SD2 - uge 7, tirsdag

Anne Petersen

7.1 (9.1 i kompenediet)

Vi loader data og kigger på det:

```
setwd("C:/Users/Anne/Dropbox/Arbejde/STATforLIFE2/uge7")
data <- read.table("bms_exercise9_1.txt", header=T)
data
```

```
##      N table harvest yield
## 1  0.5     1   Early 23.95
## 2  0.5     1    Late 26.18
## 3  0.5     2   Early 25.02
## 4  0.5     2    Late 22.71
## 5  1.0     1   Early 34.24
## 6  1.0     1    Late 31.59
## 7  1.0     2   Early 29.87
## 8  1.0     2    Late 32.70
## 9  2.0     1   Early 39.38
## 10 2.0     1    Late 32.74
## 11 2.0     2   Early 34.90
## 12 2.0     2    Late 31.41
## 13 3.0     1   Early 36.29
## 14 3.0     1    Late 43.44
## 15 3.0     2   Early 40.65
## 16 3.0     2    Late 29.77
```

1.

Vi observerer først at alle kombinationsmuligheder af `harvest`, `N` og `table` er repræsenterede. Vi bemærker desuden at de er to blokke, `table=1` og `table=2`, som muligvis kan have forskellige vækstbetingelser. Der er altså tale om et fuldstændigt blokforsøg (og formentligt et randomiseret et).

Randomiseringen kan da foretages i ét trin: For hver blok gennemføres en lodtrækning for placeringen af hver af de otte `N:harvest`-kombinationer.

2.

Vi opstiller en statistisk model:

$$\text{yield}_i = \gamma(\text{harvest}_i \times \text{N}_i) + a(\text{table}_i) + e_i$$

for $i = 1...16$ og hvor det antages at e_i'erne er iid med $e_1 \sim N(0, \sigma^2)$ og at $a(1)$ og $a(2)$ er iid med $a(i) \sim N(0, \sigma_A^2)$.

Vi vil gerne fitte denne model i R. Først omdannes variablene `N`og `table` til faktorer:

```
data$N <- factor(data$N)
data$table <- factor(data$table)

library(nlme)
model1 <- lme(yield ~ N*harvest, random=~1|table, data,
              method="ML")
```

Vi vil gerne se, om den systematiske del af modellen kan reduceres. Vi ønsker derimod ikke at reducere den tilfældige del, da vi gerne vil have mulighed for at generalisere til andre blokke og antagelsen om en tilfældig effekt således er teoretisk meningsfuld. Vi starter med at undersøge, om vekselvirkningseffekten er signifikant:

```
model2 <- lme(yield ~ N + harvest, random=~1|table, data,
              method="ML")
anova(model1, model2)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model1     1 10 98.63507 106.3610 -39.31753
## model2     2  7 94.79748 100.2056 -40.39874 1 vs 2 2.162413  0.5394
```

Vi finder, at vekselvirkningseffekten ikke er signifikant ($p = 0.5394$). Vi reducerer derfor til den additive model, og undersøger nu om hver af hovedeffekterne er signifikante:

```
model3a <- lme(yield ~ harvest, random=~1|table, method="ML", data)
model3b <- lme(yield ~ N, random=~1|table, method="ML", data)
anova(model2, model3a)
```

```
##         Model df       AIC      BIC    logLik   Test  L.Ratio p-value
## model2      1  7  94.79748 100.2056 -40.39874
## model3a     2  4 109.24122 112.3316 -50.62061 1 vs 2 20.44374   1e-04
```

```
anova(model2, model3b)
```

```
##         Model df      AIC       BIC    logLik   Test  L.Ratio p-value
## model2      1  7 94.79748 100.20560 -40.39874
## model3b     2  6 94.10442  98.73996 -41.05221 1 vs 2 1.306945  0.2529
```

Vi finder en signifikant effekt af `N` ($p = 0.0001$), men ikke af `harvest` ($p = 0.2529$). Vi reducerer derfor til `model3b` og undersøger om der fortsat er en signifikant effekt af `N` i denne model:

```
model4a <- lme(yield ~ 1, random=~1|table, method="ML", data)
anova(model3b, model4a)
```

```
##         Model df       AIC       BIC    logLik   Test  L.Ratio p-value
## model3b     1  6  94.10442  98.73996 -41.05221
## model4a     2  3 107.59825 109.91602 -50.79913 1 vs 2 19.49383   2e-04
```

