Introducerende Statistik og Dataanalyse med R

Ensidet og tosidet variansanalyse

Jens Ledet Jensen

I dag

Generelle lineære model præsenteret via

ensidet variansanalyse (one-way anova)

tosidet variansanalyse (two-way anova)

Vigtige begreber:

- Faktor
- F-test, testtabel

Ser på 184 Hornede tudseøgler: 154 levende, 30 spist af tornskade
Respons: længde af horn

ID	L
1	13.9
1	14.6
1	17.5
⋮	
1	28.1
1	30.2
2	15.2
2	15.6
2	17.7
⋮	
2	24.1
2	27.9

Model: $X_{ji} \sim N(\mu_j, \sigma^2)$
Teste: $\mu_1 = \mu_2$

ID deler data op i to grupper

Faktor: variabel der bruges til at dele data op

En faktors værdier kaldes *faktorniveauer*

I R angives at en variabel skal bruges som faktor ved at anvende funktionen factor:

fakID=factor(ID) (liste med tekststrenge)

```
> tal=c(1,1,2,2,2,3,3)
> class(tal)
[1] "numeric"
>
> fak=factor(tal)
> class(fak)
[1] "factor"
>
> fak
[1] 1 1 2 2 2 3 3
Levels: 1 2 3
```

Indgangene i *fak* kaldes faktorværdierne (tekststrenge)

De mulige værdier kaldes *niveauer* (levels)

Generel lineær model i R:

lm(modelformel), modelformel:

respons $\sim$ sum af faktorer og regressionsvariable

Model: $X_i \sim N(\xi_i, \sigma^2)$, $i = 1, \ldots, n$, uafhængige

modelformel angiver hvor $(\xi_1, \ldots, \xi_n)$ kan variere

modelformel definerer indirekte parametrene i modellen

Tudseøgler: $(\xi_1, \ldots, \xi_n) = (\underbrace{\mu_1, \mu_1, \ldots, \mu_1}_{154}, \underbrace{\mu_2, \mu_2, \ldots, \mu_2}_{30})$

Tudseøgler: L $\sim$ fakID

Faktor og modelformel er omtalt

Næste: Data til one-way anova: effekt af lys på bladenes hårdhed

Data med tre grupper: Effekt af lys. Oecologia (1991)

Jeg benytter simulerede data der passer med oplysninger i artiklen

Tre grupper: skygge , lys mellem blade og fuld sol

Måler egenskab (hårdhed, kiloPascal) ved blade fra inga oerstediana benth (4-10 m, sydamerika)

Notation: x: målte hårdhed

Oecologia (1991) 86:552–560

Oecologia

The effects of light on foliar chemistry, growth and susceptibility of seedlings of a canopy tree to an attine ant

x_{ji}: målinger

$j = 1, 2, 3$ angiver lysgruppe

$i = 1, \ldots, n_j$ angiver prøvenummer

$n = n_1 + n_2 + n_3$

Nr (i)	$j = 1$	$j = 2$	$j = 3$
1	206.2	235.7	278.6
2	274.9	217.1	279.9
3	202.6	208.9	266.2
4	199.6	240.5	259.4
5	220.4	239.2	256.7
6	230.1	269.9	204.0
7	201.9	253.5	220.6
8	224.8	237.1	302.1
9	194.8	231.8	217.5
10	264.7	156.2	204.9

x_i: målinger, $i = 1, \ldots, n$

Gruppe: faktor

Gruppe$_i$ angiver lysgruppe hørende til måling i

Nr	Gruppe	Hårdhed
1	1	206.2
2	1	274.9
	$\vdots$	
10	1	264.7
11	2	235.7
12	2	217.1
	$\vdots$	
29	3	217.5
30	3	204.9

Enten: $X_{ji} \sim N(\mu_j, \sigma_j^2)$

Eller: $X_i \sim N(\mu_{Gruppe_i}, \sigma^2_{Gruppe_i})$

Data er normalfordelt, hver lysgruppe har sin egen midddelværdi (μ_j) og sin egen spredning (σ_j)

