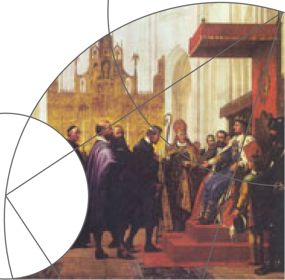
Statistisk Dataanalyse 1: Introduktion til lineær regression og ensidet variansanalyse

Anders Tolver
Institut for Matematiske Fag



Undervisning under COVID-19

Coronavirus og COVID-19 påvirker stadig undervisningen på KU-SCIENCE.

- Orienter dig på KUnet særligt omkring **Coronavirus** under Studieinformation.
- Vær opmærksom på reglerne, hvis du eller dine nære kontakter bliver smittede.

Fysisk undervisning på Campus er underlagt restriktioner

- Læs informationen i modulet om COVID19 på Absalon.
- Hold afstand og hjælp til med at rengøre kontaktflader.
- Møde kun op, hvis du har tilmeldt dig fysiske timer via Absalon.

Forelæsningen streames og optages

- Møder du fysisk op til forelæsningen, kan du komme til at optræde med billede og/eller lyd på optagelsen.
- Optagelsen lægges på kursets Absalon side og er kun tilgængelig for personer knyttet til kurset.
- Optagelsen benyttes til undervisningsbrug og er foretaget med hjemmel i databeskyttelseslovens artikel 6, stk. 1, litra c) og e) samt universitetsloven.
- Optagelsen vil være tilgængelig i 1 år, og du kan altid henvende dig til Anders Tolver, hvis du ønsker at blive fjernet fra en del af optagelsen.

Dagens program

Morgen (8:45-10:15):

- Lineær regression og korrelation
- Ensidet variansanalyse

Vi vender tilbage til lineær regression og ensidet variansanalyse mange gange de kommende uger.

Idag: primært introduktion/motivation

I eftermiddag (ca. kl. 15:00) uploades video som berører:

- Evt. hængepartier fra i morges
- R Markdown
- Om at arbejde med datasæt i R

isdals-pakken

Data til bogens eksempler og opgaver ligger i R-pakken *isdals*.

- Installér pakken via *Tools*-menuen.

  Kun første gang du skal bruge pakken.
- Load pakken, enten ved at sætte hak ved den i listen over pakker eller med kommandoen

  ```
  library(isdals)
  ```

  Dette skal ske I hver R-session hvor du skal bruge pakken.
- Load data med kommandoen `data(navn-på-datasættet)`

  Dette skal ske i hver R-session hvor du skal bruge datasættet.

Du bliver ledt gennem dette i opgave HS.2 i dag.

