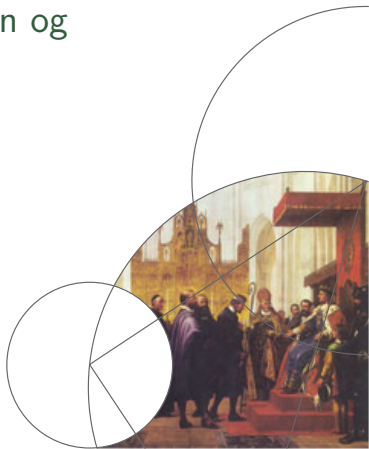




Det Natur- og Biovidenskabelige Fakultet

# Statistisk Dataanalyse 1: Introduktion til lineær regression og ensidet variansanalyse

Anders Tolver  
Institut for Matematiske Fag



# Dagens program

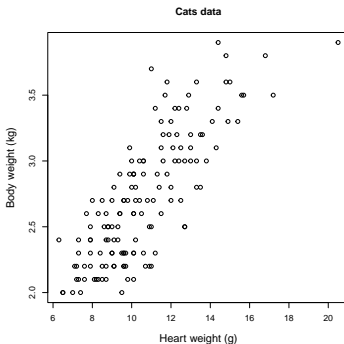
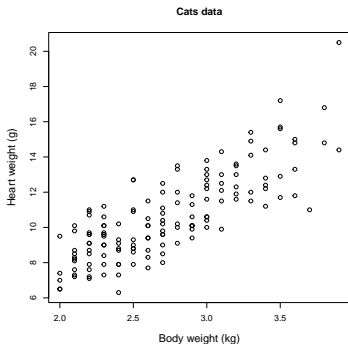
- Praktiske oplysninger og spørgsmål
- Sammenhæng ml. kontinuerte variable
- Lineær regression
- Ensidet variansanalyse (ANOVA)



# Sammenhæng mellem to kontinuerte variable (lineær association)



# Kropsvægt og hjertevægt for 144 katte



**Overvej:** Hvorfor tænker vi straks, at der ses en klar sammenhæng ml. kropsvægt og hjertevægt?

## Hvad betyder sammenhæng (association)

Ofte indsamles data bestående af **par  $(x, y)$  af kvantitative, kontinuerte variable** med henblik på at undersøge om der er sammenhæng ml.  $x$  og  $y$ .

Hvad mener vi med **sammenhæng** (eng: association)?

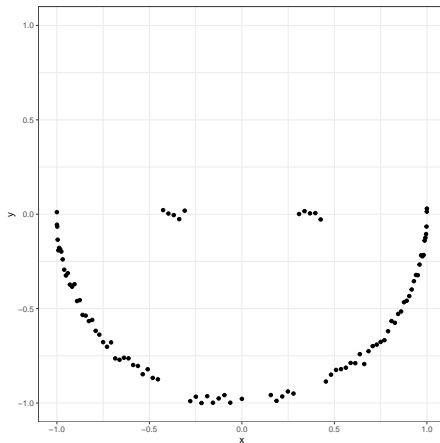
- Intuitivt/visuelt  $\rightarrow$  se på scatterplot
- Intuitivt: hvis  $x$  er stor, så er  $y$  typisk stor
- Intuitivt: jeg kan bedre gætte værdien af  $y$ , hvis jeg kender  $x$

Hvad skal vi se på?

- Kan vi lave (objektiv) mål for sammenhæng  $\rightarrow$  **korrelationskoefficient**
- Hvordan kan modellere og udnytte (matematisk) sammenhæng  $\rightarrow$  prædiktion



# Er der en sammenhæng?



**Overvej:** Prøv at argumentere både for og imod at der er en sammenhæng ml.  $x$  og  $y$ ?

# Korrelationskoefficienten

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

$$\hat{\rho} = \frac{\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{sd_x \cdot sd_y}$$

(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  $y$ :

$$\hat{\rho} = \frac{\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{sd_x \cdot sd_y}$$