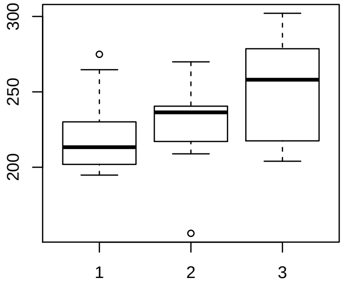
Data indlæses i R:

haardhed: vektor med målte hårdhedsværdier

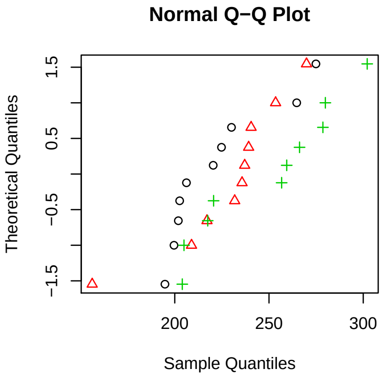
gr: vector med gruppenummer

Gruppe: faktor dannet ud fra *gr*

boxplot(haardhed∼Gruppe)



qqnormFlere(haardhed,gr)



Normalfordelingsmodel ser rimelig ud, cirka samme varians (test senere)

Data: log(længden) af blomsterblad for tre irisarter
Opgave: Opstil en statistisk model for data og lav grafiske undersøgelser

```
logLaeng=log(iris[,1])

art=iris[,5]
class(art)

boxplot()

source("../source/Rfunktioner.txt")
qqnormFlere()
```

Figurer viser at ...
Model: $\text{LogLaeng}_i \sim N(\mu_{\text{art}_i}, \sigma^2)$, $i = 1, \ldots, 150$

Data er præsenteret (hårdhed af blade for tre lysgrupper)

Næste: test for at middelværdier er ens (når varianser er ens)

Gruppe	1	2	3
Gennemsnit	222	229	249
Empirisk spredning	27.8	30.7	34.8
n	10	10	10

Hvordan tester vi tre middelværdier ens ?

Model: $\text{Haardhed}_i \sim N(\mu_{\text{Gruppe}_i}, \sigma^2)$

Modelformel: haardhed~Gruppe (Gruppe er en faktor)

Dette giver model med 3 middelværdiparametre μ_1, μ_2 og μ_3

En faktor giver et bidrag (en parameter) for hvert niveau af faktor

R bruger et generelt niveau (intercept) og forskel til dette niveau:

$$\text{intercept} = \mu_1, \qquad \text{Gruppe2} = \mu_2 - \mu_1, \qquad \text{Gruppe3} = \mu_3 - \mu_1$$

Kørsel i R: Obs: Gruppe er en faktor ikke en regressionsvariabel

```
lmUD=lm(haardhed~Gruppe)

summary(lmUD)
```

Output fra summary

```
Call:
lm(formula = haardhed ~ Gruppe)

Residuals:
   Min     1Q Median     3Q    Max 
-72.79 -20.10   4.76  15.79  53.11 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  222.000      9.874  22.483   <2e-16 ***
Gruppe2        6.990     13.964   0.501   0.6207    
Gruppe3       26.990     13.964   1.933   0.0638 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.22 on 27 degrees of freedom
Multiple R-squared:  0.1297,Adjusted R-squared:  0.06527 
F-statistic: 2.013 on 2 and 27 DF,  p-value: 0.1532
```

Model M: $E(\text{Haardhed}_i) = \xi_i = \mu_{\text{Gruppe}_i}$

$\hat{\mu}_1 = 222.00$

$\hat{\mu}_2 = 6.99 + 222.00 = 228.99$

$\hat{\mu}_3 = 26.99 + 222.00 = 248.99$

Skøn over spredning σ: $s(M) = 31.22$ (residual standard error)

$\hat{\mu}_j =$ gennemsnit af Haardhed$_i$ over den j'te gruppe

$s^2 = \frac{1}{n-3} \sum_i (\text{Haardhed}_i - \hat{\xi}_i)^2$, $\hat{\xi}_i = \hat{\mu}_{\text{Gruppe}_i}$

Opgave: find skøn over parametre i model for blomsterlængde i iris-data

logLaeng=log(iris[,1])

art=iris[,5]

summary(lm(logLaeng∼art))

Fra R-kørsel finder vi at skøn over de tre middelværdiparametre er ... og skøn over spredning σ er ...