Der er stadig en signifikant effekt af `N` ($p = 0.0002$). Vi undersøger om det er meningsfuldt at antage linearitet, dvs. at bruge `N` som numerisk kovariat i stedet for faktor:

```
model4b <- lme(yield ~ as.numeric(as.character(N)),
               random=~1|table, method="ML", data)
anova(model3b, model4b)
```

```
##         Model df      AIC      BIC    logLik   Test L.Ratio p-value
## model3b     1  6 94.10442 98.73996 -41.05221
## model4b     2  4 95.52553 98.61589 -43.76277 1 vs 2 5.42111  0.0665
```

Vi finder $p = 0.0665$ og reducerer altså til modellen med lineær effekt af `N`. Vores slutmodel er altså `model4b`, dvs. følgende model:

$$\text{yield}_i = \alpha + \beta \cdot \text{N}_i + a(\text{table}_i) + e_i$$

med notation og antagelser som ovenfor. Vi bestemmer parameterestimater og tilhørende konfidensintervaller for modellen (og husker først at genfitte modellen med REML-estimation):

```
model4b_REML <- lme(yield ~ as.numeric(as.character(N)),
                    random=~1|table, method="REML", data)
intervals(model4b_REML)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                                 lower      est.     upper
## (Intercept)                 20.162119 24.683220 29.204322
## as.numeric(as.character(N))  2.430244  4.611864  6.793484
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: table
##                       lower     est.    upper
## sd((Intercept)) 0.05267183 1.222007 28.35105
##
##  Within-group standard error:
##    lower     est.    upper
## 2.640892 3.878352 5.695657
```

Vi finder følgende parameterestimater (95% konfidensintervaller i parantes):

$$\hat{\alpha} = 24.68\,[20.16, 29.20],\ \hat{\beta} = 4.61\,[2.43, 6.79]$$

$$\hat{\sigma}_A = 1.22\,[0.05, 28.35],\ s = 3.88\,[2.64, 5.69]$$

Vi konkluderer altså, at der ikke blev fundet en signifikant effekt af `harvest`, men at der var en signifikant, lineær sammenhæng mellem `yield` og `N`, således at større mængde nitrat er sammenhængende med større udbytte. Vi ser desuden, at den del af variansen, som skyldes blokfaktoren udgør

$$\frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + s^2} = \frac{1.22^2}{1.22^2 + 3.88^2} = 0.0903$$

dvs. kun omtrent 9%. Blokkene forklarer altså ikke en særlig stor mængde af variationen i data, hvilket tyder på, at blokke er relativt homogene.

Opgave 7.2 (9.2 i kompendiet)

Vi indlæser først data og betragter de 10 første observationer:

```
eggdata <- read.table("bms_exercise9_2.txt", header=T)
head(eggdata, 10)
```

```
##    day method taste
## 1    1      1   9.7
## 2    1      2   8.7
## 3    1      4   5.4
## 4    1      5   5.0
## 5    2      2   9.6
## 6    2      3   8.8
## 7    2      6   5.5
## 8    2     10   3.6
## 9    3      2   9.0
## 10   3      4   7.3
```

1.

Vi bemærker først, at vi har to faktorer i forsøget: `day` (niveauer: 1, ..., 15) og `method` (niveauer: 1, ..., 10). Vi anser `day` som en blokfaktor, da vi forestiller os, at der kan være variation i forsøgspersonernes smagspræferencer mellem dage, selvom det ikke er det, vi er interesserede i at undersøge. Vi har derfor $\nu_B = 15$. Det vi er interesserede i at undersøge effekten af, er derimod `method`, som altså er vores behandlingsvariabel. Vi har altså $\nu_T = 10$. Bemærk, at $\nu_B > \nu_T$. Vi ser, at der afprøves 4 behandlinger pr. blok. Altså er $r_B = 4$. Da $r_B < \nu_T$, er der tale om et ufuldsændigt blokdesign. Vi ser desuden, at hver behandling optræder i alt $r_T = 6$ gange. Vi vil gerne undersøge, om hvert par af `method`-niveauer optræder lige mange gange i forsøget, dvs. om der er tale om et balanceret ufuldstændigt blokdesign (BIBD). Vi vil derfor udfylde den såkaldte coincidence matrice. Dette gøres let i hånden ved at kigge på data og tælle hvor mange gange hver kombination forekommer. Alternativt kan man skrive sin egen funktion i R, som gør det automatisk:

```
coincMatrix <- function(treatment, block, data) {
  nBlock <- length(levels(factor(data[, block])))
  treats <- levels(factor(data[, treatment]))
  nTreat <- length(treats)
  crossTab <- table(data[, block], data[, treatment])
  if (any(crossTab > 1)) {
    stop("At least one combination of block and treatment
         occurs more than once. Not valid BIBD")
  }
  m <- matrix(0, nTreat, nTreat, dimnames=list(treats,
                                               treats))
  for (i in 1:nBlock) {
    useLevs <- names(crossTab[i, crossTab[i, ] != 0])
    for (j in 1:length(useLevs)) {
      m[useLevs[j], useLevs] <- m[useLevs[j], useLevs] + 1
    }
  }
  m
}
coincMatrix("method", "day", eggdata)
```

```
##    1 2 3 4 5 6 7 8 9 10
## 1  6 2 2 2 2 2 2 2 2  2
## 2  2 6 2 2 2 2 2 2 2  2
## 3  2 2 6 2 2 2 2 2 2  2
## 4  2 2 2 6 2 2 2 2 2  2
## 5  2 2 2 2 6 2 2 2 2  2
## 6  2 2 2 2 2 6 2 2 2  2
## 7  2 2 2 2 2 2 6 2 2  2
## 8  2 2 2 2 2 2 2 6 2  2
## 9  2 2 2 2 2 2 2 2 6  2
## 10 2 2 2 2 2 2 2 2 2  6
```

