)` til at undersøge om der er en påviselig effekt af behandlingen på udbyttet.

i) Indlæs fra ugeplanen for uge 1 datasættet `fosfor.txt`, som er samme datasæt som tidligere i opgaven, bortset fra at der her er supplerende oplysninger om hvordan de ni behadnlinger fremkommer som kombination af mængden af tilført fosfor i 1981 og 1982. Gem datasættet som en dataframe i R under et passende navn.

j) Vi ønsker nu at undersøge om udbyttet kunne tænkes kun at afhænge af fosfortilførslen i 1982 (variablen `p82`). Tilpas derfor en `model82` som modellen for ensidet variansanalyse med udbytte som funktion af fosfortilførslen i 1982. (se bort fra vektoren `blok`).

k) Skriv `anova(model82, model)`, hvor `model` er resultatet fra spørgsmål (f). Hvad konkluderer du heraf?

Opgave 1.4

Ved et potteforsøg med salat registreredes tørvægten af 16 potter efter høst. Her var 4 gødet med 0.5, 4 med 1.0, 4 med 2.0 og 4 med 3.0 gange normal mængde N-gødning. Data (dw = dry weight, N= mængde N) indlæses i de første få linier af nedenstående R-program som fortsætter med en variansanalyse af data.

```
> dw= scan()
1: 23.95 25.02 26.18 22.71 34.24 29.87 31.59 32.70
9: 39.28 34.90 32.74 31.41 36.29 40.65 43.44 29.77
17:
Read 16 items
> N= scan()
1: 0.5 0.5 0.5 0.5 1.0 1.0 1.0 1.0 2.0 2.0 2.0 2.0 3.0 3.0 3.0 3.0
17:
Read 16 items

> res1<-lm(dw ~factor(N))
> res2<-lm(dw ~1)
> anova(res2,res1)

Analysis of Variance Table

Model 1: dw ~ 1
Model 2: dw ~ factor(N)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     15 534.82
2     12 158.81  3    376.01 9.4706 0.001735 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Opskriv den statistiske model svarende til ovenstående R-kørsel.

b) Nøgletallene fra variansanalysetabellen er (afrundet) 9.47, 0.0017 og 158.81 (med df= 12). Forklar hvad disse tal står for og specielt, hvad du af de førstnævnte to tal konkluderer om betydningen af N-mængden. Kvadratroden af 158.81/12 har en speciel betydning, hvilken?

R-kørslen er fortsat med

```
> summary(res1)

Call:
lm(formula = dw ~ factor(N))

Residuals:
    Min      1Q  Median      3Q     Max
-7.7675 -1.7769 -0.0963  1.8212  5.9025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.465      1.819  13.450 1.34e-08 ***
factor(N)1     7.635      2.572   2.968  0.01174 *
factor(N)2    10.118      2.572   3.933  0.00199 **
factor(N)3    13.073      2.572   5.082  0.00027 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.638 on 12 degrees of freedom
Multiple R-squared:  0.7031,        Adjusted R-squared:  0.6288
F-statistic: 9.471 on 3 and 12 DF,  p-value: 0.001735
```

c) I modellen er estimeret en middelværdi (prædikteret værdi) for hver af de fire N-grupper. Hvordan finder du de fire værdier ud fra udskriften?

d) Hvordan udtrykker du omvendt tallet 7.635 ud fra de nævnte fire estimerede middelværdier? Hvad er derved betydningen af tallet?

e) Beregn et estimat og et konfidensinterval for forskellen på middel-tørvægt mellem 1.0 N og 0.5 N. Hertil skal du bruge at 0.975-fraktilen i t-fordelingen med 12 frihedsgrader som du kan beregne i R med `qt(0.975, df=12)`. Hvad er LSD-værdien for sammenligning af to grupper, og hvordan bruges den?

Med `lm(..)` funktionen i R kan man også bede om at få tilpasset en model uden intercept (konstantled). Det er gjort nedenfor:

```
> res1refit<-lm(dw ~factor(N)-1)
> summary(res1refit)

Call:
lm(formula = dw ~ factor(N) - 1)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.7675 -1.7769 -0.0963  1.8212  5.9025