Sceneskift

Data er analyseret via lm og summary. Næste: forstå output

Model: $X_i \sim N(\xi_i(M), \sigma^2)$, $i = 1, \ldots, n$

Model M med $d(M)$ parametre:

parametre i $\xi_i(M)$ findes ved at minimere $\sum_i \left(x_i - \xi_i(M)\right)^2$

variansskøn $s^2(M) = \frac{1}{n-d(M)} \sum_i \left(x_i - \hat{\xi}_i(M)\right)^2 = \mathrm{RSE}^2$

θ: en parameter i middelværdimodel M

$\hat{\theta} \sim N(\theta, \sigma^2 C)$, C er kendt ud fra middelværdimodel

Teste $\theta = \theta_0$: $T = \frac{\hat{\theta}-\theta_0}{\mathrm{sd}_s(\hat{\theta})} \sim t(\mathrm{df}(M))$

standard error $\mathrm{sd}_s(\hat{\theta}) = s(M)\sqrt{C}$

Konfidensinterval (95%): $\hat{\theta} \pm t_0 \cdot \mathrm{sd}_s(\hat{\theta}) = [\hat{\theta} - t_0 \cdot \mathrm{sd}_s(\hat{\theta}), \hat{\theta} + t_0 \cdot \mathrm{sd}_s(\hat{\theta})]$

$t_0 = t_{\mathrm{inv}}(0.975, \mathrm{df}(M))$, $\mathrm{df}(M) = n - d(M)$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  222.000      9.874  22.483   <2e-16
Gruppe2        6.990     13.964   0.501   0.6207
Gruppe3       26.990     13.964   1.933   0.0638
```

$\hat{\mu}_1 = 222.0, \quad \hat{\mu}_2 - \hat{\mu}_1 = 6.990, \quad \hat{\mu}_3 - \hat{\mu}_1 = 26.990$

Std.Error: standard error

t value: t-teststørrelse for hypotesen $\theta = 0$

Pr(>|t|): p-værdi $= 2(1 - t_{\text{cdf}}(|\text{t value}|, \text{df}(M)))$

Teste $\mu_3 - \mu_1 = 0$ eller $\mu_3 = \mu_1$: p-værdi $= 0.064$

```
Residual standard error: 31.22 on 27 degrees of freedom
Multiple R-squared:  0.1297,Adjusted R-squared:  0.06527
F-statistic: 2.013 on 2 and 27 DF,  p-value: 0.1532
```

Degrees of freedom: $\mathrm{df}(M) = n - d(M)$

Residual standard error: $s(M) = \sqrt{\frac{1}{n-d(M)} \sum_i (x_i - \hat{\xi}_i)^2}$

R-squared: $R^2 = 1 - \sum_i r_i^2 / \sum_i (x_i - \bar{x})^2$, $r_i = x_i - \hat{\xi}_i = \text{residual}$

lmUD=lm(...),   sumUD=summary(lmUD)

sumUD$sigma giver $s(M)$
sumUD$df[2] giver df($M$)
lmUD$residuals giver residualer $r_i = x_i - \hat{\xi}_i$
lmUD$fitted.values giver de forventede værdier $\hat{\xi}_i$

confint(lmUD) giver konfidensintervaller

