

## Eksamens i Statistik 1, vejledende besvarelse

### 9. april 2015

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `april15.R`.

### Opgave 1

- Vi regner først på middelværdi og varians for  $X_i$ :

$$E_\theta X_i = \sum_{x=1}^{\infty} \theta x \left(\frac{1}{2}\right)^{x+1} = \frac{\theta}{2} \sum_{x=1}^{\infty} x \left(\frac{1}{2}\right)^x = \frac{\theta}{2} \frac{\frac{1}{2}}{(1-\frac{1}{2})^2} = \theta$$

$$E_\theta X_i^2 = \sum_{x=1}^{\infty} \theta x^2 \left(\frac{1}{2}\right)^{x+1} = \frac{\theta}{2} \sum_{x=1}^{\infty} x^2 \left(\frac{1}{2}\right)^x = \frac{\theta}{2} \frac{\frac{1}{2} \cdot \frac{3}{2}}{(1-\frac{1}{2})^3} = 3\theta$$

$$V_\theta X_i = E_\theta X_i^2 - (E_\theta X_i)^2 = 3\theta - \theta^2$$

hvor vi har benyttet formlerne for de uendelige summer der er givet i opgaven.

Da  $\tilde{\theta}$  er gennemsnit af  $n$  iid.  $X_i$ 'er fås umiddelbart at

$$E_\theta \tilde{\theta} = \theta, \quad V_\theta \tilde{\theta} = \frac{1}{n}(3\theta - \theta^2)$$

Specielt er  $\tilde{\theta}$  central da middelværdien er lig den sande parameter.

- Likelihoodfunktionen er  $L_x(\theta) = \prod_{i=1}^{\infty} f_\theta(x_i)$ .

De  $x_i$  der er 0, bidrager hver med værdien  $1 - \frac{\theta}{2}$ , og da der er  $n_0$  af disse bidrager de tilsammen med faktoren  $(1 - \frac{\theta}{2})^{n_0}$ . De  $x_i$  der ikke er 0, bidrager med  $\theta(\frac{1}{2})^{x+1}$ , men ledet  $(\frac{1}{2})^{x+1}$  afhænger ikke af  $\theta$  og er derfor blot en proportionalitetsfaktor. Der er  $n - n_0$  af disse  $x_i$  der derfor i alt bidrager med  $\theta^{n-n_0}$ . I alt får formlen fra opgaven.

Logaritme og efterfølgende differentiation giver følgende funktioner der alle er defineret for  $\theta \in (0, 2)$ :

$$\begin{aligned} \ell_x(\theta) &= -n_0 \log\left(1 - \frac{\theta}{2}\right) - (n - n_0) \log \theta \\ S_x(\theta) &= \frac{n_0}{2 - \theta} - \frac{n - n_0}{\theta} \\ I_x(\theta) &= \frac{n_0}{(2 - \theta)^2} + \frac{n - n_0}{\theta^2}. \end{aligned}$$

- Vi løser først scoreligningen og ser at løsningen er entydig:

$$S_x(\theta) = 0 \Leftrightarrow \frac{n_0}{2 - \theta} = \frac{n - n_0}{\theta} \Leftrightarrow n\theta = 2(n - n_0) \Leftrightarrow \theta = \frac{2(n - n_0)}{n}$$

Bemærk at løsningen ligger i  $(0, 2)$  når  $0 < n_0 < n$ . Da  $I_x(\theta) > 0$  for alle  $\theta \in (0, 2)$ , er løsningen et entydigt minimum for  $\ell_x$ , således at

$$\hat{\theta} = \frac{2(n - N_0)}{n}$$

som ønsket.

(Argument for påstanden efter spørgsmål 3: Hvis  $n_0 = 0$ , er  $L_x(\theta) = \theta^n$  der har maksimum  $[0, 2]$  for  $\theta = 2$ . Hvis  $n_0 = n$ , er  $L_x(\theta) = (1 - \frac{\theta}{2})^n$  der har der har maksimum  $[0, 2]$  for  $\theta = 0$ . Begge dele passer med formel (1). Bemærk i øvrigt at  $f_\theta$  faktisk også er en tæthed når  $\theta \in \{0, 2\}$ .)

4.  $N_0$  er binomialfordelt med antalsparameter  $n$  og sansynlighedsparameter  $1 - \frac{\theta}{2}$ . Desuden er  $n - N_0$  er binomialfordelt med antalsparameter  $n$  og sansynlighedsparameter  $\frac{\theta}{2}$ .

Det følger at

$$\begin{aligned} E_\theta \hat{\theta} &= \frac{2}{n} \cdot \frac{n\theta}{2} = \theta \\ V_\theta \hat{\theta} &= \frac{4}{n^2} \cdot n \frac{\theta}{2} \left(1 - \frac{\theta}{2}\right) = \frac{1}{n}(2\theta - \theta^2) \end{aligned}$$

Vi ser at  $\hat{\theta}$  er central. De to estimatorer  $\hat{\theta}$  og  $\tilde{\theta}$  er altså begge centrale, så vi vil foretrække den med mindst varians. Da

$$V_\theta \hat{\theta} = \frac{1}{n}(2\theta - \theta^2) < \frac{1}{n}(3\theta - \theta^2) = V_\theta \tilde{\theta}$$

foretrækker vi således  $\hat{\theta}$  (MLE).

5. Vi får

$$\hat{\theta} = 1.6, \quad \log L_x(1) = -2.079, \quad \log L_x(\hat{\theta}) = 0.812.$$

Vi får derefter at den observerede værdi af  $LR$  er

$$LR(x) = 2 \left( \log L_x(\hat{\theta}) - \log L_x(1) \right) = 5.78.$$

Værdien skal vurderes i  $\chi^2$ -fordelingen med 1 frihedsgrad. Dette giver  $p$ -værdien

$$p = P(LR \geq 5.78) = 0.016.$$

Hypotesen afvises, og det tyder altså på at  $\theta \neq 1$ .

