

02402: Introduktion til Statistik

Forelæsning 11: En-vejs variansanalyse, ANOVA

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

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.

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

Overview

- 1 **Introduktion**
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Envejs variansanalyse – eksempel

Group A	Group B	Group 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 middelværdien) på grupperne A, B and C?

Variansanalyse (ANOVA) kan anvendes til analysen, såfremt observationerne i hver gruppe kan antages at være normalforde.

Envejs variansanalyse – eksempel i R

```
# Input data
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)

## Define treatment groups
treatm <- factor(c(1, 1, 1, 1,
                  2, 2, 2, 2,
                  3, 3, 3, 3))

## Plot data by treatment groups
par(mfrow = c(1,2))
plot(y ~ as.numeric(treatm), xlab = "Treatment", ylab = "y")
boxplot(y ~ treatm, xlab = "Treatment", ylab = "y")
```

Eksempel: Komøg og antibiotika

Nedbrydning af komøg: hvor meget organisk materiale er tilbage?

Control	α -cypermethrin	Ivermectin	Spiramycin
2.43	3.00	3.03	2.80
2.63	3.02	2.81	2.85
2.56	2.87	3.06	2.84
2.76	2.96	3.11	2.93
2.70	2.77	2.94	
2.54	2.75	3.06	

Komøg og antibiotika – eksempel i R

```
dung <- c(2.43, 2.63, 2.56, 2.76, 2.70, 2.54, 3.00, 3.02, 2.87, 2.96,  
         2.77, 2.75, 3.03, 2.81, 3.06, 3.11, 2.94, 3.06, 2.80, 2.85,  
         2.84, 2.93)  
treat <- factor(c(rep("control", 6), rep("a-cyperm", 6),  
                 rep("iverm", 6), rep("spiram", 4)))  
  
plot(dung ~ as.numeric(treat), xlab = "Treatment", ylab = "y")  
boxplot(dung ~ treat, xlab = "Treatment", ylab = "y")  
  
model <- lm(dung ~ treat)  
anova(model)
```


Overview

- 1 Introduktion
- 2 Model og hypotese**
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Envejs variansanalyse, model

- Modellen kan opskrives som

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

hvor det antages ε_{ij} er i.i.d. med

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

- μ er samlet middelværdi
- α_i angiver effekt af gruppe (behandling) i .
- Y_{ij} : den j 'te måling i gruppe i (j går fra 1 til n_i).

Envejs variansanalyse, hypotese

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

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

- Hypotesen er så givet ved:

$$H_0 : \alpha_i = 0 \quad \text{for alle } i$$

med alternativ hypotese

$$H_1 : \alpha_i \neq 0 \quad \text{for mindst et } i$$

Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen**
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Envejs variansanalyse, dekomposition og ANOVA-tabellen

- Med modellen

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

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 niveauer.
- Metoden kaldes variensanalyse, fordi testningen foregår ved at sammenligne varianser.

Formler for kvadratafvigelsesummer

- Kvadratafvigelsesummer ("den totale varians")

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

Formler for kvadratafvigelsessummer

- Kvadratafvigelsessum ("den totale varians")

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

- Kvadratafvigelsessummen for residualerne ("variens tilbage efter model")

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

Formler for kvadratafvigelsessummer

- Kvadratafvigelsessum ("den totale varians")

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

- Kvadratafvigelsessummen for residualerne ("variens tilbage efter model")

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

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

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

Envejs variansanalyse, estimater

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = \bar{y}_i - \bar{y}$
- $\hat{\sigma}^2 = \frac{SSE}{n-k}$

```
# \bar{y}
mean(y)

## [1] 5.233

# \bar{y}_i: use 'tapply'
tapply(y, treatm, mean)

##      1      2      3
## 3.025 5.900 6.775

# SSE: use anova(..)
```

Variansanalysetabel

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares
Treatment	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$
Residual	$n - k$	SSE	$MSE = \frac{SSE}{n-k}$
Total	$n - 1$	SST	

```
# One-way ANOVA using anova() and lm()
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
```

Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)**
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Envejs variansanalyse, F-test

- Vi har altså: (Theorem 8.2)

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

- og kan finde teststørrelsen:

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

hvor

- k er antal nivåer af faktoren,
 - n er antal observationer.
- Vælg et signifikansniveau α , og beregn teststørrelsen F .
 - Sammenlign teststørrelsen med den relevante fraktil i F -fordelingen:

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

F-fordelingen og F-testet

```
# Remember, this is "under H0" (i.e. we compute as if H0 is true)

# Number of groups
k <- 3

# Total number of observations
n <- 12

# Sequence for plot
xseq <- seq(0, 10, by = 0.1)

# Plot density of the F-distribution
plot(xseq, df(xseq, df1 = k-1, df2 = n-k), type = "l", xlab = "x", ylab = "f(x)")

# Plot critical value for significance level 5%
cr <- qf(0.95, df1 = k-1, df2 = n-k)
abline(v = cr, col = "red")
```