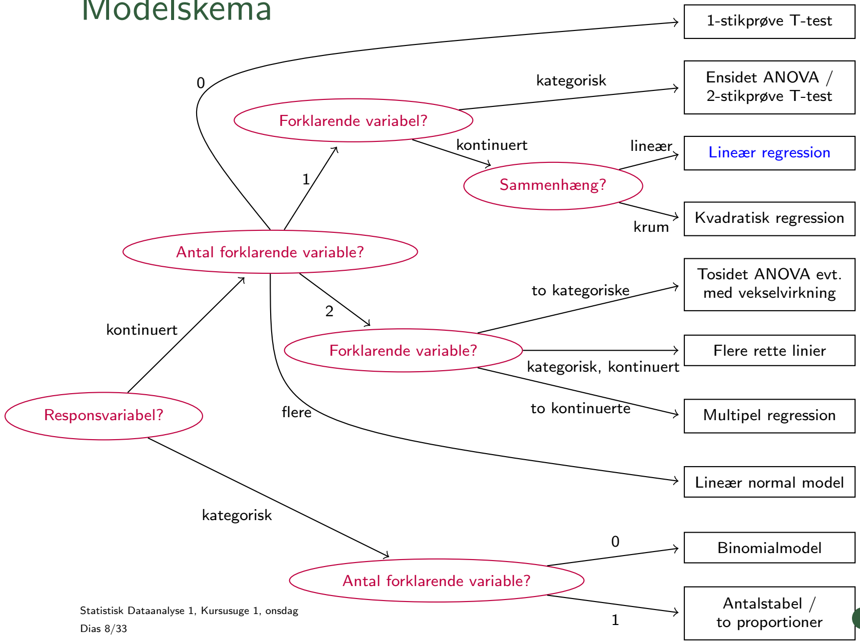
Lineær regression og korrelation

Eksempel: Kattes krops- og hjertevægt

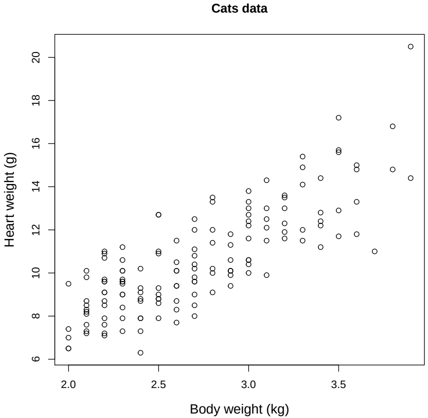
Data: Kropsvægt i kg, vægt af hjerte i gram for 144 katte.
Glem alt om kattenes køn i dag.

Ønsker at **prædiktere** (forudsige) hjertevægt udfra kropsvægt: Brug **Hwt** som **responsvariabel**, **Bwt** som **forklarende variabel**.

Modelskema


1-stikprøve T-test
0
kategorisk
Ensidet ANOVA / 2-stikprøve T-test
Forklarende variabel?
kontinuert
lineær
Lineær regression
1
Sammenhæng?
Kvadratisk regression
krum
Antal forklarende variable?
Tosidet ANOVA evt. med vekselvirkning
to kategoriske
2
kontinuert
Forklarende variable?
Flere rette linier
kategorisk, kontinuert
Responsvariabel?
flere
to kontinuerte
Multipel regression
Lineær normal model
kategorisk
0
Binomialmodel
Antal forklarende variable?
1
Antalstabel / to proportioner

Giver lineær regression overhovedet mening her?



Det ser faktisk ud til at punkterne varierer omkring en ret linie, så lineær regression giver mening.

Lineær regression

Ligning for ret linie med skæring (intercept) α og hældning (slope) β:

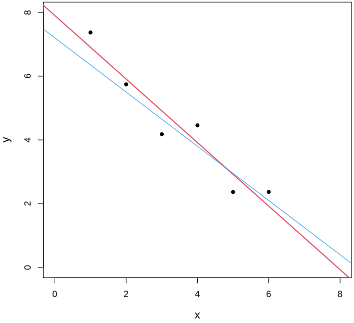
$$y = \alpha + \beta \cdot x$$

Vores opgave er at finde den rette linie der "'passer bedst"' med data.

Altså: Find de værdier af α og β der passer bedst.

Legetøjsdata

Dette er nogle andre data! To gode forslag til rette linier, men hvilken linie er bedst?

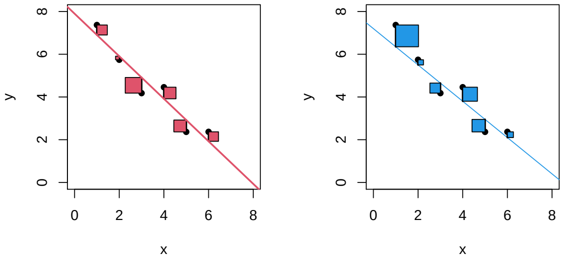


Bliver nødt til at have en objektiv metode: **Mindste kvadraters metode (least squares)**

Mindste kvadraters metode (least squares)

For alle rette linier i hele verden:

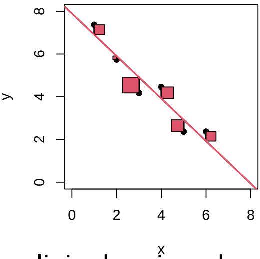
- Lodret afstand mellem punkter og linie, $r_i = y_i - \alpha - \beta x_i$
- Kvadrér disse afstande, r_i^2, og beregn $r_1^2 + \cdots + r_n^2$

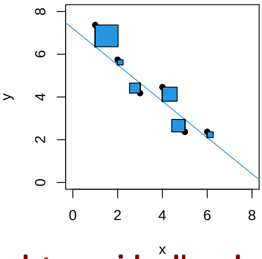


Mindste kvadraters metode (least squares)

For alle rette linier i hele verden:

- Lodret afstand mellem punkter og linie, $r_i = y_i - \alpha - \beta x_i$
- Kvadrér disse afstande, r_i^2, og beregn $r_1^2 + \cdots + r_n^2$





Find den linie der giver den **mindste residualkvadratsum.**

Formlerne

Det viser sig at den bedste rette linie er givet ved følgende formler:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\alpha} &= \bar{y} - \hat{\beta} \cdot \bar{x} \end{aligned}$$

hvor $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ og $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ er gennemsnittene.

