Eksamen i Statistisk Dataanalyse 2
(NMAB14002U)

14. april 2016

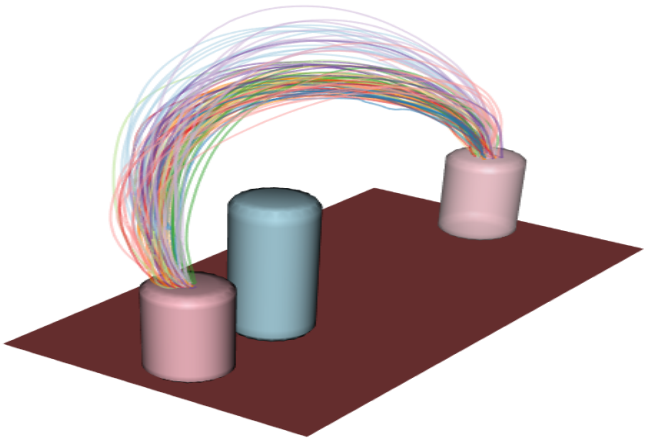
Alle sædvanlige hjælpemidler, herunder bøger, noter, R-programmer og lommeregner samt brug af programmet R på egen PC, er tilladt. Det er *ikke* tilladt at benytte PC til nogle former for aktivitet, som involverer opkobling til et netværk eller kommunikation med andre. Det er tilladt at skrive med blyant. Opgavesættet består af 9 sider med i alt 3 opgaver, der indgår med vægtningen 40 %, 30 % og 30 % i bedømmelsen.

Til besvarelse af opgave 1 har du fået udleveret en USB-nøgle med et datasæt, som du skal indlæse og anvende i R på din egen PC for at kunne besvare opgaven. Til opgave 3 er der vedlagt noget R-udskrift, som kan benyttes i besvarelsen (det er ikke sikkert at alle dele af udskriften skal benyttes). Husk at det er vigtigt at specificere de statistiske modeller og hypoteser du bruger, og at komme med konklusioner på analyserne.

Opgave 1 (4 spørgsmål)

For at opnå en bedre forståelse af menneskets bevægeapparat betragtes data fra et eksperiment, hvor man har bedt 10 forskellige forsøgspersoner flytte en cylindrisk genstand fra en fast startposition til en fast slutposition over en cylindrisk forhindring (-se figur 1).

Figur 1: Skematisk tegning af nogle kurver, der beskriver en cylinder der flyttes fra startposition til slutposition over en forhindring.



Vi interesserer os i denne opgave for den maksimale vertikale hastighed (v i m/s) undervejs i bevægelsen. Hver forsøgsperson har udført eksperimentet med 15 forskellige forsøgsopstillinger:

der anvendes tre forskellige højder af forhindringen (givet ved variablen `size`), og forhindringen kan placeres i 5 forskellige afstande fra startpositionen (givet ved variablen `position`). Alle $3 \cdot 5 = 15$ kombinationer afprøves for hver forsøgsperson. Som en ekstra krølle på halen, har hver forsøgsperson gentaget eksperimentet 10 gange med hver af de 15 forhindringer.

Data til opgaven er udleveret på vedlagte USB-nøgle under filnavnet `vdata.txt` og for at besvare opgaven fuldstændigt, vil det være nødvendigt at køre udvalgte R-kommandoer på din egen medbragte computer. Data kan f.eks. indlæses ved brug af kommandoen

```
vdata <- read.table(file.choose(), header = T)
```

hvor du vælger filen `vdata.txt`. De første linjer i datasættet er organiseret som vist nedenfor.

```
head(vdata, 16)
```

```
##    subj eksp size position         v
## 1     1    1    S       15 1.2539921
## 2     1    1    S       15 1.1453259
## 3     1    1    S       15 1.1464599
## 4     1    1    S       15 1.1430789
## 5     1    1    S       15 1.2008527
## 6     1    1    S       15 1.1864415
## 7     1    1    S       15 1.2837330
## 8     1    1    S       15 1.2273574
## 9     1    1    S       15 1.2648388
## 10    1    1    S       15 1.2002877
## 11    2    1    S       15 0.8210083
## 12    2    1    S       15 0.9170812
## 13    2    1    S       15 0.9591325
## 14    2    1    S       15 0.8153299
## 15    2    1    S       15 0.9099289
## 16    2    1    S       15 0.8915508
```

Data til opgaven er venligst stillet til rådighed af Lars Lau Rakêt, Institut for Matematiske Fag, KU og Britta Grimme, Institut für Neuroinformatik, Ruhr University Bochum, Germany. Ved besvarelsen af delspørgsmål 1.-2. nedenfor skal du opfatte variablen `position` som en faktor.

