Kursusintroduktion og repetition: 1-sidet ANOVA + lineær regression

Statistisk Dataanalyse 2

Anders Tolver

Velkommen

Kursusansvarlig/forelæser: Anders Tolver

Øvelseslærere: Adrian Fibach Balk-Møller og Aleksander Søltoft

Lokaler:

Tirsdage fra kl. 8:15-12:00: A1-09.01 (2-03), Dyrlægevej 88

Torsdage fra kl. 8:15-12:00: A1-09.01 (2-03), Dyrlægevej 88
Torsdage fra kl. 13:00-16:30: A1-01.18, Bülowsvej 17

Forelæsninger idag

- Praktiske oplysninger
- Lidt faglig intro
- Repetition af R
- Repetition af ensidet ANOVA og lineær regressionsanalyse

Praktisk info: Absalon

Vi benytter KUs kursusadministrationssystem Absalon til

- beskeder med general info om kurset
- Vigtigt: sørg for at beskeder videresendes til en e-mail-adresse, du checker jævnligt!
- I Absalon findes linket Kursusoversigt som giver diverse praktiske oplysninger samt en løbende beskrivelse af kursusaktiviteter og -materiale.

God skik: skriv direkte til mig <tolver@math.ku.dk>, hvis du opdager fejl, mangler eller uklarheder vedrørende kursusafviklingen.

I praksis kan næsten alt materiale (præsentationer, R-programmer, datasæt, opgaver, løsninger) tilgås uden om Absalon og vil kun sjældent blive uploadet direkte i Absalon.

Praktisk info: overordnet kursusplan

Ugestruktur, kursusuge 1-7

- Tirsdag
  - 8:15-10:00 Forelæsninger
  - 10:00-12:00 Regneøvelser (Adrian+Aleksander)
- Torsdag
  - 8:15-10:00 Forelæsninger
  - 10:00-12:00 Regneøvelser (Adrian+Aleksander)
  - 13:00-15:00 Selvstændigt arbejde med en "case"
  - 15:15-16:30 (ca) Opsamling på casearbejde og ugens emner

Uge 8 benyttes til repetition (intet ny pensum).

Ca. en uge før eksamen afholdes en spørgetime.

Eksamen er annonceret til 9/11-2017, men det er dit ansvar at holde dig orienteret omkring dette!

Arbejdspres (normering): 22 timer om ugen!

Maksimalt udbytte kræver 10 timers hjemmearbejde per uge.

Praktisk info: regneøvelser

Der er tilmeldt omkring 30 studerende til Statistisk Dataanalyse 2.

Der er p.t. to instruktorer (=Adrian+Aleksander).

I praksis: hvis jeg har krudt til det, vil jeg aflaste instruktorerne ved også at være tilstede i starten øvelsestimerne efter behov.

- Opgaver med og uden computer (-husk dog computer, altid!!!)
- God idé at arbejde i små grupper, men sørg for at alle i gruppen har en computer

Fortrolighed med R opnås kun ved, at man selv trykker på knapperne — ikke blot ved at se andre gøre det!

Praktisk info: afleveringsopgaver og eksamen

Fire afleveringsopgaver

- Efter uge 1, 3, 5 og 7. Afleveres senest tirsdag kl. 10 i efterfølgende uge (dvs. uge 2, 4, 6, 8)
- Frivillige, men det anbefales stærkt at lave dem:
  - god eksamenstræning
  - mulighed for feedback

Eksamen (-med brug af PC)

- Annonceret til d. 9/11-2017 (men check selv!)
- Eksamen er skriftlig med en varighed på 4 timer og alle sædvanlige hjælpemidler er tilladt.
- Det er påkrævet, at man ved eksamen medbringer en computer med R-installeret for at kunne løse alle opgaver.
- I vil under kurset komme til at se eksempler på opgaver som viser, hvordan R kunne tænkes brugt i forbindelse med eksamensopgaverne.

Specielt omkring brug af PC ved eksamen

Hvorfor brug af PC ved eksamen?

- Tanken er ikke at ændre på sværhedsgraden af eksamen.
- Tanken er at eksamenssituationen i højere grad skal afspejle den måde undervisningen alligevel foregår på!

