02323 Introduktion til statistik

Uge 10: Ensidet variansanalyse - ANOVA

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

- Praktiske informationer
- Opsummering
- Motiverende eksempel

1. Variansanalyse (ANOVA)
   - Variation "dekomposition"
   - Estimering af parametre
   - ANOVA tabellen
2. F-test
3. Post hoc sammenligninger
   - Konfidensinterval for forskel mellem to grupper
   - t-test for forskel mellem to grupper
4. Modelkontrol
5. Et gennemregnet eksempel – fra bogen

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Øvelser

Vi fortsætter "klasseundervisning" i fire lokaler:

I bygning 324:

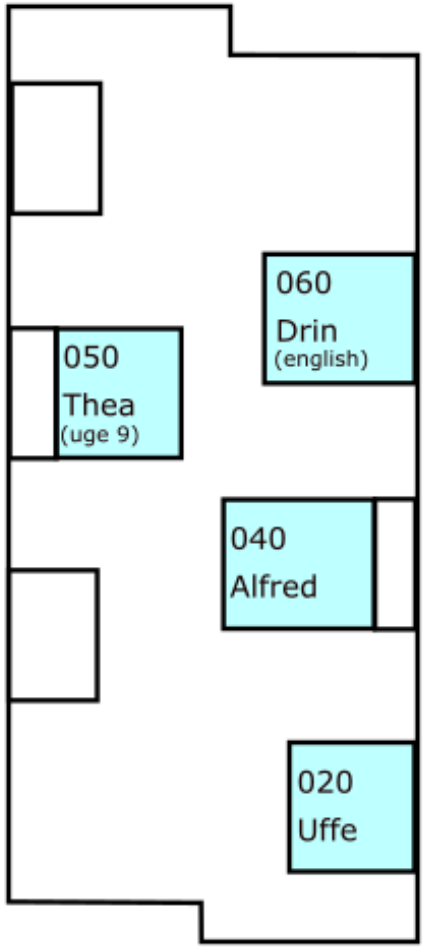
Lokale 020: Uffe (taler dansk)

Lokale 040: Alfred (taler dansk)

Lokale 060: Drin (taler engelsk)

Lokale 050: Thea (gennemgår øvelser fra uge 9)

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



Bygning 324

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Opsummering - den generelle lineære model

De sidste par gange har vi talt om den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i$$

- hvor $x_{1i}, x_{2i}, \ldots, x_{pi}$ er *kvantitative forklarende variable*.
- Vi antager $\varepsilon_i \sim N(0, \sigma^2)$ (og i.i.d.).

Vi har også (tidligere i kurset) talt om en statistisk model for 2 grupper:

$$Y_i = \mu_1 + \delta x_i + \varepsilon_i$$

- her er x_i en *binær (0/1) forklarende variabel*.
- Her antages $\varepsilon_i \sim N(0, \sigma^2)$ (og i.i.d.), dvs dette er "pooled" eksemplet med éns varians.

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering
- Motiverende eksempel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel: 3 grupper

Data med *kvalitativ forklarende variabel*:

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?

- Gå til Python notebook "ANOVA_1.ipynb" i VS Code
  1) ANOVA: Introduction and concept

02402 Statistik (Polyteknisk grundlag)

Variansanalyse (ANOVA)

- Variation "dekomposition"
- Estimering af parametre
- ANOVA tabellen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

"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 vigtig for udviklingen i statistik.

- I dag: Et inddelingskriterium (ensidet ANOVA)
- *I kursus 02402 - Næste uge: To inddelingskriterier (tosidet ANOVA)*
- Inddelingskriterium = faktor
- Første faktor kaldes typisk *treatment*, anden faktor *block*

ANOVA: 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 den samlede middelværdi
- α_i angiver effekten af gruppe (treatment) i
- Y_{ij} er måling j i gruppe i (j går fra 1 til n_i)

Matrice notation

Model for "flere grupper":

$$Y_i = \mu + \alpha_1 \cdot x_{1i} + \alpha_2 \cdot x_{2i} + \alpha_3 \cdot x_{3i} + \varepsilon_i$$