Formlerne

Det viser sig at den bedste rette linie er givet ved følgende formler:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

hvor $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ og $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ er gennemsnittene.

Bemærk:

- Fortegnet på $\hat{\beta}$
- Regressionslinien går gennem $(\bar{x}, \bar{y})$

I praksis skal vi ikke bruge formlerne — det lader vi R klare!

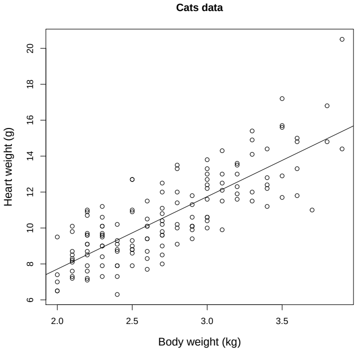
R

```
library(MASS)
data(cats)
plot(Hwt ~ Bwt, data=cats)
lm(Hwt ~ Bwt, data=cats)

##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Coefficients:
## (Intercept)          Bwt
##     -0.3567       4.0341

abline(-0.3567, 4.0341)
```

Eksempel: Kattes krops- og hjertevægt

Regressionslinien — den bedste rette linie — for kattene:

$$\mathrm{Hwt} = -0.3567 + 4.0341 \cdot \mathrm{Bwt}$$



Fortolkning af parametrene?

Fortolkning!

Mindste kvadraters metode giver estimeret regressionslinie:

$$y = \hat{\alpha} + \hat{\beta} \cdot x$$

Fortolkning!

Mindste kvadraters metode giver estimeret regressionslinie:

$$y = \hat{\alpha} + \hat{\beta} \cdot x$$

Fortolkning:

- Modellen/linien fortæller, hvad vi vil forvente for et givet x:
$$\hat{y} = \hat{\alpha} + \hat{\beta} \cdot x$$
- $\hat{\beta}$: For to enheder med en forskel i x-værdi på Δx, vil vi forvente en forskel på $\Delta y = \hat{\beta} \cdot \Delta x$.
- $\hat{\alpha}$: den forventede y-værdi for $x = 0$ (hvis det giver mening).

Usikkerhed

Har endnu intet sagt om usikkerheden på estimaterne!

- Hvor meget kan vi stole på estimaterne?
- Hvor meget anderledes kunne estimaterne blive hvis vi kiggede på andre katte fra samme population?
- Er der overhovedet en sammenhæng?

Coming up: Standard errors, konfidensintervaller, hypotesetest, prædiktionsintervaller, modelkontrol.

Brug af lineær regression

Hvornår kan vi bruge lineær regression?

- Begge variable skal være **kvantitative**
- Der skal være et **"'naturligt"' valg af hhv. respons og forklarende variabel** (hvad er x hhv. y?)
- Der skal være **tilnærmelsesvis lineær sammenhæng**.
- Et par antagelser mere som vi vender tilbage til...
- Pas på hvis der er ekstremt store/små værdier af x eller y. Kan trække meget i linien.

Brug af lineær regression

Hvornår kan vi bruge lineær regression?

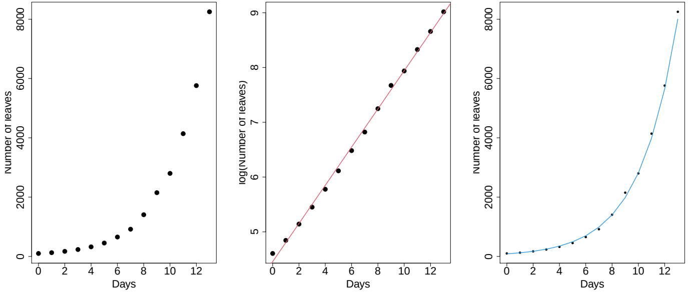
- Begge variable skal være **kvantitative**
- Der skal være et **"'naturligt"' valg af hhv. respons og forklarende variabel** (hvad er x hhv. y?)
- Der skal være **tilnærmelsesvis lineær sammenhæng**.
- Et par antagelser mere som vi vender tilbage til...
- Pas på hvis der er ekstremt store/små værdier af x eller y. Kan trække meget i linien.

