Introduktion til Statistik

Forelæsning 11: Tovejs variansanalyse, ANOVA

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 009
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2017

Kapitel 8: Tovejs variansanalyse (tovejs ANOVA)

k UAFHÆNGIGE grupper og blokdesign der giver to faktorer

- Test om middelværdi for om mindst en gruppe er forskellig de andre andres
- Model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

Specifikke metoder, tovejs variansanalyse:

- ANOVA-tabel: $SST = SS(Tr) + SS(Bl) + SSE$
  - SST, $SS(Tr)$ og $SS(Bl)$ beregnes som ved envejs ANOVA
  - $SSE = SST - SS(Tr) - SS(Bl)$
- F-test
- Post hoc test(s): Parvise t-test med poolet varians estimat
  - Hvis planlagt på forhånd, så uden Bonferroni korrektion
  - Hvis alle sammenligninger udføres, så med Bonferroni korrektion

Chapter 8: Two-way Analysis of Variance

k INDEPENDENT treatments and block design give two factors

- Test if mean for at least one group is different from the others
- Model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

Specific methods, two-way analysis of variance:

- ANOVA-table: $SST = SS(Tr) + SS(Bl) + SSE$
  - SST, $SS(Tr)$ and $SS(Bl)$ calculated as in one-way ANOVA
  - $SSE = SST - SS(Tr) - SS(Bl)$
- F-test
- Post hoc test(s): pairwise t-test with pooled variance estimate
  - If planned on beforehand, then without Bonferroni correction
  - If all samples are compared, then with Bonferroni correction

Oversigt

Tovejs variansanalyse - eksempel

- Samme data som for envejs, dog ved vi nu at forsøget var inddelt i blokke

	Behandling A	Behandling B	Behandling C
Blok 1	2.8	5.5	5.8
Blok 2	3.6	6.3	8.3
Blok 3	3.4	6.1	6.9
Blok 4	2.3	5.7	6.1

- f.eks:
  - tre *grupper* på fire *blokke*
  - eller tre *behandlinger* på fire *personer*
  - eller tre *afgrøder* på fire *marker* (deraf blokke)
  - eller anden lignende opdeling

- *Envejs* ANOVA: *Completely randomized design*
- *Tovejs* ANOVA: *Randomized block design*

Tovejs variansanalyse - eksempel

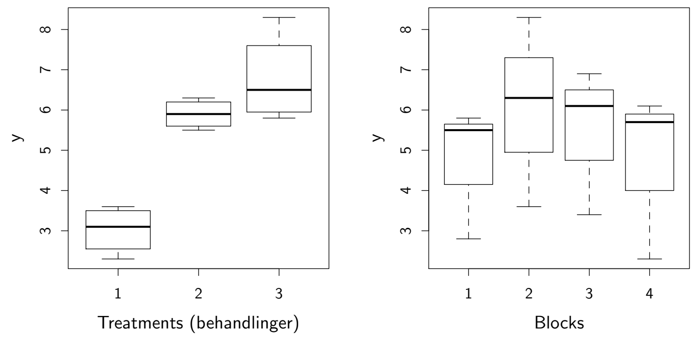
- Samme data som for envejs, dog ved vi nu at forsøget var udført på fire personer

	Behandling A	Behandling B	Behandling C
Person 1	2.8	5.5	5.8
Person 2	3.6	6.3	8.3
Person 3	3.4	6.1	6.9
Person 4	2.3	5.7	6.1

- Besvarer:
  - Er der signifikant forskel på middelværdien af behandling A, B og C?
- Variansanalyse (ANOVA) kan anvendes til analysen såfremt observationerne efter gruppering (residualerne) kan antages at være i.i.d. normalfordelte (dog med mange samples dækker CLT)


```
## Observationer
y <- c(2.8, 3.6, 3.4, 2.3,
       5.5, 6.3, 6.1, 5.7,
       5.8, 8.3, 6.9, 6.1)

## Behandlinger (grupper, afgrøder, ...)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

## Blokke (personer, marker, ...)
block <- factor(c(1, 2, 3, 4,
                  1, 2, 3, 4,
                  1, 2, 3, 4))

## Til formler senere
(k <- length(unique(treatm)))
(l <- length(unique(block)))

## Plots
par(mfrow=c(1,2))

## Punkterne inddelt ved behandlinger
plot(treatm, y, xlab="Treatments", ylab="y")
## Punkterne inddelt ved blokke
plot(block, y, xlab="Blocks", ylab="y")


## Plot box-plots inddelt ved behandlinger
plot(treatm, y, xlab="Treatments", ylab="y")
## Plot box-plots inddelt ved blokke
plot(block, y, xlab="Blocks", ylab="y")
```