1. Opskriv en statistisk model, som du vil tage som udgangspunkt for en statistisk analyse af data med henblik på at undersøge, hvordan den maksimale hastighed `v` afhænger af forsøgsopstillingen (givet ved variablene `size` og `position`). Du bedes argumentere for dit valg af model.

2. Foretag modelreduktion med henblik på at afgøre, hvordan den maksimale hastighed `v` afhænger af forhindringen, og angiv parameterestimater samt 95 %-konfidensintervaller for middelværdi- og variansparametre i slutmodellen. Forklar i ord hvad modellen udtrykker.

I det følgende fortsættes analysen ovenfor med henblik på at undersøge muligheden for at lade forhindringens placering (givet ved variablen `position`) indgå som en numerisk variabel.

3. Tag udgangspunkt i din slutmodel fra delspørgsmål 2. Udfør et test eller flere test med henblik på at finde ud af, om det er rimelig at antage, at forhindringens afstand til startpositionen har en lineær sammenhæng med den maksimale hastighed `v`. Du bedes tydeligt gøre rede for, hvilke modeller du tester imod hinanden.

4. Benyt din statistiske analyse fra delspørgsmål 1.-3. til at beregne et estimat og et 95 %-konfidensinterval for forskellen i den forventede maksimale hastighed (`v`) når der benyttes en stor forhindring (`size = T`) henholdsvis en middelstor forhindring (`size = M`) der placeres 15 cm (`position = 15`) fra startpositionen. Besvar det samme spørgsmål, hvis forhindringerne placeres 45 cm (`position = 45`) fra startpositionen.

Opgave 2 (3 spørgsmål)

Ved et dyrkningsforsøg interesserer man sig for udbyttet af 7 forskellige kartoffelsorter (givet ved faktoren `sort` med niveauer $A, B, \ldots, F$). Til forsøget råder man over 7 marker, der hver især er delt i 4 parceller. Nedenfor vises et forslag til en forsøgsplan

G	E	C	A	A	E	B	F	B	C	D	E	G	A
D	F	G	E	C	F	D	A	F	G	C	B	D	B

1. Argumenter for at forsøgsplanen ikke er et balanceret ufuldstændigt blokforsøg (BIBD). Modificer forsøgsplanen ved at bytte om på to af behandlingerne fra mark 2 og mark 7 (talt fra venstre), så forsøgsplanen bliver til et BIBD.

Ved et andet forsøg råder man over 6 bede, der hver skal beplantes med 4 hindbærbuske. Forsøget har til formål at sammenligne forskellige metoder til at optimere udbyttet af hindbær. Ved forsøget afprøves to forskellige beskæringsmetoder givet ved faktoren `B` med niveauer `1=ingen beskæring` og `2=beskæring`, samt to forskellige måder at gøde buskene på givet ved faktoren `G` med niveauer `1=NPK` og `2=havekompost`. Endelig indgår to forskellige hindbærsorter i forsøget givet ved faktoren `S` med niveauerne `1=Autumn Bliss, 2=Glen Prosen`.

2. Giv et forslag til en forsøgsplan for allokering af behandlingerne `B` og `G`, hvis 3 bede (=12 forsøgsenheder) skal beplantes med sorten `Autumn Bliss`, og 3 bede beplantes med sorten `Glen Prosen`. Opskriv en statistisk model, som du ville benytte til analyse af data fra forsøget og tegn et tilhørende faktordiagram. Forklar hvordan randomiseringen bør foretages.

Blandt haveentusiaster er der mange som mener, at det er vigtigt for en god bestøvning (og et stort udbytte), at forskellige hindbærsorter kan bestøve hinanden.

