02402 Statistik (Polyteknisk grundlag)

Uge 12: Tosidet variansanalyse - ANOVA

Nicolai Siim Larsen
DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

1. Introduktion: Regneeksempel og TV-data fra B&O
2. Modellen
3. Beregning: Variansdekomposition og ANOVA-tabellen
4. Hypotesetest (F-test)
5. Post hoc sammenligninger
6. Modekontrol
7. Et gennemregnet eksempel fra bogen

Variansanalyse - ANOVA

"ANalysis Of VAriance"(ANOVA) blev introduceret 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 (ensidet ANOVA)
- Denne uge: To inddelingskriterier (tosidet ANOVA)
- Inddelingskriterium = faktor
- Første faktor kaldes typisk *treatment*, anden faktor *block*

Dagsorden

1. Introduktion: Regneeksempel og TV-data fra B&O
2. Modellen
3. Beregning: Variansdekomposition og ANOVA-tabellen
4. Hypotesetest (F-test)
5. Post hoc sammenligninger
6. Modekontrol
7. Et gennemregnet eksempel fra bogen

Udvikling af fjernsyn hos Bang & Olufsen

Lyd- og billedkvalitet vurderes af mennesker:

Bang & Olufsen data i R

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

# Alle 8 dommere bedømte de 12 kombinationer 2 gange.
# Her er pointene for første bedømmelse af et bestemt
# billede på tre forskellige TV
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

Tosidet variansanalyse - Eksempel

- Samme data som for den ensidet ANOVA, men nu videt det, 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

  - Tre *grupper* fordelt på fire *blokke*
  - Tre *behandlinger* fordelt på fire *personer*
- *Ensidet* eller *tosidet* ANOVA
- *Fuldstændigt randomiseret forsøg* eller *Randomiseret blokforsøg*

Tosidet variansanalyse - Eksempel

- Samme data som for ensidet ANOVA, men nu vides det, 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

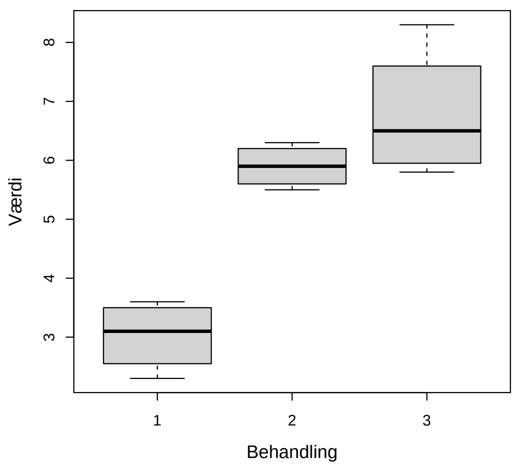
- Er der forskel (i middelværdierne) på grupperne A, B og C?
- Variansanalyse (ANOVA) kan anvendes til analysen, såfremt observationerne i hver celle kan antages at være normalfordelte, eller hvis der er tilstrækkeligt mange observationer (CLT).

Eksemplet i 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)

# Behandling (gruppe)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

# Blok (person)
block <- factor(c(1, 2, 3, 4,
                  1, 2, 3, 4,
                  1, 2, 3, 4))

# Antal behandlinger og blokke
k <- length(unique(treatm))
l <- length(unique(block))

# Boxplot fordelt på behandlinger
plot(treatm, y, xlab = "Gruppe", ylab = "y")

# Boxplot fordelt på blokke
plot(block, y, xlab = "Blok", ylab="y")
```

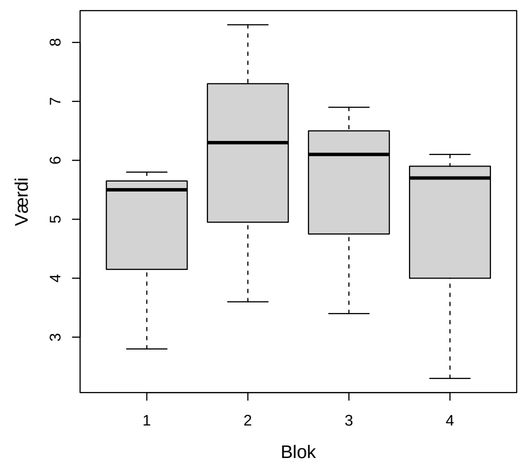

Eksemplet i R

```
# Boxplot fordelt på behandlinger
plot(treatm, y, xlab = "Behandling", ylab = "Værdi")
```



Eksemplet i R

```
# Boxplot fordelt på blokke
plot(block, y, xlab = "Blok", ylab="Værdi")
```



Dagsorden

1 Introduktion: Regneeksempel og TV-data fra B&O
2 Modellen
3 Beregning: Variansdekomposition og ANOVA-tabellen
4 Hypotesetest (F-test)
5 Post hoc sammenligninger
6 Modekontrol
7 Et gennemregnet eksempel fra bogen

Tosidet variansanalyse - Model

- Modellen:

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

hvor fejlene er uafhængige og ensfordelte med

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

  - μ er den samlede middelværdi
  - α_i angiver effekten for behandling $i \in \{1, \dots, k\}$
  - β_j angiver effekten for blok $j \in \{1, \dots, l\}$
  - Der er k behandlinger og l blokke

Tosidet variansanalyse - Estimation

- Man beregner parameterestimaterne $\hat{\mu}$, $\hat{\alpha}_i$ og $\hat{\beta}_j$ ved:

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

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

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

```r
# Samlet gennemsnit
mu_hat <- mean(y)