6. Det udfyldte skema ser således ud (varierer naturligvis en smule fra simulation til simulation):

$n$	$\theta$	Relativ hyppighed hvormed hypotesen forkastes
50	1	0.0668
250	1	0.0498
50	1.2	0.3372
250	1.2	0.8894
50	1.4	0.8590

Det bemærkes at

- testets faktiske størrelse (niveau) er meget tæt på de ønskede 5% når  $n = 250$ , men lidt for stort når  $n = 50$ .
- styrken af testet som forventet stiger når  $n$  vokser og når afstanden mellem den sande værdi og hypoteseværdien stiger.

## Opgave 2

1. Da  $(X_1, X_2)^T$  og  $X_3$  er uafhængige og hver for sig normalfordelte, bliver fordelingen af  $X$  også en normalfordeling. Middelværdi og varians er

$$EX = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad VX = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

Da  $\det(VX) = 0$ , er  $VX$  ikke invertibel, så fordelingen af  $X$  er en singulær normalfordeling.

2. Vi har  $Y = CX$  hvor

$$C = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$$

Det følger at

$$Y \sim N(0, CVX C^T),$$

og når man regner på variansmatricen får man

$$VX = \begin{pmatrix} 18 & 0 \\ 0 & 18 \end{pmatrix}$$

Denne matrix er invertibel, så fordelingen af  $Y$  er en regulær normalfordeling. Desuden er  $Y_1$  og  $Y_2$  uafhængige.

3. Betragt så

$$Z = X_1 + c \cdot X_2 = (1 \ c) \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

der er normalfordelt på  $\mathbb{R}$  med middelværdi 0 og varians

$$VZ = (1 \ c) \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ c \end{pmatrix} = 1 + 4c^2 + 4c$$

Kravet er at  $Z$  er konstant med sandsynlighed 1, altså at  $VZ = 0$ . Men

$$4c^2 + 4c + 1 = 0 \Leftrightarrow c = -\frac{1}{2}.$$

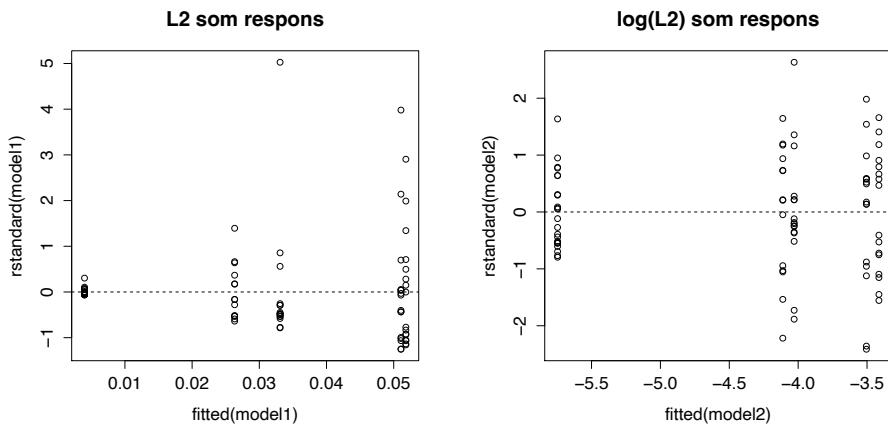
Der gælder altså at  $X_1 - X_2/2$  er 0 med sandsynlighed 1.

## Opgave 3

- Modellerne fitteres for eksempel med følgende kommandoer:

```
model1 <- lm(L2 ~ grp, data=L2Data)
model2 <- lm(log(L2) ~ grp, data=L2Data)
```

Residualplottene er vist nedenfor. Der er tydelige problemer med varianshomogenitet når  $L_2$  bruges som respons. Residualplottet ser meget bedre ud når  $\log(L_2)$  benyttes som respons.



- Det nemmeste er at fitte modellen uden referencegruppe:

```
model2A <- lm(L2 ~ grp - 1, data=L2Data)
```

Så kan  $\hat{\alpha}_{\text{Rask}}$  og  $\text{SE}(\hat{\alpha}_{\text{Rask}})$  aflæses direkte fra summary:

$$\hat{\alpha}_{\text{Rask}} = -5.7465, \quad \text{SE}(\hat{\alpha}_{\text{Rask}}) = 0.2081$$

Fordelingen af  $\hat{\alpha}_{\text{Rask}}$  er  $N(\alpha_{\text{Rask}}, \sigma^2/23)$  da der er 23 raske heste i studiet.

- Hypotesen om ens fordeling i alle grupper kan skrives som  $H : \xi \in L_1$  svarende til den konstante faktor eller som

$$H : \alpha_{\text{Rask}} = \alpha_{\text{HF}} = \alpha_{\text{HB}} = \alpha_{\text{VF}} = \alpha_{\text{VB}}$$

Det testes med det sædvanlige  $F$ -test. I dette tilfælde fås  $f = 18.195$  og en  $p$ -værdi på  $1.157 \cdot 10^{-10}$ . Hypotesen afvises altså klart, så der *er* forskel på middelværdierne.

Hypotesen om at de fire halthedsgrupper ikke adskiller sig fra hinanden er

$$H : \alpha_{\text{HF}} = \alpha_{\text{HB}} = \alpha_{\text{VF}} = \alpha_{\text{VB}}$$

Dette svarer til at slå de fire grupper sammen. Vi får  $f = 1.96$  og en  $p$ -værdi på 0.13. Vi kan derfor ikke afvise hypotesen, og der er ikke noget i data der tyder på at fordelingen af  $\log(L_2)$  er forskellig i de fire halthedsgrupper.

- Eftersom der ikke er forskel på de fire halthedegrupper, slås de sammen, og det giver mening at tale om forskellen i middelværdi mellem halte og raske heste.

Forskellen estimeres til 1.99 med 95% konfidensinterval 1.50–2.49. Forskellen er signifikant forskellig fra 0 ( $p$ -værdien er  $10^{-12}$ ).