Hvordan bliver PCen integreret ved eksamen?

- Der vil blive udleveret et datasæt på en USB-nøgle.
- Nogle eksamensspørgsmål vil kræve, at man indlæser datasættet, kører nogle (velkendte) R-kommandoer og benytter output i eksamensbesvarelsen som stadig skrives i hånden.
- I lighed med tidligere, kan man også blive bedt om at kommentere/benytte R-output, som er trykt i eksamenssættet.

Læringsmål

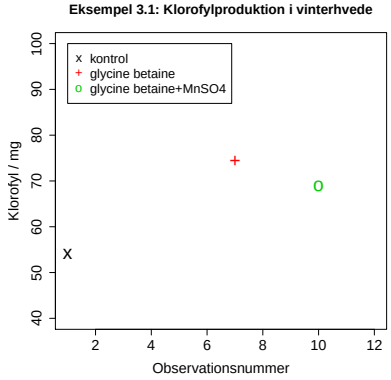
Kursusmål

- Kursets målsætning er, at de studerende skal lære at forstå og håndtere analyser af data som involverer flere forskellige variationskilder (både systematiske og tilfældige) samt gentagne målinger.

For en udtømmende opfyldelse af kursets mål skal den studerende efter endt kursusforløb være i stand til

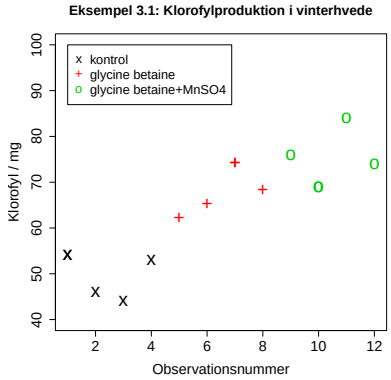
- Viden (-fire punkter fra kursusdatabasen: læs selv!)
- Færdigheder (-tre punkter fra kursusdatabasen: læs selv!)
- Kompetencer (-to punkter fra kursusdatabasen)
  - at anvende modellerne til analyse af data, herunder vælge en passende model og kontrollere modellens forudsætninger
  - at formulere videnskabelige spørgsmål som statistiske hypoteser samt besvare spørgsmålene ud fra resultaterne af de statistiske analyser

Hvad går statistik ud på?



- Virker behandlingen og hvilken er i givet fald bedst?
- Føler du dig sikker på din konklusion?

Hvad går statistik ud på?



- Der er variation mellem klorofylmålingerne - også når vi betragter forsøgsenheder, som har fået samme behandling.
- Statistik er et værktøj til matematisk beskrivelse af variation.

Emneoversigt på Statistisk Dataanlyse 2

Alle modeller på SD2 har formen:

$$\text{måling} \to Y = \text{“systematisk variation”} + \text{“tilfældig variation”}$$

Systematisk variation

- Faktorer (kategoriske): køn, behandling, dosisgruppe, tidsgruppe [Kompendium kap. 2, 3]
- Covariater (kontinuerte): alder, vægt, startmåling, dosis, tid [Kompendium kap. 4, 6]

Tilfældig variation

- Tilfældige faktorer: person (f.eks. laborant / patient), mark, blok, laboratorium [Kompendium kap. 7, 8, 10]
- Residual varians: resterende uforklaret variation (målestøj)

Øvrige emner

- Modelkontrol: validering af modellens matematiske forudsætninger [Kompendium kap. 5]
- Forsøgsplanlægning: hvordan tilrettelægges et godt forsøg? [Kompendium kap. 9]

Dagens eksempel

Koagulationstider i sekunder for blodprøver fra 15 rotter, der har fået en af tre forskellige fodertyper (A, B, C):

Foder	Koagulationstid (sek.)				
A	63	67	71	64	65
B	68	66	71	67	68
C	56	62	60	61	63

Fuldstændigt randomiseret forsøg: rotter tilfældigt allokeret til de tre grupper

Forsøgsenheder: rotterne, $i = 1, \ldots, 15$.

Responsvariabel: Y, dvs. vi har observationer $Y_1, \ldots, Y_{15}$

Faktoren: `foder` inddeler forsøgsenhederne i grupper.