```
                  2.5 %    97.5 %
(Intercept) 201.739813 242.26019
Gruppe2     -21.662231  35.64223
Gruppe3      -1.662231  55.64223
```

95%-konfidensinterval for $\delta = \mu_3 - \mu_1$: $[-1.662231, 55.64223] \approx [-2, 56]$

(OBS: vi vidste godt at nul ligger i konfidensintervallet!)

Konfidensinterval for spredning σ: Webbog afsnit 2.6

erstat s^2 med $s^2(M)$ og df med $df(M)$

Output fra summary(lm(..)) er beskrevet

Næste: Teste alle tre middelværdier ens

Parametertabel: kan kun teste parameter lig med nul

$$\text{eksempel: teste } \delta = \mu_3 - \mu_1 = 0\text{: t value} = 1.933, \text{ Pr}(>|\text{t}|) = 0.0638$$

Ønsker: test for hypotesen $H: \ \mu_1 = \mu_2 = \mu_3$

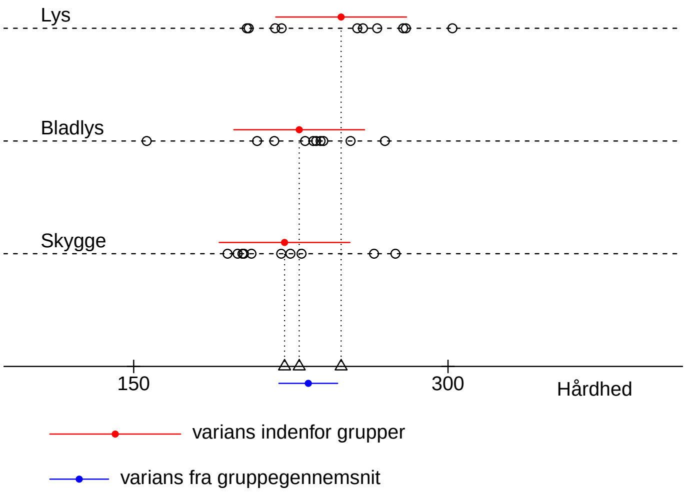
Teste reduktion fra model M_1 til model M_2 (generel notation her):

$$M_1: \ X_{ji} \sim N(\mu_j, \sigma^2), \qquad M_2: \ X_{ji} \sim N(\mu, \sigma^2)$$

eller

$$X_i \sim N(\mu_{\text{Gruppe}_i}, \sigma^2), \qquad X_i \sim N(\mu, \sigma^2)$$

Metode: sammenligne variation mellem grupper med variation indenfor grupper



Under hypotesen $\mu_1 = \mu_2 = \mu_3$ må variation mellem grupper ikke være for stor i forhold til indenfor grupper

Her: dobbeltindeksnotation (enkeltindeks senere)

x_{ji}: i'te måling i den j'te gruppe, $j = 1, \ldots, k$

$$\bar{x}_j = \frac{1}{n_j} \sum_i x_{ji}, \quad \bar{x} = \frac{1}{n} \sum_{ji} x_{ji}, \quad n = \sum_j n_j$$

Variation indenfor gruppe: $s^2(M_1) = \frac{1}{n-k} \sum_{ji} \left(x_{ij} - \bar{x}_j\right)^2$

Variation mellem grupper: $s^2(M_1, M_2) = \frac{1}{k-1} \sum_j n_j \left(\bar{x}_j - \bar{x}\right)^2$

Teststørrelse $F = \frac{s^2(M_1, M_2)}{s^2(M_1)} = \begin{cases} \text{lille} & \text{ikke kritisk for } H \\ \text{stor} & \text{kritisk for } H \end{cases}$

Teori: $s^2(M_1) \sim \sigma^2 \chi^2(n-k)/(n-k)$, $s^2(M_1, M_2) \sim \sigma^2 \chi^2(k-1)/(k-1)$

$$F \sim F(3-1, n-3), \quad p\text{-værdi} = 1 - F_{\text{cdf}}(F, k-1, n-k)$$

I model $X_i \sim N(\mu, \sigma^2)$ ved vi at

$\bar{X}$ og $\sum_i (X_i - \bar{X})^2 \sim \sigma^2 \chi^2(n-1)$ er uafhængige

I model $X_{ji} \sim N(\mu, \sigma^2)$ med $n_1 = \cdots = n_k$ har vi derfor

$\sum_j n_j (\bar{X}_j - \bar{X})^2 \sim \sigma^2 \chi^2(k-1)$, $\bar{X} = \sum_j \bar{X}_j / k$

uafhængig af $\sum_{j,i} (X_{ji} - \bar{X}_j)^2 \sim \sigma^2 \chi^2(n-k)$

Benyt funktionen anova i R:

anova(lm(haardhed∼1),lm(haardhed∼Gruppe))

