# Kursus 02323: Introduktion til Statistik

# Forelæsning 11: Envejs variansanalyse, ANOVA

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2021



## Kapitel 8: Envejs variansanalyse (envejs ANOVA)

**$k$ UAFHÆNGIGE grupper**

- Test om middelværdi for mindst en gruppe er forskellig fra de andre gruppers middelværdi
- Model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

**Specifikke metoder, envejs variansanalyse:**

- ANOVA-tabel: $SST = SS(Tr) + SSE$
- $F$-test
- Post hoc test(s): Parvise $t$-test med poolet varians estimat
  - Hvis planlagt på forhånd, så uden Bonferroni korrektion
  - Hvis alle sammenligninger udføres, så med Bonferroni korrektion



## Chapter 8: One-way Analysis of Variance

**$k$ INDEPENDENT samples (groups)**

- Test if the mean of at least one of the groups is different from the mean of the other groups
- Model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

**Specific methods, one-way analysis of variance:**

- ANOVA-table: $SST = SS(Tr) + SSE$
- $F$-test
- Post hoc test(s): pairwise $t$-test with pooled variance estimate
  - If planned on beforehand, then without Bonferroni correction
  - If all samples are compared, then with Bonferroni correction



## 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 på grupperne A, B og C?
  (dvs. forskel i middelværdi på population A, B og C?)
- Variansanalyse (ANOVA) kan anvendes til analysen såfremt observationerne i hver gruppe kan antages at være normalfordelte (vigtigt når man har få observationer, men jo flere man observationer man har des mindre vigtigt ifølge CLT)



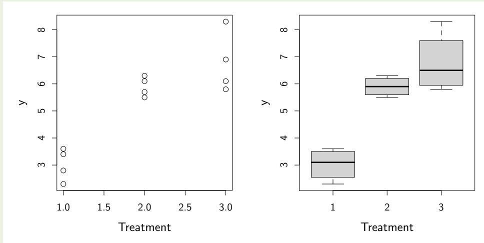


# Envejs variansanalyse - eksempel

```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 som punkter
plot(as.numeric(treatm), y, xlab="Treatment", ylab="y")
## Plot som box-plot
plot(treatm, y, xlab="Treatment", ylab="y")
```








# Envejs variansanalyse, model og hypotese

Opstil en model

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

hvor det antages, at

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

- $\mu$ er samlet middelværdi
- $\alpha_i$ angiver effekt af gruppe (behandling) $i$
- $j$ tæller målinger i grupperne, fra 1 til $n_i$ i hver gruppe





# Envejs variansanalyse, model og hypotese

Hypotese

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

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

- så vi opsætter hypotesen

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

## Envejs variansanalyse, opspaltning og ANOVA tabellen

Med modellen

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

kan den totale variation i $Y$ opspaltes

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

hvor

- $SST$: Kvadratafvigelsessum (“*den totale varians*”)
- $SSE$: Kvadratafvigelsessum af residualer (“*varians tilbage efter model*”)
- $SS(Tr)$: Kvadratafvigelsessum af gruppering (“*varians forklaret af model*”)

- “Envejs” hentyder til, at der kun er én faktor (én opdeling) i forsøget, på i alt $k$ niveauer
- 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*”)

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

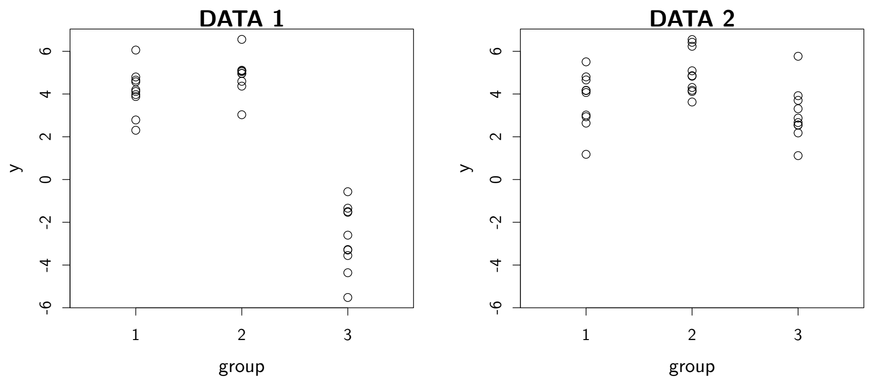
- Kvadratafvigelsessum af gruppering (“*varians forklaret af model*”)

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





## Spørgsmål den totale varians (SST) Socrative.com, room: PBAC



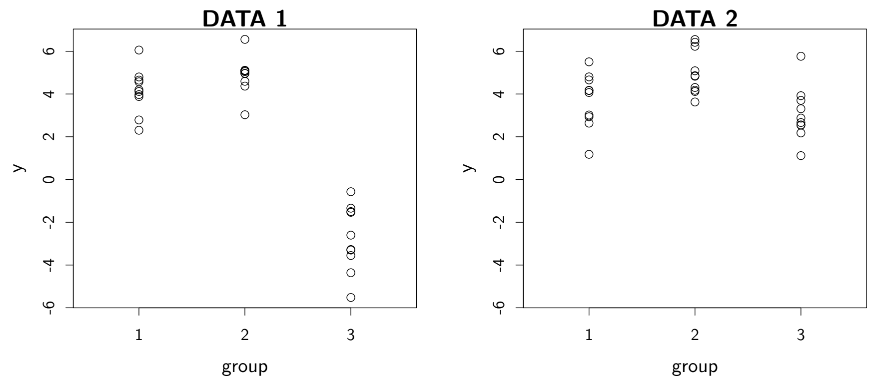

For hvilken data er SST (totale variation) størst?