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

Korrelationskoefficienten er

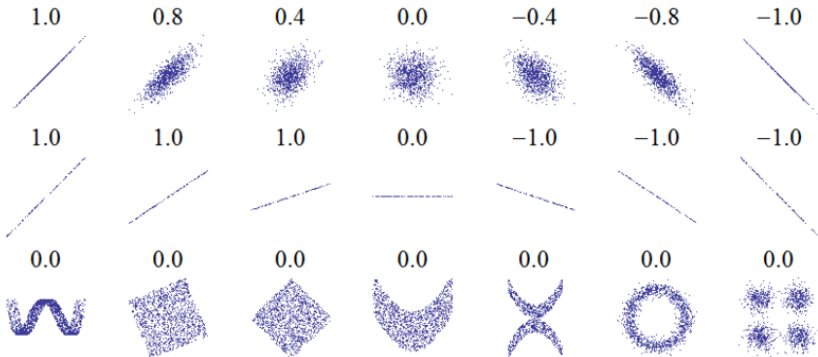
- altid mellem  $-1$  og  $+1$
- 0 hvis der ikke er nogen (lineær) information om  $y$  i  $x$ , eller omvendt
- $\pm 1$  hvis observationerne ligger perfekt på en linje med positiv/negativ hældning

**Intuition:** Måler om punkter over gennemsnit for  $x$  har en lige så stor tendens til at ligge over/under gennemsnit for  $y$ ?





# Korrelationskoefficienten: eksempler



# Sammenhæng, korrelation eller effekt?

**Sammenhæng:** kendskab til  $x$  forbedrer muligheder for at udtale os om  $y$  (eller omvendt!).

**Korrelation** (=lineær association) *væsentlig* forskellig fra 0: vi kan bruge lineær funktion til at udtale os om  $y$  på baggrund af  $x$  (eller omvendt!)

Ved **lineær regression** forsøger man at beskrive  $y$  ud fra  $x$  ved en lineær funktion

$$a + b \cdot x.$$

Kræver valg af **respons**  $y$  og **forklarende variabel**  $x$ .

Mere skal til for at konkludere at  $x$  har **(kausal) effekt** på  $y$ .



# Lineær regression



## Eksempel: Kattes krops- og hjertevægt

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

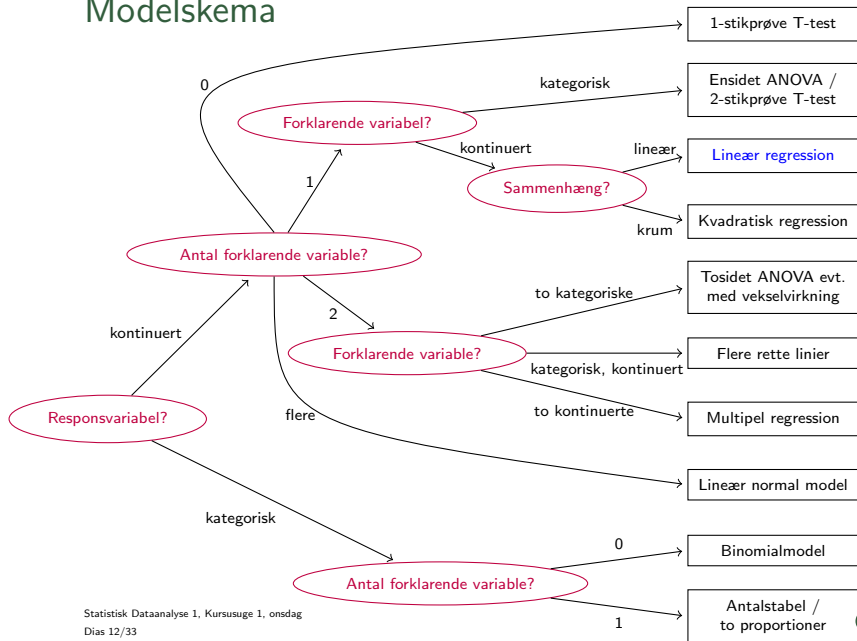
Ønsker at **prædiktere** (forudsige) hjertevægt ud fra  
kropsvægt: Brug

- **Hwt** som **responsvariabel**
- **Bwt** som **forklarende variabel**

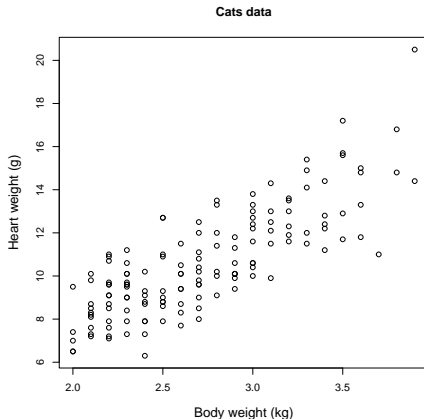
**Overvej:** Hvorfor virker det mest naturligt med  $x$  som  
forklarende (også kaldet **uafhængig**) variabel?



