Kursus 02402/02323 Introducerende Statistik

Forelæsning 11: Tovejs variansanalyse, ANOVA

Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: perbb@dtu.dk

Oversigt

Oversigt

Udvikling af TV hos Bang & Olufsen

Lyd- og billedkvalitet måles med det menneskelige måleinstrument:

Vi har udviklet et værktøj, som bla. bruges af B&O til variansanalyse: PanelCheck (*Viser Panelcheck programmet med TV data*)

Bang & Olufsen data i R:

```
##  # Getting the Bang and Olufsen data from the lmerTest-package:
library(lmerTest) # (Udviklet af os)
data(TVbo)

# Each of 8 assessors scored each of 12 combinations 2 times
# Let's look at only a single picture and one of the two reps:
# And let us look at the sharpness
TVbosubset <- subset(TVbo,Picture==1 & Repeat==1)[,c(1, 2, 9)]

sharp <- matrix(TVbosubset$Sharpness, nrow=8, byrow=T)
colnames(sharp) <- c("TV3", "TV2", "TV1")
rownames(sharp) <- c("Person 1", "Person 2", "Person 3",
                     "Person 4", "Person 5", "Person 6",
                     "Person 7", "Person 8")

library(xtable)
xtable(sharp)
```


Bang & Olufsen data i R:

	TV3	TV2	TV1
Person 1	9.30	4.70	6.60
Person 2	10.20	7.00	8.80
Person 3	11.50	9.50	8.00
Person 4	11.90	6.60	8.20
Person 5	10.70	4.20	5.40
Person 6	10.90	9.10	7.10
Person 7	8.50	5.00	6.30
Person 8	12.60	8.90	10.70

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


Oversigt

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
  - j tæller målinger i grupperne, fra 1 til n_i for behandling i

Estimater af parametrene i modellen

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

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

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

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

```r
## 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)
```

Oversigt

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")

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

- 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

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

```
## 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

```
## 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			

```
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
```


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

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

  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

- 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

```
## 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, ...) {qqnorm(y, main="",...);
  qqline(y)}
## Kan vi se et afvigende qq-norm plot?
wallyplot(fit$residuals, FUN = qqwrap)
```

Oversigt