Spørgsmål signifikant effekt Socrative.com, room: PBAC

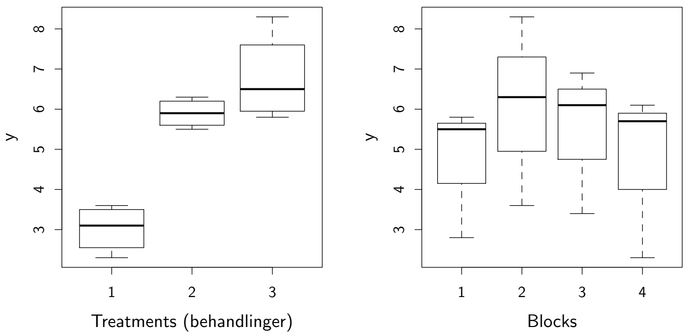


Tror du at vi vil påvise en signifikant forskel på (mindst en af) behandlingerne?

A: Ja B: Nej C: Ved ikke

Svar A: Ja, ses tydeligt på plottet (dog kunne der være få målinger), og der er en signifikant effekt af behandlingerne

Spørgsmål signifikant effekt Socrative.com, room: PBAC



Tror du at vi vil påvise en signifikant forskel på blokkene (personer)?

A: Ja B: Nej C: Ved ikke

Svar B: Nej, der ses ikke umiddelbart en signifikant effekt for personer, MEN der er en signifikant effekt! (dette ses ikke på plottet, kan først 'ses' efter blokkene har forklaret en del af variansen)

Tovejs variansanalyse, model

- Opstil en model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

hvor afvigelserne

$$\varepsilon_{ij} \sim N(0, \sigma^2) \text{ og i.i.d.}$$

- μ er middelværdi for alle målinger
- α_i angiver effekt for behandling i
- β_j angiver niveau for blok j
- der er k behandlinger og l blokke
- j tæller fra 1 til l (målinger for behandling i)

Estimater af parametrene i modellen

- Vi kan beregne estimater af parametrene ($\hat{\mu}$ og $\hat{\alpha}_i$, og $\hat{\beta}_j$)

$$\hat{\mu} = \bar{y} = \frac{1}{k \cdot l} \sum_{i=1}^{k} \sum_{j=1}^{l} y_{ij}$$

$$\hat{\alpha}_i = \left(\frac{1}{l} \sum_{j=1}^{l} y_{ij} \right) - \hat{\mu}$$

$$\hat{\beta}_j = \left(\frac{1}{k} \sum_{i=1}^{k} y_{ij} \right) - \hat{\mu}$$

```
## Sample mean
(muHat <- mean(y))
## Sample mean for hver behandling
(alphaHat <- tapply(y, treatm, mean) - muHat)
## Sample mean for hver blok
(betaHat <- tapply(y, block, mean) - muHat)
```


Tovejs variansanalyse, opspaltning og ANOVA tabellen

- Med modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- kan den totale variation i data opspaltes:

$$SST = SS(Tr) + SS(Bl) + SSE$$

- 'Tovejs' hentyder til, at der er to faktorer i forsøget
- Metoden kaldes variansanalyse, fordi testningen foregår ved at sammenligne varianser

Formler for kvadratafvigelsessummer

- Kvadratafvigelsessum ("den totale varians") (samme som for envejs)

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij} - \hat{\mu})^2$$

- Kvadratafvigelsessum for behandling ("Varians forklaret ved gruppering i behandlinger")

$$SS(Tr) = l \cdot \sum_{i=1}^{k} \hat{\alpha}_i^2$$

Formler for kvadratafvigelsessummer

- Kvadratafvigelsessum for blokke (personer) ("Varians forklaret ved gruppering i blokke")

$$SS(Bl) = k \cdot \sum_{j=1}^{l} \hat{\beta}_j^2$$

- Kvadratafvigelsessum af residualer ("Varians tilbage efter model")