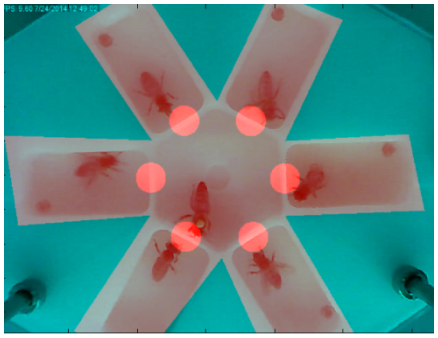
3. Giv et forslag til en forsøgsplan, hvor hver kombination af faktorerne `B`, `G` og `S` optræder lige mange gange på de i alt 24 forsøgsenheder, og hvor der på hvert bed skal være to buske af hvert sort, således at der er optimale forhold for bestøvning. Husk at begrunde dit svar.

Opgave 3 (3 spørgsmål)

Ved et eksperiment ønsker man at undersøge om arbejderbier, som inficeres med *nosema ceranae*, bliver mindre kontaktsøgende overfor bidronningen (formodentlig i et forsøg på at beskytte denne).

Ved eksperimentet anvendes en forsøgsomstilling, der består af en såkaldt *arena* (-se figur 2), hvor bidronningen placeres i et midterkammer, der er forbundet med 6 tunneller, som hver indeholder en arbejderbi.

Figur 2: En arena indeholdende en dronning og seks arbejderbier, hvor dronningen er i kontakt med en af arbejderbierne.



I grænseområdet mellem midterkammer og tunneler (markeret med cirkler på figuren) kan arbejderbien overføre næring fra en sukkerblanding til dronningen. For at stimulere dronningen til at opsøge kontakt med arbejderbierne, så har dronningen ikke selv adgang til føde. Ved forsøget holdes hver arena under observation i 1 time, og det optælles hvor lang `tid` (målt i sekunder), hver enkelt arbejderbi var i kontakt med dronningen.

Forsøget er blevet gentaget 61 gange (forsøgsnummeret er givet ved faktoren `arena`) hver gang med 6 nye arbejderbier, således at datasættet består af $6 \cdot 61 = 366$ datalinjer svarende til en datalinje per arbejderbi. De 366 bier stammer fra 5 forskellige kolonier (givet ved faktoren `colony`). Ud af de 366 arbejderbier var 249 inficeret med *nosema ceranae* (givet ved værdien `yes` af variablen `exposed`). Endvidere har man for hver arbejderbi registreret alderen, `age`, målt i dage.

Data til opgaven er venligst stillet til rådighed af Antoine Lecocq fra Department of Plants and Environmental Sciences, KU. De første datalinjer ser ud som følger

```
bees <- read.table(file = "bees.txt", header = T)
head(bees, 20)
```

```
##    colony arena age exposed tid
## 1       A    A1  15     yes  27
## 2       A    A1  15     yes  20
## 3       A    A1  15     yes  25
## 4       A    A1  15      no 177
## 5       A    A1  15     yes 543
## 6       A    A1  15     yes 259
## 7       B   A10   9      no  11
```

```
## 8       B   A10   9      no 962
## 9       B   A10   9      no  88
## 10      B   A10   9      no   5
## 11      B   A10   9      no  93
## 12      B   A10   9      no 624
## 13      B   A11  13      no  40
## 14      B   A11  13      no  14
## 15      B   A11  13      no  84
## 16      B   A11  13      no 113
## 17      B   A11  13      no 154
## 18      B   A11  13      no  88
## 19      B   A12  13      no 895
## 20      B   A12  13      no  54
```

Besvar følgende 3 delspørgsmål ved brug af R-udskriften sidst i opgavesættet. Bemærk at der kan være dele af R-udskriften, som ikke skal benyttes.

