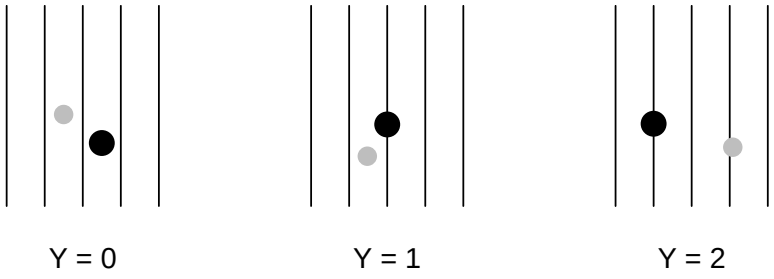
Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Statistik 2
4 timer skriftlig eksamen
22. juni 2017

4 timers skriftlig prøve. Alle hjælpemidler tilladt (inkl. computer uden netforbindelse). Opgavesættet består af tre opgaver med 16 delopgaver. Ved bedømmelsen vægtes alle delopgaver ens. Til besvarelsen af opgave 3 har du fået udleveret et datasæt på en USB-nøgle.

Opgave 1

På et (uendelig stort) trægulv med parallelle planker (uden mellemrum) kastes tilfældigt to papbrikker. Papbrikkerne er ikke nødvendigvis ens, men har dog en form, så de hver højst kan berøre to planker i gulvet (dvs. krydse en streg i gulvet). Lad Y angive antallet af papbrikker der berører to planker. De mulige udfald kan realiseres ved situationerne indikeret på følgende figur.



Sandsynligheden for at Y antager værdien j kaldes p_j, hvor $p_0+p_1+p_2 = 1$, og fordelingen af Y har tæthed $f(y) = \prod_{j=0}^{2} p_j^{1_{(y=j)}}$ mht. tællemålet på $\{0, 1, 2\}$. Lad nu $Y_1, Y_2, \ldots, Y_n$ være uafhængige og identisk fordelte med samme fordeling som Y og betragt estimatoren $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} 1_{(Y_i=2)}$, der angiver andelen

af forsøg, hvor vi observerer at begge to brikker krydser en streg, når de 2 papbrikker kastes n gange uafhængigt af hinanden.

1. Argumenter for at estimatoren $\hat{\theta}_n$ er konsistent for p_2 og asymptotisk normalfordelt med asymptotisk varians $\frac{p_2(1-p_2)}{n}$.

2. Vis at fordelingen af Y kan opfattes som en todimensional eksponentiel familie, hvis man parametriserer med $\theta = (\theta_1, \theta_2)$, hvor

$$\theta_1 = \log\left(\frac{p_1}{1-p_1-p_2}\right) \quad \text{og} \quad \theta_2 = \log\left(\frac{p_2}{1-p_1-p_2}\right).$$

   Angiv normeringskonstanten og den kanoniske stikprøvefunktion.

Vi antager i delspørgsmål 3.-5., at der benyttes to helt ens papbrikker. I dette tilfælde vil Y følge en binomialfordeling med antalsparameter 2 og sandsynlighedsparameter p, hvor p angiver sandsynligheden for at hver enkelt brik krydser en streg. Dette kan benyttes ved besvarelse af resten af opgaven.

3. Under binomialfordelingsantagelsen vil $P_p(Y = 2)$ være lig med p^2. Giv et forslag til, hvordan man kan konstruere en konsistent estimator for p baseret på $\hat{\theta}_n$ fra delspørgsmål 1. Angiv den asymptotiske fordeling af din estimator.

4. Opskriv likelihoodfunktion, loglikelihoodfunktion og scorefunktion for p svarende til uafhængige identiske fordelte observationer $Y_1, \ldots, Y_n$.

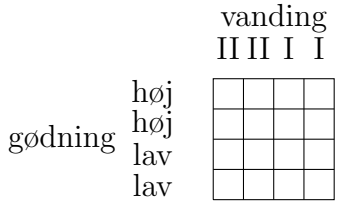
5. Gør rede for, at der findes en entydig maksimaliseringsestimator for p, og angiv dens asymptotiske fordeling.

I virkeligheden interesserer man sig for at estimere bredden B af plankerne i gulvet så præcist som muligt ud fra n kast med to genstande med kendt størrelse, hvor man for hvert kast alene observerer om der er nul, en eller to genstande, som krydser en streg. Alle planker antages at have samme bredde. Lad os betragte situationen, hvor man har adgang til cirkulære papbrikker med kendt diameter $D < B$. Det kan uden bevis benyttes, at en tilfældigt kastet cirkulær papbrik vil krydse en streg i gulvet med sandsynlighed $p = \frac{D}{B}$.

