02402: Introduktion til Statistik

Forelæsning 12: Tovejs variansanalyse, ANOVA

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overview

Analysis of Variance

"ANalysis Of VAriance"(ANOVA) introduceredes af R.A. Fisher for ca. 100 år siden som en systematisk måde at analysere grupper på og har siden da været helt centralt i statistik og anvendelser deraf.

- Sidste uge: Et inddelingskriterium (one-way ANOVA)
- Denne uge: To inddelingskriterier (two-way ANOVA)
- Inddelingskriterium = faktor
- Første faktor kaldes typisk *treatment*, anden faktor *block*.

Overview

Udvikling af fjernsyn hos Bang & Olufsen

Lyd- og billedkvalitet måles med det menneskelige måleinstrument:

Bang & Olufsen data i R

```r
# Get the B&O data from the lmerTest-package
library(lmerTest)
data(TVbo)

# Each of 8 assessors scored each of 12 combinations 2 times.
# Take a look at the sharpness scores for one single picture
# and one of the two repetitions
TVbo_sub <- subset(TVbo, Picture == 1 & Repeat == 1)[, c(1, 2, 9)]
sharp <- matrix(TVbo_sub$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 ANOVA, dog ved vi nu at forsøget var inddelt i blokke:

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

  - Dvs. tre *grupper* på fire *blokke*,
  - eller tre *behandlinger* på fire *personer*,
  - eller tre *afgrøder* på fire *marker* (deraf blokke),
  - eller lignende...
- *Envejs* vs. *tovejs* ANOVA
- *Completely randomized design* vs. *Randomized block design*

Tovejs variansanalyse – eksempel

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

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

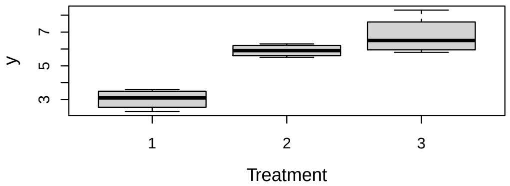
- Besvar: Er der signifikant forskel (i middel) på grupperne A, B og C?
- Variansanalyse (ANOVA) kan anvendes til analysen såfremt observationerne i hver gruppe kan antages at være normalfordelte eller hvis der er tilstrækkeligt mange observationer (CLT)

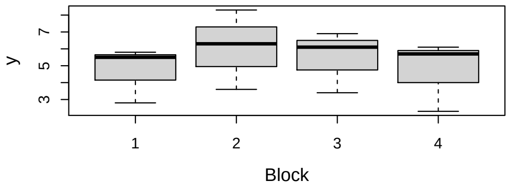
Eksemplet i R

```
# Observations
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)

# Treatments (groups, varieties)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

# Blocks (persons, fields)
block <- factor(c(1, 2, 3, 4,
                  1, 2, 3, 4,
                  1, 2, 3, 4))

# No. of treatments and no. of blocks (for later formulas)
(k <- length(unique(treatm)))
(l <- length(unique(block)))

# Box plots by treatment
plot(treatm, y, xlab = "Treatment", ylab = "y")

# Box plots by  block
plot(block, y, xlab = "Block", ylab="y")
```


Eksemplet i R

```
# Box plots by treatment
plot(treatm, y, xlab = "Treatment", ylab = "y")

# Box plots by  block
plot(block, y, xlab = "Block", ylab="y")
```





Overview

Tovejs variansanalyse, model

- Vi opstiller modellen:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

hvor afvigelserne er i.i.d. med

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

  - μ er den samlede middelværdi
  - α_i angiver effekten for behandling i
  - β_j angiver niveau/effekt for blok j
  - Der er k behandlinger og l blokke.

Estimater for parametrene i modellen

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

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

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

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

```r
# Sample mean
(mu_hat <- mean(y))

# Sample mean deviation for each treatment
(alpha_hat <- tapply(y, treatm, mean) - mu_hat)

# Sample mean deviation for each block
(beta_hat <- tapply(y, block, mean) - mu_hat)
```


Overview

Tovejs ANOVA, dekomposition og ANOVA-tabellen, Theorem 8.20

- Med modellen

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

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 (eller “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 *behandling*”)

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

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 (“residualvarinansen efter modelfit”)

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

Overview

Tovejs ANOVA: hypotese om forskellig effekt af *behandling*, Theorem 8.22

- Vi vil nu sammenligne (flere end to) middelværdier α_i i modellen

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

- Hypotesen om *ingen forskel/effekt som følge af behandling* kan formuleres ved:

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

- Under $H_{0,Tr}$ da gælder at teststørrelsen

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

er F-fordelt med $k-1$ og $(k-1)(l-1)$ frihedsgrader.

Tovejs ANOVA: hypotese om forskelligt niveau for blokke/personer, Theorem 8.22

- Vi vil nu sammenligne (flere end to) middelværdier β_j i modellen

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

- Hypotesen om *ingen forskel/effekt som følge af blokke* kan formuleres ved:

$$H_{0,Bl}: \quad \beta_j = 0 \quad \text{for alle } j$$
$$H_{1,Bl}: \quad \beta_j \neq 0 \quad \text{for mindst et } j$$

- Under $H_{0,Bl}$ da gælder at teststørrelsen

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

er F-fordelt med $l-1$ og $(k-1)(l-1)$ frihedsgrader.

F-fordeling og hypotese for behandlinger