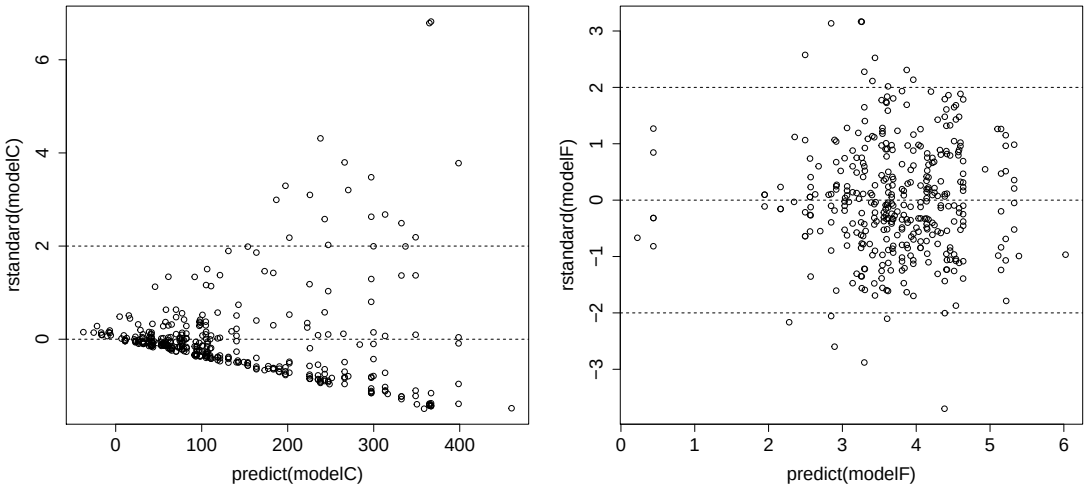
1. Opskriv en statistisk model svarende til den af modellerne `modelA-modelF` fra R-udskriften, som du finder mest velegnet til at analysere data. Husk at begrunde dit valg af model.

2. Angiv estimater for samtlige parametre, der indgår i modellen fra delspørgsmål 1. Benyt relevante dele af R-udskriften til at diskutere, hvilke variable i modellen der bidrager væsentligt til at forklare variationen i data.

