Case 1: Andenårsvirkning af fosfortilførsel

Statistisk Dataanalyse 2

Anders Tolver

Institut for Matematiske Fag

Uge 1, torsdag d. 6/2-2014

Første del: additive model for blokforsøg

27 forsøgsenheder (parceller):

- Y_i: udbyttet for i-te forsøgsenhed, $i = 1, \ldots, 27$.

To faktorer:

- `fosfor`: med 9 niveauer som beskriver kombinationen af fosfortilførslen i 1981 og 1982
- `blok`: med 3 niveauer som beskriver hvilken blok pågældende forsøgsenhed tilhører.

Der lægges op til at benytte følgende udgangsmodel:

$$Y_i = \alpha(\texttt{fosfor}_i) + \beta(\texttt{blok}_i) + e_i,$$

hvor $e_1, \ldots, e_{27}$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$.

Første del: test for effekt af fosforbehandling

Hypotese, $H_0 : \alpha(1) = \ldots = \alpha(9) = 0$ svarende til modellen

$$Y_i = \beta(\texttt{blok}_i) + e_i, \quad e_i \sim N(0, \sigma^2).$$

Test af hypotesen i R:

```
> ### additive model with 'fosfor' and 'blok'
> model1<-lm(data$udbytte~factor(data$fosfor)+factor(data$blok))
> ### oneway ANOVA with 'blok' but without 'fosfor'
> model.blok<-lm(data$udbytte~factor(data$blok))

> anova(model.blok,model1)

Analysis of Variance Table


  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     24 21378.4
2     16  3469.1  8   17909.3 10.325 4.977e-05 ***
```

Vi konkluderer, at fosfortilførslen har en stærkt signifikant effekt på udbyttet ($F = 10.325, p < 0.0001$).

Første del: test for effekt af andenårs fosfortilførslen

Hyp., $H_0 : \alpha(1) = \alpha(2) = \alpha(3), \alpha(4) = \alpha(5) = \alpha(6), \alpha(7) = \alpha(8) = \alpha(9)$ svarende til modellen

$$Y_i = \gamma(\texttt{p81}_i) + \beta(\texttt{blok}_i) + e_i, \quad e_i \sim N(0, \sigma^2).$$

Test hypotesen i R:

```
> data$p81<-factor(data$p81) ## turn ´p81´ into a factor
> ### additive model with 'p81' and 'blok'
> model2<-lm(data$udbytte~data$p81+factor(data$blok))

> anova(model2,model1) ### test for effekt of 'p82'

  Analysis of Variance Table


    Res.Df     RSS Df Sum of Sq      F    Pr(>F)
  1     22 13935.8
  2     16  3469.1  6   10466.7 8.0456 0.0004028 ***
```

Vi konkluderer, at andenårstilførslen af fosfor har en stærkt signifikant effekt på udbyttet ($F = 8.0456, p = 0.0004$).

Første del: test for effekt af `p81` og `blok`

Både førsteårstilførslen ($F = 6.49, p = 0.0013$) og blokken ($F = 6.80, p = 0.0073$) har signifikant indflydelse på udbyttet.

```
> data$p82<-factor(data$p82)
> model2<-lm(data$udbytte~data$p82+data$blok) ### additive model wi

> anova(model2,model1) ### test for effect of 'p81'

Analysis of Variance Table

  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     22 11915.6
2     16  3469.1  6    8446.4 6.4927 0.001278 **

> model3<-lm(data$udbytte~data$fosfor) ### model without 'blok'-effe
> anova(model3,model1) ### test for effect of 'blok'

  Analysis of Variance Table

  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     18 6417.3
2     16 3469.1  2    2948.2 6.7988 0.007293 **
```

Anden del: additive ANOVA (på papir)

9 forsøgsenheder (gennemsnit):

- Y_i: gennemsnitsudbytte over de tre blokke for hver behandlingsgruppe, $i = 1, \ldots, 9$.