Ensidet variansanalyse: afhænger koag.-tiden af fodertypen?

Eksempel (fortsat)

Niveauerne for faktoren `foder` er A, B og C.

Værdierne knyttet til forsøgsenhederne er

$$\begin{aligned} \texttt{foder}_1 = \texttt{foder}_2 = \texttt{foder}_3 = \texttt{foder}_4 = \texttt{foder}_5 &= \texttt{A} \\ \texttt{foder}_6 = \texttt{foder}_7 = \texttt{foder}_8 = \texttt{foder}_9 = \texttt{foder}_{10} &= \texttt{B} \\ \texttt{foder}_{11} = \texttt{foder}_{12} = \texttt{foder}_{13} = \texttt{foder}_{14} = \texttt{foder}_{15} &= \texttt{C} \end{aligned}$$

`foder` er balanceret da $n_{\texttt{foder}}(A) = n_{\texttt{foder}}(B) = n_{\texttt{foder}}(C) = 5$.

Ensidet variansanalyse

Koagulationstiderne igen:

Foder	Koagulationstid (sek.)				
A	63	67	71	64	65
B	68	66	71	67	68
C	56	62	60	61	63

Modellen for ensidet variansanalyse, model A

$$A: \; Y_i = \alpha(\texttt{foder}_i) + e_i, \; e_i \sim N(0, \sigma^2)$$

Parametre i modellen: $\alpha(\texttt{A}), \alpha(\texttt{B}), \alpha(\texttt{C})$ og σ^2.

Interesseret i om der er en effekt af foder. Relevant hypotese?

Ensidet variansanalyse: model

Modellen for ensidet variansanalyse, model A

$$A: \; Y_i = \alpha(\texttt{foder}_i) + e_i, \; i = 1, \ldots, 15$$

hvor $e_1, \ldots, e_{15}$ er uafhængige og $N(0, \sigma^2)$-fordelte.

Antagelser:

- middelværdi afhænger (kun) af `foder`.
- uafhængighed
- varianshomogenitet, dvs. samme σ^2 for all i
- normalfordelt

Kontrol af modelantagelser: kapitel 5 (uge 2), se også opgave 1.2(i).

Ensidet variansanalyse: estimation

Middelværdiparametre estimeres ved gruppegennemsnit, fx.

$$\hat{\alpha}(\texttt{A}) = \bar{Y}_A$$

Residualkvadratsum (variation indenfor grupper):

$$\mathrm{SS}_e = \sum_{i=1}^{N} (Y_i - \bar{Y}_{\texttt{foder}_i})^2$$

Mean square error eller residual mean square

$$s^2 = \hat{\sigma}^2 = \mathrm{MS}_e = \frac{1}{N-k} \mathrm{SS}_e$$

hvor N er antal obs. og k er antal grupper, her $N = 15$ og $k = 3$.

Ensidet variansanalyse: hypotese

Hypotesen om ens middelværdier i de tre grupper,

$$\mathrm{H}_0 : \alpha(A) = \alpha(B) = \alpha(C) = \alpha$$

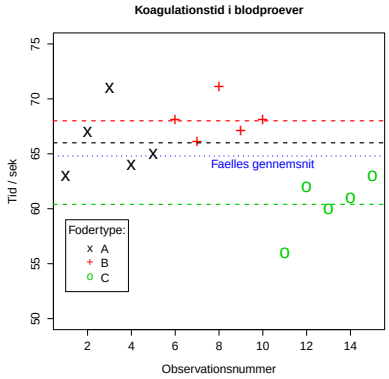
svarer til model B (med samme antagelser på e'erne som før):

$$B: \; Y_i = \alpha + e_i$$

Residualkvadratsum i model B (total variation):

$$\mathrm{SS}_e^0 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

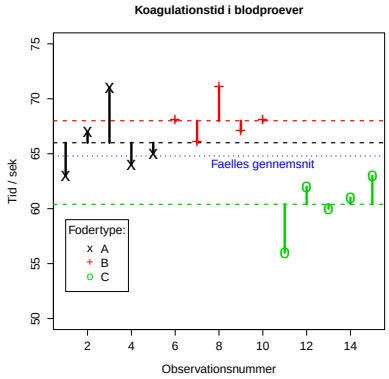
Ensidet variansanlyse



Forklar ud fra tegningen hvad følgende størrelser udtrykker:

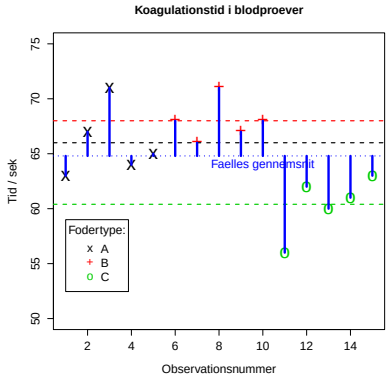
- $\hat{\alpha}(\mathrm{A}), \hat{\alpha}(\mathrm{B}), \hat{\alpha}(\mathrm{C}), \hat{\alpha}$

Ensidet variansanlyse



- Hvad kaldes de lodrette linjestykker?
- Ved beregning af hvilken størrelse skal man bruge længderne af de lodrette linjestykker?

Ensidet variansanlyse



- Hvad kaldes de lodrette linjestykker?
- Ved beregning af hvilken størrelse skal man bruge længderne af de lodrette linjestykker?

Test af hypotese

Teststørrelse,

$$F_{AB} = \frac{(SS_e^0 - SS_e)/(k-1)}{SS_e/(N-k)}$$

Tæller måler "forskellen" mellem de to modeller — nævneren normerer den korrekt i forhold til model A.

F_{AB} skal vurderes i $F(k-1, N-k)$-fordelingen.

Bemærk:

$$F_{AB} = \frac{MS_{\texttt{foder}}}{MS_e} = \frac{\sum_{i=1}^N (\bar{Y}_{\texttt{foder}_i} - \bar{Y})^2/(k-1)}{\sum_{i=1}^N (Y_i - \bar{Y}_{\texttt{foder}_i})^2/(N-k)}$$

Dvs. variation mellem grupper målt ift. variation indenfor grupper.

Ensidet variansanalyse i R

NB: `foder` bruges som faktor! Det sker automatisk her — hvorfor?

```
data<-read.table(file="../data/koagul.txt",header=T)
modelA <- lm(data$tid ~ data$foder )
modelB <- lm(data$tid ~ 1)
```

Test for hypotesen om ens middelværdier:

```
anova(modelB, modelA)
```

Fit af model uden referenceniveau (uden intercept):

```
modelA2 = lm(data$tid ~ data$foder - 1)
summary(modelA2)
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## data$foderA      66.0   1.177568 56.04771 6.862449e-16
## data$foderB      68.0   1.177568 57.74613 4.802121e-16
## data$foderC      60.4   1.177568 51.29215 1.980510e-15
```


```
Residual standard error: 2.633 on 12 degrees of freedom
```


Parvise sammenligninger

LSD-værdi (least significant difference)

$$t_{0.975,\texttt{df}} \cdot s \cdot \sqrt{1/n_1 + 1/n_2}$$

hvor

- `df` er residual degrees of freedom (-her 12)
- $t_{0.975,12} = 2.179$ er 97.5%−fraktilen i en t−fordeling med 12 frihedsgrader (-fås i R som `qt(0.975,12)`)
- s er estimatet for residual spredningen
- n_1, n_2 er antal observationer i de grupper, der sammenlignes

Da forsøget er balanceret fås her (altid): $LSD = 3.63$. Hvorfor?

Er der forskel på koagulationstiden for `foder=B` og `foder=C`?