Kan også skrives på matriceform:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots \\ 1 & x_{12} & x_{22} & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots \end{bmatrix} \qquad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Man må indføre "dummy variable":

$$\{x_{11}, x_{12}, \dots, x_{1n}\} = \{1,1,1,1,\dots,0,0,0,0,\dots,0,0,0,0\}$$
$$\{x_{21}, x_{22}, \dots, x_{2n}\} = \{0,0,0,0,\dots,1,1,1,1,\dots,0,0,0,0\}$$
$$\{x_{31}, x_{32}, \dots, x_{3n}\} = \{0,0,0,0,\dots,0,0,0,0,\dots,1,1,1,1\}$$

Koncept for sammenligning (hypotesetest)

- Vi vil nu sammenligne **middelværdier** $(\mu_i = \mu + \alpha_i)$ i modellen:

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

- *Nulhypotesen* er givet ved:

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

- *Modhypotesen* (alternativhypotesen) er givet ved:

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

Vi skal opstille én samlet test: "F-test"

(dvs ikke det samme som at teste alle $\alpha_i = 0$ én ad gangen)

Kahoot! (x3)

02402 Statistik (Polyteknisk grundlag)

- Variation "dekomposition"
- Estimering af parametre
- ANOVA tabellen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Variation "dekomposition"

Med modellen

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

kan den "totale variation" i data *opspaltes*:

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

SST: "Total variation"

SS(Tr): Variation *imellem* grupperne

SSE: Variation *inden for* grupperne

Formler for kvadratafvigelsessummer

- Den samlede variation

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

Formler for kvadratafvigelsessummer

- Den samlede variation

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

- Variation mellem grupperne (variation forklaret af modellen)

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

Formler for kvadratafvigelsessummer

- Den samlede variation

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

- Variation mellem grupperne (variation forklaret af modellen)

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

Variation inden for grupperne (variation tilbage efter model, dvs. af residualerne)

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

Dekomponering af variation i data

Theorem 8.2 Variability decomposition

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

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2}_{\text{SS(Tr)}}, \tag{8-6}$$

where

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{k}\sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}. \tag{8-7}$$

Expressed in short form

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

Bevis i bogen (s. 312)

Proof

Add and subtract $\bar{y}_i$ in SST to get

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2 \qquad (8\text{-}17)\\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(y_{ij}-\bar{y}_i)^2+(\bar{y}_i-\bar{y})^2+2(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})\right]\\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})\\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{y})^2+2\sum_{i=1}^{k}(\bar{y}_i-\bar{y})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i),
\end{aligned}$$

now observe that $\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)=0$, and the proof is completed.

Kahoot!
(x1)

02402 Statistik (Polyteknisk grundlag)

- Variation "dekomposition"
- Estimering af parametre
- ANOVA tabellen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Parameterestimater

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

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

Hvad er $\bar{y}$:

_____

Hvad er $\bar{y}_i$:

_____

- Gå til Python notebook "ANOVA_1.ipynb" i VS Code
  2) ANOVA: Estimate parameters μ, α_i and σ^2

MSE - sammenvægtet varians inden for grupperne

Theorem 8.4 Within group variability

The sum of squared errors SSE divided by $n-k$, also called the residual mean square $MSE = SSE/(n-k)$ is the weighted average of the sample variances from each group

$$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{n-k}, \tag{8-14}$$

where s_i^2 is the variance within the ith group

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2. \tag{8-15}$$

When $k = 2$, that is, we are in the two-sample case presented in Section 3.2, the result here is a copy of the pooled variance expression in Method 3.52

$$\text{For } k = 2: \; MSE = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2}. \tag{8-16}$$

02402 Statistik (Polyteknisk grundlag)

- Variation "dekomposition"
- Estimering af parametre
- 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	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$
Total	$n-1$	SST	

n: Antal observationer
k: Antal grupper

- Gå til Python notebook "ANOVA_1.ipynb" i VS Code
  3) ANOVA: The ANOVA table with Python "ols"

02402 Statistik (Polyteknisk grundlag)

F-test

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

F-test

- Vi har (Sætning 8.2)

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

- Herfra kan man udlede teststørrelsen:

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{MS(Tr)}{MSE} = \frac{\text{"between group variation"}}{\text{"within group variation"}},$$

  hvor

  - k er antal grupper, n er antal observationer.