Det er **forbudt at fortolke regressionslinien (langt) udenfor observationsintervallet** (ekstrapolation).

Hvad hvis sammenhængen ikke er lineær?

Sommetider kan man transformere sig til lineær sammenhæng.

Eksempel 2.4: Hvis (x, y) sammenhængen er eksponentiel, så er $(x, \log(y))$-sammenhængen lineær.



Hvis det er uklart hvad der skal være x og y?

Sommetider har man datapar (x, y) som er **ligeværdige** i den forstand at det ikke er klart at den ene forklarer den anden.

Eksempel: Hovedomkreds og maveomkreds hos nyfødte.

- Begge dele er mål for babyens størrelse
- Ikke specielt naturligt at modellere den ene som funktion af den anden—medmindre man er interesseret i prædiktion.
- Interessant at sige noget om graden af sammenhæng

I disse tilfælde beregner man ofte **korrelationskoefficienten.**

Korrelationskoefficienten

Korrelationskoefficienten måler graden af **lineær sammenhæng** mellem x og y:

$$\hat{\rho} = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\mathrm{sd}_x \cdot \mathrm{sd}_y}$$

Kan tænke på $\hat{\rho}$ som hældningen i en lineær regression hvor man bruger standardiserede versioner af x og y.

Korrelationskoefficienten

Korrelationskoefficienten måler graden af **lineær sammenhæng** mellem x og y:

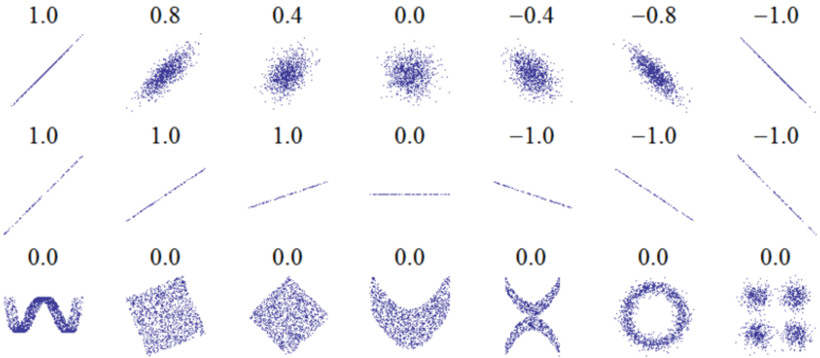
$$\hat{\rho} = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\mathrm{sd}_x \cdot \mathrm{sd}_y}$$

Kan tænke på $\hat{\rho}$ som hældningen i en lineær regression hvor man bruger standardiserede versioner af x og y.

Korrelationskoefficienten er

- altid mellem -1 og $+1$
- 0 hvis der ikke er nogen (lineær) information om y i x, eller omvendt
- ± 1 hvis observationerne ligger perfekt på en linje med positiv/negativ hældning

Korrelationskoefficienten: eksempler



Ensidet variansanalyse

Eksempel 3.2: Nedbrydning af organisk materiale

Data

- Fem typer antibiotika og en kontrolbehandling.
- 36 kvier inddelt i seks grupper. Foder tilsat antibiotikum.
- Gødning gravet ned i poser og mængden af organisk materiale målt efter 8 uger.
- For spiramycin: Kun fire brugbare målinger.

Eksempel 3.2: Nedbrydning af organisk materiale

Data

- Fem typer antibiotika og en kontrolbehandling.
- 36 kvier inddelt i seks grupper. Foder tilsat antibiotikum.
- Gødning gravet ned i poser og mængden af organisk materiale målt efter 8 uger.
- For spiramycin: Kun fire brugbare målinger.

Formål

- Påvirker antibiotika nedbrydningen af organisk materiale?
- Hvis kontrolmålingerne ligger lavere end de andre, tyder det på at antibiotika hæmmer nedbrydningen.