```
Model 1: haardhed ~ 1
Model 2: haardhed ~ Gruppe
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     29 30249
2     27 26325  2    3924.4 2.0125 0.1532
```

"Model 1: haardhed∼1" model M_2, alle har samme middelværdi

"Model 2: haardhed∼Gruppe" model M_1, hver gruppe har sin egen middelværdi

$s^2(M_1) = \frac{26325}{27} = 975.0, \quad s^2(M_1, M_2) = \frac{3924.4}{2} = 1962.2$

$F = 1962.2/975.0 = 2.0125, \quad 1 - F_{\text{cdf}}(2.0125, 2, 27) = 0.1532$

Opgave: Undersøg om de tre arter af iris har samme middelværdi af log(længde)

```
logLaeng=log(iris[,1])

art=iris[,5]

anova(lm(logLaeng~ ),lm(logLaeng~ ))
```

Fra R-kørsel finder vi F-teststørrelsen $F = ..$, som vurderes i en $F(.,.)$-fordeling. Da p-værdien er ...

Vi har lavet test for samme middelværdi af hårdhed af blade for tre lysgrupper under forudsætning om samme varians

Næste: test for at varianser er ens (skal laves før ovenstående test)

Gruppe	1	2	3
Gennemsnit	222	229	249
Empirisk spredning	27.8	30.7	34.8
n	10	10	10

Hvordan tester vi tre varianser ens ?

Variansskøn inden for gruppe j: s_j^2 med df_j frihedsgrader, $j = 1, \ldots, k$

Fælles variansskøn: $s^2 = \frac{\sum_{j=1}^k df_j s_j^2}{df}$, $\quad df = \sum_j df_j$

Hypotese: samme varians i de k grupper:

Likelihood ratio test (opgave 4.5)

Sammenligner $\log\Big(\sum_j \frac{df_j}{df} s_j^2\Big)$ med $\sum_j \frac{df_j}{df} \log(s_j^2)$:

Teststørrelse: $\mathrm{Ba} = \frac{1}{C}\{df \cdot \log(s^2) - \sum_{j=1}^k df_j \cdot \log(s_j^2)\}$

$C = 1 + \frac{1}{3(k-1)}\{\sum_{j=1}^k \frac{1}{df_j} - \frac{1}{df}\}$

p–værdi $= 1 - \chi^2_{\mathrm{cdf}}(\mathrm{Ba}, k-1)$ (approksimativt)

Taylorudvikling: $\log(1+x) \approx x - \frac{1}{2}x^2, \;\; x = \frac{s^2}{\sigma^2} - 1$, lille

$V \sim \chi^2(f)/f$: $\;E(V) = 1, \;\; \mathrm{Var}(V) = \frac{2}{f}$

$$E\big(df \cdot \log(s^2) - \textstyle\sum_{j=1}^{k} df_j \cdot \log(s_j^2)\big) = (k-1)C + \text{restled}$$

Hvorfor Ba $\approx \chi^2(k-1)$?

tester fra k parametre ned til 1 parameter

Gruppe	1	2	3
Gennemsnit	222	229	249
Empirisk spredning	27.8	30.7	34.8
n	10	10	10

$k = 3$ grupper, $\quad df = 9 + 9 + 9 = 27$

$$s^2 = (9 \cdot 27.8^2 + 9 \cdot 30.7^2 + 9 \cdot 34.8^2)/27 = 975.4567$$

$$C = 1 + \tfrac{1}{3(3-1)}\big(1/9 + 1/9 + 1/9 - 1/27\big) = 1.0494$$

$$Ba = \frac{1}{1.0494}\big(27 \cdot \log(975.4567) - 9 \cdot \log(27.8^2) - 9 \cdot \log(30.7^2)$$

$$-9 \cdot \log(34.8^2)\big) = 0.4300$$

$$p\text{-værdi} = 1 - \chi^2_{\mathrm{cdf}}(0.4300, 3-1) = 0.81$$

R: 1-pchisq(0.4300,3)

Konklusion: data strider ikke mod samme varians af hårdhed i de 3 lysgrupper

```
bartlett.test(haardhed,gr)