Hvad vil man konkludere når F er meget stor/lille?

F-test

Under nulhypotesen følger F en F-fordeling:

Theorem 8.6

Under the null hypothesis

$$H_0: \quad \alpha_i = 0, \quad i = 1, 2, \ldots, k, \tag{8-18}$$

the test statistic

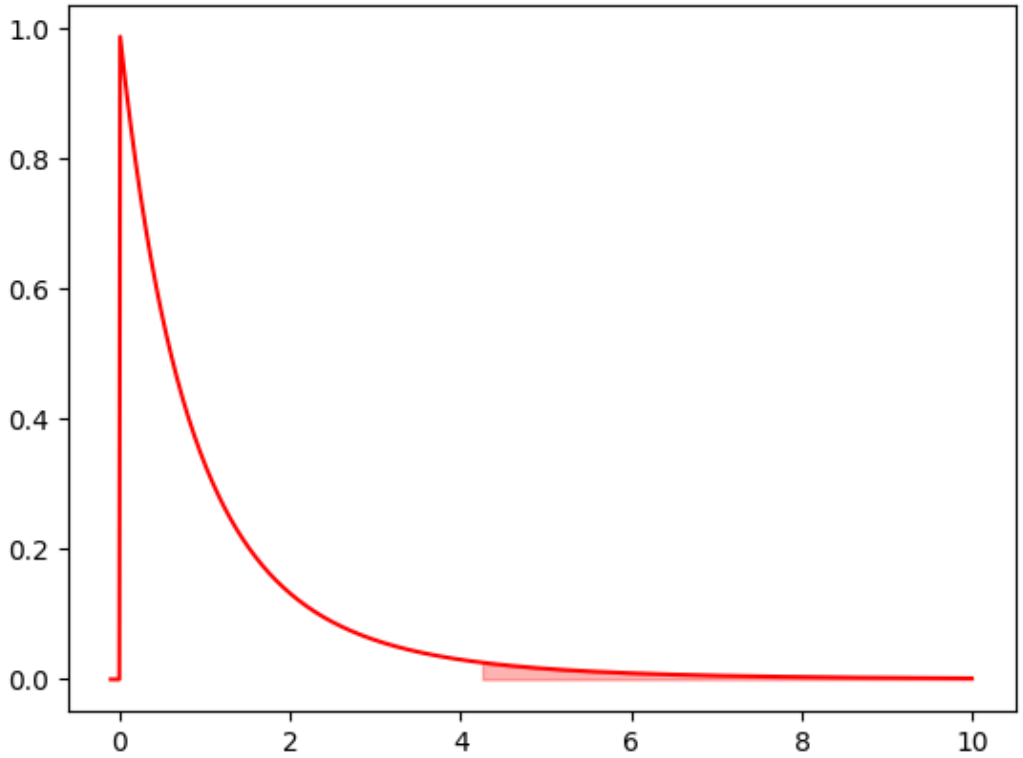
$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)}, \tag{8-19}$$

follows an F-distribution with $k-1$ and $n-k$ degrees of freedom.

Se Ex 2.97 + Appendix i bogen om F-fordeling

F-fordelingen og F-testen

Plot af F-fordeling med $n = 12$ og $k = 3$:

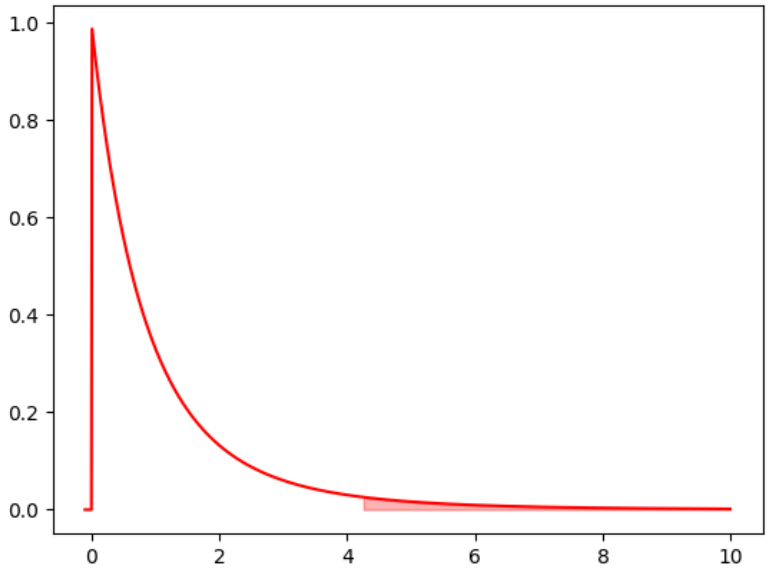


De 5% *mest ekstreme* værdier ligger over 4.26.

Tænkepause (tegn og fortæl)

F-testen er en én-sidet test!

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



Hvad er "F_{obs}" og "F_{crit}"?

Hvordan finder vi p-værdien? (og hvordan gør man med Python)

Variansanalyseskema inkl. F-test

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

- Gå til Python notebook "ANOVA_1.ipynb" i VS Code

  4) ANOVA: F-test (+KAHOOT x4)

Post hoc sammenligninger

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

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Post hoc samenligninger

Hvad mener vi med "Post hoc sammenligninger"?

Har vi allerede lært at sammenligne to grupper?

Hvis vi har MANGE grupper og sammenligner dem parvist, hvad sker der så med risikoen for at lave en Type I fejl?

02402 Statistik (Polyteknisk grundlag)

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

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Post hoc konfidensinterval for parvis forskel på grupper

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).

Genbesøg også Theorem 3.54

Bemærk de færre frihedsgrader, da der estimeres flere parametre i beregningen af $MSE = SSE/(n-k) = s_p^2$ (det sammenvejede variansestimat).

Post hoc beregning af "Lest Significant Difference"

Remark 8.13 Least Significant Difference (LSD) values

If there is the same number of observations in each treatment group $m = n_1 = \ldots = n_k$ the LSD value for a particular significance level

$$LSD_\alpha = t_{1-\alpha/2}\sqrt{2 \cdot MSE/m} \tag{8-28}$$

will have the same value for all the possible comparisons made.
The LSD value is particularly useful as a "measuring stick" with which we can go and compare all the observed means directly: the observed means with difference higher than the LSD are significantly different on the α-level. When used for all of the comparisons, as suggested, one should as level use the Bonferroni corrected version $LSD_{\alpha_{\text{Bonferroni}}}$ (see Remark 8.14 below for an elaborated explanation).

02402 Statistik (Polyteknisk grundlag)

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

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Post hoc t-test for parvis forskel på grupper

Method 8.10 Post hoc pairwise hypothesis tests

A single pre-planned level α hypothesis tests

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

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).

Bonferroni korrektion af α

Fik du fat på det med "Bonferroni korrektion" af α?

Hvad gør man og hvorfor gør man det?

(Læs selv 8.14 i bogen)

02402 Statistik (Polyteknisk grundlag)

Modelkontrol

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Modelkontrol (analyse af residualer)

Method 8.15 Normality control in one-way ANOVA

To control for the normality assumptions in one-way ANOVA we perform a q-q plot on the pooled set of n estimated residuals

$$e_{ij} = y_{ij} - \bar{y}_i, \; j = 1, \ldots, n_i, \; i = 1 \ldots, k. \tag{8-32}$$

Modelkontrol

Vores model:

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

Hvad er vores antagelser?

Hvad kan vi gøre for at tjekke om antagelserne er OK?

- Gå til Python notebook "ANOVA_1.ipynb" i VS Code
  5) ANOVA: Model control (+KAHOOT x1)

Et gennemregnet eksempel – fra bogen

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Et gennemregnet eksempel – fra bogen

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

Dagsorden

- Praktiske informationer
- Opsummering
- Motiverende eksempel

1. Variansanalyse (ANOVA)
   - Variation "dekomposition"
   - Estimering af parametre
   - ANOVA tabellen
2. F-test
3. Post hoc sammenligninger
   - Konfidensinterval for forskel mellem to grupper
   - t-test for forskel mellem to grupper
4. Modelkontrol
5. Et gennemregnet eksempel – fra bogen