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu})^2$$

Tovejs ANOVA: Hypotese om forskellig effekt af behandling

- Vi vil nu sammenligne (flere end to) middelværdier $\mu + \alpha_i$ i modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- Opstil hypotesen

$$\begin{aligned} H_{0,\text{Tr}}: &\quad \alpha_i = 0 \quad \text{for alle } i \\ H_{1,\text{Tr}}: &\quad \alpha_i \neq 0 \quad \text{for mindst et } i \end{aligned}$$

- Under $H_{0,\text{Tr}}$ følger

$$F_{\text{Tr}} = \frac{SS(\text{Tr})/(k-1)}{SSE/((k-1)(l-1))}$$

en F-distribution med $k-1$ og $(k-1)(l-1)$ frihedsgrader

Tovejs ANOVA: Hypotese om forskelligt niveau for personer (blokke)

- Vi vil nu sammenligne (flere end to) middelværdier $\mu + \beta_i$ i modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- Opstil hypotesen

$$\begin{aligned} H_{0,\text{Bl}}: &\quad \beta_i = 0 \quad \text{for alle } i \\ H_{1,\text{Bl}}: &\quad \beta_i \neq 0 \quad \text{for mindst et } i \end{aligned}$$

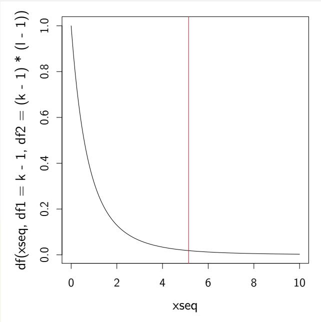
- Under $H_{0,\text{Bl}}$ følger

$$F_{\text{Bl}} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}$$

en F-distribution med $l-1$ og $(k-1)(l-1)$ frihedsgrader

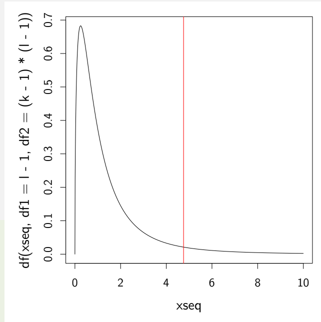
Eksempel: F-fordeling og hypotese for behandlinger

```
## Husk, dette er under H0 (altså vi regner som om H0 er sand):
## Sekvens til plot
xseq <- seq(0, 10, by=0.01)
## Plot F fordelingens tæthedsfunktion
plot(xseq, df(xseq, df1=k-1, df2=(k-1)*(l-1)), type="l")
## Kritisk værdi for signifikans niveau 5 pct.
cr <- qf(0.95, df1=k-1, df2=(k-1)*(l-1))
## Tegn den i plottet
abline(v=cr, col="red")
## Test statistikkens værdi:
## Værdien
(Ftr <- (SSTr/(k-1)) / (SSE/((k-1)*(l-1))))
## p-værdien er da
(1 - pf(Ftr, df1=k-1, df2=(k-1)*(l-1)))
```



F-fordeling og hypotese for blokke

```
## Husk, dette er under H0 (altså vi regner som om H0 er sand):
## Sekvens til plot
xseq <- seq(0, 10, by=0.01)
## Plot F fordelingens tæthedsfunktion
plot(xseq, df(xseq, df1=l-1, df2=(k-1)*(l-1)), type="l")
## Kritisk værdi for signifikans niveau 5 pct.
cr <- qf(0.95, df1=l-1, df2=(k-1)*(l-1))
## Tegn den i plottet
abline(v=cr, col="red")
## Test statistikkens værdi:
## Værdien
(Fbl <- (SSBl/(l-1)) / (SSE/((k-1)*(l-1))))
## p-værdien er da
(1 - pf(Fbl, df1=l-1, df2=(k-1)*(l-1)))
```



