Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 11: Envejs variansanalyse, ANOVA

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: jkmo@dtu.dk

Oversigt

Oversigt

F-fordelingen

Hvis $Q_1 \sim \chi^2(n_1)$ og $Q_2 \sim \chi^2(n_2)$, og Q_1 og Q_2 er uafhængige så følger

$$F = \frac{Q_1/n_1}{Q_2/n_2} \tag{1}$$

en F-fordeling med n_1 og n_2 frihedsgrader.
Tæthedsfunktionen for en F-fordeling er givet ved

$$f_F(x) = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)\left(1+\frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}; \quad x \geq 0 \tag{2}$$

hvor

$$B(\nu_1, \nu_2) = \frac{\Gamma(\nu_1)\Gamma(\nu_2)}{\Gamma(\nu_1+\nu_2)} \tag{3}$$

er Beta-funktionen.

F-fordelingen som stikpøvefordeling

Lad $X_1, \ldots, X_{n_1}$ være i.i.d. $N(\mu_1, \sigma_1^2)$ og lad $Y_1, \ldots, Y_{n_2}$ være i.i.d. $N(\mu_2, \sigma_2^2)$ så gælder at

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1) \tag{4}$$

hvor S_1^2 og S_2^2 er stikprøve variansen for X hhv. Y.

Eksempel

Lad $X_1, \ldots, X_{10}$ være i.i.d. $N(\mu_1, \sigma^2)$ og lad $Y_1, \ldots, Y_{10}$ være i.i.d. $N(\mu_2, \sigma^2)$ find

$$P(S_1^2/S_2^2 > 2) \tag{5}$$

hvor S_1^2 og S_2^2 er stikprøve variansen for X hhv. Y.

Oversigt

Motiverende eksempel - energiforbrug

Forskel på energiforbrug?

I et ernæringsstudie ønsker man at undersøge om der er en forskel i energiforbrug for forskellige typer (moderat fysisk krævende) arbejde. In the study, the energy usage of 9 nurses from hospital A and 9 (other) nurses from hospital B have been measured. The measurements are given in the following table in mega Joule (MJ):

Stikprøve fra hver hospital, $n_1 = n_2 = 9$:

Hospital A	Hospital B
7.53	9.21
7.48	11.51
8.08	12.79
8.09	11.85
10.15	9.97
8.40	8.79
10.88	9.69
6.13	9.68
7.90	9.19

The pooled two-sample t-test statistic

Beregning af den poolede teststørrelse (Metode 3.63 og 3.64)

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0 : \ \delta = \delta_0$$

the pooled two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

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

Sammenvægtet (Pooled) variance set up

Assume that $X_1 \sim N(\mu_1, \sigma)$ and $X_2 \sim N(\mu_2, \sigma)$. Then the pooled two-sample statistic seen as a random variable (Theorem 3.54, Example 2.85 og Exercise 2.16):

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}} \tag{6}$$

follows, under the null hypothesis and under the assumption that $\sigma_1^2 = \sigma_2^2$, a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if the two population distributions are normal.

Sammenvægtet (Pooled) variance set up

Atag at $X_1 \sim N(\mu, \sigma)$ and $X_2 \sim N(\mu, \sigma)$, og $n_1 = n_2 = n$. Hvad er fordelingen af

$$T^2 = \frac{(\bar{X}_1 - \bar{X}_2)^2}{S_p^2/n + S_p^2/n} \tag{7}$$

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.

Envejs variansanalyse - eksempel

```
## 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")
```