Data

Data er tilgængelige i datasættet `antibio` i `isdals`-pakken.

```
library(isdals)
data(antibio)
head(antibio, n=7)

##       type  org
## 1 Ivermect 3.03
## 2 Ivermect 2.81
## 3 Ivermect 3.06
## 4 Ivermect 3.11
## 5 Ivermect 2.94
## 6 Ivermect 3.06
## 7  Alfacyp 3.00
```

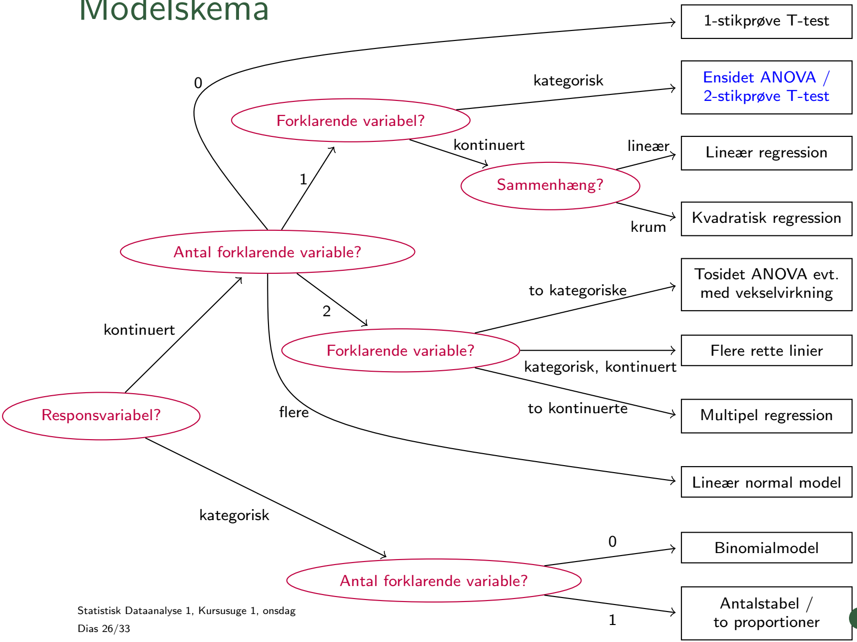
Data

Data er tilgængelige i datasættet `antibio` i `isdals`-pakken.

```
library(isdals)
data(antibio)
head(antibio, n=7)

##       type  org
## 1 Ivermect 3.03
## 2 Ivermect 2.81
## 3 Ivermect 3.06
## 4 Ivermect 3.11
## 5 Ivermect 2.94
## 6 Ivermect 3.06
## 7  Alfacyp 3.00
```

To variable: `type` og `org`. Datatyper? Responsvariabel? Forklarende variabel?

Modelskema


1-stikprøve T-test
0
kategorisk
Ensidet ANOVA / 2-stikprøve T-test
Forklarende variabel?
kontinuert
lineær
Lineær regression
1
Sammenhæng?
Kvadratisk regression
krum
Antal forklarende variable?
Tosidet ANOVA evt. med vekselvirkning
to kategoriske
2
kontinuert
Forklarende variable?
Flere rette linier
kategorisk, kontinuert
Responsvariabel?
flere
to kontinuerte
Multipel regression
Lineær normal model
kategorisk
0
Binomialmodel
Antal forklarende variable?
Antalstabel / to proportioner
1

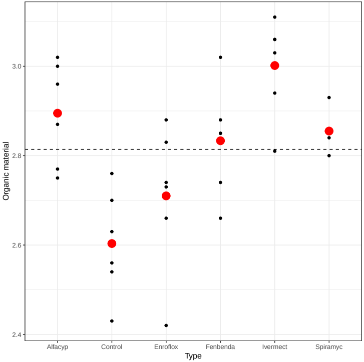
Hvorfor hedder det ensidet variansanalyse?

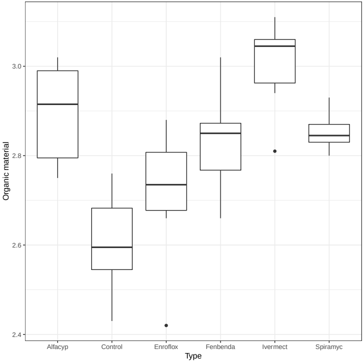
- **Ensidet:** Fordi der kun er en enkelt forklarende variabel
- **Variansanalyse:** Fordi forskelle mellem grupper påvises ved at sammenligne forskellige kilder til variation

  Variansanalyse = Analysis of variance = ANOVA.