6. Argumenter for, at maksimaliseringsestimatoren er asymptotisk normalfordelt med asymptotisk varians $\frac{1}{2n} \cdot \frac{B^2(B-D)}{D}$, når binomialfordelingsmodellen reparametriseres med den ukendte bredde B.

Opgave 2

Vi betragter et dyrkningsforsøg med gulerødder udført på 16 forsøgsplots organiseret i et kvadrat som anført på figuren.



Planterne i de to første vandrette rækker modtager en høj dosis gødning, mens planter i de to sidste vandrette rækker modtager en lav dosis gødning. Planter i de to første lodrette søjler vandes to gange om dagen, mens planter i de to sidste lodrette søjler kun vandes en gang om dagen. Det samlede udbytte på hvert af de 16 forsøgsplot angives med vektoren $X = (X_i)_{i \in I}$, og man kan tænke på eksperimentet som et tofaktorforsøg med faktorerne `G` (gødning) og `V` (vanding). Observationerne X_i antages at være uafhængige og normalfordelte $\mathcal{N}(\xi_i, \sigma^2)$. Desværre ødelægges alle gulerødderne fra forsøgsplottet i øverste venstre hjørne, så målingen af udbyttet herfra indgår ikke i analyseresultaterne nedenfor.

1. Brug R-udskriften sidst i opgaven til at teste den additive hypotese op mod vekselvirkningsmodellen.

Det oplyses, at

$$\begin{aligned} ||X||^2 &= 1089.907, \quad ||P_{G\times V}X||^2 = 1076.128, \quad ||P_G X||^2 = 1055.808 \\ ||P_V X||^2 &= 1074.725, \quad ||P_1 X||^2 = 1055.714 \end{aligned}$$

2. Angiv dimensionen af det additive underrum $L_G + L_V$ og udregn størrelsen

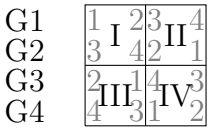
$$\frac{(||P_{G\times V}X||^2 - ||P_G X||^2 - ||P_V X||^2 + ||P_1 X||^2)/(\dim(L_{G\times V}) - \dim(L_G + L_V))}{(||X||^2 - ||P_{G\times V}X||^2)/(N - \dim(L_{G\times V}))},$$

   hvor N er antallet af observationer. Sammenlign med resultaterne fra delspørgsmål 1. og kommenter.

3. Estimer parametrene i den additive model hvor begge faktorerne indgår, og angiv estimatorernes simultane fordeling. Forklar hvordan parametrene i R-udskriften skal fortolkes. Det er ikke et krav, at du angiver estimatet for σ^2.

Næste år udvides dyrkningsforsøget, således at der anvendes 4 forskellige niveauer af faktorerne gødning (`G`) og vanding (`V`). Desuden anvendes nu fire sorter `1, 2, 3, 4`, således at de 16 forsøgsplot tilplantes som anført på figuren nedenfor (til venstre).

G1	1	2	3	4
G2	3	4	2	1
G3	2	1	4	3
G4	4	3	1	2
	A	B	C	D



Alle planter i hver vandret række modtager samme mængde gødning med labels `G1,G2,G3,G4`, mens planter i hver søjle modtager samme daglige mængde vand med labels `A,B,C,D`. Endelig benyttes fire forskellige typer jord, således at jordtypen for de fire forsøgsplot i hvert hjørne alle har samme værdi af faktoren `J` med labels `I, II, III, IV` (illustreret skematisk på figuren til højre). Bemærk at man nu bør tænke på eksperimentet som et firefaktorforsøg med fire faktorer `S` (sort), `G` (gødning), `V` (vanding) og `J` (jordtype), som hver har fire labels. Det er nærliggende at omtale forsøgsplanen som et *sudoku*-design, selvom dette næppe er en veletableret terminologi.