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
factor(N)0.5   24.465      1.819   13.45 1.34e-08 ***
factor(N)1     32.100      1.819   17.65 5.98e-10 ***
factor(N)2     34.582      1.819   19.01 2.52e-10 ***
factor(N)3     37.538      1.819   20.64 9.67e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.638 on 12 degrees of freedom
Multiple R-squared:  0.9907,        Adjusted R-squared:  0.9876
F-statistic: 319.9 on 4 and 12 DF,  p-value: 4.464e-12
```

f) Sammenlign med udskriften fra før og forklar forskelle og sammenhænge mellem de to udskrifter.

Opgave 1.5

I Alexander, R.M. (1971) er vist kropsvægt og vægt af flyvemuskler for 14 fuglearter. I nedenstående R-program er indlæst $X = \ln W$ og $Y = \ln F$ hvor W er kropsvægt (g) og F er vægt af flyvemusklerne (g). Dette er gjort ved at markere og kopiere tre søjler fra et Excel-ark og dernæst i R skrive

```
> birds= read.table("clipboard", header=T)
```

fulgt af en udskrift af data med R-kommandoen

```
> birds
   species    X     Y
1        1 1.31 -0.89
2        2 2.35  0.68
3        3 3.03  1.73
4        4 3.35  1.68
5        5 3.72  2.35
6        6 4.55  3.40
7        7 5.34  3.40
8        8 5.60  3.61
9        9 6.12  4.19
10      10 6.59  5.08
11      11 7.01  5.55
12      12 7.90  5.91
13      13 8.69  6.91
14      14 9.05  7.01
```

(Data findes endvidere i tekst-format i filen `opg1-5.txt` som findes på ugeplanen for kursusuge 1.)

a) Opskriv en statistisk model for Y som funktion af X.

R-kørslen er fortsat med

```
>  attach(birds)
>  res = lm(Y ~ X)
>  summary(res)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58948 -0.22142 -0.01972  0.28246  0.54417

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.57670    0.21872  -7.209 1.07e-05 ***
X            0.97418    0.03767  25.859 6.82e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3247 on 12 degrees of freedom
Multiple R-Squared: 0.9824,     Adjusted R-squared: 0.9809
F-statistic: 668.7 on 1 and 12 DF,  p-value: 6.823e-12
```

b) Hvad er estimaterne for parametrene i modellen? Beregn også konfidensintervaller for de parametre der specificerer sammenhængen mellem X og Y. Hertil skal du bruge at 0.975-fraktilen i t-fordelingen med 12 frihedsgrader som du kan beregne i R med `qt(0.975, df=12)`.

c) Omskriv modellen til en model for sammenhængen mellem W og F. Stemmer resultatet rimeligt med en formodning om at flyvemusklerne udgør en fast procentdel af kropsvægten uanset artens størrelse? Brug eventuelt følgende R-udskrift:

```
>  qt(0.975, df=12)
[1] 2.178813
>  t= (0.97418-1)/0.03767
>  t
[1] -0.685426
>  pt(t, df=12)
[1] 0.2530504
```

En kolibri-art har kropsvægt 2.7 gram. Vi ønsker at se, om denne art passer ind i mønsteret for de øvrige og får følgende fra R:

```
>  predict(res, newdata= data.frame(X=log(2.7)), interval="prediction")
            fit       lwr       upr
[1,] -0.6090932 -1.423310 0.2051234
```

d) Vægten af flyvemusklerne for denne kolibri-art er faktisk 1.0 gram. Hvad konkluderer du heraf?

e) Hvilke(n) tegning(er) ville du ønske at se for at kontrollere antagelserne bag den benyttede model, og hvordan producerer du dem i R?

Exercise 1.6

Exercise 1.3 fra BMS (kompendiet).

Exercise 1.7

Exercise 1.8 fra BMS (kompendiet).

Exercise 1.8

Exercise 2.1, kun første to spørgsmål, fra BMS (kompendiet). I forbindelse med første spørgsmål skal du desuden angive hvilke *niveauer* der er for den enkelte faktor. Vær sikker på at du forstår forskellen på begreberne "faktor" og "niveau".

Opgave 1.9: Dataudvælgelse i dataframes.

Indlæs datasættet fra filen `fosfor.txt` fra uge-hjemmesiden som en dataframe i R. Problemstillingen er beskrevet i opgave 1.3. Vi skal i denne opgave prøve nogle typisk forekommende beregninger på datasættet. Metoderne hører til i den lidt mere avancerede ende af R-delen fra Statistisk Dataanalyse 1.

a) Få R til at optælle antallet af observationer med hver behandling (Brug `table(..)`).

b) Få R til at optælle antallet af observationer med hver kombination af fosfortilførslen i 1981 og 1982. (Brug igen `table(..)`, men denne gang skal den producere en tosidet tabel)

c) Udtræk (som en ny dataframe) den del af datasættet som havde fosfortilførsel 0 i 1982. (Brug `subset(..)`, jf. R-vejl., s. 9).

d) Beregn gennemsnit og stikprøvespredning af de 9 udbytter med fosfortilførsel 0 i 1982. Tegn også et boxplot over dem.

e) Tegn tre boxplots ved siden af hinanden i samme diagram over udbytterne opdelt efter fosfortilførslen 1982. (R-vejl., afsn. 4.1.1).

f) Undertiden har man store datasæt og vil gerne nøjes med at se nogle få af linierne. Brug `head(..)` til at udskrive de første 10 linier af datasættet.

g) Udskriv rækkenumrene på de observationer i datasættet, der havde fosfortilførsel 0 i 1982. Bemærk at du kan få dem som numrene på de elementer i vektoren `p82` som har værdien 0.

h) Udskriv rækkenumrene på de observationer i datasættet, der havde fosfortilførsel 0 i 1982 *og* i 1981.

i) Lav en ny dataframe som udelader en bestemt observation (række) fra det oprindelige datasæt. Lad os sige observation nr. 8.

Referencer

Alexander, R.M. (1971). *Size and shape. Studies in Biology no 29.* Edward Arnold.

(BMS) Bibby, B.M., Martinussen, T. & Skovgaard, I.M. (2010). *Experimental Design in the Agricultural Sciences.*

(R). Martinussen, T., Skovgaard, I. & Sørensen, H. (2007). *A note on R: Linear models with random effects.* Findes under Absalon på notatet "Undervisningsmateriale".

Hjælp til visse af spørgsmålene

Opg1.4a Bemærk at der står `factor(N)` i modelopskrivningen, hvilket betyder at N opfattes som en faktor med fire niveauer, idet N antager 4 forskellige værdier. Når en faktor indgår i modelopskrivningen i `lm(..)`, indgår den som i en variansanalyse, altså med et bidrag til middelværdien som afhænger af niveauet (gruppen).

Opg1.4b Tænk på hvilken hypotese der testes i en ensidet variansanalyse og find den tilhørende (F)-teststørrelse og P-værdi. Hvad konkluderer du generelt em en hypotese hvis P-værdien er meget lille? $s = \sqrt{MSe}$ er estimatet for spredningen σ; spredningen hører til en fordeling ("population") — hvilken?

Opg1.4c R benytter i denne kørsel en lidt usædvanlig opskrivning af den ensidede variansanalysemodel, nemlig med et konstantled, μ, så middelværdien for den k'te gruppe skrives $\mu + \alpha_k$. Det er estimaterne for μ, α_1, ..., α_4 der er udskrevet; dog er et af α'erne er udeladt idet det er sat til 0 ("benyttet som reference-niveau). Du kan have glæde af allerede her at sammenligne med udskriften senere i opgaven hvor modellen er opskrevet mere direkte (uden konstantled).

Opg1.4e LSD-værdien kan beregnes "manuelt" som vist side 27 i notesættet, men du har faktisk også beregnet det ved at beregne et konfidensinterval for forskellen mellem to gruppers middelværdi. Se ud fra formlen side 27 om det gør nogen forskel hvilke to grupper der sammenlignes.

Opg1.4f De fire estimater herfra er de fire gruppe-gennemsnit. De kan omregnes til de estimater du så tidligere.

Opg1.5a Model for lineær regressionsanalyse (idet `X` ikke er en faktor).

Opg1.5b Husk den generelle måde at opskrive konfidensintervaller for parametre i varians- og regressionsanalyse (og mere generelt i lineære modeller):

$$\text{estimat} \pm \text{t-fraktil} \cdot \text{se(est)}$$

Opg1.5c Hvad skal værdien af hældningen fra regressionsanalysemodellen være for at $F = kW$ for en eller anden konstant k. (Tag logaritmen på begge sider af denne sammenhæng.)

Opg1.5d R-kommandoen har beregnet et prædiktionsinterval for $Y = \ln F$ for en observation med $W = 2.7$. Hvad betyder et prædiktionsinterval?

Opg1.5e Det ville være dejligt at se en figur med rådata for at se om en lineær sammenhæng er rimelig, en der undersøger antagelsen om konstant spredning (residualplot) og en der undersøger antagelsen om normalfordeling (qq-plot„ eller eventuelt et histogram, over residualerne).

Opg1.9f Husk på hvordan du kan finde hjælp til en funktion i R.

Opg1.9g Brug `which(..)`, hvor prikkerne skal erstattes af betingelsen.

Opg1.9i `fosfor[-8,]`