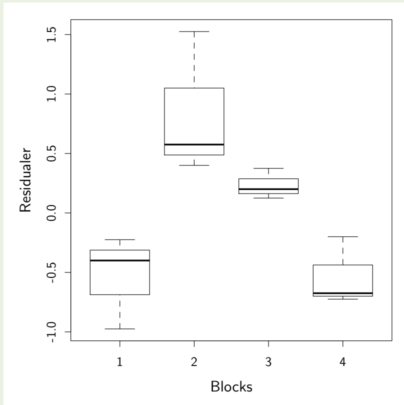
Variansanalysetabel

Variations-kilde	Friheds-grader	Kvadrat-afvi. sum	Gns. kvadratafv. sum	Test-størrelse F	p-værdi
Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
Behandling	$k-1$	$SS(Tr)$	$MS(Tr)=\frac{SS(Tr)}{k-1}$	$F_{Tr}=\frac{MS(Tr)}{MSE}$	$P(F>F_{Tr})$
Block	$l-1$	$SS(Bl)$	$MS(Bl)=\frac{SS(Bl)}{l-1}$	$F_{Bl}=\frac{MS(Bl)}{MSE}$	$P(F>F_{Bl})$
Residual	$(k-1)(l-1)$	SSE	$MSE=\frac{SSE}{(k-1)(l-1)}$		
Total	$n-1$	SST			

```
anova(lm(y ~ treatm + block))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## treatm     2  30.79   15.40   74.40 5.8e-05 ***
## block      3   3.95    1.32    6.37   0.027 *
## Residuals  6   1.24    0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Prøv at se sammenhængen med blokke efter varians af behandlinger er forklaret

```
## Se sammenhængen mellem blokke og residualerne efter behandlingerne
fit <- lm(y ~ treatm)
plot(block, fit$residuals, xlab="Blocks", ylab="Residualer")
```

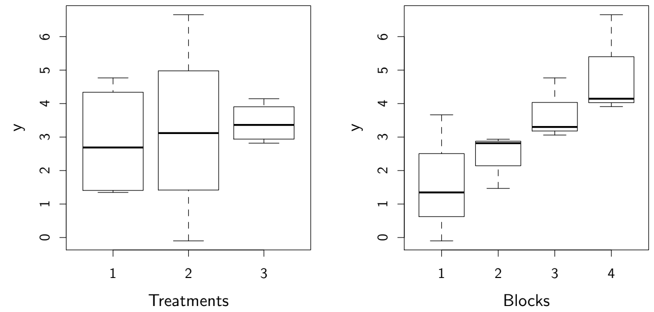


QUIZ lidt om ANOVA og hypotesetest

Simuler data fra to-vejs model (behandlinger og blokke):

```
## Sæt først behandlingernes middelværdier: ens
alpha <- c(4, 4, 4)
## Sæt først blokkenes middelværdier: ens
beta <- c(-1, -1, -1, -1)
## Antal behandlinger og antal blokke
k <- length(alpha)
l <- length(beta)
## Simuler med normalfordelte afvigelser
y <- rep(alpha, each=l) + rep(beta, k) + rnorm(k*l, sd=2)
## Indsæt i dataframe
D <- data.frame(y, treatm=factor(rep(1:k, each=l)), block=factor(rep(1:l, k)))
D
## Plots
par(mfrow=c(1,2))
## Plot box-plots inddelt ved behandlinger
plot(D$treatm, D$y, xlab="Treatments", ylab="y", type='p')
## Plot box-plots inddelt ved blokke
plot(D$block, D$y, xlab="Blocks", ylab="y", type='p')
```


ANOVA og hypotesetest quiz Socrative.com, room: PBAC

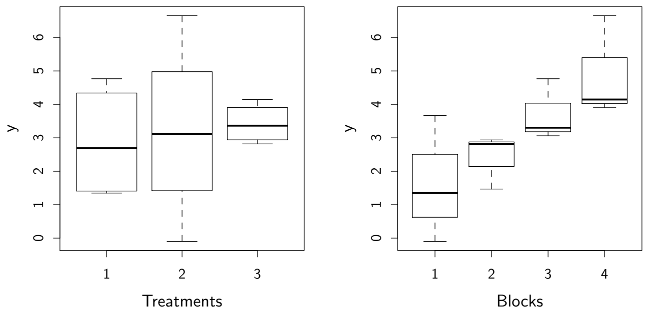


Burde vi nu påvise en signifikant effekt her ($\alpha = 0.05$)?

A: Ja B: Nej C: Ved ikke

Svar B: Nej, der er ikke forskel på middelværdierne, så vi burde ikke påvise en signifikant effekt

ANOVA og hypotesetest quiz Socrative.com, room: PBAC



Hver gang vi gentager eksperimentet og testen nu, hvad er da sandsynligheden for vi påviser en signifikant effekt ved signifikansniveau $\alpha = 0.05$?

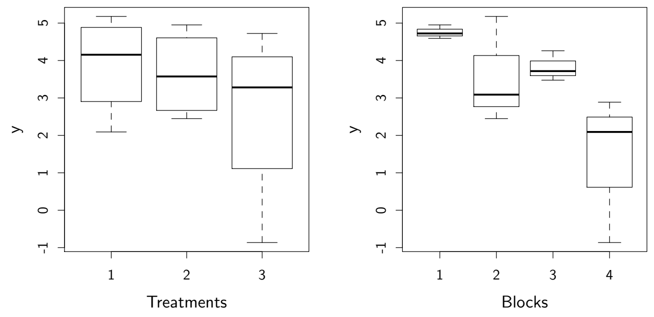
A: 1% B: 5% C: 95% D: 99% E: Ved ikke

Svar B: 5% $= \alpha$ er sandsynlighed for at påvise signifikant effekt, når der faktisk ikke er nogen effekt (Type I fejl)

Undersøg hvor ofte man laver en Type I fejl