3. Et tidligere forsøg med 10 dage gamle ikke-eksponerede bier (`exposed=no`) viste, at arbejderbierne typisk tilbragte 1 % af deres tid med dronningen. Diskuter om resultaterne fra det nye forsøg (som du har regnet på her) er i overensstemmelse med de tidligere resultater.

```r
### nogle statistiske modeller
library(lme4)
library(nlme)
modelA <- lmer(tid ~ exposed + age + (1|colony) + (1|arena)
               , data = bees)
modelB <- lme(tid ~ exposed + age, random = ~ 1 | arena, data = bees)
modelC <- lm(tid ~ exposed + age + arena, data = bees)
modelD <- lmer(log(tid) ~ exposed + age + (1|colony) + (1|arena)
               , data = bees)
modelE <- lme(log(tid) ~ exposed + age, random = ~ 1 | arena
              , data = bees)
modelF <- lm(log(tid) ~ exposed + age + arena, data = bees)
```

```
### Nogle figurer ...
plot(predict(modelC),rstandard(modelC))
plot(predict(modelF),rstandard(modelF))
```




```
### udpluk af summary for udvalgte modeller
summary(modelA)
```

```
##              Estimate Std. Error    t value
## (Intercept) 307.63637  91.175272  3.3741207
## exposedyes  -10.65667  31.433123 -0.3390267
## age         -13.41915   6.610983 -2.0298270
```

```
VarCorr(modelA)

##  Groups   Name        Std.Dev.
##  arena    (Intercept)   0.000
##  colony   (Intercept)  90.481
##  Residual             268.370

confint(modelA)

##                  2.5 %     97.5 %
## .sig01         0.00000  48.184244
## .sig02        26.14682 188.095645
## .sigma       249.34563 288.622627
## (Intercept) 117.18329 487.768550
## exposedyes  -71.83591  51.203032
## age         -26.27817   1.402519
```

```r
summary(modelB)
```

```
##                  Value Std.Error  DF     t-value    p-value
## (Intercept) 155.640154 70.484099 303  2.20815979 0.02798181
## exposedyes   -2.003368 30.983896 303 -0.06465837 0.94848865
## age          -1.169069  5.182885 303 -0.22556335 0.82169311
```

```r
VarCorr(modelB)

## arena = pdLogChol(1)
##             Variance     StdDev
## (Intercept) 2.049888e-03   0.04527568
## Residual    7.563597e+04 275.01993736

intervals(modelB)

## Approximate 95% confidence intervals
##
##  Fixed effects:
##                 lower       est.      upper
## (Intercept)  16.93984 155.640154 294.340463
## exposedyes  -62.97423  -2.003368  58.967490
## age         -11.36808  -1.169069   9.029937
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: arena
##                        lower       est.        upper
## sd((Intercept)) 9.98036e-26 0.04527568 2.053921e+22
##
##  Within-group standard error:
##    lower     est.    upper
## 255.7249 275.0199 295.7708
```

```r
summary(modelD)
```

```
##                Estimate Std. Error    t value
## (Intercept)  4.92910818 0.56085586  8.7885472
## exposedyes  -0.11270663 0.19010037 -0.5928796
## age         -0.08902916 0.04234406 -2.1025182
```

```
VarCorr(modelD)
```

```
##  Groups   Name        Std.Dev.
##  arena    (Intercept) 0.45401
##  colony   (Intercept) 0.44039
##  Residual             1.50899
```

```
confint(modelD)
```

```
##                  2.5 %        97.5 %
## .sig01       0.1142797  0.6936590145
## .sig02       0.0000000  0.9573694857
## .sigma       1.3950722  1.6363496014
## (Intercept)  3.7696522  6.0181229475
## exposedyes  -0.4844662  0.2763936397
## age         -0.1704116 -0.0008033664
```

```
summary(modelE)
```

```
##                   Value  Std.Error  DF    t-value      p-value
## (Intercept)  4.53281371 0.47640192 303  9.5146839 5.954684e-19
## exposedyes  -0.02848505 0.18800468 303 -0.1515125 8.796723e-01
## age         -0.06433223 0.03585581 303 -1.7941924 7.377928e-02
```

```
VarCorr(modelE)
```

```
## arena = pdLogChol(1)
##             Variance  StdDev
## (Intercept) 0.2640256 0.5138342
## Residual    2.3057697 1.5184761
```

```
intervals(modelE)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                   lower        est.       upper
## (Intercept)  3.5953385  4.53281371 5.470288892
## exposedyes  -0.3984452 -0.02848505 0.341475083
## age         -0.1348902 -0.06433223 0.006225704
## attr(,"label")
## [1] "Fixed effects:"
##
```

```
##  Random Effects:
##   Level: arena
##                      lower      est.     upper
## sd((Intercept)) 0.3233856 0.5138342 0.8164419
##
##  Within-group standard error:
##    lower     est.    upper
## 1.402283 1.518476 1.644297
```

```
### estimable anvendt paa forskellige modeller
library(gmodels)
est1 <- c(10, 0, 10)
est2 <- c(0, 0, 10)
est3 <- c(1, 0, 10)
est4 <- c(10, 0, 1)
est5 <- c(0, 0, 1)
est6 <- c(1, 0, 1)
est <- rbind(est1, est2, est3, est4, est5, est6)
```

```
estimable(modelB, est, conf.int = 0.95)
```

```
##       Estimate Std. Error t value   DF Pr(>|t|)  Lower.CI  Upper.CI
## est1 1544.7108   656.7679  2.3520 3030   0.0187  256.9550 2832.4667
## est2  -11.6907    51.8289 -0.2256 3030   0.8216 -113.3140   89.9326
## est3  143.9495    28.9817  4.9669  303   0.0000   86.9186  200.9804
## est4 1555.2325   700.0096  2.2217  303   0.0270  177.7368 2932.7282
## est5   -1.1691     5.1829 -0.2256  303   0.8217  -11.3681    9.0299
## est6  154.4711    65.6768  2.3520  303   0.0193   25.2307  283.7115
```

```
estimable(modelE, est, conf.int = 0.95)
```

```
##      Estimate Std. Error t value   DF Pr(>|t|) Lower.CI Upper.CI
## est1  44.6848     4.4291 10.0888 3030   0.0000  36.0004  53.3692
## est2  -0.6433     0.3586 -1.7942 3030   0.0729  -1.3464   0.0597
## est3   3.8895     0.1865 20.8497  303   0.0000   3.5224   4.2566
## est4  45.2638     4.7304  9.5688  303   0.0000  35.9553  54.5723
## est5  -0.0643     0.0359 -1.7942  303   0.0738  -0.1349   0.0062
## est6   4.4685     0.4429 10.0888  303   0.0000   3.5969   5.3401
```