4. Find minimum ($\wedge$) mellem hvert par af de fire faktorer `S, G, V` og `J`.

5. Angiv dimensionen af det additive underrum $L_S + L_G + L_V + L_J$.

```
### data

data

##      G  V     y
## 1  hoj II    NA
## 2  hoj II  9.92
## 3  hoj  I  8.74
## 4  hoj  I  6.24
## 5  hoj II  9.93
## 6  hoj II 10.45
## 7  hoj  I  6.04
## 8  hoj  I  8.00
## 9  lav II  8.73
## 10 lav II 10.56
## 11 lav  I  7.34
## 12 lav  I  8.57
## 13 lav II  8.25
## 14 lav II  9.31
## 15 lav  I  5.57
## 16 lav  I  8.19

### fit af modeller

mod1 <- lm(y ~ G * V)
mod2 <- lm(y ~ G + V)
mod3 <- lm(y ~ G)

### test

anova(mod2, mod1)

## Analysis of Variance Table
##
## Model 1: y ~ G + V
## Model 2: y ~ G * V
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     12 14.797
## 2     11 13.779  1    1.0177 0.8124 0.3867
```

```r
anova(mod3, mod2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ G
## Model 2: y ~ G + V
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     13 34.099
## 2     12 14.797  1    19.302 15.653 0.001906 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
### estimater

summary(mod1)$coefficients
```

```
##             Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  10.1000  0.6461834 15.6302367 7.384091e-09
## Glav         -0.8875  0.8548203 -1.0382299 3.214470e-01
## VI           -2.8450  0.8548203 -3.3281848 6.732883e-03
## Glav:VI       1.0500  1.1649238  0.9013465 3.867137e-01
```

```r
summary(mod2)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.7769231  0.5334385 18.328118 3.855168e-10
## Glav        -0.3221154  0.5761794 -0.559054 5.864070e-01
## VI          -2.2796154  0.5761794 -3.956433 1.905544e-03
```

```r
A <- model.matrix(mod2) ### designmatrix
t(A) %*% A
```

```
##             (Intercept) Glav VI
## (Intercept)          15    8  8
## Glav                  8    8  4
## VI                    8    4  8
```

```r
solve(t(A) %*% A)
```

```
##             (Intercept)        Glav          VI
## (Intercept)   0.2307692 -0.15384615 -0.15384615
## Glav         -0.1538462  0.26923077  0.01923077
## VI           -0.1538462  0.01923077  0.26923077
```

Opgave 3

Accelerationsmålinger fra heste kan benyttes til udregning af en symmetriscore `S`. En lav værdi er udtryk for, at hesten bevæger sig meget symmetrisk, når den traver. Ved et eksperiment ønsker man at undersøge, hvordan symmetriscoren påvirkes af, om en hest traver ligeud (`M`) eller mod højre (`H`) eller venstre (`V`) i cirkler med lille diameter (=`8` meter) eller stor diameter (=`16` meter). Der indgår 5 måleserier for hver af de 27 heste i eksperimentet.

Data til opgaven er gjort tilgængelige i filen `stat2juni2017opg3.txt` som er udleveret på vedlagte USB nøgle. Data er stillet til rådighed af Maj Halling Thomsen, Institut for Produktionsdyr og Heste, KU. I datasættet indgår faktorerne `hest`, `retning` og `diameter` samt symmetriscoren `S`.

Det kan benyttes, at $\mathbb{G}$ (nedenfor) er afsluttet over for dannelse af minimum.

$$\begin{aligned}\mathbb{G} \;=\; \{&\texttt{hest} \times \texttt{retning}, \quad \texttt{diameter} \times \texttt{retning}, \quad \texttt{hest},\\ &\texttt{retning}, \quad \texttt{diameter}, \quad \texttt{retning} \wedge \texttt{diameter}, \quad \texttt{1}\}.\end{aligned}$$

Delspørgsmål 2.-5. kan besvares uafhængigt af delspørgsmål 1.

1. Forklar hvad faktoren $\texttt{retning} \wedge \texttt{diameter}$ beskriver. Organiser faktorerne fra $\mathbb{G}$ i et faktorstrukturdiagram og tilføj dimensionerne $\dim(L_G)$ og $\dim(V_G)$ for $G \in \mathbb{G}$. (Du skal ikke tilføje $||P_G X||^2$ og $||Q_G X||^2$!)

2. Opstil en passende varianskomponentmodel til analyse af data.

3. Estimer modellen fra 2. i R og angiv estimater for: i) variansparametrene, ii) kovariansmatricen svarende til alle observationer (=symmetrimålinger) taget på samme hest.

4. Undersøg hvordan faktorerne `diameter` og `retning` påvirker symmetriscoren.

5. Man kunne få den ide, at symmetriscoren er en lineær funktion af den reciprokke diameter ($\frac{1}{\texttt{diameter}}$). Her skal trav langs en ret linje fortolkes som en diameter på $+\infty$ og dermed en reciprok diameter på 0. Giv et forslag til, hvordan man kan teste denne hypotese og udfør dette test.

```
### nyttig R-kode
acc_data <- read.table(file = "stat2juni2017opg3.txt", header = T)
acc_data$invdiam <- 1/acc_data$diameter
acc_data$invdiam[acc_data$diameter == 0] <- 0
```