A: DATA1 B: DATA2 C: Omtrent lige stor D: Ved ikke





## Spørgsmål: residual variansen (SSE) Socrative.com, room: PBAC




For hvilken data er SSE (residual variationen) størst?

A: DATA1 B: DATA2 C: Omtrent lige stor D: Ved ikke

# Envejs variansanalyse, F-test

- Vi har altså
$$SST = SS(Tr) + SSE$$
- og <u>under $H_0 : \alpha_i = 0$ for alle $i$</u> *(dvs. ingen forskel i middelværdi), da vil* teststatistikken
$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)}$$
følge en $F$-fordeling, hvor
  - $k$ er antal nivauer af faktoren (antal grupper)
  - $n$ er antal observationer
- Signifikansniveau $\alpha$ vælges og teststatistikken $F_{\text{obs}}$ beregnes
- Teststatistikken sammenlignes med en fraktil i $F$ fordelingen
$$F \sim F_\alpha(k-1, n-k)$$



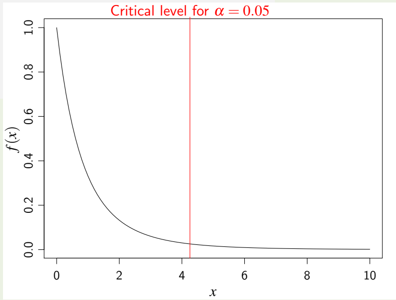


# F-fordeling

```
## 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", xlab="x", ylab="f(x)")
## 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
(Fobs <- (SSTr/(k-1)) / (SSE/(n-k)))
## p-værdien er da
(1 - pf(Fobs, 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 |
| *Gruppering* | $k-1$ | $SS(Tr)$ | $MS(Tr) = \frac{SS(Tr)}{k-1}$ | $F_{\text{obs}} = \frac{MS(Tr)}{MSE}$ | $P(F > F_{\text{obs}})$ |
| *Residual* | $n-k$ | $SSE$ | $MSE = \frac{SSE}{n-k}$ | | |
| *Total* | $n-1$ | $SST$ | | | |

```
## Alt dette beregnes med lm() og anova()

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
```





# Spørgsmål ANOVA table Socrative.com, room: PBAC

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

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatm     3   37.6   12.54    4.51  0.024 *
## Residuals 12   33.3    2.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Hvad er den totale variation SST?**

A: 12.54 B: 37.6 C: 70.9 D: Ved ikke

## Spørgsmål ANOVA table Socrative.com, room: PBAC

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

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatm     3   37.6   12.54    4.51  0.024 *
## Residuals 12   33.3    2.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Husk antagelsen om normalfordelte afvigelser $\varepsilon_{ij} \sim N(0, \sigma^2)$

Hvad er $\hat{\sigma}^2$?

A: $\frac{33.3}{12}$ B: $\frac{37.6}{3}$ C: 4.51 D: Ved ikke





## Spørgsmål ANOVA table Socrative.com, room: PBAC

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

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatm     3   37.6   12.54    4.51  0.024 *
## Residuals 12   33.3    2.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusionen på 5% signifikansniveau test af: $H_0: \alpha_i = 0$ for alle $i$?

A: $H_0$ accepteres B: $H_0$ afvises C: Ved ikke





## Post hoc konfidensinterval

*Enkelt forudplanlagt* konfidensinterval for forskel på to grupper

- En enkelt forudplanlagt sammenligning af forskelle på gruppe $i$ 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

- Forskel fra Welch two-sample test: Alle observationer er anvendt i beregningen af $MSE = SSE/(n-k) = s_p^2$ (i.e. pooled varians estimat med alle observationer)

*Mange* konfidensintervaller

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





## Post hoc parvis hypotesetest

*Enkelt forudplanlagt* t-test for forskel på grupper

- En enkelt forudplanlagt hypotesetest på $\alpha$ signifikansniveau om forskel af gruppe $i$ og $j$

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

  udføres ved

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

  og

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

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

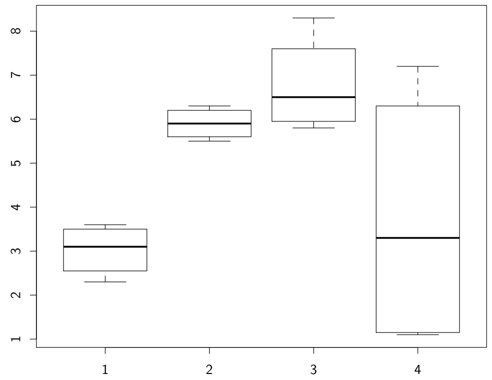
*Mange* t-tests

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

# 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
library(MESS)
qqwrap <- function(x, y, ...) {qqnorm(y, main="",...); qqline(y)}
## Kan vi se et afvigende qq-norm plot?
wallyplot(fit1$residuals, FUN = qqwrap)
```