To faktorer:

- p81: med 3 niveauer som beskriver fosfortilførslen i 1981
- p82: med 3 niveauer som beskriver fosfortilførslen i 1982

Der lægges op til at benytte følgende udgangsmodel:

$$Y_i = \alpha(\text{p81}_i) + \beta(\text{p82}_i) + e_i,$$

hvor $e_1, \ldots, e_9$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$.

Anden del: additive ANOVA (i R)

```
> data$p81fac<-factor(data$p81) ### factor on 3 levels
> data$p82fac<-factor(data$p82) ### factor on 3 levels
> mod1<-lm(data$yield.mean~data$p81fac+data$p82fac)
> coef(mod1)

  (Intercept) data$p81fac30 data$p81fac60 data$p82fac20 data$p82fac4
    339.65556      20.66667      40.66667      20.56667      45.766
```

Tabel over estimaterne. Hvordan beregnes disse?

	p82		
p81	0	20	40
0	339.6556	360.2222	385.4222
30	360.3222	380.8889	406.0889
60	380.3222	400.8889	426.0889

Anden del: model med lineær effekt af p82

Man kan f.eks. fitte følgende model:

$$Y_i = \alpha(\texttt{P81}) + \beta(\texttt{P81}) \cdot \texttt{p82} + e_i, \quad e_i \sim N(0, \sigma^2).$$

NB: benyt p81 som faktor og p82 som numerisk variabel!!!

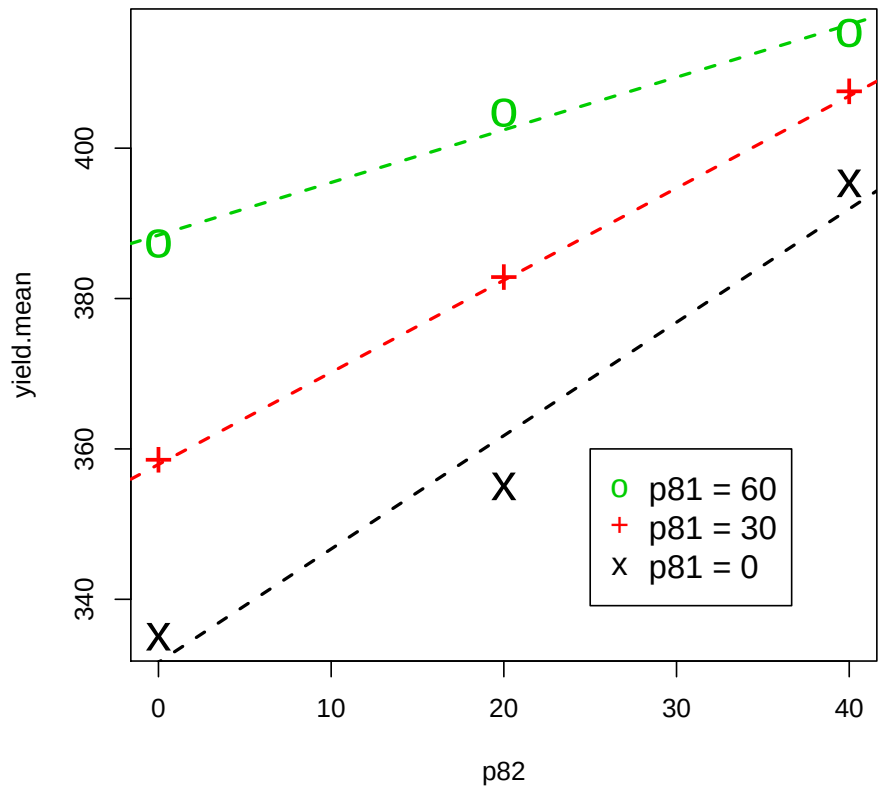
```
> data$p81fac<-factor(data$p81) ### factor on 3 levels
> mod2<-lm(yield.mean~p81fac*p82,data)
> coef(mod2)
```

```
 (Intercept)     p81fac30     p81fac60          p82 p81fac30:p82 p81
   331.61667     26.31667     56.81667      1.50750     -0.28250
```

p81	0	30	60
$\alpha(..)$	331.6167	331.6167+26.3167	331.6167+56.8167
$\beta(..)$	1.5075	1.5075-0.2825	1.5075-0.8075

Forklar hvordan parameterestimaterne skal fortolkes?