## Eksamens i Statistik 1, vejledende besvarelse

### 14. april 2016

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `april16.R`.

#### Opgave 1

- Likelihoodfunktionen og log-likelihoodfunktionen (på nær en additiv konstant) for en observation  $x$  er givet ved

$$\begin{aligned} L_x(\beta, \sigma^2) &= \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - e^{\beta t_i})^2\right) \\ &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - e^{\beta t_i})^2\right) \\ \ell_x(\beta, \sigma^2) &= -\log L_x(\beta, \sigma^2) = \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum (x_i - e^{\beta t_i})^2 \end{aligned}$$

hvor  $(\beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ , og produktet samt summerne er fra  $i = 1$  til  $n$ .

Scorefunktionen fås ved differentiation mht.  $\beta$  hhv.  $\sigma^2$ :

$$S_x(\beta, \sigma^2) = \left( -\frac{1}{\sigma^2} \sum (x_i - e^{\beta t_i}) e^{\beta t_i} t_i , \quad \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} \sum (x_i - e^{\beta t_i})^2 \right)$$

Indsættelse af de foreslæde værdier giver følgende:

```
> n <- 6
> s2 <- 1.606481
> b <- 1.15625068
> sum((x-exp(b*t))*exp(b*t)*t) # Første koordinat i S(b,s2)
[1] 6.034892e-05
> n/2/s2 - sum((x-exp(b*t))^2) / 2 / s2^2 # Anden koordinat i S(b,s2)
[1] 3.051034e-07
```

Altså er  $S_x(1.15625068, 1.606481)$  lig nul pånær afrunding.

- Elementet på plads (1,2) i den stokastiske version af informationsfunktionen er

$$I_{X,12}(\beta, \sigma^2) = \frac{1}{\sigma^4} \sum (X_i - e^{\beta t_i}) e^{\beta t_i} t_i$$

der har middelværdi nul eftersom  $EX_i = e^{\beta t_i}$ . Således er element (1,2) i Fisherinformationsmatrixen lig nul.

Elementet på plads (1,1) i den stokastiske version af informationsfunktionen er

$$I_{X,11} = -\frac{1}{\sigma^2} \sum \left\{ -e^{2\beta t_i} t_i^2 + (X_i - e^{\beta t_i}) e^{\beta t_i} t_i^2 \right\}$$

Her har sidste led middelværdi nul, så vi får

$$i(\beta, \sigma^2)_{11} = EI_{X,11} = \frac{1}{\sigma^2} \sum t_i^2 e^{2\beta t_i}$$

3. Det sædvanlige asymptotiske resultat giver at  $\hat{\beta}$  er asymptotisk normalfordelt med middelværdi  $\beta$  og varians lig element  $(1, 1)$  i den inverse Fisherinformation.

Eftersom Fisherinformationen er en diagonalmatrix, er den asymptotiske varians altså

$$\left[ i(\beta, \sigma^2)^{-1} \right]_{11} = \frac{1}{i(\beta, \sigma^2)_{11}} = \frac{\sigma^2}{\sum t_i^2 e^{2\beta t_i}}$$

Ved at indsætte estimaterne, fås den estimerede spredning

$$SE(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum t_i^2 e^{2\hat{\beta} t_i}} = 0.0121,$$

og det falske 95% Wald konfidensinterval er så

$$\hat{\beta} \pm 1.96 \cdot SE(\hat{\beta}) = 1.156 \pm 0.024 = (1.132, 1.180).$$

4. Den alternative model er en lineær normal model for  $Z'$ erne. Designmatricen  $A$  er en  $1 \times n$  matrix med  $t_i$  på plads  $i$ . Vi får derfor

$$\hat{\gamma} = (A^T A)^{-1} A^T Z = \frac{\sum t_i z_i}{\sum t_i^2}, \quad \text{Var}(\hat{\gamma}) = \tau^2 (A^T A)^{-1} = \frac{\tau^2}{\sum t_i^2}, \quad SE(\hat{\gamma}) = \frac{\tilde{\tau}}{\sqrt{\sum t_i^2}}.$$

For det givne datasæt får vi

$$\hat{\gamma} = 1.1454, \quad \tilde{\tau}^2 = 0.2874, \quad SE(\hat{\gamma}) = 0.1205.$$

Bemerk evt. at den alternative model er en lineær regression med respons  $x$  og forklarende variabel  $t$ , men uden intercept. Modellen kan fittes med kommandoen `lm(z~t-1)`. Fra summary for denne model kan estimat og estimeret spredning aflæses direkte.

## Opgave 2

1. Vi har  $Y = CX$  hvor

$$C = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 2 & -3 \end{pmatrix}$$

Det følger at

$$Y \sim N(0, CVX C^T),$$

og når man regner på variansmatricen får man

$$VY = \begin{pmatrix} 10 & -24 \\ -24 & 63 \end{pmatrix}$$

Denne matrix har determinant 54 og er dermed invertibel, så fordelingen af  $Y$  er en regulær normalfordeling på  $\mathbb{R}^2$ .

2. Betragt så

$$\begin{pmatrix} Z \\ X_1 + X_3 \end{pmatrix} = \begin{pmatrix} c_1 & c_3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$$

der er normalfordelt med på  $\mathbb{R}^2$  med middelværdi 0 og variansmatrix

$$\Gamma = \begin{pmatrix} c_1 & c_3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 & 1 \\ c_3 & 1 \end{pmatrix} = \begin{pmatrix} c_1^2 + 4c_3^2 + 2c_1c_3 & 2c_1 + 5c_3 \\ 2c_1 + 5c_3 & 7 \end{pmatrix}$$

Kravene om at  $VZ = 21$  og at  $Z$  og  $X_1 + X_3$  er uafhængige er derfor opfyldt hvis og kun hvis

$$c_1^2 + 4c_3^2 + 2c_1c_3 = 21, \quad 2c_1 + 5c_3 = 0.$$

Fra den sidste betingelse får  $c_1 = -\frac{5}{2}c_3$ , der indsættes i første betingelse:

$$\frac{25}{4}c_3^2 + 4c_3^2 - 5c_3^2 = 21 \Leftrightarrow \frac{21}{4}c_3^2 = 21 \Leftrightarrow c_3^2 = 4 \Leftrightarrow c_3 = \pm 2$$

For  $c_3 = -2$  fås  $c_1 = 5$ , for  $c_3 = 2$  fås  $c_1 = -5$ , så betingelserne er opfyldt for  $(c_1, c_3) = (5, -2)$  og  $(c_1, c_3) = (-5, 2)$ .