```
## Antal gentageleser
nRep <- 10000
signifEff <- logical(nRep)
##
for(i in 1:nRep){
  print(i)
  ## Simuler med normalfordelte afvigelser
  D$y <- rep(alpha, each=l) + rep(beta, k) + rnorm(k*l, sd=2)
  ## Er der påvist en signifikant effekt?
  ans <- anova(lm(y ~ treatm + block, data=D))
  signifEff[i] <- ans[1,"Pr(>F)"] < 0.05
}
## Ved hvor stor en andel blev der påvist signifikant effekt?
sum(signifEff)/nRep

## Faktisk burde treatm fjernes når den er ikke-signifikant
```

ANOVA og hypotesetest quiz Socrative.com, room: PBAC



Vil vi sjældnere lave fejl hvis standardafvigelsen på afvigelserne ($\varepsilon_i \sim N(0,\sigma^2)$) gøres mindre?

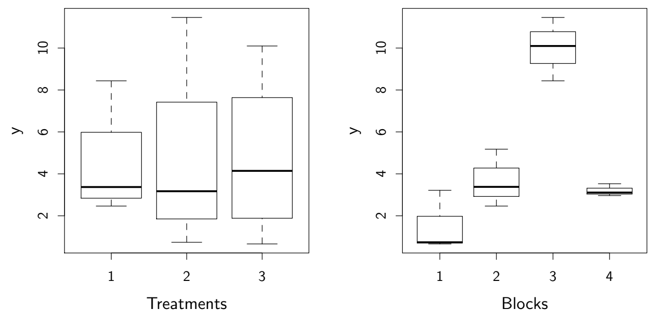
A: Ja B: Nej C: Ved ikke

Svar B: Nej, når der ikke er nogen effekt er det kun signifikansniveaut α, der bestemmer sandsynligheden for at tage fejl (Type I fejl)

Ændre middelværdi for en blok, så der nu simuleres med en tydelig effekt

```
## Ændre middelværdi for en blok, så der nu simuleres med en tydelig effekt

## Sæt først behandlingernes middelværdier: ens
alpha <- c(4, 4, 4)
## Sæt først blokkenes middelværdier: sæt en højere
beta <- c(-1, -1, 5, -1)
## Simuler med normalfordelte afvigelser
D$y <- rep(alpha, each=l) + rep(beta, k) + rnorm(k*l, sd=2)