Alternativ til LSD

```
pairwise.t.test(data$tid,data$foder,p.adj="none")
```

Konklusion

Slutmodel (ensidet variansanalyse)

$$Y_i = \alpha(\texttt{foder}_i) + e_i, \quad e_i \sim N(0, \sigma^2)$$

Effekt af foder påvist med stor sikkerhed ($F = 11.2, p = 0.0018$)

$$\hat{\alpha}(\texttt{A}) = 66.0, \quad \hat{\alpha}(\texttt{B}) = 68.0, \quad \hat{\alpha}(\texttt{C}) = 60.4, \quad s = \hat{\sigma} = 2.633$$

Skal suppleres med en/flere af følgende størrelser:

- s.e. på esitmaterne. Her: $s/\sqrt{5} = 1.178$ for alle 3 $\hat{\alpha}$'er.
- LSD-værdi: Her $LSD = 3.63$ (-se foregående slide)
- 95%−konfidensintervaller, f.eks.

$$\hat{\alpha}(\texttt{A}) : 66.0 \pm 2.179 \cdot 1.178 = 66.0 \pm 2.57$$

```
confint(modelA2) ### <- giver konfidensintervaller i R
```

Variansanalyseskema (modelskema)

Særligt i forb. med flersidet variansanalyse (>1 faktor) kan man med fordel lave et variansanalyseskema svarende til de forskellige modeller.

For ensidet variansanalyse ser skemaet således ud

Model	Faktorer	Middelværdi	SS_e	DF_e
A	`foder`	α(`foder`)	83.2	12
B	0	α	238.4	14

SS_e : residual kvadratsum
DF_e : residual frihedsgradsantal

DF_e er forskellen mellem antal observationer (-her 15) og antallet af parametre til beskrivelse af middelværdistrukturen i modellen (model A=3 / model B=1).

Variansanalyseskema (testskema)

Særligt i forb. med flersidet variansanalyse (>1 faktor) kan man med fordel lave et variansanalyseskema svarende til de forskellige tests.

For ensidet variansanalyse er der kun et relevant test og testskemaet ser således ud

Test	Faktor	F	df	$p-$værdi
$A \to B$	`foder`	11.2	2	0.0018

Teststørrelsens opbygning

$$F = \frac{\overbrace{(238.4-83.2)}^{\text{forskel i } SS_e} / \overbrace{(14-12)}^{\text{forskel i } DF_e}}{\underbrace{83.2}_{SS_e \text{ stor model}} \quad / \quad \underbrace{12}_{DF_e \text{ stor model}}} = 11.19 \sim F(2,12) \leftarrow \text{f-fordeling}$$

```
> 1-pf(11.19,2,12) ###  <- giver p-værdi i R
[1] 0.001808224
```

Parvise sammenligninger vs test af beh. effekt

Advarsel

Der findes ingen generel regel som siger, at der er en behandlingseffekt præcis hvis der findes mindst et par af behandlinger som er signifikant forskellige!

På de følgende sider vises gruopvækkende eksempler på dette.

Lad os først opsummere for eksemplet med koagulationstider

- Er der en effekt af behandlingen?
- Findes der et par af behandlinger, som er signifikant forskellige?
- Er alle par af behandlinger signifikant forskellige?

Parvise sammenligninger vs test af beh. effekt

Samme forsøgsdesign som i eksemplet med koagulation blot er der ændret lidt på observationerne.
Test skema

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data$foder	2	155.20	77.60	2.80	0.1006
Residuals	12	332.80	27.73		

Parvise sammenligninger (p-værdier)

	A	B
B	0.56	
C	0.12	0.04

- Er der en effekt af behandlingen?
- Hvad konkluderer du ud fra de parvise sammenligninger?

Parvise sammenligninger vs test af beh. effekt

Samme forsøgsdesign som i eksemplet med koagulation blot er der ændret lidt på observationerne.

Test skema

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data$foder	2	144.40	72.20	8.61	0.0048
Residuals	12	100.67	8.39		

Parvise sammenligninger (p-værdier)

	A	B
B	0.06	
C	0.06	0.00

- Er der en effekt af behandlingen?
- Hvad konkluderer du ud fra de parvise sammenligninger?

Repetition: lineær regression (kort)

Eksempel fra SD1-noter/bog:
9 sammenhørende målinger af stearinsyreindhold, `acid`, i fedt (målt i %) og fordøjeligheden, `dig` (målt i %).