## Opgave 3

1. Lad  $X_1, \dots, X_{50}$  være de stokastiske variable svarende til log-serinmængden. I modelA antages  $X_1, \dots, X_n$  at være uafhængige, og  $X_i \sim N(\xi_i, \sigma^2)$  hvor

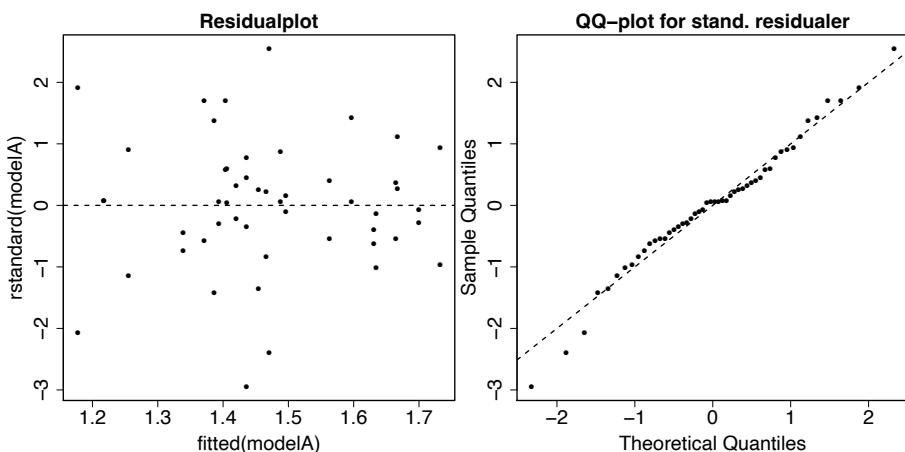
$$\xi_i = \alpha_{f(i)} + \beta_{f(i)} \cdot t_i$$

hvor  $t_i$  er hydrolysetiden,  $f(i)$  angiver fodertypen for observation  $i$ ,  $\alpha$ 'erne er ukendte interceptparametre og  $\beta$ 'erne er ukendte hældningsparametre. Der er fem  $\alpha$ 'er og fem  $\beta$ 'er svarende til at der både intercept og hælding er forskellige for de fem fodertyper.

Residualplot og normalfordelings-QQ-plot for de standardiserede residualer er vist nedenfor. Vi noterer følgende:

- Punkterne i residualplottet fordeler sig ligeligt om  $x$ -aksen og med cirka lige stor lodret variation i alle dele af plottet (fra venstre mod højre). Der er således ingen tegn på misspecifikation af middelværdi eller på variansinhomogenitet.
- Punkterne i QQ-plottet varierer omkring den rette linie ( $y = x$ ), så normalfordelingsantagelsen synes at være fornuftig. Der er heller ikke kritiske outliers.

Vi har desuden ikke grund til at tvivle på antagelsen om uafhængighed. Samlet ser ser modelA altså ud til at være en fornuftig model til data.



2. Hypotesen er

$$H : \beta_{\text{byg}} = \beta_{\text{fiskemel}} = \beta_{\text{majss}} = \beta_{\text{kb.mel}} = \beta_{\text{soja}}$$

Den statistiske model under hypotesen er præcis `modelB`.

Hypotesen testes derfor ved at sammenligne modellerne `modelA` og `modelB`. Testet udføres som et  $F$ -test.  $F$ -teststørrelsen er 2.07 der skal vurderes i  $F$ -fordelingen med (4, 40) frihedsgrader. Dette giver  $p$ -værdien 0.10, så vi kan ikke forkaste hypotesen.

Data tyder således ikke på at hældningerne er forskellige, og det er rimeligt at fortsætte analysen vha. `modelB` hvor hældningerne er ens.

3. Vi får følgende estimerater og konfidensintervaller:

- Forskellen mellem forventet log-serinmængde for byg og majs, dvs.  $\delta_{bm} = \alpha_{majs} - \alpha_{byg}$ . Estimat og konfidensinterval kommer direkte fra parametreriseringen i `modelB`:

$$\hat{\delta}_{bm} = 0.158, \quad 95\% \text{ KI } (0.149, 0.167)$$

- Forskellen mellem forventet log-serinmængde for majs og soja, dvs.  $\delta_{ms} = \alpha_{soja} - \alpha_{majs}$ . Estimat og konfidensinterval kan fx fås ved at ændre referencegruppe. Vi får

$$\hat{\delta}_{ms} = 0.071, \quad 95\% \text{ KI } (0.062, 0.080)$$

4. Forskellen mellem forventet log-serinmængde ved 20 og 50 timers hydrolysetid (for fastholdt fodertype), dvs.  $30\beta$  hvor  $\beta$  er den fælles hældning. Estimat og KI fås ved at gange estimat og KI for  $\beta$  med 30:

$$30\hat{\beta} = -0.122, \quad 95\% \text{ KI } (-0.126, -0.118)$$

Den prædikterede værdi for log-serin efter 40 timer for soja er

$$1.535 + 0.229 - 0.00407 \cdot 40 = 1.602$$

hvor estimererne er taget fra `modelB`. 95% Prædiktionsintervallet fås nemmest vha. `predict`-funktionen og viser sig at være (1.581, 1.623).