Vi ser, at alle par optræder netop $\lambda = 2$ gange og altså er der tale om et BIBD.

Randomisering bør da foretages i to trin:

1. Først trækkes der lod om hvilken dag, der skal svare til hvilken blok
2. Dernæst trækkes der lod om hvilken rækkefølge, behandlingerne ('method') testes for hver blok ('day')

2.

Vi vil nu opstille en statistisk model. Vi lader `day` indgå som tilfældig effekt jf. ovenstående og får derved følgende model:

$$\text{taste}_i = \alpha(\text{method}_i) + b(\text{day}_i) + e_i$$

for $i = 1, ..., 60$ og hvor det antages at e_i'erne er iid med $e_1 \sim N(0, \sigma^2)$ og $b(1), ..., b(15)$ er iid med $b(i) \sim N(0, \sigma_B^2)$.

Vi fitter denne model i R (og omdanner først relevante variable til faktorer):

```
eggdata$day <- factor(eggdata$day)
eggdata$method <- factor(eggdata$method)
modelA <- lme(taste ~ method, random=~1|day, method="ML", eggdata)
```

Vi tester for signifikant virkning af `method`:

```
modelB <- lme(taste ~ 1, random=~1|day, method="ML", eggdata)
anova(modelA, modelB)
```

```
##        Model df      AIC      BIC     logLik   Test  L.Ratio p-value
## modelA     1 12 161.8287 186.9608  -68.91432
## modelB     2  3 283.2185 289.5015 -138.60924 1 vs 2 139.3898  <.0001
```

Vi finder LR=139.34 og $p < 0.0001$. Altså er der en signifikant virkning af `method`. `modelA` er dermed vores slutmodel, idet vi ligesom i ovenstående ikke ønsker at reducere den tilfældige del af modellen af teorestiske årsager. Vi estimater og dertilhørende konfidensintervaller for modellens parametre:

```
modelA_REML <- lme(taste ~ method-1, random=~1|day, method="REML",
                   eggdata)
intervals(modelA_REML)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##             lower     est.     upper
## method1  9.105064 9.800023 10.494981
## method2  8.944943 9.639902 10.334860
## method3  8.278131 8.973089  9.668048
## method4  7.080014 7.774972  8.469931
## method5  6.984183 7.679141  8.374100
## method6  5.076348 5.771307  6.466265
## method7  4.480686 5.175644  5.870603
## method8  3.500356 4.195315  4.890273
## method9  2.888924 3.583882  4.278841
## method10 1.978433 2.673391  3.368350
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: day
##                       lower      est.    upper
## sd((Intercept)) 0.3452485 0.5915873 1.013692
##
##  Within-group standard error:
##     lower      est.     upper
## 0.5612195 0.7081006 0.8934230
```

Vi finder altså følgende parameterestimater (95% konfidensinterval i parantes):

$$\hat{\alpha}(1) = 9.800023\,(9.105064, 10.494981)$$

$$\hat{\alpha}(2) = 9.639902\,(8.944943, 10.334860)$$

$$\hat{\alpha}(3) = 8.973089\,(8.278131, 9.668048)$$

$$\hat{\alpha}(4) = 7.774972\,(7.080014, 8.469931)$$

$$\hat{\alpha}(5) = 7.679141\,(6.984183, 8.374100)$$

$$\hat{\alpha}(6) = 5.771307\,(5.076348, 6.466265)$$

$$\hat{\alpha}(7) = 5.175644\,(4.480686, 5.870603)$$

$$\hat{\alpha}(8) = 4.195315\,(3.500356, 4.890273)$$

$$\hat{\alpha}(9) = 3.583882\,(2.888924, 4.278841)$$

$$\hat{\alpha}(10) = 2.673391\,(1.978433, 3.368350)$$

$$\hat{\sigma}_A = 0.5916,\ s = 0.7081$$

Vi konkluderer, at der er en signifikant effekt af `method`, dvs., at forskellene mellem de forskellige typer æggepulver er signifikante. Endvidere ses det, at metode 1 vurderes til at have bedst smag, metode 2 vurderes til at have næstbedst smag og så videre ned til metode 10, som vurderes til at have dårligst smag.