Anden del: lav en god figur



Hvad kan vi konkludere om andenårsvirkningen af fosfortilførslen?

Andel del: model med lineær effekt af p81 og p82

Man kan f.eks. fitte følgende model:

$$Y_i = \alpha + \beta \cdot \text{p81} + \gamma \cdot \text{p82} + e_i, \quad e_i \sim N(0, \sigma^2).$$

NB: benyt p81 og p82 som numeriske variable - *ikke* som faktorer!!!

```
> mod3= lm(yield.mean ~ p81 + p82)
> mod3

Coefficients:
(Intercept)          p81          p82
   338.9944       0.6778       1.1442
```

Effekten af andenårsvirkningen i forhold til førsteårsvirkningen kan f.eks. kvantificeres gennem

$$r_{21} = \frac{\hat{\gamma}}{\hat{\beta}} = \frac{1.1442}{0.6778} = 1.689.$$

Variation er vigtig!

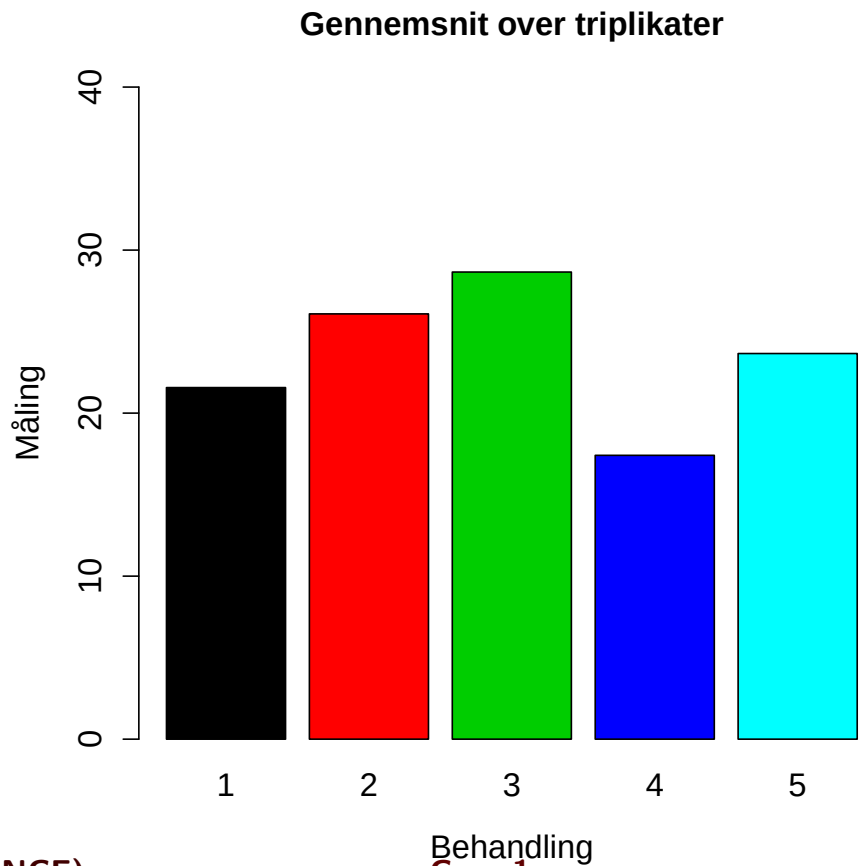
Forsøgsdesign: 5 forskellige behandlinger ønskes afprøvet. Der er råd til at lave 3 gentagelser per behandling (triplikater).

Formål: Undersøg om der er forskelle mellem behandlingerne.