Bartlett test of homogeneity of variances
data:  haardhed and gr
Bartlett's K-squared = 0.43003, df = 2, p-value = 0.8065
```

Hvis kun variansskøn s_j^2 er til rådighed: selv kode test

Med outputs fra lm: bartlett.test(list(lmUD1,lmUD2,lmUD3))

Opgave: Undersøg om der er samme varians på log(længde) i iris data

logLaeng=log(iris[,1])

art=iris[,5]

bartlett.test()

Fra R-kørsel ses, at p-værdien i Bartletts test for ens varianser (afsnit 6.5 i webbogen), hypotesen $\sigma^2_{\text{setosa}} = \sigma^2_{\text{versicolor}} = \sigma^2_{\text{virginica}}$, er ..., hvorfor data ...

One-way anova er færdigbehandlet

Næste: two-way anova

Alanin i lymfevæsken af tusindben:

køn: han / hun,   art: art1 / art2 / art3

kqn	art	Ala
han	1	21.5
han	1	19.6
han	1	20.9
han	1	22.8
han	2	14.5
han	2	17.4
han	2	15.0
han	2	17.8
han	3	16.0
han	3	20.3
han	3	18.5
han	3	19.3
hun	1	14.8
hun	1	15.6
hun	1	13.5
hun	1	16.4
hun	2	12.1
hun	2	11.4
hun	2	12.7
hun	2	14.5
hun	3	14.4
hun	3	14.7
hun	3	13.8
hun	3	12.0

kqn: køn

Ala: koncentration af Alanin

Model: $\text{Ala}_i \sim N(\mu_{\text{kqn}_i,\text{art}_i}, \sigma^2)$, $i = 1, \ldots, n$

Parametre: $\mu_{\text{han},1}, \mu_{\text{han},1}, \ldots, \mu_{\text{hun},3}, \sigma^2$

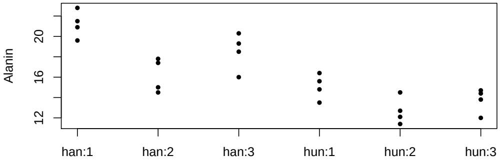
Seks grupper med hver sin middelværdi

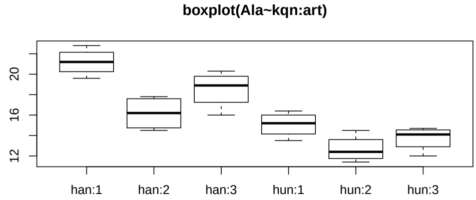
Teste additivitet: $\mu_{\text{kqn},\text{art}} = \eta_{\text{kqn}} + \zeta_{\text{art}}$

bidrag fra køn plus bidrag fra art

Fortolkning: ...

Kig på data


Alanin
12
16
20
han:1
han:2
han:3
hun:1
hun:2
hun:3


boxplot(Ala~kqn:art)
12
16
20
han:1
han:2
han:3
hun:1
hun:2
hun:3

kqn inddeler i 2 grupper

art inddeler i 3 grupper

kqn*art inddeler i $2 \cdot 3 = 6$ grupper

- kqn*art bruges i modelformel
- udenfor modelformel bruges kqn:art

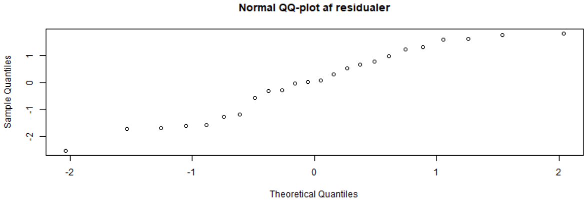
Model M_0: $\text{Ala}_i \sim N(\mu_{\text{kqn}_i,\text{art}_i}, \sigma^2_{\text{kqn}_i,\text{art}_i})$, $i = 1, \ldots, n$

Hypotese om fælles varians: $\sigma^2_{\text{han},1} = \sigma^2_{\text{han},2} = \cdots = \sigma^2_{\text{hun},3}$

Bartletts test: bartlett.test(Ala,kqn:art) (kqn og art er faktorer)

Ba = 0.81732, df = 5, p-value = 0.9759

Data strider ikke mod samme varians i de 6 grupper



Model M_1: hver af de 6 kombinationer af køn og art har sin egen middelværdi, $d(M_1) = 6$

køn	art 1	art 2	art 3
han	$\mu_{\text{han},1}$	$\mu_{\text{han},2}$	$\mu_{\text{han},3}$
hun	$\mu_{\text{hun},1}$	$\mu_{\text{hun},2}$	$\mu_{\text{hun},3}$

Model M_2 (additive model): middelværdi består af et bidrag fra køn plus et bidrag fra art, $d(M_2) = 2 + 3 - 1 = 4$

køn	art 1	art 2	art 3
han	$\eta_{\text{han}} + \zeta_1$	$\eta_{\text{han}} + \zeta_2$	$\eta_{\text{han}} + \zeta_3$
hun	$\eta_{\text{hun}} + \zeta_1$	$\eta_{\text{hun}} + \zeta_2$	$\eta_{\text{hun}} + \zeta_3$

Forskel mellem arter under additive model:

køn	art 2– art 1	art 3– art 1
han	$\zeta_2 - \zeta_1$	$\zeta_3 - \zeta_1$
hun	$\zeta_2 - \zeta_1$	$\zeta_3 - \zeta_1$

$$\delta_2 = \zeta_2 - \zeta_1, \;\; \delta_3 = \zeta_3 - \zeta_1$$

Forskel mellem arter er den samme for de to køn

Forskel mellem køn:

køn	art 1	art 2	art 3
hun–han	$\eta_{\text{hun}} - \eta_{\text{han}}$	$\eta_{\text{hun}} - \eta_{\text{han}}$	$\eta_{\text{hun}} - \eta_{\text{han}}$

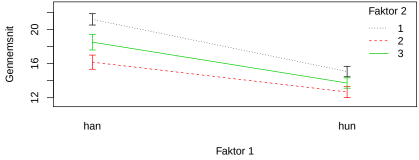
$$\delta_{\text{hun}} = \eta_{\text{hun}} - \eta_{\text{han}}$$

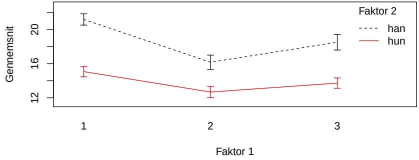
Forskel mellem køn er den samme for det tre arter

køn	art 1	art 2	art 3
han	μ	$\mu + \delta_2$	$\mu + \delta_3$
hun	$\mu + \delta_{\text{hun}}$	$\mu + \delta_2 + \delta_{\text{hun}}$	$\mu + \delta_3 + \delta_{\text{hun}}$

4 parametre!

Gennemsnit plus minus standard error:





Middelværdi $\eta_{\text{kqn}} + \zeta_{\text{art}}$ giver parallelle kurver
R: additivitetsPlot(kqn,art,Ala), additivitetsPlot(art,kqn,Ala) eller interaction.plot

Tilvækst i tænder på grise, 3 grupper mht vitaminDosis, 2 grupper mht fodringsMetode