5. For fodertype  $j$  er  $\gamma_j = \alpha_j + 50\beta$ . Således er

$$\bar{\gamma} = \frac{1}{5} \sum_{j=1}^5 \gamma_j = \frac{1}{5} \sum_{j=1}^5 \alpha_j + 50 \cdot \beta.$$

Hvis vi sætter  $\theta = (\alpha_1, \dots, \alpha_5, \beta)$  og  $\psi^T = (1/5, 1/5, 1/5, 1/5, 1/5, 50)$ , har vi altså

$$\bar{\gamma} = \psi^T \theta.$$

Vi benytter eksempel 10.30, og får

$$\hat{\gamma} = \psi^T \hat{\theta}, \quad \text{Var}(\hat{\gamma}) = \psi^T \text{Var}(\hat{\theta}) \psi$$

Den ønskede parametrerisering opnås med kommandoen

```
lm(log(serin) ~ foder+tid-1, data=hydrolyse)
```

og estimat  $\hat{\theta}$  og variansmatrix  $\text{Var}(\hat{\theta})$  fås fx. med `coef` og `vcov`. Vi får  $\hat{\gamma} = 1.39236$  med estimeret spredning  $\text{SE}(\hat{\gamma}) = 0.00185$ .

Det tilhørende 95% konfidensinterval er

$$1.39236 \pm 2.015 \cdot 0.00185 = 1.39236 \pm 0.00373 = (1.38863, 1.39609)$$

hvor vi har benyttet at 97.5% fraktilen i  $t_{44}$  fordelingen er 2.015.

## Opgave 4

1. Scorefunktionen og informationsfunktionen er

$$S_x(\theta) = -\frac{n}{\theta} + \frac{\sum x_i}{1-\theta}, \quad I_x(\theta) = \frac{n}{\theta^2} + \frac{\sum x_i}{(1-\theta)^2}.$$

Vi løser scoreligningen

$$S_x(\theta) = 0 \Leftrightarrow \theta \sum x_i = n - n\theta \Leftrightarrow \theta(n + \sum x_i) = n \Leftrightarrow \theta = \frac{n}{n + \sum x_i},$$

og bemærker desuden at  $I_x(\theta) > 0$  for alle  $\theta \in (0, 1)$ . Løsningen ligger i  $(0, 1]$  og er kun 1 hvis  $\sum x_i = 0$  (hvilket sker med  $\theta^n$ ). Hvis vi tillader en estimator på randen, er løsningen til scoreligningen således en entydigt bestemt MLE, dvs.

$$\hat{\theta} = \frac{n}{n + \sum x_i}.$$

2. For den givne observation er  $\sum x_i = 33$ , så ML estimatet er

$$\hat{\theta} = \frac{12}{12 + 33} = 0.267.$$

Likelihood ratio teststørrelsen er

$$LR(x) = 2\ell_x(0.4) - 2\ell_x(\hat{\theta}) = 3.51$$

der skal vurderes i  $\chi^2$  fordelingen med en frihedsgrad. Dette giver  $p$ -værdien 0.06. Med et signifikansniveau på 5% kan vi således ikke afvise hypotesen.

3. Det udfyldte skema ser således ud (varierer naturligvis lidt fra simulation til simulation):

$n$	$\theta$	Relativ hyppighed
20	0.4	0.052
20	0.5	0.255
20	0.6	0.713
40	0.4	0.049
40	0.5	0.458
40	0.6	0.953

Vi ser at

- Det faktiske niveau for testet er tæt på de ønskede 5%, både for  $n = 20$  og  $n = 40$
- Styrken vokser når  $n$  vokser (som forventet)
- Styrken vokser når afstanden mellem det sande  $\theta$  og 0.4 vokser (som forventet)

## Eksamens i Statistik 1, vejledende besvarelse

### 30. juni 2016

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `juni16.R`.

### Opgave 1

- På grund af uafhængighed er likelihoodfunktionen givet ved

$$L_x(\beta) = \prod \frac{(\beta t_i)^{x_i}}{x_i!} e^{-\beta t_i} \propto \beta^{S_x} e^{-\beta S_t}, \quad \beta > 0$$

hvor  $S_x$  og  $S_t$  er summerne fra opgaveteksten. Minus-log-likelihoodfunktionen, scorefunktion og informationsfunktion bliver derfor

$$\begin{aligned}\ell_x(\beta) &= -\log L_x(\beta) = -S_x \log \beta + \beta S_t, \quad \beta > 0 \\ S_x(\beta) &= \ell'_x(\beta) = -\frac{S_x}{\beta} + S_t, \quad \beta > 0 \\ I_x(\beta) &= \ell''_x(\beta) = \frac{S_x}{\beta^2}, \quad \beta > 0\end{aligned}$$

Notationen er faktisk temmelig uheldig her, eftersom  $S_x$  både dækker over scorefunktionen og summern af  $x$ 'erne.

Endelig fås Fisherinformationen

$$I(\beta) = E_\beta I_X(\beta) = \frac{E_\beta S_x}{\beta^2} = \frac{E_\beta \sum X_i}{\beta^2} = \frac{\sum \beta t_i}{\beta^2} = \frac{\sum t_i}{\beta} = \frac{S_t}{\beta}.$$

- Vi betragter først estimatoren

$$\tilde{\beta} = \frac{1}{n} \sum_{i=1}^n \frac{X_i}{t_i} \quad \text{og}$$

Vi bestemmer middelværdi og varians:

$$\begin{aligned}E_\beta \tilde{\beta} &= \frac{1}{n} \sum \frac{\beta t_i}{t_i} = \beta \\ V_\beta \tilde{\beta} &= \frac{1}{n^2} \sum V_\beta \left( \frac{X_i}{t_i} \right) = \frac{1}{n^2} \sum \frac{\beta t_i}{t_i^2} = \frac{\beta}{n^2} \sum \frac{1}{t_i}\end{aligned}$$

Da  $E_\beta \tilde{\beta} = \beta$  for alle  $\beta$ , er  $\tilde{\beta}$  central en central estimator for  $\beta$ .