Model for lineær regression:

$$\texttt{dig}_i = \alpha + \beta \cdot \texttt{acid}_i + e_i, \quad e_i \sim N(0, \sigma^2)$$

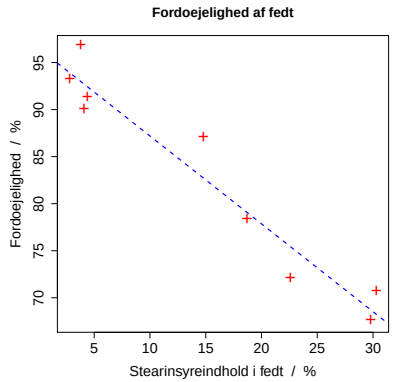
Interesserede i effekt af `acid`. Relevant hypotese?

```
data<-read.table(file='../data/linreg.txt',header=T)
modellin = lm(data$dig~data$acid)
summary(modellin)
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 96.5333636 1.67517720  57.62576 1.243407e-10
## data$acid   -0.9337416 0.09262067 -10.08135 2.028095e-05
```

Hvad skal vi konkludere?

Repetition: lineær regression (kort)



- Slutmodel: $\texttt{dig}_i = \alpha + \beta \cdot \texttt{acid}_i + e_i, \quad e_i \sim N(0, \sigma^2)$
- Stearinsyreindholdet i fedt har signifikant effekt på fordøjelsen.
- Estimater: $\hat{\alpha} = 96.533, \hat{\beta} = -0.934, \hat{\sigma}^2 = 2.97^2 = 8.82$

1-sidet ANOVA: klorofyl i vinterhvede (selvstudie)

Model for ensidet variansanalyse:

$$Y_i = \alpha(\texttt{beh}_i) + e_i, \quad e_i \sim N(0, \sigma^2)$$

Parametre i modellen:
$\alpha(\texttt{kontrol}), \alpha(\texttt{glycine}), \alpha(\texttt{glycine} + MnS0_4), \sigma^2$

Interesserede i effekt af behandlingen.

- Hvad beskriver parametrene i modellen?
- Hvordan udtrykkes den statistiske hypotese vha. parametrene i modellen?

1-sidet ANOVA: i R (selvstudie)

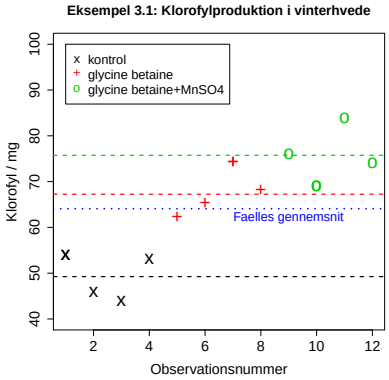
Indlæsning af data, fit af modeller og test af nulhypotese

```
data<-read.table(file="../data/chloro.txt",header=T)
model0<-lm(data$chloro~factor(data$treat))
model1<-lm(data$chloro~1)
anova(model1,model0)

## Analysis of Variance Table
##
## Model 1: data$chloro ~ 1
## Model 2: data$chloro ~ factor(data$treat)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     11 1734.92
## 2      9  270.25  2    1464.7 24.389 0.0002324 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvad skal vi konkludere?

1-sidet ANOVA: klorofyl i vinterhvede (selvstudie)



- Testet sammenligner variationen mellem gruppegennemsnit med variationen inden for grupper

1-sidet ANOVA: klorofyl i vinterhvede (selvstudie)

Konklusion: Slutmodel (ensidet variansanalyse)

$$Y_i = \alpha(\mathrm{beh}_i) + e_i, \quad e_i \sim N(0, \sigma^2)$$

Behandlingerne giver signifikante forskelle i klorofylindholdet og parameterestimaterne bliver ...

```
model0a<-lm(data$chloro~factor(data$treat))
model0b<-lm(data$chloro~factor(data$treat)-1)
model0c<-lm(data$chloro~relevel(factor(data$treat),ref="2"))
model0d<-lm(data$chloro~relevel(factor(data$treat),ref="3"))
```

Udskriv estimaterne for din yndlingsparametrisering...

```
summary(model0b) ### <- vaelg din yndlingsmodel!
```

- Fire forskellige måder at fitte *samme* model på
- Forskellen ligger i, hvordan R afrapporterer parameterestimater!
- Alle metoder er rigtige men forklar i ord, at du ved, hvad du gør!