```
len=ToothGrowth[,1]

M=ToothGrowth[,2]

D=factor(ToothGrowth[,3])

source("Rfunktioner.txt")

additivitetsPlot(M,D,len)

additivitetsPlot(D,M,len)
```

Vi har kigget på data og den additive model

Næste: analyse i R

kqn, art: begge faktorer: Se webbog afsnit 4.1

Modelformel: kqn*art giver ny faktor der deler op efter både kqn og art
Direkte i kommandovindue: kqn:art

R modelformel:
kqn*art samme som kqn+art+kqn*art samme som kqn+art+kqn:art

Parametrisering i R:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} =$$

$$\begin{bmatrix} a\, a\, a \\ a\, a\, a \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ d{-}a & d{-}a & d{-}a \end{bmatrix} + \begin{bmatrix} 0 & b{-}a & c{-}a \\ 0 & b{-}a & c{-}a \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & e{-}b{-}d{+}a & f{-}c{-}d{+}a \end{bmatrix}$$

Intercept kqnhun art2 art3 kqnhun:art2 kqnhun:art3

R: kqn + art + kqn:art

summary(lm(Ala∼kqn*art)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2000     0.7267  29.175  < 2e-16 ***
kqnhun       -6.1250     1.0276  -5.960 1.22e-05 ***
art2         -5.0250     1.0276  -4.890 0.000118 ***
art3         -2.6750     1.0276  -2.603 0.017983 *
kqnhun:art2   2.6250     1.4533   1.806 0.087631 .
kqnhun:art3   1.3250     1.4533   0.912 0.373967

Residual standard error: 1.453 on 18 degrees of freedom
```