# Modelskema



# 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)  $\alpha$  og hældning (slope)  $\beta$ :

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

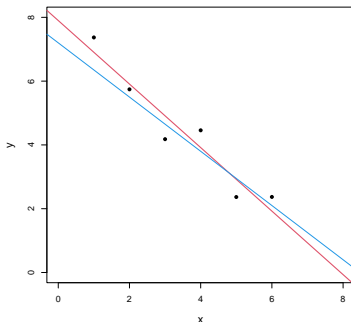
Vores opgave er at finde den rette linie der ”’passer bedst’” med data.

Altså: Find de værdier af  $\alpha$  og  $\beta$  der passer bedst.



## Legetøjsdata

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



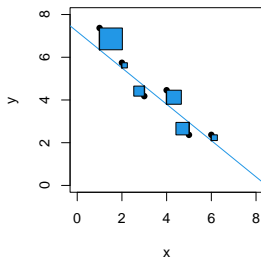
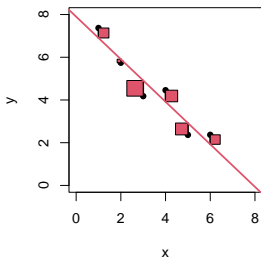
Bliver nødt til at have en objektiv metode: **Mindste kvadraters metode (least squares)**



# Mindste kvadraters metode (least squares)

For alle mulige linjer kan vi se på:

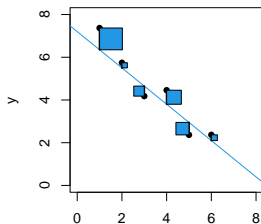
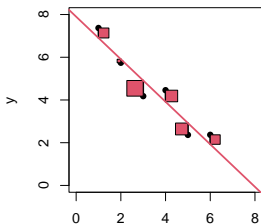
- Lodret afstand mellem punkter og linie,  $r_i = y_i - \alpha - \beta x_i$
- Kvadrér disse afstande,  $r_i^2$ , og beregn  $r_1^2 + \dots + r_n^2$



# Mindste kvadraters metode (least squares)

For alle mulige linjer kan vi se på:

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



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:

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

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.

Bemærk:

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

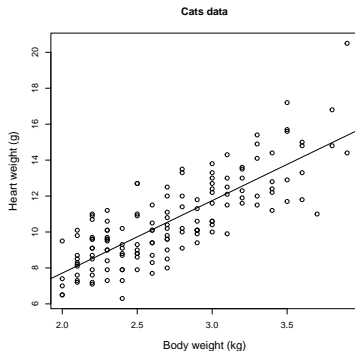
I praksis skal vi ikke bruge formlerne — det lader vi R klare!



# Eksempel: Kattes krops- og hjertevægt

**Regressionslinien** — den bedste rette linie — for kattene:

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



## 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:

- Model/linjen 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  $\Delta x$ , 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).

Advarsel: pas på med at

- ekstrapolere til ekstreme  $x$ -værdier:  $\hat{\alpha} + \hat{\beta} \cdot 10$
- udtale dig om den forventede ændring i hjertevægt, hvis vi feder alle katte op til de har taget 1 kg på



# Usikkerhed

Har endnu intet sagt om usikkerheden på estimerne!

- Hvor meget kan vi stole på estimerne?
- Hvor meget anderledes kunne estimerne blive hvis vi kiggede ny sample af 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?

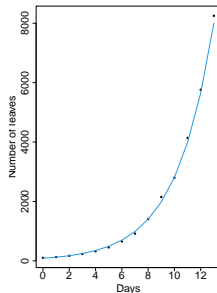
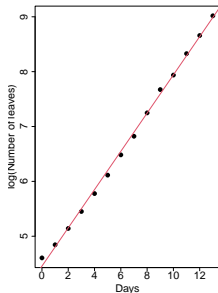
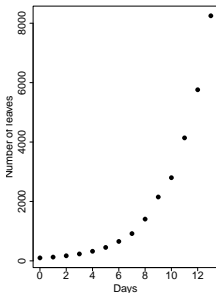
- Begge variable skal være **kvantitative**
- Der skal være et **”naturligt” valg af hhv. respons og forklarende variabel** (hvad er  $x$  hhv.  $y$ ?)
- Der skal være **tilnærmelsesvis lineær sammenhæng**.
- Et par antagelser mere som vi vender tilbage til...
- Pas på hvis der er ekstremt store/små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