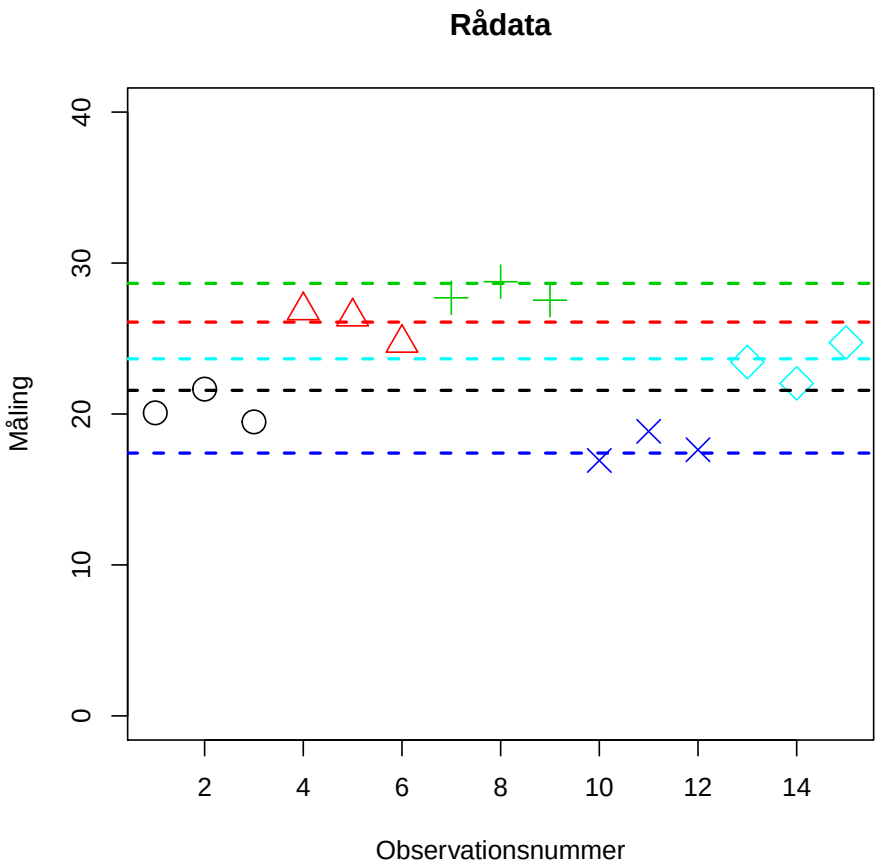
På de følgende slides vises eksempler på, hvordan man kunne forestille sig at behandle data.

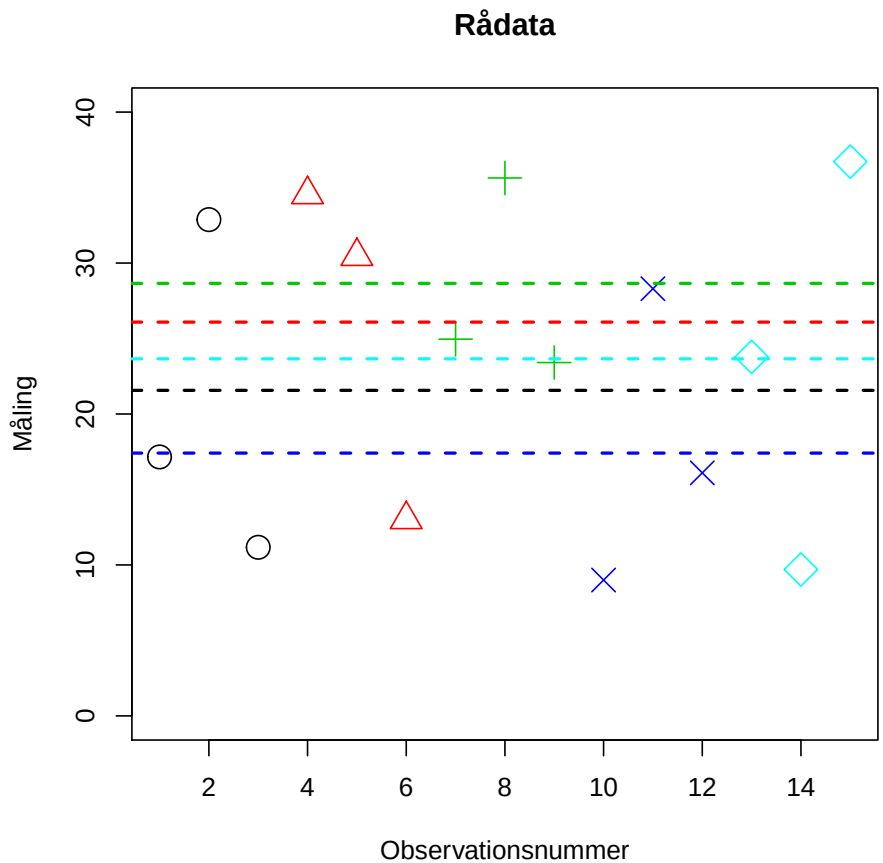
Eksempel 1: hvad er problemet?