```
# Plot density of relevant F-distribution. Remember that this is "under H0"
# (computed as if H0 were true)
xseq <- seq(0, 10, by = 0.1)
plot(xseq, df(xseq, df1 = k-1, df2 = (k-1)*(l-1)), type = "l")

# Show critical value (5% signif. level) for test of treatment hypothesis
critical_value <- qf(0.95, df1 = k-1, df2 = (k-1)*(l-1))
abline(v = critical_value, col = "red")

# Compute value of the  test statistic
(FTr <- (SSTr/(k-1)) / (SSE/((k-1)*(l-1))))

# Compute p-value for the test
1 - pf(FTr, df1 = k-1, df2 = (k-1)*(l-1))
```


F-fordeling og hypotese for blokke

```
# Plot density of relevant F-distribution. Remember that this is "under H0"
# (computed as if H0 were true)
xseq <- seq(0, 10, by = 0.1)
plot(xseq, df(xseq, df1 = l-1, df2 = (k-1)*(l-1)), type = "l")

# Show critical value (5% signif. level) for test of treatment hypothesis
critical_value <- qf(0.95, df1 = l-1, df2 = (k-1)*(l-1))
abline(v = critical_value, col = "red")

# Compute value of the  test statistic
(FBl <- (SSBl/(l-1)) / (SSE/((k-1)*(l-1))))

# Compute p-value for the test
1 - pf(FBl, df1 = l-1, df2 = (k-1)*(l-1))
```


Variansanalysetabel

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\mathrm{Tr}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\mathrm{Tr}})$
Block	$l-1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{\mathrm{Bl}} = \frac{MS(Bl)}{MSE}$	$P(F > F_{\mathrm{Bl}})$
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
```

Overview

Post hoc konfidensintervaller

- Som ved envejs ANOVA (brug metode 8.9 og 8.10), men skift $n-k$ frihedsgrader ud med $(k-1)(l-1)$ (og brug MSE fra tovejs ANOVA).
- Gøres med enten behandlinger eller blokke.
- En enkelt forudplanlagt sammenligning af forskelle på behandling i og j findes ved:

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

hvor $t_{1-\alpha/2}$ er fra t-fordelingen med $(k-1)(l-1)$ frihedsgrader.

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

Post hoc parvis hypotesetest

- For en enkelt forudplanlagt hypotesetest på niveau α

$$H_0: \ \alpha_i = \alpha_j, \ \ H_1: \ \alpha_i \neq \alpha_j$$

udføres ved:

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

and:

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

hvor t-fordelingen med $(k-1)(l-1)$ frihedsgrader anvendes.

- Hvis alle $M = k(k-1)/2$ kombinationer af parvise konfidensintervaller udregnes, så bruges det korrigerede signifikansniveau: $\alpha_{\text{Bonferroni}} = \alpha/M$.

Overview

Modelkontrol: Varianshomogenitet

Se på box-plots om spredningen af *residualer* ser ud til at afhænge af gruppen

```
# Save the  fitted model
fit <- lm(y ~ treatm + block)

# Make box plots of residuals
par(mfrow = c(1,2))
plot(treatm, fit$residuals, xlab = "Treatment")
plot(block, fit$residuals, xlab = "Block")
```


Modelkontrol: Normalfordelingsantagelse

Se normalfordelings-QQ-plottet af residualerne:

```
# Normal QQ-plot of the residuals
qqnorm(fit$residuals)
qqline(fit$residuals)
```


Overview

Et gennemregnet eksempel – fra bogen

Introduction to Statistics perbb

Agendas eNotes Course Material Podcast Forum Quiz Admin Logout

Example 8.26 Car tires

In a study of 3 different types of tires ("treatment") effect on the fuel economy, drives of 1000 km in 4 different cars ("blocks") were carried out. The results are listed in the following table in km/l.

	Car 1	Car 2	Car 3	Car 4	*Mean*
Tire 1	22.5	24.3	24.9	22.4	22.525
Tire 2	21.5	21.3	23.9	18.4	21.275
Tire 3	22.2	21.9	21.7	17.9	20.925
Mean	21.400	22.167	23.167	19.567	21.575

Let us analyse these data with a two-way ANOVA model, but first some explorative plotting:

Tovejs variansanalyse – manglende eller ekstra observationer

Tovejs ANOVA i introstat er meget "pæn" – vi har én observation pr. række og søjle. Men i praksis:

- Ofte mangler der observationer i en gruppe.
- Ofte mangler der observationer i en gruppe.
- ... eller vi har mere end én observation nogen steder.
- ANOVA fungerer som regel fint alligevel!
  - Frihedsgraderne kan være anderledes: det kan R håndtere.

	Group A	Group B	Group C
Block 1	2.8	5.5	5.8
Block 2	NA	6.3	8.3
Block 3	3.4	6.1	NA
Block 4	2.3	5.7	6.1

Overview