## Plots
par(mfrow=c(1,2))
## Plot box-plots inddelt ved behandlinger
plot(D$treatm, D$y, xlab="Treatments", ylab="y", type='p')
## Plot box-plots inddelt ved blokke
plot(D$block, D$y, xlab="Blocks", ylab="y", type='p')
```


ANOVA og hypotesetest quiz Socrative.com, room: PBAC

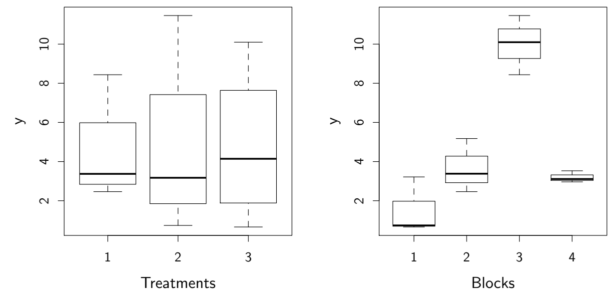


Een middelværdi i `beta` er sat til 5, bør vi da påvise en signifikant effekt?

A: Ja B: Nej C: Ved ikke

Svar A: Ja, nu er der forskel på middelværdier (der er en effekt) og derfor bør vi påvise en signifikant effekt

ANOVA og hypotesetest quiz Socrative.com, room: PBAC



Påvirker standardafvigelsen på fejlene nu hvor ofte vi ikke får påvist en signifikant effekt?

A: Ja B: Nej C: Ved ikke

Svar A: Ja, nu er der en effekt, derfor kan vi lave en Type II fejl (dvs. ikke påvise effekten, selvom den er der). Sandsynligheden for at lave en Type II fejl, er $1-\beta$ (hvor β er testens styrke: Sandsynligheden for at påvise en signifikant effekt, når den er der). FORDI, hvis σ bliver mindre, så detekteres effekten nemmere (tænk bare på, at spredningen i box-plottet bliver mindre, så ses effekten tydeligere).

Post hoc konfidensinterval

- Som ved envejs, skift $(n-k)$ frihedsgrader ud med $(k-1)(l-1)$ (og brug MSE fra tovejs).
- Gøres med enten behandlinger eller blokke
- En enkelt forudplanlagt sammenligning af forskelle på behandling i og j findes ved

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

  hvor $t_{1-\alpha/2}$ er fra t-fordelingen med $(k-1)(l-1)$ frihedsgrader.
- Hvis alle kombinationer af parvise konfidensintervaller brug formlen M gange, men med $\alpha_{\text{Bonferroni}} = \alpha/M$

Post hoc parvis hypotesetest

- En enkelt forudplanlagt hypotesetest på α signifikansniveau om forskel af behandling i og j

$$H_0:\ \mu_i = \mu_j,\ \ H_1:\ \mu_i \neq \mu_j$$

  udføres ved

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$$

  og

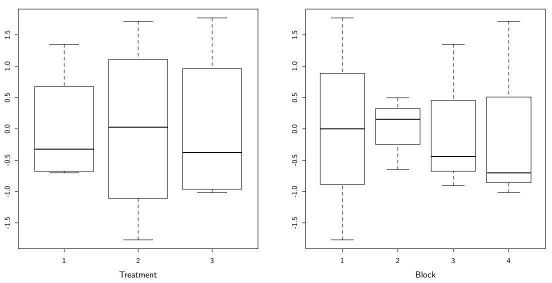
$$p\text{-value} = 2P(t > |t_{\text{obs}}|)$$

  hvor t-fordelingen med $(k-1)(l-1)$ frihedsgrader anvendes
- Hvis alle $M = k(k-1)/2$ kombinationer af hypotesetests: korrigeret signifikans niveau $\alpha_{\text{Bonferroni}} = \alpha/M$

Varians homogenitet

Se på box-plot om spredning af residualer ser ud til at afhænge af gruppen

```
## Gem fittet
fit <- lm(y ~ treatm + block)
## Box plot
par(mfrow=c(1,2))
plot(treatm, fit$residuals, y, xlab="Treatment")
## Box plot
plot(block, fit$residuals, xlab="Block")
```



Normalfordelingsantagelse

Se på qq-normal plot

```
## qq-normal plot af residualer
qqnorm(fit$residuals)
qqline(fit$residuals)

## Eller med et Wally plot
require(MESS)
qqwrap <- function(x, y, ...){
  stdy <- (y-mean(y))/sd(y)
  qqnorm(stdy, main="", ...)
  qqline(stdy)
}
## Kan vi se et afvigende qq-norm plot?
wallyplot(fit$residuals, FUN = qqwrap)
```