Men  $V_\beta \tilde{\beta} \neq I^{-1}(\beta)$ , så den nedre grænse for variansen for en central estimator fra Cramer-Rao opnås ikke. Vi kan strengt taget (endnu) ikke udelukke at estimatoren  $\tilde{\beta}$  har den mindste varians blandt centrale estimatorer, men Cramer-Rao fortæller os det ikke.

Vi betragter så estimatoren

$$\check{\beta} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n t_i}.$$

og bestemmer middelværdi og varians:

$$\begin{aligned} E_{\beta} \check{\beta} &= \frac{\sum \beta t_i}{\sum t_i} = \beta \\ V_{\beta} \check{\beta} &= \frac{\sum \beta t_i}{(\sum t_i)^2} = \frac{\beta}{\sum t_i} = \frac{\beta}{S_t} = i^{-1}(\beta) \end{aligned}$$

Vi ser at  $\check{\beta}$  er central og variansminimal (jf. Cramer-Rao). Vi kan således nu også slutte at  $\check{\beta}$  ikke er variansminimal.

3. Betragt  $(x_1, \dots, x_n)$  med  $S_x \neq 0$ . Vi løser først scoreligningen:

$$S_x(\beta) = 0 \Leftrightarrow \frac{S_x}{\beta} = S_t \Leftrightarrow \beta = \frac{S_x}{S_t}$$

Da  $I_x(\beta) > 0$  for alle  $\beta$  er løsningen til scoreligningen faktisk et globalt minimum for  $\ell_x$  hvis  $S_x \neq 0$ . Desuden er løsningen positiv når  $S_x > 0$ .

Hvis summen  $S_x = 0$ , er alle  $x_i = 0$ . Så er

$$L_x(\beta) = \prod e^{-\beta t_i} = e^{-\beta S_t}$$

der ikke har maksimum på  $(0, \infty)$ . Men  $L_x$  er aftagende så det er naturligt at udvide parametermængden og sætte  $\hat{\beta} = 0$ .

Konklusion: ML estimatoren defineres naturligt som

$$\hat{\beta} = \frac{S_x}{S_t} = \check{\beta}.$$

4. For de givne data er MLE  $\hat{\beta} = \check{\beta} = 3.818$  med standard error

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\beta}}{\sum t_i}} = 0.833,$$

så Wald 95% konfidensintervallet er

$$\hat{\beta} \pm 1.96 \cdot SD(\hat{\beta}) = 3.818 \pm 1.96 \cdot 0.833 = (2.185, 5.451)$$

Betrægt så hypotesen  $H : \beta = 2$ . Likelihood ratio teststørrelsen er

$$LR(x) = 2(\ell_x(2) - \ell_x(\hat{\beta})) = 2(-S_x \log 2 + 2S_t + S_x \log \hat{\beta} - \hat{\beta} S_t) = 7.16$$

Denne skal vurderes i  $\chi^2$  fordelingen med 1 frihedsgrad:

$$p = P(LR(X) \geq 7.16) = 0.007$$

så hypotesen forkastes. Der er altså evidens i data på at  $\beta$  er større end 2.

5. Jeg fik følgende skema:

Estimator	$n$	$\beta$	Teoretiske værdier		Simulation	
			middelværdi	spredning	gennemsnit	spredning
$\check{\beta}$	10	3	3	0.937	2.969	0.924
	20	3	3	0.735	2.981	0.730
$\hat{\beta}$	10	3	3	0.739	2.965	0.738
	20	3	3	0.534	2.991	0.527

Vi ser at de teoretiske og simulerede værdier stemmer pænt overens. Specielt er begge estimatorer centrale og  $\hat{\beta} = \check{\beta}$  har mindre spredning end  $\check{\beta}$ . Desuden bliver spredningerne mindre når sample size  $n$  vokser. Med andre ord: Alt er som forventet.

## Opgave 2

1.  $X$  er normalfordelt på  $\mathbb{R}^2$  med middelværdi 0 og varians

$$\text{V}X = \begin{pmatrix} 1 & 3 & 0 & 0 \\ 3 & 9 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 3 & 9 \end{pmatrix}.$$

Variansmatricen har determinant nul og er derfor singulær. Fordelingen af  $X$  er således en singulær normalfordeling på  $\mathbb{R}^4$ .

2.  $X_2$  og  $X_4$  er begge  $N(0, 9)$ -fordelte, så de har samme fordeling.

Alligevel har  $X_1 + X_2$  og  $X_1 + X_4$  forskellige fordelinger. Dette skyldes at  $X_1$  og  $X_2$  er afhængige — faktisk er  $X_2 = 3X_1$  — mens  $X_1$  og  $X_4$  er uafhængige.

Mere specifikt kan vi finde den simultane fordeling af  $X_1 + X_2$  og  $X_1 + X_4$ :

$$\begin{pmatrix} X_1 + X_2 \\ X_1 + X_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}$$

der er normalfordelt med middelværdi 0 og varians

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 & 0 & 0 \\ 3 & 9 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 3 & 9 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 16 & 4 \\ 4 & 10 \end{pmatrix}.$$

Altså er  $X_1 + X_2 \sim N(0, 16)$ , mens  $X_1 + X_4 \sim N(0, 10)$ .

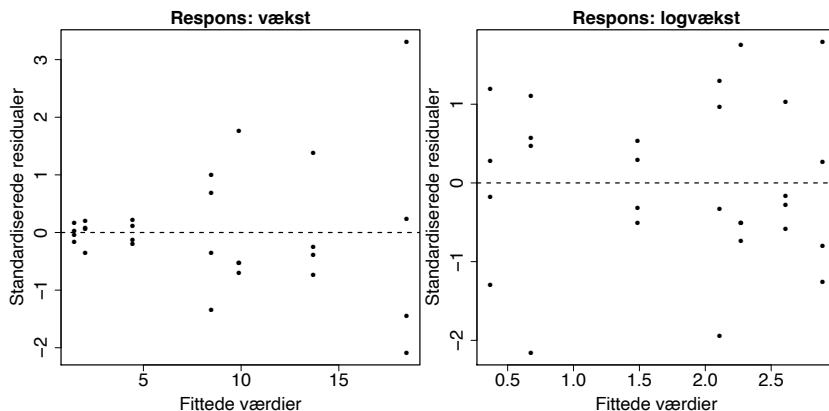
Eftersom denne variansmatrix er regulær (determinanten er 144), er den simultane fordeling af  $X_1 + X_2$  og  $X_1 + X_4$  en regulær normalfordeling og „fylder derfor hele  $\mathbb{R}^2$ “. Der findes således ikke et underrum  $U$  som beskrevet i spørgsmålet.

