Kursus 02402/02323 Introducerende Statistik

Forelæsning 10: Envejs variansanalyse, ANOVA

Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Bygning 324, Rum 220
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: perbb@dtu.dk

Oversigt

Envejs variansanalyse - eksempel

Gruppe A	Gruppe B	Gruppe C
2.8	5.5	5.8
3.6	6.3	8.3
3.4	6.1	6.9
2.3	5.7	6.1

Er der 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.

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:

```r
# Getting the Bang and Olufsen data from the lmerTest-package:
library(lmerTest) # (Udviklet af os)
data(TVbo)
head(TVbo)
# Defining the factor identifying the 12 TVset and Picture combs:

TVbo$TVPic <- factor(TVbo$TVset:TVbo$Picture)
# Each of 8 assessors scored each of 12 combinations 2 times
# Averaging the two replicates for each Assessor and TVpic:
library(doBy)
TVbonoise <- summaryBy(Noise ~ Assessor + TVPic, data = TVbo,
                       keep.names = T)
# One-way ANOVA of the Noise: (Not the correct analysis!!)
anova(lm(Noise ~ TVPic, data = TVbonoise))
# Two-way ANOVA of the Noise: (Much better analysis - next week)
anova(lm(Noise ~ Assessor + TVPic, data = TVbonoise))
```


Envejs variansanalyse - eksempel

```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)

## Grupper (behandlinger)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

## Plot
par(mfrow=c(1,2))
plot(as.numeric(treatm), y, xlab="Treatment", ylab="y")
##
plot(treatm, y, xlab="Treatment", ylab="y")
```


Envejs variansanalyse, model

- Opstil en model
$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$
hvor det antages, at
$$\epsilon_{ij} \sim N(0, \sigma^2)$$
- μ er samlet middelværdi
- α_i angiver effekt af gruppe (behandling) i
- j tæller målinger i grupperne, fra 1 til n_i i hver gruppe

Envejs variansanalyse, hypotese

- Vi vil nu sammenligne (flere end to) middelværdier $\mu + \alpha_i$ i modellen
$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$
- dvs. vi kan specificere hypotesen:
$$\begin{aligned} H_0 &: \quad \alpha_i = 0 \quad \text{for alle } i \\ H_1 &: \quad \alpha_i \neq 0 \quad \text{for mindst et } i \end{aligned}$$

Envejs variansanalyse, opspaltning og ANOVA tabellen

- Med modellen

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

- kan den totale variation i data opspaltes:

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

- 'Envejs' hentyder til, at der kun er én faktor i forsøget, på i alt k nivauer
- Metoden kaldes variansanalyse, fordi testningen foregår ved at sammenligne varianser

Formler for kvadratafvigelsessummer

- Kvadratafvigelsessum ("den totale varians")

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

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

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

- Kvadratafvigelsessum af behandling ("Varians forklaret af model")

$$SS(Tr) = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

Variansanalysetabel

Variations-kilde	Friheds-grader	Kvadrat-afvig. 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	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

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

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## treatm     2   30.8   15.40    26.7 0.00017 ***
## Residuals  9    5.2    0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Eksempel

```
## Antal grupper
k <- 3
## Antal i hver gruppe
ni <- 10
## Simuler data fra model med 3 means
yModel1 <- rep( c(4, 5, -3), each=ni) + rnorm(ni*k, sd=1)
## Simuler data fra model med 3 andre means
yModel2 <- rep( c(1, 3, 1), each=ni) + rnorm(ni*k, sd=1)
## 3 grupper
group <- rep(1:k, each=ni)
## Plot dem
par(mfrow=c(1,2))
plot(group, yModel1, ylim=range(yModel1,yModel2))
plot(group, yModel2, ylim=range(yModel1,yModel2))

## Beregn SST: total varians, hvilken er højst?
(SST1 <- sum( (yModel1 - mean(yModel1))^2 ))
(SST2 <- sum( (yModel2 - mean(yModel2))^2 ))

## Beregn SSE: total residual variation, hvilken er højst?
(SSE1 <- sum(tapply(yModel1, group, function(x){ sum((x - mean(x))^2) })))
(SSE2 <- sum(tapply(yModel2, group, function(x){ sum((x - mean(x))^2) })))
```

Envejs variansanalyse, F-test

- Vi har altså: (Theorem 8.2)