I behøver ikke læse detaljerne i bogen nu. Vi vender tilbage senere...

Hvordan ser data ud?





- Hvad kan vi se?
- Kan vi konkludere at der er forskel på grupperne?

Between-group og within-group variation

Alle observationer er ikke ens! Men hvorfor ikke?

- Fordi der (potentielt) er forskel på behandlingerne $\rightarrow$ **between-group variation**
- Fordi der er biologisk variation, ikke-ens respons selv hvis gødningen behandles ens $\rightarrow$ **within-group variation**

Between-group og within-group variation

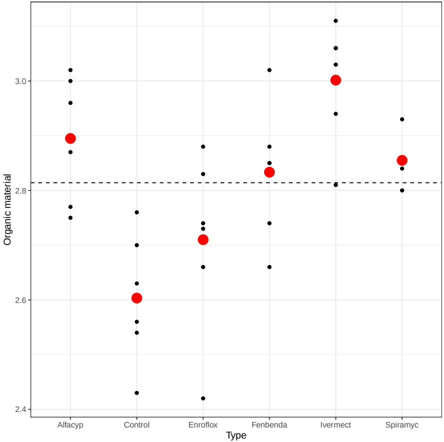
Alle observationer er ikke ens! Men hvorfor ikke?

- Fordi der (potentielt) er forskel på behandlingerne → **between-group variation**
- Fordi der er biologisk variation, ikke-ens respons selv hvis gødningen behandles ens → **within-group variation**

Hvis between-group variation er stor ift. within-group variation, er det tegn på at der er forskel på grupperne.

Der er formler i bogen, for $\mathrm{SS}_{\text{between}}$ og $\mathrm{SS}_{\text{within}}$ i bogen, men det er vigtigere at forstå den grafiske betydning.

Between-group og within-group variation



- **Between-group variation:** Forskel mellem de forskellige grupper. Gruppegennemsnit vs totalgennemsnit.
- **Within-group variation:** Forskel mellem obs. fra samme gruppe. Punkter vs gruppegennemsnit

Gruppegennemsnit og -spredninger

Gruppegennemsnit (og -spredninger) er vigtige:

Behandling	n_j	$\bar{y}_j$	s_j
Control	6	2.603	0.119
α-cyperm.	6	2.895	0.117
Enrofloxacin	6	2.710	0.162
Fenbendaz.	6	2.833	0.124
Ivermectin	6	3.002	0.109
Spiramycin	4	2.855	0.054

Gruppegennemsnit og -spredninger

Gruppegennemsnit (og -spredninger) er vigtige:

Behandling	n_j	$\bar{y}_j$	s_j
Control	6	2.603	0.119
α-cyperm.	6	2.895	0.117
Enrofloxacin	6	2.710	0.162
Fenbendaz.	6	2.833	0.124
Ivermectin	6	3.002	0.109
Spiramycin	4	2.855	0.054

Gennemsnittene kan beregnes i R på følgende måde:

```
data(antibio)
lm(org ~ type-1, data=antibio)
```

Hvad mon der sker hvis vi ikke skriver −1? Se opgave HS.4!

Usikkerhed

Gennemsnittene er **estimater for populationsgennemsnit**, dvs. gennemsnit af responsen hvis vi testede behandlingerne på alle kvier i verden.

Usikkerhed

Gennemsnittene er **estimater for populationsgennemsnit**, dvs. gennemsnit af responsen hvis vi testede behandlingerne på alle kvier i verden.

Har endnu intet sagt om usikkerheden på gennemsnittene:

- Hvor meget kan vi stole på estimaterne?
- Hvor meget anderledes kunne estimaterne blive hvis vi kiggede på andre kvier fra samme population?
- Er der forskel på behandlingerne?

Coming up: Standard errors, konfidensintervaller, hypotesetest, prædiktionsintervaller, modelkontrol.

Opsummering – til eget brug

- Hvornår er det rimeligt at benytte lineær regression?
- Hvad er fortolkningen af parametrene i en lineær regression?
- Hvad er princippet i at bestemme den bedste rette linie?
- Hvad måler korrelationskoefficienten?
- Hvad er formålet i en ensidet variansanalyse?
- Hvilke typer variation er der når vi har data fra flere grupper?
- Kan vi konkludere om der er forskel på grupperne på baggrund af plots og/eller tabel med gruppegennemsnit?