Oversigt

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

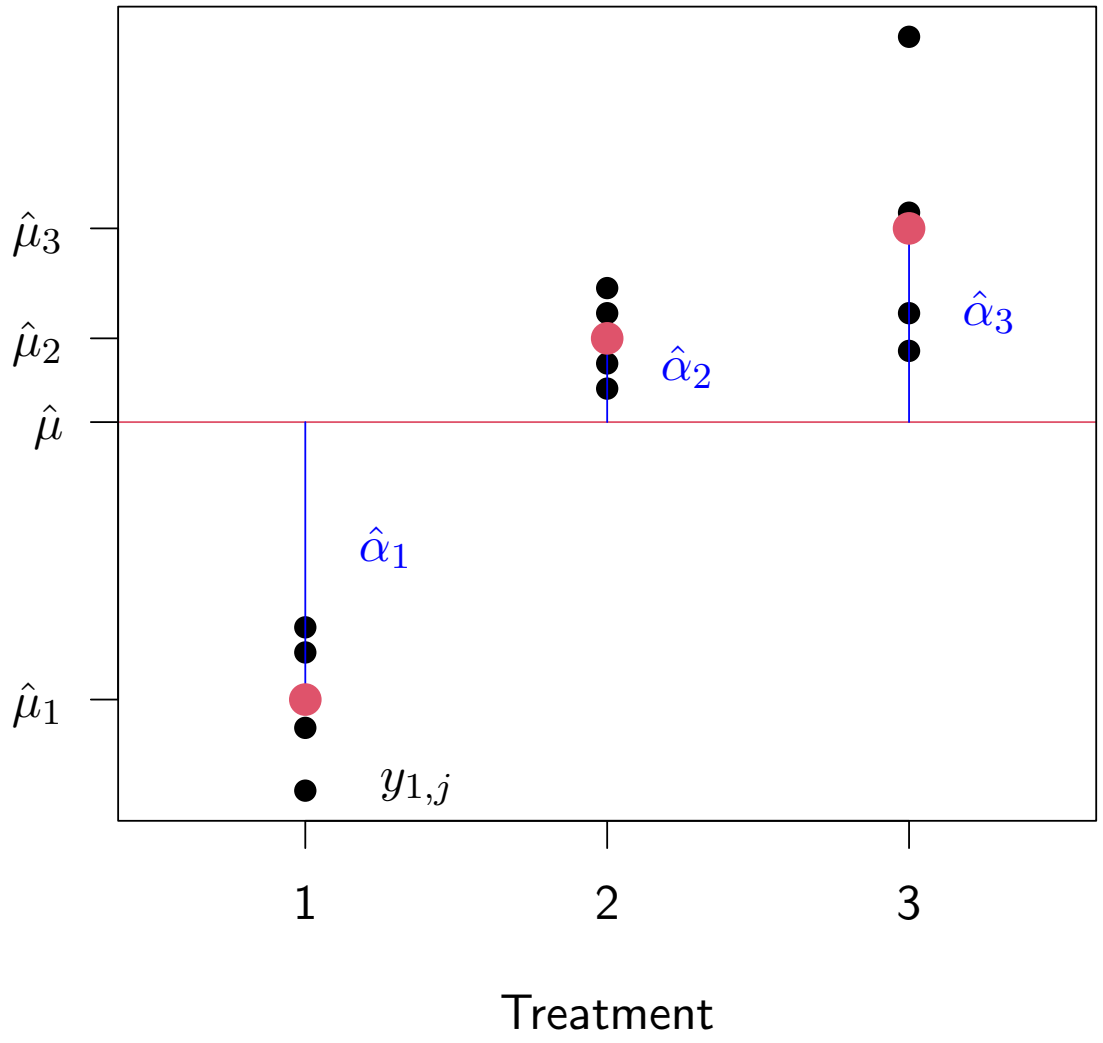
- Vi vil nu sammenligne (flere end to) middelværdier $\mu + \alpha_i$ i modellen

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

  hvor $\sum n_i \alpha_i = 0$.

- $\hat{\mu} = \bar{y}$, $\alpha_i = \bar{y}_i - \bar{y}$
- dvs. vi kan specificere hypotesen:

$$\begin{aligned} H_0: &\quad \alpha_i = 0 \quad \text{for alle } i \\ H_1: &\quad \alpha_i \neq 0 \quad \text{for mindst et } i \end{aligned}$$


$\hat{\mu}_3$
$\hat{\mu}_2$
$\hat{\mu}$
$\hat{\mu}_1$
$\hat{\alpha}_1$
$\hat{\alpha}_2$
$\hat{\alpha}_3$
$y_{1,j}$
1
2
3
Treatment

Oversigt

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")

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

Oversigt

Envejs variansanalyse, F-test

- Vi har altså:

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

- 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

$$F \sim F(k-1, n-k)$$

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
```

Oversigt

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

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{8}$$

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 $\alpha_{\text{Bonferroni}} = \alpha/M$

Oversigt

Varians homogenitet

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

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

Normalfordelingsantagelse

Se på qq-normal plot

```
## 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)
```

Oversigt

Eksempel 1 (Skive fjord)

Et af vandmiljøplanernes hovedmål var at reducere kvælstoftilførslen. Undersøg om der er sket en reduktion i løbet af de 4 perioder defineret ved vandmiljøplanerne. Betragt indledningsvis kun tilførslen i September måned.

Oversigt