For at sammenligne metoderne parvis, udregnes LSD-værdien. Bemærk, at da der er tale om et balanceret forsøg, er det kun nødvendigt at udregne én sådan værdi. Husk, at

$$\text{LSD} = t_{0.975,\text{df}} \cdot \text{SE}(\alpha(j) - alpha(i))$$

for $i \neq j$. Hvis vi fitter en model med et intercept, kan vi aflæse SE-størrelsen direkte:

```
modelA_REML_INTERCEPT <- lme(taste ~ method, random=~1|day,
                             method="REML", eggdata)
summary(modelA_REML_INTERCEPT)$tTable
```

```
##                  Value Std.Error DF     t-value      p-value
## (Intercept)  9.8000225 0.3426658 36  28.5993568 2.385234e-26
## method2     -0.1601210 0.4364793 36  -0.3668466 7.158810e-01
## method3     -0.8269335 0.4364793 36  -1.8945535 6.620632e-02
## method4     -2.0250504 0.4364793 36  -4.6395103 4.495777e-05
## method5     -2.1208810 0.4364793 36  -4.8590639 2.309396e-05
## method6     -4.0287157 0.4364793 36  -9.2300260 5.051561e-11
## method7     -4.6243780 0.4364793 36 -10.5947236 1.297837e-12
## method8     -5.6047076 0.4364793 36 -12.8407167 5.267721e-15
## method9     -6.2161401 0.4364793 36 -14.2415446 2.303385e-16
## method10    -7.1266310 0.4364793 36 -16.3275332 3.197818e-18
```

```
(SE <- summary(modelA_REML_INTERCEPT)$tTable[2,2])
```

```
## [1] 0.4364793
```

Og vi kan nu bestemme LSD-værdien:

```
t <- qt(0.975, 6*10-10-15+1)
(LSD <- SE*t)
```

```
## [1] 0.8852211
```

(Bemærk: df = antal observationer - antal niveauer for method - antal niveauer for day + 1)

Vi ser altså, at metoder, som har smagsvurderingsforskelle større end LSD = 0.89 er signifikante. Vi ser dermed at metode 1 har signifikant forskellig smag sammenlignet med metode 3 (og 4, . . . , 10), metode 2 er signifikant forskellig fra metode 4-10 og at metode 3 er signifikant forskellig fra metode 4-10 osv. De fulde resultater kan ses af R-kørslen nedenfor:

```
ests <- fixef(modelA_REML)
abs(ests[1] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##    FALSE    FALSE    FALSE     TRUE     TRUE     TRUE     TRUE     TRUE
##  method9 method10
##     TRUE     TRUE
```

```
abs(ests[2] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##    FALSE    FALSE    FALSE     TRUE     TRUE     TRUE     TRUE     TRUE
##  method9 method10
##     TRUE     TRUE
```

```r
abs(ests[3] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##    FALSE    FALSE    FALSE     TRUE     TRUE     TRUE     TRUE     TRUE
##  method9 method10
##     TRUE     TRUE
```

```r
abs(ests[4] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##     TRUE     TRUE     TRUE    FALSE    FALSE     TRUE     TRUE     TRUE
##  method9 method10
##     TRUE     TRUE
```

```r
abs(ests[5] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##     TRUE     TRUE     TRUE    FALSE    FALSE     TRUE     TRUE     TRUE
##  method9 method10
##     TRUE     TRUE
```

```r
abs(ests[6] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##     TRUE     TRUE     TRUE     TRUE     TRUE    FALSE    FALSE     TRUE
##  method9 method10
##     TRUE     TRUE
```

```r
abs(ests[7] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##     TRUE     TRUE     TRUE     TRUE     TRUE    FALSE    FALSE     TRUE
##  method9 method10
##     TRUE     TRUE
```

```r
abs(ests[8] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE    FALSE
##  method9 method10
##    FALSE     TRUE
```

```r
abs(ests[9] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE    FALSE
##  method9 method10
##    FALSE     TRUE
```

```r
abs(ests[10] - ests) > LSD
```

```
##  method1  method2  method3  method4  method5  method6  method7  method8
##     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
##  method9 method10
##     TRUE    FALSE
```