"Nu skal du høre: jeg har lavet det her forsøg og udregnet gennemsnit over triplikater. Der er en klar forskel på behandlingerne, men jeg ville gerne underbygge det med noget statistik."



Eksempel 1: hvordan ser rådata ud?

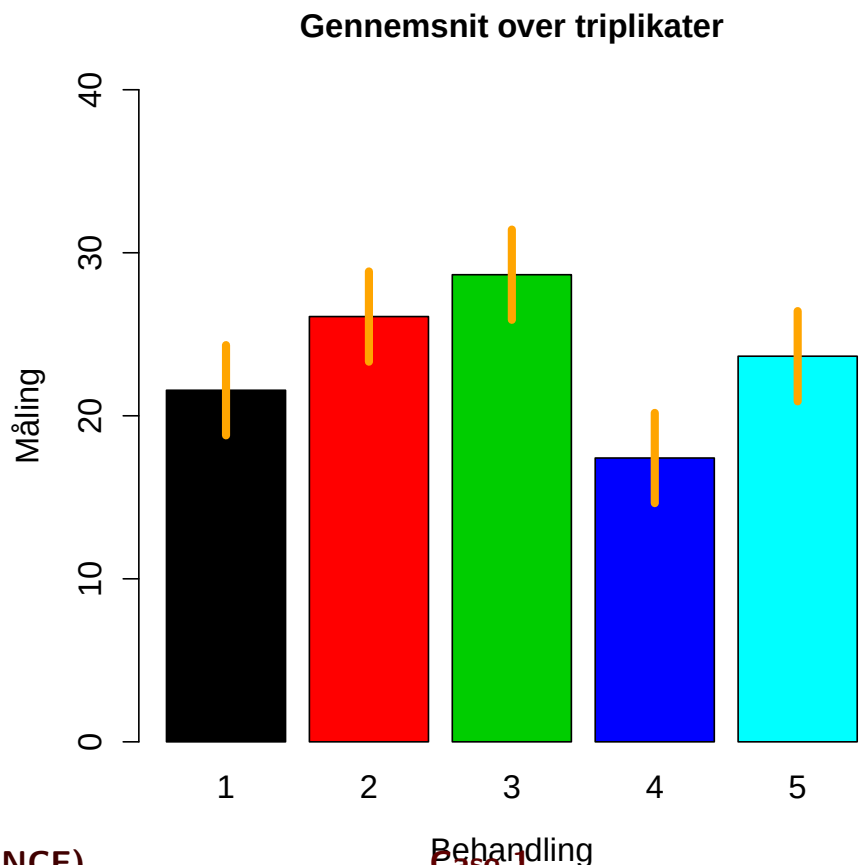




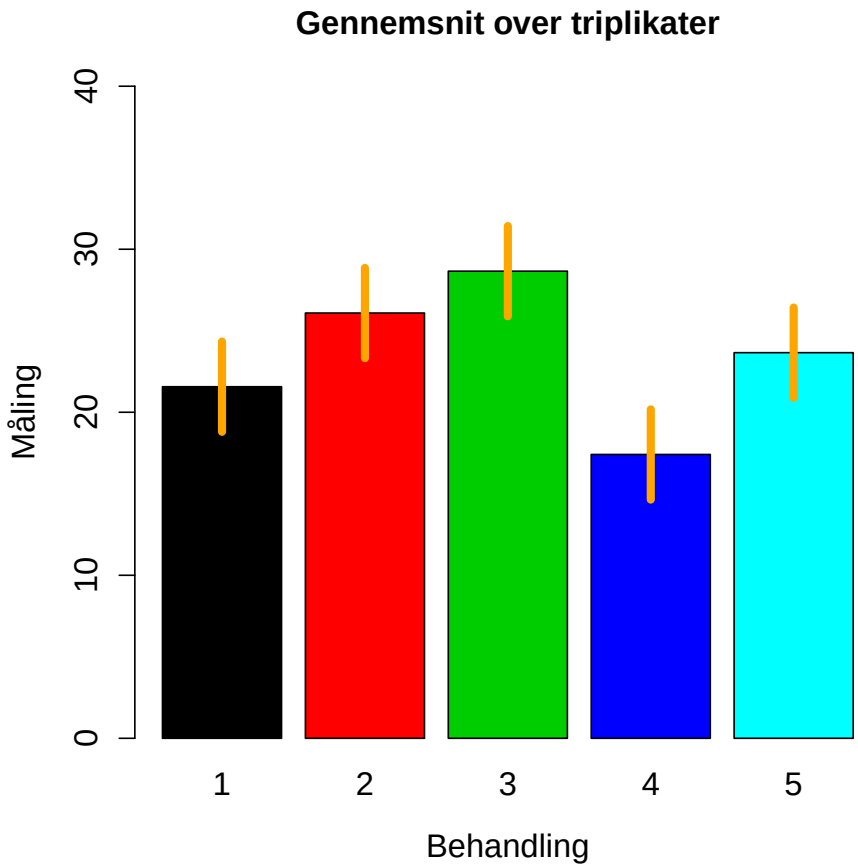
- Gennemsnit siger intet om variationen inden for behandlingsgrupper.
- Grp.-variation har indflydelse på vurdering af behandlingsforskelle.

Eksempel 2: her går det bedre!

"Nu skal du høre: jeg har lavet det her forsøg og udregnet gennemsnit over triplikater med error bars. Der er en klar forskel på behandlingerne, men jeg ville gerne underbygge det med noget statistik."



Eksempel 2: her går det bedre!



- Er der signifikant forskel på behandlingerne?
- Pas på: Ved vurdering af parvise forskelle, kan man ikke bare *lægge error bars sammen*!!!

Eksempel 3: her kører det bare!

```
> mod0<-lm(y~beh-1)
> mod1<-lm(y~1)

> anova(mod1,mod0)
Analysis of Variance Table

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     14 268.599
2     10  46.292  4   222.307 12.006 0.0007813 ***