## Opgave 3

1. Modellerne fitteres med kommandoerne

```
m0 <- lm(vaekst ~ dosisGrp, data=rotteData)
m1 <- lm(logvaekst ~ dosisGrp, data=rotteData)
```

Residualplottene ses nedenfor. Der er oplagt problemer med variansomogenitet når vækst bruges som respons: variansen vokser når de fittede værdier vokser. Når logvækst benyttes som respons, ser variansen derimod ud til at være konstant.



2. Forskellen i EX mellem dosis 3 og dosis 0 aflæses direkte fra summary af model m1:

$$\text{Estimat: } 2.522, \quad 95\% \text{ KI: } (2.195, 2.849)$$

Forskellen i EX mellem dosis 3 og dosis 2 kan fx fås ved at genfitte modellen med dosisgruppe d2 som reference. Man får:

$$\text{Estimat: } 0.6196, \quad 95\% \text{ KI: } (0.2923, 0.9468)$$

3. Den lineære regressionsmodel fitteres med kommandoen

```
m2 <- lm(logvaekst ~ dosis, data=rotteData)
```

Forskellen i forventet log-vækst ved dosis 3 og dosis 0 er  $3\beta$  hvor  $\beta$  er hældningen i modellen. Vi får:

$$\text{Estimat: } 2.617, \quad 95\% \text{ KI: } (2.295, 2.939)$$

Forventet log-vækst ved dosis 0 er netop interceptparametren i modellen. Aflæsning giver

$$\text{Estimat: } 0.4642, \quad 95\% \text{ KI: } (0.2708, 0.6575)$$

4. Den kvadratiske regressionsmodel fitteres med kommandoerne

```
rotteData$dosisKvd <- rotteData$dosis^2
m3 <- lm(logvaekst ~ dosis + dosisKvd, data=rotteData)
```

Modellen siger at

$$EX = \beta_0 + \beta_1 d + \beta_2 d^2$$

så den ønskede forskel er

$$\delta = \beta_0 + 3\beta_1 + 9\beta_2 - \beta_0 = 3\beta_1 + 9\beta_2$$

der kan skrives som  $C\beta$  hvor  $C = (0 \ 3 \ 9)$ .

Estimatet for  $\delta$  er  $\hat{\delta} = C\hat{\beta} = 2.617$  og den tilhørende standard error er

$$\text{SE}(\hat{\delta}) = \sqrt{C \text{Var}(\hat{\beta}) C^T} = 0.135$$

Konfidensintervallet bliver således

$$\hat{\delta} \pm t_{0.975, 26} \cdot \text{SE}(\hat{\delta}) = (2.339, 2.896)$$

Bemærk at estimerne i spørgsmål 3 og 4 er ens, men det er „tilfældigt“ (specielt for disse data, ikke noget strukturelt ved modellerne).

5. Der gælder

$$L_{\text{linreg}} \subset L_{\text{kvadreg}} \subset L_{\text{anova}}$$

og begge mængdeinklusioner betyder „ægte underrum“.

Modelreduktionen starter dermed i den ensidede variansanalyse,  $\xi \in L_{\text{anova}}$  hvor  $\xi$  betegner middelværdivektoren.

Hypotesen  $H : \xi \in L_{\text{kvadreg}}$  testes med et  $F$ -test. Vi får

$$f = 1.95, \quad p = 0.14 \quad (\text{beregnet i } F(4, 21))$$

så hypotesen kan ikke afvises.

Vi bruger derfor den kvadratiske regressionsmodel. Hypotesen  $H : \xi \in L_{\text{kvadreg}}$  eller  $H : \beta_2 = 0$  kan testes med et  $t$ -test (eller et  $F$ -test). Vi får

$$t = -3.12, \quad p = 0.004 \quad (\text{beregnet i } t(25))$$

så hypotesen afvises.

Slutmodellen er således den kvadratiske regressionsmodel, hvor middelværdien beskrives som et andengradspolynomium i dosis.

6. Vi bruger den kvadratiske regressionsmodel. Prædiktionsintervallet for log-vækst kan beregnes vha. kommandoerne

```
> newData <- data.frame(dosis=2.25, dosisKvd=2.25^2)
> predict(m3, newData, interval="p")
```

Man får prædiktionen 2.498 og 95% prædiktionsintervallet på (1.99, 3.01) for log-vækst og dermed et 95% prædiktionsinterval for vækst på (7.3, 20.2). En vægtøgning på 18 gram er således ikke usædvanligt.

# Reeksamen i Statistik 2

25. august 2016

Eksamensvarer 4 timer. Alle hjælpemidler er tilladt under eksamen, også computer, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant.

Eksamenssættet består af tre opgaver med i alt 22 delspørgsmål. De tre opgaver vægtes ens. Data til opgave 3 ligger i filen `benzin.txt` på en USB-stick. Sticken skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den kan altså ikke indgå som en del af besvarelsen.

## Opgave 1

- Betræt fordelingen med tæthed

$$f_p(x) = (x+1)p^x(1-p)^2 \quad \text{for } x \in \mathbb{N}_0$$

mht tællemålet på  $\mathbb{N}_0$ . Fordelingen afhænger af parameteren  $p \in (0, 1)$ .

Du kan uden bevis benytte at  $f_p$  er en tæthed.