# Effekt af behandlinger
alpha_hat <- tapply(y, treatm, mean) - mu_hat

# Effekt af blokke
beta_hat <- tapply(y, block, mean) - mu_hat
```


Dagsorden

1 Introduktion: Regneeksempel og TV-data fra B&O
2 Modellen
3 Beregning: Variansdekomposition og ANOVA-tabellen
4 Hypotesetest (F-test)
5 Post hoc sammenligninger
6 Modekontrol
7 Et gennemregnet eksempel fra bogen

Tosidet variansanalyse - Dekomposition og variansanalyseskema - Sætning 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:

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

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

Formler for kvadratafvigelsessummer

- Den samlede variation (samme som for en ensidet analyse)

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

- Variation mellem behandlingerne/grupperne (Variation forklaret af *behandlingerne*)

$$SS(Tr) = l \cdot \sum_{i=1}^{k} (\bar{y}_{i.} - \hat{\mu})^2 = l \cdot \sum_{i=1}^{k} \hat{\alpha}_i^2$$

Formler for kvadratafvigelsessummer

- Variation mellem blokkene/personerne (Variation forklaret af *blokkene*)

$$SS(Bl) = k \cdot \sum_{j=1}^{l} (\bar{y}_{.j} - \hat{\mu})^2 = k \cdot \sum_{j=1}^{l} \hat{\beta}_j^2$$

- Variation af residualerne (Variation som modellen ikke forklarer)

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

Dagsorden

1. Introduktion: Regneeksempel og TV-data fra B&O
2. Modellen
3. Beregning: Variansdekomposition og ANOVA-tabellen
4. Hypotesetest (F-test)
5. Post hoc sammenligninger
6. Modekontrol
7. Et gennemregnet eksempel fra bogen

Tosidet variansanalyse - Hypotese om forskellige *behandlingseffekter* - Sætning 8.22

- Man ønsker at sammenligne behandlingseffekterne (middelværdierne α_i) i modellen

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

- Nulhypotesen om *ingen forskel/effekt mellem behandlingerne* kan formuleres som:

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

- Under $H_{0,Tr}$ 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.

Tosidet variansanalyse - Hypotese om forskellige *blokeffekter* - Sætning 8.22

- Man ønsker at sammenligne blokeffekterne (middelværdierne β_j) i modellen

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

- Nulhypotesen om *ingen forskel/effekt mellem blokkene* kan formuleres som:

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

- Under $H_{0,Bl}$ 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-fordelingen og hypotesen for behandlinger

```r
# 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-fordelingen og hypotesen for blokke

```r
# 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))
```

Variansanalyseskema

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


Dagsorden

1. Introduktion: Regneeksempel og TV-data fra B&O
2. Modellen
3. Beregning: Variansdekomposition og ANOVA-tabellen
4. Hypotesetest (F-test)
5. Post hoc sammenligninger
6. Modekontrol
7. Et gennemregnet eksempel fra bogen

Post hoc konfidensintervaller

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

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

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

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

  på signifikansniveauet α beregnes teststørrelsen 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-værdien ved:

$$p = 2P(T > |t_{\text{obs}}|),$$

  hvor T følger en t-fordeling med $(k-1)(l-1)$ frihedsgrader.
- Hvis M kombinationer af parvise konfidensintervaller udregnes, så bruges det korrigerede signifikansniveau: $\alpha_{\text{Bonferroni}} = \alpha/M$.

Dagsorden

1. Introduktion: Regneeksempel og TV-data fra B&O
2. Modellen
3. Beregning: Variansdekomposition og ANOVA-tabellen
4. Hypotesetest (F-test)
5. Post hoc sammenligninger
6. **Modekontrol**
7. Et gennemregnet eksempel fra bogen

Modelkontrol - Varianshomogenitet

Se på box plots om spredningen af *residualerne* ser ud til at afhænge af grupperne eller blokkene

```
# Estimer modellen
fit <- lm(y ~ treatm + block)

# Tjek box plots af residualerne
par(mfrow = c(1,2))
plot(treatm, fit$residuals, xlab = "Behandling")
plot(block, fit$residuals, xlab = "Blok")
```


Modelkontrol - Normalfordelingsantagelse

Se normalfordelings-QQ-plottet af residualerne:

```
# Normal QQ-plot for residualerne
qqnorm(fit$residuals)
qqline(fit$residuals)
```


Dagsorden

1. Introduktion: Regneeksempel og TV-data fra B&O
2. Modellen
3. Beregning: Variansdekomposition og ANOVA-tabellen
4. Hypotesetest (F-test)
5. Post hoc sammenligninger
6. Modekontrol
7. **Et gennemregnet eksempel fra bogen**

Et gennemregnet eksempel – fra bogen

Introduction to Statistics perbb
Agendas eNotes Course Material Podcast Forum Quiz Admin Logout

8.3.5 A complete worked through example: Car tires

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:

```
## Collecting the data in a data frame
```


Tosidet variansanalyse - Manglende eller flere observationer

Tosidet ANOVA i kurset her er meget "pæn" – vi har præcis én observation pr. række og søjle. Men i praksis:

- Mangler der ofte observationer i en gruppe.
- Man har mere end én observation nogle steder.
- Modellen kan nemt tilpasses. (R håndterer det meste)

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

Dagsorden

1. Introduktion: Regneeksempel og TV-data fra B&O
2. Modellen
3. Beregning: Variansdekomposition og ANOVA-tabellen
4. Hypotesetest (F-test)
5. Post hoc sammenligninger
6. Modekontrol
7. Et gennemregnet eksempel fra bogen