02402 Statistik (Polyteknisk grundlag)

Uge 11: Tosidet variansanalyse (Two-way ANOVA)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

- Praktiske informationer
- Opsummering om ANOVA
- Motiverende eksempel

1. Tosidet Vaniansanalyse (Two-way ANOVA)
    - Parameterestimater og "dekomposition" af variationen
    - ANOVA-tabellen
2. F-test
3. Post hoc sammenligninger
    - Konfidensinterval for forskel mellem to grupper/blokke
    - t-test for forskel mellem to grupper/blokke
4. Modelkontrol
5. Et gennemregnet eksempel fra bogen

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering om ANOVA
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Øvelser

Vi fortsætter "klasseundervisning" i fire lokaler:

I bygning 324:

Lokale 020: Uffe (taler dansk)

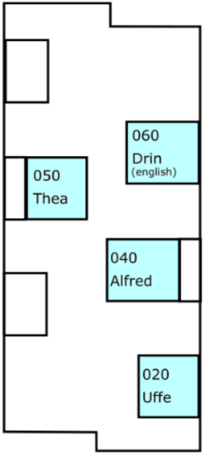
Lokale 040: Alfred (taler dansk)

Lokale 060: Drin (taler engelsk)

Lokale 050: Thea (taler dansk)

Alle gennemgår øvelser fra denne uge.

Der er også hjælpelærere til stede i 324 (foyer/stuen) og 306 (1.sal, nord)



Bygning 324

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering om ANOVA
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Introduktion

- Sidste uge: Et inddelingskriterium (ensidet ANOVA)

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

- **Denne uge: To inddelingskriterier (tosidet ANOVA)**

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

- Inddelingskriterium = faktor (eller kategorisk variabel)
- Første faktor kaldes typisk *treatment*, anden faktor *block*

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering om ANOVA
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Tosidet variansanalyse - Eksempel

- Samme data som for den 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

  - Fx tre *behandlinger* fordelt på fire *personer*

Tosidet variansanalyse - Eksempel

- Samme data som for den 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

  - Fx tre *behandlinger* fordelt på fire *personer*

- *Ensidet* eller *tosidet* ANOVA

- *Fuldstændigt randomiseret forsøg* eller *Randomiseret blokforsøg*

Tosidet variansanalyse - Eksempel

	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?

Tosidet variansanalyse - Eksempel

	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).
- OBS: I dette kursus har vi altid kun én observation i hver celle (dvs med samme α og samme β), når vi laver tosidet ANOVA.

Eksempel i Python

- Gå til Python notebook "ANOVA_2.ipynb" i VS Code
    1) Introduction and data example
       (+ KAHOOT x2)

Lidt notation

Vi skal regne forskellige gennemsnit i data:

	B_1	$\dots$	B_l
A_1	y_{11}	$\dots$	$y_{1,l}$
$\vdots$	$\vdots$	$\dots$	$\vdots$
A_k	$y_{k,1}$	$\dots$	$y_{k,l}$

Hvad er $\bar{\bar{y}}$?

Hvad er $\bar{y}_{i.}$ $(i = 1, 2, ..., k)$?

Hvad er $\bar{y}_{.j}$ $(j = 1, 2, ..., l)$?

Eksempel

	Gruppe A	Gruppe B	Gruppe C	Mean:
Blok 1	2.8	5.5	5.8	**4.700**
Blok 2	3.6	6.3	8.3	**6.067**
Blok 3	3.4	6.1	6.9	**5.467**
Blok 4	2.3	5.7	6.1	**4.700**
Mean:	**3.025**	**5.900**	**6.775**	**5.253**

02402 Statistik (Polyteknisk grundlag)

Tosidet Variansanalyse (Two-way ANOVA)

- Parameterestimater og "dekomposition" af variationen
- ANOVA-tabellen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Two-way ANOVA: 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 gruppe (Treatment) $i \in \{1, ..., k\}$
  - β_j angiver effekten for blok $j \in \{1, ..., l\}$
  - Der er k grupper og l blokke

Eksempel

Man ønsker at undersøge arbejdstiden som studerende bruger på nogle forskellige kurser.
Desuden ønsker man at undersøge om arbejdstiden har ændret sig over tid.
Derfor spørger man nogle studerende fra 5 forskellige årgange på 5 forskellige kurser (én person for hvert kursus, hvert år), hvor meget arbejdstid de har brugt på de forskellige kurser.

Hvad er vores data? Hvad er interessevariabel og hvad er forklarende variable?

Hvad vil være en relevant statistisk model?

Kahoot!