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

- og kan finde teststørrelsen:

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)}$$

  hvor
  - k er antal nivauer af faktoren
  - n er antal observationer

- Signifikansniveau α vælges og teststørrelsen F beregnes

- Teststørrelsen sammenlignes med en fraktil (percentile) i F fordelingen

$$F \sim F_\alpha(k-1, n-k) \text{ (Theorem 8.6)}$$

F-fordeling

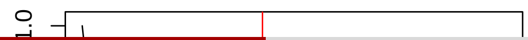
```
## Husk, dette er under H0 (altså vi regner som om H0 er sand):
## Antal grupper
k <- 3
## Antal punkter
n <- 12
## Sekvens til plot
xseq <- seq(0, 10, by=0.1)
## Plot F fordelingens tæthedsfunktion
plot(xseq, df(xseq, df1=k-1, df2=n-k), type="l")
## Kritisk værdi for signifikans niveau 5 %
cr <- qf(0.95, df1=k-1, df2=n-k)
## Tegn den i plottet
abline(v=cr, col="red")
## Test statistikkens værdi:
## Værdien
(F <- (SSTr/(k-1)) / (SSE/(n-k)))
## p-værdien er da
(1 - pf(F, df1=k-1, df2=n-k))
```


Variansanalysetabel

Variations-kilde	Friheds-grader	Kvadrat-afvig. 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	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

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

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## treatm     2   30.8   15.40    26.7 0.00017 ***
## Residuals  9    5.2    0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indenfor-gruppe variabilitet og relation til 2-gruppe t-test (Theorem 8.4)

The residual sum of squares SSE divided by $n-k$, also called Residual mean square $MSE = SSE/(n-k)$ er den gennemsnitlige varians inden for grupperne:

$$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{n-k} \tag{1}$$

$$s_i^2 = \frac{1}{n_i-1}\sum_{i=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

Hvis $k=2$:(cf. Method 3.63)

$$\text{For } k=2: \; MSE = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2}$$

$$\text{For } k=2: \; F_{\text{obs}} = t_{\text{obs}}^2$$

where t_{obs} is the pooled version coming from Methods 3.63 and 3.64.

Post hoc konfidensinterval

- 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 $n-k$ frihedsgrader.

- Læg mærke til færre frihedsgrader, da der er estimeret flere parametre i beregningen af $MSE = SSE/(n-k) = s_p^2$ (i.e. pooled varians estimat)

- Hvis alle $M = k(k-1)/2$ kombinationer af parvise konfidensintervaller udregnes brug formlen M gange, men hver gang med $\alpha_{\text{Bonferroni}} = \alpha/M$

Post hoc parvis hypotesetest

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

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

  udføres ved

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}} \tag{2}$$

  og

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

  hvor t-fordelingen med $n-k$ frihedsgrader anvendes

- Hvis alle $M = k(k-1)/2$ kombinationer af hypotesetests, bruges det korrigerede signifikans niveau $\alpha_{\text{Bonferroni}} = \alpha/M$

Varians homogenitet

Se på box-plot om spredning ser meget forskellig ud for hver gruppe

```r
## Box plot
plot(treatm,y)
```


Normalfordelingsantagelse

Se på qq-normal plot

```r
## qq-normal plot af residualer
fit1 <- lm(y ~ treatm)
qqnorm(fit1$residuals)
qqline(fit1$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(fit1$residuals, FUN = qqwrap)
```

Next week: Two-way ANOVA

```
# Getting the Bang and Olufsen data from the lmerTest-package:
library(lmerTest) # (Udviklet af os)
data(TVbo)
head(TVbo)
# Defining the factor identifying the 12 TVset and Picture combs:

TVbo$TVPic <- factor(TVbo$TVset:TVbo$Picture)
# Each of 8 assessors scored each of 12 combinations 2 times
# Averaging the two replicates for each Assessor and TVpic:
library(doBy)
TVbonoise <- summaryBy(Noise ~ Assessor + TVPic, data = TVbo,
                       keep.names = T)
# One-way ANOVA of the Noise: (Not the correct analysis!!)
anova(lm(Noise ~ TVPic, data = TVbonoise))
# Two-way ANOVA of the Noise: (Much better analysis - next week)
anova(lm(Noise ~ Assessor + TVPic, data = TVbonoise))
```