Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 12: Tovejs variansanalyse, ANOVA

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: jkmo@dtu.dk

Oversigt

Oversigt

Motiverende eksempel - sovemedicin

Forskel på sovemedicin?

I et studie er man interesseret i at sammenligne 2 sovemidler A og B. For 10 testpersoner har man fået følgende resultater, der er givet i forlænget søvntid (i timer) (Forskellen på effekten af de to midler er angivet):

Stikprøve, $n = 10$:

person	A	B	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

Parret setup og analyse = one-sample analyse

```
x1=c(.7,-1.6,-.2,-1.2,-1,3.4,3.7,.8,0,2)
x2=c(1.9,.8,1.1,.1,-.1,4.4,5.5,1.6,4.6,3.4)
dif=x2-x1
t.test(dif)

##
##  One Sample t-test
##
## data:  dif
## t = 4.67, df = 9, p-value = 0.0012
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.86133 2.47867
## sample estimates:
## mean of x
##      1.67
```

Parret setup og analyse = one-sample analyse

```
t.test(x2, x1, paired=TRUE)

##
##  Paired t-test
##
## data:  x2 and x1
## t = 4.67, df = 9, p-value = 0.0012
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  0.86133 2.47867
## sample estimates:
## mean of the differences
##                    1.67
```

Tovejs variansanalyse - eksempel

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

	Gruppe A	Gruppe B	Gruppe 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

  - dvs. tre *grupper* på fire *blokke*
  - el. tre *behandlinger* på fire *personer*
  - el. tre *afgrøder* på fire *marker* (deraf blokke)
  - el. lign.

- *Envejs* vs. *tovejs* ANOVA

- *Completely randomized design* vs. *Randomized block design*

Tovejs variansanalyse - eksempel

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

	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

- Besvar: Er der signifikant forskel (i middel) på grupperne A, B og C?

- Variansanalyse (ANOVA) kan anvendes til analysen såfremt observationerne i hver gruppe kan antages at være normalfordelte (dog med mange samples dækker CLT)

```r
##################################
## Input data og plot

## 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)))

## [1] 3

(l <- length(unique(block)))

## [1] 4
```

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Tovejs variansanalyse, model

- Opstil en model

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

hvor afvigelsen

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

  - μ er middelværdi for alle målinger
  - α_i angiver effekt for behandling i
  - β_j angiver niveau for blok i
  - der er k behandlinger og l blokke

Estimater af parametrene i modellen

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

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

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

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

```
########################################
## Beregn estimater af parametrene i modellen
## Sample mean
 (muHat <- mean(y))

## [1] 5.2333

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

##        1        2        3
## -2.20833  0.66667  1.54167

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

##        1        2        3        4
## -0.53333  0.83333  0.23333 -0.53333
```

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Tovejs variansanalyse, opspaltning og ANOVA tabellen

- Med modellen

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

- 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 af behandlingdel af modellen")

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

Formler for kvadratafvigelsessummer

- Kvadratafvigelsessum for blokke (personer) ("Varians forklaret af blokdel af modellen")

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

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

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

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Tovejs ANOVA: hypotese om forskellig effekt af behandling

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

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

- Opstil hypotesen

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

- Under $H_{0,Tr}$ følger

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

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

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

- Opstil hypotesen

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

- Under $H_{0,Bl}$ følger

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

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

F-fordeling og hypotese for behandlinger

```
####################################
## Plot F fordeling og se kritisk værdi for behandlinger

## Husk, dette er under H0 (altså vi regner som om H0 er sand):
## Sekvens til plot
xseq <- seq(0, 10, by=0.1)
## 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

```
###################################
## Plot F fordeling og se kritisk værdi for blokke