(x3)

02402 Statistik (Polyteknisk grundlag)

- Parameterestimater og "dekomposition" af variationen
- ANOVA-tabellen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Parameterestimater

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

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

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

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

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

$\hat{\alpha}_i$ og $\hat{\beta}_j$ kaldes også de *marginale* effekter, ved at være i en bestemt treatment gruppe eller i en bestemt block.

Variation "dekomposition"

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*:

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

SST: "Total variation"

SS(Tr): Variation *imellem* "Treatment" grupperne

SS(Bl): Variation *imellem* "Block" grupperne

SSE: Variation *inden for* grupperne

Formler for kvadratafvigelsessummer

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

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

- *Total Sum of Squares*

Formler for kvadratafvigelsessummer

- Variation mellem (Treatment) grupperne (variation *forklaret* af α_i'erne):

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

- *Treatment Sum of Squares*
- *"Variation between treatment groups"*

Formler for kvadratafvigelsessummer

- Variation mellem blokkene (Variation *forklaret* af β_j'erne)

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

- *Block Sum of Squares*
- *"Variation between blocks"*

SST, SS(TR) og SS(Bl) i Python

Vi beregner først gennemsnit:

	Gruppe A	Gruppe B	Gruppe C	Mean:
Blok 1	2.8	5.5	5.8	**4.700**
Blok 2	3.6	6.3	8.3	**6.067**
Blok 3	3.4	6.1	6.9	**5.467**
Blok 4	2.3	5.7	6.1	**4.700**
Mean:	**3.025**	**5.900**	**6.775**	**5.253**

Herefter indsættes i formler for SST, SS(Tr) og Ss(Bl).

- Gå til Python notebook "ANOVA_2.ipynb" i VS Code
    2) SST, SS(Tr) and SS(Bl)

Variation af residualerne

- Variation *tilbage efter model*, dvs. af *residualerne*

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

- *Sum of Squared Errors*

Dekomponering af variation i data

Theorem 8.20 Variation decomposition

The total sum of squares (SST) can be decomposed into sum of squared errors (SSE), treatment sum of squares ($SS(Tr)$), and a block sum of squares ($SS(Bl)$)

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij}-\hat{\mu})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij}-\hat{\alpha}_i-\hat{\beta}_j-\hat{\mu})^2}_{\text{SSE}} + l\cdot\underbrace{\sum_{i=1}^{k}\hat{\alpha}_i^2}_{\text{SS(Tr)}} + k\cdot\underbrace{\sum_{j=1}^{l}\hat{\beta}_j^2}_{\text{SS(Bl)}}$$

$$= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{\bar{y}})^2}_{\text{SSE}} + \underbrace{l\cdot\sum_{i=1}^{k}(\bar{y}_{i\cdot}-\bar{\bar{y}})^2}_{\text{SS(Tr)}} + \underbrace{k\cdot\sum_{j=1}^{l}(\bar{y}_{\cdot j}-\bar{\bar{y}})^2}_{SS(Bl)}, \tag{8-41}$$

Expressed in short form

$$SST = SS(Tr) + SS(Bl) + SSE. \tag{8-42}$$

MSE

Estimering af residual variansen:

$$\hat{\sigma}^2 = \text{"}MSE\text{"} = \frac{SSE}{(k-1)(l-1)}$$

Nu divideres med $(k-1)(l-1)$ (i stedet for $(n-k)$, for one way ANOVA).

MSE er den varians der er *tilbage* efter man har *korrigeret for marginale effekter*.

Eksempel i Python

- Gå til Python notebook "ANOVA_2.ipynb" i VS Code
    3) Estimate parameters μ, α_i, β_j and σ^2 (MSE)

02402 Statistik (Polyteknisk grundlag)

- Parameterestimater og "dekomposition" af variationen
- ANOVA-tabellen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Variansanalyseskema

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

n: Antal observationer i alt

k: Antal grupper (*Treatment*)

l: Antal blokke (*Block*)

02402 Statistik (Polyteknisk grundlag)

F-test

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Koncept for sammenligning (hypotesetest)

- Man ønsker at sammenligne **middelværdier** (*effekterne* af gruppe eller blok) 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 (og modhypotesen) om ingen forskel mellem **(Treatment) grupperne** kan formuleres som:

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

- Nulhypotesen (og modhypotesen) om ingen forskel mellem **blokkene** kan formuleres som:

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

F-test

Vi opstiller nu to forskellige F-test (én for *Treatment* og én for *Block*):

- Under $H_{0,Tr}$ (α_i'er $= 0$) gælder, at teststørrelsen

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

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

- Under $H_{0,Bl}$ (β_j'er $= 0$) gælder, at teststørrelsen

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

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

F-test

Theorem 8.22

Under the null hypothesis

$$H_{0,Tr}: \quad \alpha_i = 0, \quad i = 1, 2, \ldots, k, \tag{8-44}$$

the test statistic

$$F_{Tr} = \frac{SS(Tr)/(k-1)}{SSE/((k-1)(l-1))}, \tag{8-45}$$

follows an F-distribution with $k-1$ and $(k-1)(l-1)$ degrees of freedom. Further, under the null hypothesis

$$H_{0,Bl}: \quad \beta_j = 0, \quad j = 1, 2, \ldots, l, \tag{8-46}$$

the test statistic

Kahoot!
(x3)

$$F_{Bl} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}, \tag{8-47}$$

follows an F-distribution with $l-1$ and $(k-1)(l-1)$ degrees of freedom.