> summary(mod0)
     Estimate Std. Error t value Pr(>|t|)
beh1   21.567      1.242   17.36 8.51e-09 ***
beh2   26.087      1.242   21.00 1.33e-09 ***
beh3   28.654      1.242   23.07 5.30e-10 ***
beh4   17.410      1.242   14.02 6.70e-08 ***
beh5   23.655      1.242   19.04 3.46e-09 ***

Residual standard error: 2.152 on 10 degrees of freedom
```

Eksempel 3: her kører det bare!

Der er en signifikant behandlingseffekt: $F = 12.006, p < 0.001$

Estimater og konfidensintervaller for slutmodel (1-sidet ANOVA):

$$\hat{\alpha}(\mathtt{A}) = 21.57 \quad [18.80, 24.33], \quad \hat{\sigma} = 2.152.$$

```
> confint(mod0)
         2.5 %   97.5 %
beh1 18.79909 24.33467
...
```

LSD-værdi for parvise sammenligninger:

$$t_{0.975,df} \cdot s \cdot \sqrt{1/n_1 + 1/n_2} = 2.228 \cdot 2.152 \cdot \sqrt{1/3 + 1/3} = 3.92$$

To behandlinger er signifikant forskellige, hvis deres gennemsnit afviger med mere end $LSD = 3.92$.

Vigtigt: To error bars overlapper, hvis forskellen mellem gennemsnit er mindre end 5.54. Det er forkert at lave parvise sammenligninger på denne måde!