# 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.



# 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  Alfacy 3.00
```



# Data

Data er tilgængelige i datasættet antibio i isdals-pakken.

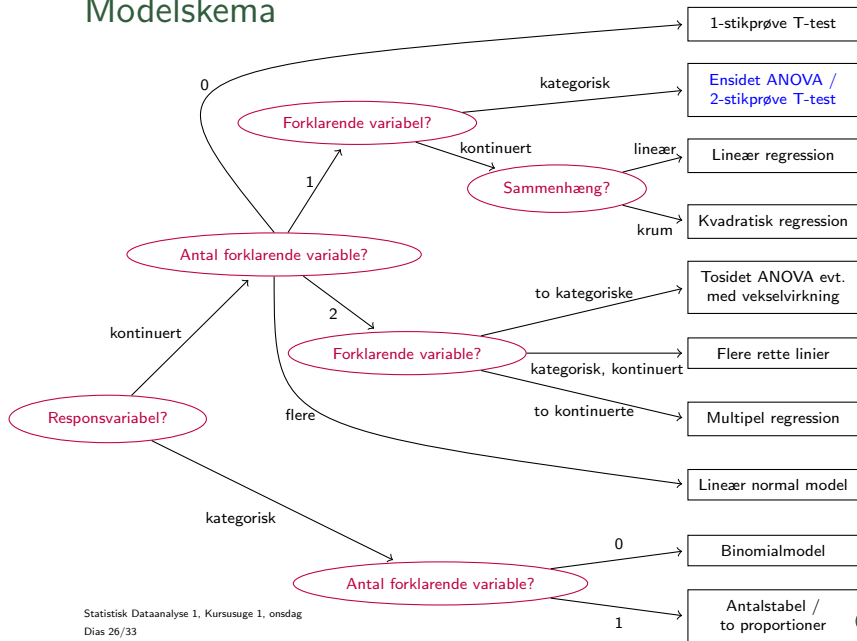
```
library(isdals)
data(antibio)
head(antibio, n=7)
```

```
##           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  Alfacy 3.00
```

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



# Modelskema





# Hvorfor hedder det ensidet variansanalyse?

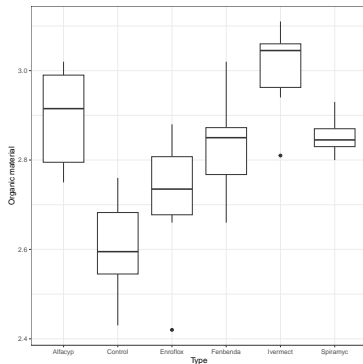
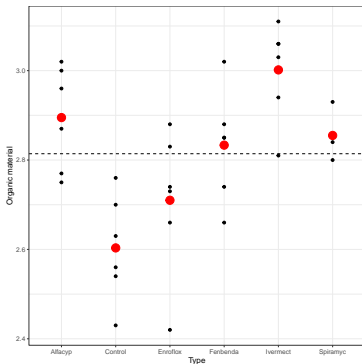
- **Ensidet:** Fordi der kun er en enkelt forklarende variabel
- **Variansanalyse:** Fordi forskelle mellem grupper påvises ved at sammenligne forskellige kilder til variation

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 → **between-group variation**
- Fordi der er biologisk variation, ikke-ens respons selv hvis gødningen behandles ens → **within-group variation**



# Between-group og within-group variation

Alle observationer er ikke ens! Men hvorfor ikke?

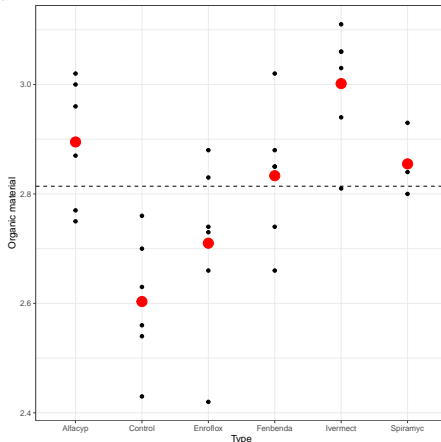
- Fordi der (potentielt) er forskel på behandlingerne → **between-group variation**
- Fordi der er biologisk variation, ikke-ens respons selv hvis gødningen behandles ens → **within-group variation**

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  $SS_{\text{between}}$  og  $SS_{\text{within}}$  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
$\alpha$ -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
$\alpha$ -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 med `summarize()` eller 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 **estimerer for populationsgennemsnit**, dvs. gennemsnit af responsen hvis vi testede behandlingerne på alle kvier i verden.





# Usikkerhed

Gennemsnittene er **estimerer 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 estimerterne?
- Hvor meget anderledes kunne estimerterne 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?