Variansanalyseskema inkl. F-test

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			

- Gå til Python notebook "ANOVA_2.ipynb" i VS Code
  4) F-test and ANOVA table (+KAHOOT x2)

Visual Studio Code

Post hoc sammenligninger

- Konfidensinterval for forskel mellem to grupper/blokke
- t-test for forskel mellem to grupper/blokke

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Post hoc sammenligninger

Sammenligne to (*Treatment*) grupper *eller* to blokke parvist.

Udklip fra bogen:

8.3.3 Post hoc comparisons

The post hoc investigation is done following the same approach and principles as for one-way ANOVA with the following differences:

1. Use the MSE and/or SSE from the two-way analysis instead of the MSE and/or SSE from the one-way analysis
2. Use $(l-1)(k-1)$ instead of $n-k$ as degrees of freedom and as denominator for SSE

02402 Statistik (Polyteknisk grundlag)

- Konfidensinterval for forskel mellem to grupper/blokke
- t-test for forskel mellem to grupper/blokke

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Post hoc konfidensinterval for parvis forskel

Method 8.9 Post hoc pairwise confidence intervals

A single pre-planned $(1-\alpha)\cdot 100\%$ confidence interval for the difference between treatment i and j is found as

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{8-22}$$

where $t_{1-\alpha/2}$ is based on the t-distribution with $n-k$ degrees of freedom.

If all $M = k(k-1)/2$ combinations of pairwise confidence intervals are calculated using the formula M times, but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$ (see Remark 8.14 below).

For Two-way ANOVA:

$\bar{y}_i$ og $\bar{y}_j$ er nu gennemsnit i to (*Treatment*) grupper *eller* to blokke

Udskift $(n-k)$ med $(k-1)(l-1)$

Brug MSE fra den tosidede variansanalyse

Man kan også beregne "Least Significant Difference (LSD)" i det tosidede tilfælde

02402 Statistik (Polyteknisk grundlag)

- Konfidensinterval for forskel mellem to grupper/blokke
- t-test for forskel mellem to grupper/blokke

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Post hoc t-test for parvis forskel

Method 8.10 Post hoc pairwise hypothesis tests

A single pre-planned level α hypothesis tests

$$H_0: \ \mu_i = \mu_j, \ \ H_1: \ \mu_i \neq \mu_j, \tag{8-23}$$

is carried out by

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, \tag{8-24}$$

and

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|), \tag{8-25}$$

where the t-distribution with $n - k$ degrees of freedom is used.

If all $M = k(k-1)/2$ combinations of pairwise hypothesis tests are carried out use the approach M times but each time with test level $\alpha_{\text{Bonferroni}} = \alpha/M$ (see Remark 8.14 below).

For Two-way ANOVA:
$\bar{y}_i$ og $\bar{y}_j$ er nu gennemsnit i to (*Treatment*) grupper *eller* to blokke
Udskift $(n-k)$ med $(k-1)(l-1)$
Brug MSE fra den tosidede variansanalyse

02402 Statistik (Polyteknisk grundlag)

Modelkontrol

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel i Python

8.3.4 Model control

Also model control runs almost exactly the same way for two-way ANOVA as for one-way:

- Use a q-q plot on residuals to check for the normality assumption
- Check variance homegenity by categorized box plots

- Gå til Python notebook "ANOVA_2.ipynb" i VS Code
  5) Model control

Visual Studio Code

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. (Python 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

Et gennemregnet eksempel fra bogen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Et gennemregnet eksempel – fra bogen

Introduction to Statistics perbb

Agendas eNotes Course Material Podcast Forum Quiz Admin Logout

ærktøj Dokumentegenskaber...

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
```

Dagsorden

- Praktiske informationer
- Opsummering om ANOVA
- Motiverende eksempel

1. Tosidet Vaniansanalyse
(Two-way ANOVA)
    - Parameterestimater og "dekomposition" af variationen
    - ANOVA-tabellen
2. F-test
3. Post hoc sammenligninger
    - Konfidensinterval for forskel mellem to grupper/blokke
    - t-test for forskel mellem to grupper/blokke
4. Modelkontrol
5. Et gennemregnet eksempel fra bogen