Gennemsnit for (han,art1): 21.20, (hun,art2): 21.20-6.125-5.0250+2.625= 12.675

Skal vi acceptere additive model svarende til:

kqnhun:art2=0 og kqnhun:art3=0 ?

anova(lm(Ala∼kqn+art),lm(Ala∼kqn*art))

```
Model 1: Ala ~ kqn + art
Model 2: Ala ~ kqn * art
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     20 44.9083
2     18 38.0175  2    6.8908 1.6313 0.22331 *
```

Model 2 (kalder jeg M_1): $\text{Ala}_i \sim N(\mu_{\text{kqn}_i,\text{art}_i}, \sigma^2)$

Model 1 (kalder jeg M_2): $\text{Ala}_i \sim N(\eta_{\text{kqn}} + \zeta_{\text{art}}, \sigma^2)$

Teste additive model: teste reduktion fra M_1 til M_2:
$F = 1.6313$, p-værdi=0.22331, data strider ikke mod reduktionen

Sammenligne "variation indefor gruppe" med "variation mellem grupper" ?

summary(lm(Ala~kqn+art))

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.5417     0.6117  33.579  < 2e-16 ***
kqnhun       -4.8083     0.6117  -7.860 1.53e-07 ***
art2         -3.7125     0.7492  -4.955 7.62e-05 ***
art3         -2.0125     0.7492  -2.686   0.0142 *
```

Intercept: $\eta_{\text{han}} + \zeta_1$

kqnhun $= \eta_{\text{hun}} - \eta_{\text{han}}$, art2 $= \zeta_2 - \zeta_1$, art3 $= \zeta_3 - \zeta_1$

Uanset art så er middelværdien for han 4.8 større end middelværdi for hun

Uanset køn så er middelværdien for art1 2.0 større end middelværdi for art3

confint(lm(Ala~kqn+art))

```
                2.5 %     97.5 %
(Intercept) 19.265582 21.8177515
kqnhun      -6.084418 -3.5322485
art2        -5.275378 -2.1496217
art3        -3.575378 -0.4496217
```

Data, lm og Testtabel er vist

Næste forelæsning: forstå testtabel og F-test generelt

Næste hvis tid: parret t-test

Parret t-test som two-way anova

	Høstudbytte		
Mark	Ny Såmaskine	Gængs Såmaskine	Forskel d
1	8.0	5.6	2.4
2	8.4	7.4	1.0
3	8.0	7.3	0.7
⋮			
9	5.6	5.5	0.1
10	6.2	5.5	0.7

$E(X_i) = \xi_i = \eta_{\text{Mark}_i} + \zeta_{\text{Maskine}_i}$, Hypotese: $\delta = \zeta_{\text{Ny}} - \zeta_{\text{Gængs}} = 0$

Parret t-test: $t = \frac{\bar{d}}{s_d/\sqrt{10}} = 3.2143$, p-værdi $= 0.0106$

Two-way anova: summary(lm(Hoest∼Mark+Maskine))

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.3850     0.4282  14.911 1.19e-07 ***
Mark2         1.1000     0.5774   1.905   0.0892 .
Mark3         0.8500     0.5774   1.472   0.1751
.
.
.
Mark10       -0.9500     0.5774  -1.645   0.1343
MaskineNy     0.8300     0.2582   3.214   0.0106 *
```

Slut for i dag

Eller: hvis tid

```
Gruppe=factor(rep(c(1,2),c(10,20)))

blodtryk=80+10*rnorm(30)+rep(c(0,5),c(10,20))
```

Benyt lm og summary til at undersøge om der er forskel i blodtryk for de to grupper