Eksempel på F-fordeling og kritisk værdi

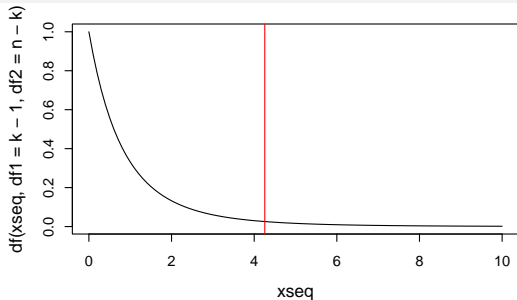
```
# Number of groups
k <- 3

# Total number of observations
n <- 12

# Sequence for plot
xseq <- seq(0, 10, by = 0.1)

# Plot density of the F-distribution
plot(xseq, df(xseq, df1 = k-1, df2 = n-k), type = "l", xlab = "x", ylab = "f(x)")

# Plot critical value for significance level 5%
cr <- qf(0.95, df1 = k-1, df2 = n-k)
abline(v = cr, col = "red")
```



Variansanalysetabel

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p -value
<i>treatment</i>	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{obs} = \frac{MS(Tr)}{MSE}$	$P(F > F_{obs})$
<i>Residual</i>	$n - k$	SSE	$MSE = \frac{SSE}{n-k}$		
<i>Total</i>	$n - 1$	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
```

Envejs ANOVA F-test “i hånden”

```

k <- 3; n <- 12 # Number of groups k, total number of observations n

# Total variation, SST
(SST <- sum( (y - mean(y))^2 ))

# Residual variance after model fit, SSE
y1 <- y[1:4]; y2 <- y[5:8]; y3 <- y[9:12]

(SSE <- sum( (y1 - mean(y1))^2 ) +
  sum( (y2 - mean(y2))^2 ) +
  sum( (y3 - mean(y3))^2 ))

# Variance explained by the model, SS(Tr)
(SSTr <- SST - SSE)

# Test statistic
(Fobs <- (SSTr/(k-1)) / (SSE/(n-k)))

# P-value
(1 - pf(Fobs, df1 = k-1, df2 = n-k))

```


Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet**
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet (Theorem 8.4)

Residualkvadratafgivelsessummen, SSE , divideret med $n - k$, også kaldet middelvadratafgivelsen $MSE = SSE/(n - k)$, er den gennemsnitlige variation inden for grupper:

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

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

KUN når $k = 2$: (jf. Method 3.52)

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

$$F_{\text{obs}} = t_{\text{obs}}^2$$

hvor t_{obs} er den sammenvejede t-teststørrelse fra Metode 3.52 og 3.53.

Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger**
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Post hoc konfidensinterval – Metode 8.9

- En enkelt *forudplanlagt* sammenligning af forskelle på behandling i og j findes ved:

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

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

- Bemærk de færre frihedsgrader, da der estimeres flere parametre i beregningen af $MSE = SSE/(n - k) = s_p^2$ (i.e. det sammenvæjede variansestimater)
- 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 – Metode 8.10

- For en enkelt forudplanlagt hypotesetest på nivo α :

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

udføres ved

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad (3)$$

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 parvise konfidensintervaller udregnes, så bruges det korrigerede signifikansniveau $\alpha_{\text{Bonferroni}} = \alpha/M$.

Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol**
- 8 Et gennemregnet eksempel – fra bogen

Varianshomogenitet

Se på box-plottet om spredningen ser (meget) forskellig ud for hver gruppe

```
# Check assumption of homogeneous variance using, e.g.,  
# a box plot.  
plot(treatm, y)
```

Normalfordelingsantagelse

Se normalfordelings-QQ-plottet af residualerne:

```
# Check normality of residuals using a normal QQ-plot  
fit1 <- lm(y ~ treatm)  
qqnorm(fit1$residuals)  
qqline(fit1$residuals)
```


Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Et gennemregnet eksempel – fra bogen

Introduction to Statistics

Agendas

▼ eNotes

Course Material

Podcast

Forum

Quiz

Admin

Dokumentegenskaber...

8.2.5 A complete worked through example: plastic types for lamps

||| Example 8.17 Plastic types for lamps

On a lamp two plastic screens are to be mounted. It is essential that these plastic screens have a good impact strength. Therefore an experiment is carried out for 5 different types of plastic. 6 samples in each plastic type are tested. The strengths of these items are determined. The following measurement data was found (strength in kJ/m^2):

Type of plastic					
I	II	III	IV	V	
44.6	52.8	53.1	51.5	48.2	
50.5	58.3	50.0	53.7	40.8	
46.3	55.4	54.4	50.5	44.5	
48.5	57.4	55.3	54.4	43.9	
45.2	58.1	50.6	47.5	45.9	
52.3	54.6	53.4	47.8	42.5	

Overview

- 1 Introduktion
- 2 Model og hypotese
- 3 Beregning – variansdekomposition og ANOVA-tabellen
- 4 Hypotetest (F-test)
- 5 Indenfor-gruppe variabilitet og sammenhæng med two-sample t-testet
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen