

Tosidet variansanalyse og faktordiagrammer

Statistisk Dataanalyse 2

Anders Tolver

Uge 1, torsdag d. 7/9-2017



Eksempel 3.2: beskrivelse af data

Datasættet fra lærebogens eksempel 3.2 består af målinger af indholdet af organisk stof (organic) i 36 forsøgsheder (mesh bags).

Forsøgshederne stammer både fra behandlet kvæg (Ivermectin) og fra ubehandlet kvæg (control).

Forsøgshederne har ligget i jorden i 8, 12 eller 16 uger.

To faktorer: **TREAT** og **TIME**



Dagens program

- Den tosidede variansanalysemodel
- Vekselvirkninger mellem to faktorer
- Den additive model i tosidet variansanalyse
- Flerfaktorforsøg
- Faktordiagrammer

Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 2/32



Eksempel 3.2: datasættet

##	TREAT	TIME	organic	nr	TREAT	TIME	organic
## 1	Ivermectin	8	3028.7	19	Control	8	2425.0
## 2	Ivermectin	8	2805.7	20	Control	8	2630.1
## 3	Ivermectin	8	3061.3	21	Control	8	2557.0
## 4	Ivermectin	8	3113.4	22	Control	8	2763.4
## 5	Ivermectin	8	2938.1	23	Control	8	2701.0
## 6	Ivermectin	8	3063.4	24	Control	8	2544.2
## 7	Ivermectin	12	2765.0	25	Control	12	2530.6
## 8	Ivermectin	12	2713.7	26	Control	12	2301.2
## 9	Ivermectin	12	2945.7	27	Control	12	2389.8
## 10	Ivermectin	12	2869.3	28	Control	12	2445.2
## 11	Ivermectin	12	2902.0	29	Control	12	2218.5
## 12	Ivermectin	12	2836.6	30	Control	12	2348.1
## 13	Ivermectin	16	2413.3	31	Control	16	1995.0
## 14	Ivermectin	16	2592.6	32	Control	16	2165.2
## 15	Ivermectin	16	2804.7	33	Control	16	1940.9
## 16	Ivermectin	16	2546.5	34	Control	16	2271.8
## 17	Ivermectin	16	2823.7	35	Control	16	2493.8
## 18	Ivermectin	16	2845.2	36	Control	16	2452.8



Eksempel 3.2: formål med forsøget

Vi kunne være interesserede i flg. spørgsmål

- Har behandlingen indflydelse på indholdet af organisk stof?
- Hvilket niveau af TREAT giver størst indhold af organisk stof?
- Afhænger indholdet af organisk stof af, hvor længe forsøgshederne har ligget i jorden?
- Hvilket niveau af TIME giver størst indhold af organisk stof?
- Afhænger effekten af behandlingen af, hvor længe forsøgshederne har ligget i jorden?
- Hvilken komb. af TREAT og TIME giver det største indhold af organisk stof?



Eksempel 3.2: oversigt over faktorer

Faktorerne TREAT og TIME er balancerede. Hvorfor?

Desuden har vi **produktfaktoren** $TREAT \times TIME$ med alle kombinationer af TREAT og TIME.

- Hvor mange forskellige niveauer har

$TREAT \times TIME$?

- Er $TREAT \times TIME$ balanceret ?



Faktorer

Forsøgsheder: $1, 2, \dots, N$.

En **faktor** inddeler forsøgshederne i et antal grupper.

Faktoren knytter en værdi til hver forsøgsheder, nemlig **niveauet** af faktoren for den pågældende forsøgsheder. Vi skriver F_j .

$n_j(F)$ er antal forsøgsh. der er på niveau j af faktoren F .

F kaldes **balanceret** hvis $n_j(F)$ er ens for alle j , dvs. hvis der er lige mange forsøgsheder i alle grupper. Skriver så n_F for antallet af forsøgsheder per niveau.

Der findes altid to trivielle inddelinger/faktorer:

Identiske faktor (I): hver forsøgsheder udgør sin egen gruppe.

Trivielle faktor (0): alle observationer betragtes som een gruppe.

For to faktorer F og G har man **produktfaktoren** $(F \times G)$:

Svarer til at observationerne grupperes efter både F og G samtidig.



Eksempel 3.2: 2-sidet ANOVA

Den (fulde) **tosidede variansanalysemodel** er givet ved

$$M: Y_i = \gamma(TREAT \times TIME_i) + e_i,$$

hvor e_1, \dots, e_{36} er uafhængige og $N(0, \sigma^2)$ -fordelte.

- middelværdi afhænger af værdien af produktfaktoren $TREAT \times TIME$
- samme σ^2 (varians) i alle grupper
- uafhængige og normalfordelte fejl

(Dette er blot den **ensidede variansanalysemodel** med $TREAT \times TIME$ som forklarende faktor!)



Eksempel 3.2: 2-sidet ANOVA

Middelværdiparametrene, $\gamma(Iver, 8), \dots, \gamma(Ubeh, 12)$, estimeres ved gruppegennemsnit over relevante forsøgsheder.

Residualkvadratsum

$$SS_e^{TREAT \times TIME} = \sum_{i=1}^{36} (Y_i - \hat{\gamma}(TREAT \times TIME_i))^2$$

Mean square error el. residual mean square

$$s^2 = \hat{\sigma}^2 = MS_e^{TREAT \times TIME} = \frac{1}{36 - 6} SS_e^{TREAT \times TIME}$$

NB: Normeringen i udtrykket for s^2 skyldes, at der er 36 observationer og 6 grupper givet ved faktoren $TREAT \times TIME$.



Interaction plots

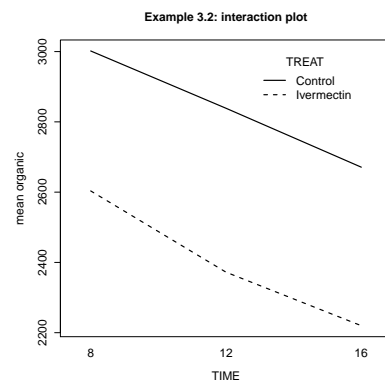
Grafisk undersøgelse af **vekselvirkning**

- For hver kombination af $TREAT \times TIME$ udregnes gruppegennemsnit.
- Gruppegennemsnit plottes op mod værdien af faktoren TIME
- Punkter hørende til samme værdi af TREAT forbindes med linjestykker
- **Parallelle kurver** taler for additiv model
- **Systematiske afvigelser** mellem kurver tyder på vekselvirkning
- Svært at eftervise vekselvirkning vha. interaction plot, men ...
- hvis et formelt statistisk test viser signifikant vekselvirkning, kan kommandoen `interaction.plot` bruges til at beskrive årsagen/retningen.



Interaction plots

Data fra eksempel 3.2: **Vekselvirkning/additiv** model ?



Eksempel 3.2: additiv model

Indledningsvis undersøges om vekselvirkningen kan fjernes så modellen reduceres til den **additive model for tosidet variansanalyse**

$$H_0 : Y_i = \alpha(TREAT_i) + \beta(TIME_i) + e_i,$$

hvor e_1, \dots, e_N er uafhængige og normalford. $N(0, \sigma^2)$.

Middelværdien i gruppen givet ved $TREAT_i = \text{Ivermectin}$ og $TIME_i = 8$ er lig med summen

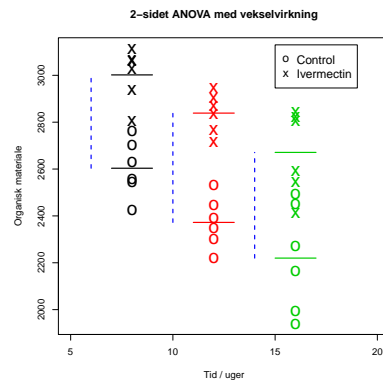
$$\alpha(\text{Ivermectin}) + \beta(8).$$

Modellen omtales også som den lineære model uden

vekselvirkning.



Eksempel 3.2: hvad udtrykker den additive model?



Ved test for reduktion til den additive model undersøges, om behandlingseffekten (-længden af de blå linjestykker) afhænger af tid!

Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 13/32



2-sidet ANOVA: reduktion

Vekselvirkning: fremgangsmåde afhænger af om vi har flere forsøgsheder (gentagelser) for hvert niveau af TREAT×TIME

gentagelser Test om modellen kan reduceres til den additive model.

ingen gentagelser Formelt test ikke muligt! Grafisk undersøgelse af vekselvirkning ved interaction plot.

Derefter: test for hovedeffekt af TREAT og TIME.

- Hvilke modeller skal fittes?
- Hvilken model, skal der testes mod?



Den additive model: Estimation

Residualkvadratsum

$$SS_e^{\text{TREAT+TIME}} = \sum_{i=1}^N (Y_i - (\hat{\alpha}(\text{TREAT}_i) + \hat{\beta}(\text{TIME}_i)))^2$$

Mean square error eller residual mean square

$$s^2 = \hat{\sigma}^2 = MS_e^{\text{TREAT+TIME}} = \frac{1}{N - k - m + 1} SS_e^{\text{TREAT+TIME}}$$

Angiver variansestimater under den additive model.



Eksempel 3.2: R

I R udføres test for vekselvirkning som følger ...

```
model1<-lm(data$organic~data$TIME:data$TREAT)
modelad<-lm(data$organic~data$TIME+data$TREAT)
anova(modelad,model1)
```

Analysis of Variance Table

```
Model 1: data$organic ~ data$TIME + data$TREAT
Model 2: data$organic ~ data$TIME:data$TREAT
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1      32 664410
2      30 656742   2     7668 0.1751 0.8402
```

Testet godkendes → **ingen vekselvirkning**

Tilsvarende analyser viser, at TREAT og TIME har marginal effekt!



Eksempel 3.2: variansanalysekema

Ved brug af R udskrifter udfyldes skemaerne

Model	Fakt.	Mv.	SS_e	df_e
1	TREAT×TIME	$\gamma(\text{TREAT} \times \text{TIME}_i)$	656742	30
2	TREAT+TIME	$\alpha(\text{TREAT}_i) + \beta(\text{TIME}_i)$	664410	32
3a	TIME	$\beta(\text{TIME}_i)$	2395959	33
3b	TREAT	$\alpha(\text{TREAT}_i)$	1432502	34
Test	Faktor	F	df	p
2 vs 1	TREAT×TIME	0.1751	2	0.84
3a vs 2	TREAT	83.397	1	0
3b vs 2	TIME	18.497	2	0

NB: Input til skema findes f.x. ved at køre summary og deviance

```
> summary(modelad)
[ ... part of output ...]
Residual standard error: 144.1 on 32 degrees of freedom
> deviance(modelad)
[1] 664410
```

Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 17/32



Additive model: konklusioner I

Slutmodel (additiv model for 2-sidet ANOVA)

$$Y_i = \alpha(\text{TREAT}_i) + \beta(\text{TIME}_i) + e_i,$$

hvor $e_1, \dots, e_{36} \sim N(0, \sigma^2)$ er uafhængige.

Parameterestimer (kan) angives således

$$\begin{aligned} \hat{\alpha}(\text{Ubeh}) + \hat{\beta}(8) &= 2583.29 & \hat{\alpha}(\text{Iver}) - \hat{\alpha}(\text{Ubeh}) &= 438.63 \\ \hat{\beta}(12) - \hat{\beta}(8) &= -197.13 & \hat{\beta}(16) - \hat{\beta}(8) &= -357.15 \end{aligned}$$

NB: For additiv model angives parameterestimer ofte ved angivelse af estimat for en referencegruppe (-her bruges (TREAT, TIME) = (Ubeh, 8)) samt forskellene til referencegruppen.

Variansestimat $s^2 = \hat{\sigma}^2 = 144.1^2$.

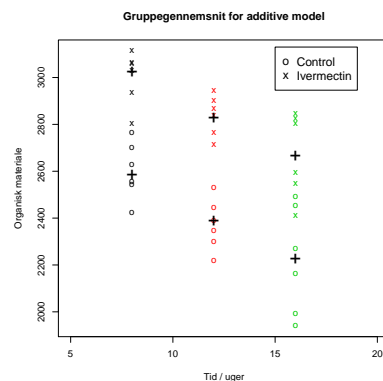
Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 18/32



Additive model: parametrisering

Struktur: gennemsnit givet ved $\text{TREAT} \times \text{TIME}$ ligger på parallelle kurver.

Punkternes beliggenhed kan beskrives ved kun 4 parametre, men parametriseringen afhænger af, hvordan modellen fittes i R.



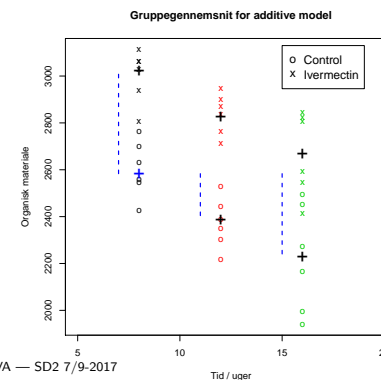
Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 19/32



Additive model: parametrisering 1

```
modelad<-lm(data$organic~data$TREAT+data$TIME)
summary(modelad)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2583.2944   48.03102  53.783871 5.933454e-33
## data$TREATIvermectin  438.6278   48.03102  9.132176 1.990792e-10
## data$TIME12      -197.1333   58.82575 -3.351140 2.076990e-03
## data$TIME16      -357.1500   58.82575 -6.071321 8.831140e-07
```



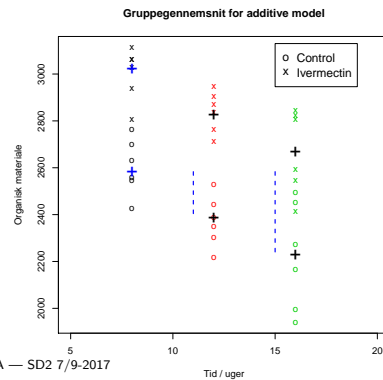
Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 20/32



Additive model: parametrisering 2

```
modelad2<-lm(data$organic~data$TREAT+data$TIME-1)
summary(modelad2)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## data$TREATControl  2583.2944   48.03102  53.783871 5.933454e-33
## data$TREATIvermectin 3021.9222   48.03102  62.916047 4.108697e-35
## data$TIME12        -197.1333   58.82575  -3.351140 2.076990e-03
## data$TIME16        -357.1500   58.82575  -6.071321 8.831140e-07
```



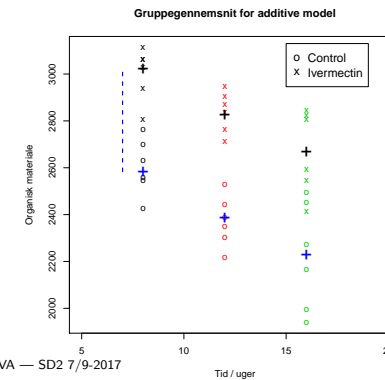
Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 21/32



Additive model: parametrisering 3

```
modelad3<-lm(data$organic~data$TIME+data$TREAT-1)
summary(modelad3)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## data$TIME8        2583.2944   48.03102  53.783871 5.933454e-33
## data$TIME12       2386.1611   48.03102  49.679579 7.307952e-32
## data$TIME16       2226.1444   48.03102  46.348052 6.540890e-31
## data$TREATIvermectin 438.6278   48.03102   9.132176 1.990792e-10
```



Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 22/32



Additive model: konf.interval / LSD

Udregn fraktil i t-fordeling: $t = t_{0.975, N-m-k+1}$, hvor
 $k = |\text{TREAT}| = 2$, $m = |\text{TIME}| = 3$, $N = k \cdot m \cdot n = 2 \cdot 3 \cdot 6$ (antal forsøg).

Et 95 %-konf. int. for $\alpha(a) + \beta(b)$ er

$$\hat{\alpha}(a) + \hat{\beta}(b) \pm t \cdot s \sqrt{\frac{k+m-1}{N}}$$

Least significant difference (LSD) for

Faktor	Kontrast	LSD
A	$\alpha(a_j) - \alpha(a_{j'})$	$t \cdot s \sqrt{\frac{2k}{N}}$
B	$\beta(b_j) - \beta(b_{j'})$	$t \cdot s \sqrt{\frac{2m}{N}}$

Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 23/32



Additive model: konklusioner II

Variansestimat og relevant t -fraktil

$$t = t_{0.975, 32} = 2.0369 \quad s = \hat{\sigma} = 144.1$$

Estimater for enkeltgrupper

TREAT	TIME		
	8	12	16
Ivermectin	3022	2825	2665
Control	2583	2386	2226

Konf.interval gruppeestimat: $\pm t \cdot s \sqrt{\frac{k+m-1}{N}} = \pm 97.8$

LSD-værdier: $LSD_{\text{TREAT}} = 97.8 \quad LSD_{\text{TIME}} = 119.8$

Anders Tolver — 2-sidet ANOVA — SD2 7/9-2017
Dias 24/32



Eksempel 3.2: resultater af forsøget

Hvordan skal vi besvare vores spørgsmål?

- Har behandlingen indflydelse på indholdet af organisk stof?
- Hvilket niveau af TREAT giver størst indhold af organisk stof?
- Afhænger indholdet af organisk stof af, hvor længe forsøgshederne har ligget i jorden?
- Hvilket niveau af TIME giver størst indhold af organisk stof?
- Afhænger effekten af behandlingen af, hvor længe forsøgshederne har ligget i jorden?
- Hvilken komb. af TREAT og TIME giver det største indhold af organisk stof?



Flerfaktorforsøg: faktordiagrammer

Faktordiagrammer benyttes til at skabe sig overblik over strukturen i et forsøgsdesign.

Dagens gennemgående eksempel indholder 3 egentlige faktorer

TREAT, TIME, TREAT × TIME

samt de trivielle faktorer

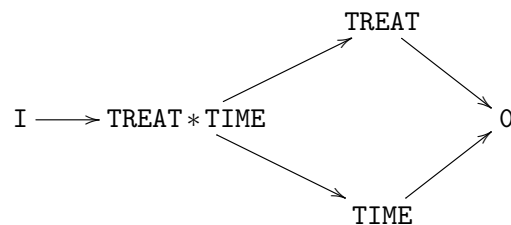
I, 0

Bemærk, at TREAT × TIME er finere end både TREAT og TIME.



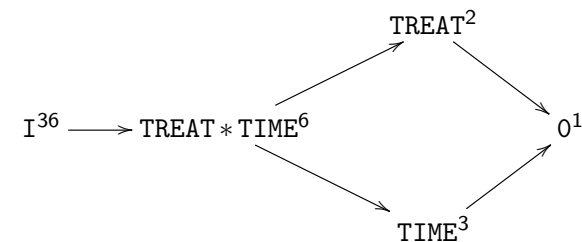
Eksempel: balanceret tofaktorforsøg

Der tegnes pile fra **finere** faktorer til **grovere**. Fineste faktor (-enhedsfaktoren I) placeres til venstre på tegningen.



Eksempel: balanceret tofaktorforsøg

Der tegnes pile fra **finere** faktorer til **grovere**. Fineste faktor (-enhedsfaktoren I) placeres til venstre på tegningen.

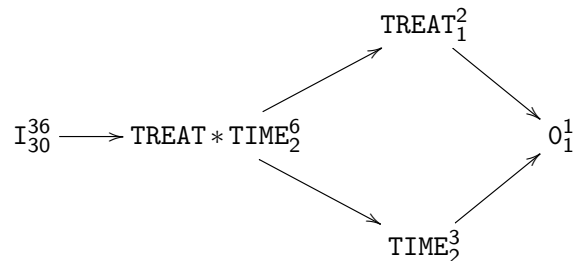


Antallet af **niveauer** skrives i øverste højre hjørne.



Eksempel: balanceret tofaktorforsøg

Der tegnes pile fra **finere** faktorer til **grovere**. Fineste faktor (-enhedsfaktoren I) placeres til venstre på tegningen.



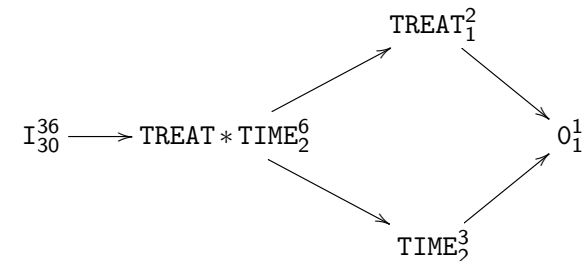
Antallet af **niveauer** skrives i øverste højre hjørne.

Antallet af **df** tilføjes i nederste højre hjørne, ved fra antallet af niveauer at fratrække **df** for grovere faktorer.



Eksempel: balanceret tofaktorforsøg

Der tegnes pile fra **finere** faktorer til **grovere**. Fineste faktor (-enhedsfaktoren I) placeres til venstre på tegningen.



Antallet af **niveauer** skrives i øverste højre hjørne.

Antallet af **df** tilføjes i nederste højre hjørne, ved fra antallet af niveauer at fratrække **df** for grovere faktorer.

Ex: Ud for $TREAT \times TIME$ skrives $6-1-2-1=2!$



Øvelse (hjemme eller senere!)

Der indgår ofte flere faktorer i et eksperiment...

Vækstforsøg med 16 planter i fire vækstkammer under forskellige gødnings- og lysforhold. Planternes vækst er målt.

	Lys 1				Lys 2			
	Kammer 1		Kammer 2		Kammer 3		Kammer 4	
Gødning	4.70	5.14	4.49	4.42	4.42	4.80	4.81	4.95
Ingen gødning	5.28	4.28	4.50	4.30	4.61	4.68	4.77	5.11

- Hvad er forsøgshederne og hvor mange er der?
- Hvad er de relevante faktorer, og hvad er deres niveauer?
- Er faktorerne balancerede? Angiv n_F for de balancerede faktorer.



Flere faktorer

Er der noget "særligt" ved faktorerne kammer og lys?

To faktorer F og G.

F er finere end G — eller F er "nested indenfor" G — hvis

- G-grupperne kan fås ved at slå F-grupper sammen
- "hvis vi kender F, så kender vi også G"

De to udsagn siger det samme!

Vi siger også **G er grovere end F**, og vi skriver $G \leq F$ eller $F \geq G$.



Trefaktorforsøg

Fra noternes eksempel 2.2.

	Lys1				Lys2			
	Kammer 1	Kammer 2	Kammer 3	Kammer 4	Kammer 1	Kammer 2	Kammer 3	Kammer 4
Gødning	*	*	*	*	*	*	*	*
Ingen gødning	*	*	*	*	*	*	*	*

Faktorer: G (Gødning), K (Kammer) og L (Lys)

Bemærk: K er finere end L!

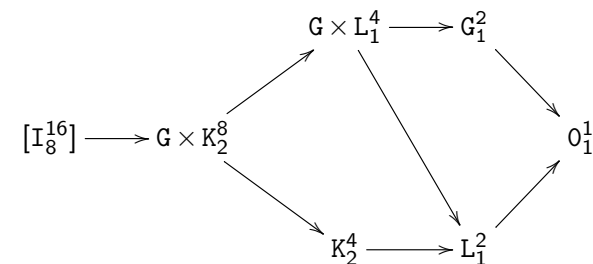
Vekselvirkninger: $G \times K$, $G \times L$, $(K \times L = K)$

Trivielle faktorer: I, 0

Lad os forsøge at tegne det tilhørende faktordiagram.



Trefaktorforsøg



Tilhørende statistiske model:

$$Y_i = \gamma(G \times K_i) + e_i, \quad e_i \sim N(0, \sigma^2).$$

G: Gødning

K: Kammer

L: Lys

NB: Vi sætter [...] omkring faktorer som ikke indgår i den systematiske del af modellen!



Spørgsmål: overvejes hjemme!

- Opskriv alle interessante hypoteser ved et tofaktorforsøg
- I hvilken rækkefølge skal modellen reduceres ved et tofaktorforsøg?
- Hvornår kan man teste for vekselvirkning i et tofaktorforsøg?
- Hvordan skal man fortolke parameterestimerne fra R-udskriften ved den additive model?
- Hvordan beregnes frihedsgradsantallet (DF) for et faktordiagram?