## Husk, dette er under H0 (altså vi regner som om H0 er sand):
## Sekvens til plot
xseq <- seq(0, 10, by=0.1)
## 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	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{Tr}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{Tr}})$
Block	$l-1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{\text{Bl}} = \frac{MS(Bl)}{MSE}$	$P(F > F_{\text{Bl}})$
Residual	$(k-1)(l-1)$	SSE	$MSE = \frac{SSE}{(k-1)(l-1)}$		
Total	$n-1$	SST			

```
##################################
## Alt dette beregnes med anova() og lm()

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

Oversigt

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

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

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

Post hoc parvis hypotesetest

- In 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

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

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 $\alpha_{\text{Bonferroni}} = \alpha/M$

Oversigt

Varians homogenitet

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

```
###################################
## Check antagelse om homogen varians af afvigelserne, ved at ana

## 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

```
################################
## Check antagelse om normalfordelte afvigelser, ved at analysere

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

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

Oversigt

Eksempel: Skive Fjord

- Undersøg om der er forskel på vandtemperaturen i forskellige år
- Undersøg om der er forskel på algekoncentrationen i forskellige år

Oversigt

The general linear model - intro

- The classical GLM leads to a unique way of describing the variations of experiments with a *continuous* variable.
- The classical GLM's include
    - Regression analysis
    - Analysis of variance - ANOVA
    - Analysis of covariance - ANCOVA
- The residuals are assumed to follow a multivariate normal distribution in the classical GLM.

The general linear model - intro

- Classical GLM's are naturally studied in the framework of the multivariate normal distribution.
- We will consider the set of n observations as a sample from a n-dimensional normal distribution.
- Under the normal distribution model, maximum-likelihood estimation of mean value parameters may be interpreted geometrically as *projection* on an appropriate subspace.

General Linear Model

- A general linear model is:

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$$

Example (Two-way ANOVA):

	B_1	B_2	B_3
A_1	y_{11}	y_{12}	y_{13}
A_2	y_{21}	y_{22}	y_{23}

Two way ANOVA (the model):

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{ i.i.d. } N(0, \sigma^2), \quad i = 1, 2, \quad j = 1, 2, 3.$$

An expanded view of this model is:

$$\begin{array}{lllllllll}
y_{11} = \mu & + \alpha_1 & & + \beta_1 & & & + \varepsilon_{11} \\
y_{21} = \mu & & + \alpha_2 & + \beta_1 & & & + \varepsilon_{21} \\
y_{12} = \mu & + \alpha_1 & & & + \beta_2 & & + \varepsilon_{12} \\
y_{22} = \mu & & + \alpha_2 & & + \beta_2 & & + \varepsilon_{22} \\
y_{13} = \mu & + \alpha_1 & & & & + \beta_3 & + \varepsilon_{13} \\
y_{23} = \mu & & + \alpha_2 & & & + \beta_3 & + \varepsilon_{23}
\end{array}$$

The exact same in matrix notation (though not identifiable):

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

The default in R would be

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

- $\mathbf{y}$ is the vector of all observations
- $\mathbf{X}$ is known as the *design matrix*
- $\boldsymbol{\beta}$ is the vector of parameters
- $\boldsymbol{\varepsilon}$ is a vector of independent $N(0, \sigma^2)$ "measurement noise"
  - The vector $\boldsymbol{\varepsilon}$ is said to follow a *multivariate normal distribution*
  - Mean vector $\mathbf{0}$
  - Covariance matrix $\sigma^2\mathbf{I}$
  - Written as: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ specifies the model, and everything can be calculated from $\mathbf{y}$ and $\mathbf{X}$.

Construction of the design matrix

In a general linear model (with both factors and covariates), it is surprisingly easy to construct the design matrix $\mathbf{X}$.

- For each factor: Add one column for each level, with ones in the rows where the corresponding observation is from that level, and zeros otherwise.
- For each covariate: Add one column with the measurements of the covariate.
- Remove linear dependencies (if necessary)

Example: linear regression:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon$$

In matrix notation:

$$\mathbf{y} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \boldsymbol{\varepsilon}$$

Oversigt