Lad  $X_1, \dots, X_n$  være uafhængige og identisk fordelte stokastiske variable med tæthed  $f_p$ , med ukendt  $p \in (0, 1)$ .

- (a) Reparametriser  $f_p$ -fordelingen ved  $\theta = \log p$  så det fremgår at det er en eksponentiel familie med kanonisk stikprøvefunktion  $t(x) = x$ .

**Solution:** Da  $p = e^\theta$  fås

$$\begin{aligned} f_p(x) &= (x+1)p^x(1-p)^2 \\ &= (x+1)e^{\theta x}(1-e^\theta)^2 \end{aligned}$$

Vi får således (se def. 2.13 i lærebogen)

$$\begin{aligned} f_\theta(x) &= (x+1)(1-e^\theta)^2e^{\theta x} \\ &= \underbrace{(x+1)}_{\text{funktion af } x} \underbrace{(1-e^\theta)^2}_{\text{funktion af } \theta} \underbrace{e^{\theta x}}_{t(x)=x} \end{aligned}$$

- (b) Identifier grundmålet. Identifier normeringskonstanten  $c(\theta)$ .

**Solution:** Lad  $\mu$  have tæthed

$$(x+1) \quad \text{for } x \in \mathbb{N}_0$$

med hensyn til tællemålet. Udtrykt ved  $\theta = \log p$  har  $X$  tæthed

$$(1 - e^\theta)^2 e^{\theta x} \quad \text{for } x \in \mathbb{N}_0$$

med hensyn til  $\mu$ . Der er altså tale om en en-dimensonal eksponentiel familie på  $\mathbb{N}_0$  med kanonisk stikprøvefunktion  $t(x) = x$ , grundmål  $\mu$  og normeringskonstant

$$c(\theta) = (1 - e^\theta)^{-2}.$$

- (c) Argumenter for at  $X_1$  har momenter af enhver orden. Find middelværdi og varians af  $X_1$ .

**Solution:** Ifølge Lemma 2.19 har  $t(X_1) = X_1$  momenter af enhver orden. Bemærk at Lemma 2.19 udtaler sig om momenter af  $t(X)$ , ikke om momenter af  $X$ . Det er derfor vigtigt at den kanoniske stikprøvefunktion er identitetsfunktionen.

Ifølge (2.15) (eller formlerne lige over, eller Lemma 2.20) er

$$E(t(X_1)) = E(X_1) = \frac{d}{d\theta} \log c(\theta) = \frac{2e^\theta}{1 - e^\theta}$$

og

$$\text{Var}(t(X_1)) = \text{Var}(X_1) = \frac{d^2}{d\theta^2} \log c(\theta) = \frac{2e^\theta}{(1 - e^\theta)^2}$$

Udtrykt ved  $p = e^\theta$  fås (ikke nødvendigt at angive i besvarelsen)

$$E(X_1) = \frac{2p}{1 - p}$$

og

$$\text{Var}(X_1) = \frac{2p}{(1 - p)^2}$$

- (d) Opskriv likelihoodligningen for  $\theta$ . Find maksimaliseringsestimatoren for  $\theta$ . Find maksimaliseringsestimatoren for  $p$ .

**Solution:** Likelihoodligningen for  $\theta$  er givet i (5.6), det følger derfor fra resultaterne i (c) at likelihoodligningen er

$$\frac{1}{n} \sum_{i=1}^n X_i = \frac{2e^\theta}{1 - e^\theta}$$

Løses denne fås maksimaliseringestimatoren for  $\theta$ :

$$\hat{\theta} = \log \left( \frac{\frac{1}{n} \sum_{i=1}^n X_i}{2 + \frac{1}{n} \sum_{i=1}^n X_i} \right)$$

Maksimaliseringestimatoren for  $p$ :

$$\hat{p} = e^{\hat{\theta}} = \frac{\frac{1}{n} \sum_{i=1}^n X_i}{2 + \frac{1}{n} \sum_{i=1}^n X_i}$$

- (e) Er maksimaliseringestimatoren for  $\theta$  veldefineret med sandsynlighed 1? Er maksimaliseringestimatoren for  $\theta$  asymptotisk veldefineret? Begrund dine svar.

**Solution:** Nej, maksimaliseringestimatoren for  $\theta$  er ikke veldefineret med sandsynlighed 1. Der er positiv sandsynlighed for at  $X_i = 0$  for  $i = 1, \dots, n$ . I det tilfælde er estimatoren ikke veldefineret, da man ikke kan tage logaritmen til 0.

Ja, maksimaliseringestimatoren for  $\theta$  er asymptotisk veldefineret. Sandsynligheden for at  $X_i = 0$  for  $i = 1, \dots, n$  går mod 0 når  $n \rightarrow \infty$ , og derfor vil argumentet til logaritmen være et tal mellem 0 og 1 (da  $X_i \geq 0$ ). Da  $p \in (0, 1)$  er  $\theta < 0$ , og vi ser at  $\hat{\theta} < 0$  med sandsynlighed gående mod 1, og maksimaliseringestimatoren er således asymptotisk veldefineret.

- (f) Gør rede for at  $\hat{\theta}$  er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling, parametrizeret ved  $\theta$ .

**Solution:** Da  $\text{Var}(X_1) > 0$  benytter vi direkte nederste formel på side 187, og får at

$$\hat{\theta} \stackrel{as}{\sim} N \left( \theta, \frac{1}{n} \frac{(1 - e^\theta)^2}{2e^\theta} \right)$$

Man kan også gå en lille omvej, og bruge at ifølge CLT, Sætning 5.11, er

$$\frac{1}{n} \sum_{i=1}^n X_i \stackrel{as}{\sim} N \left( \frac{2e^\theta}{1 - e^\theta}, \frac{1}{n} \frac{2e^\theta}{(1 - e^\theta)^2} \right)$$

Vi har at

$$\hat{\theta} = f \left( \frac{1}{n} \sum_{i=1}^n X_i \right) \quad \text{hvor} \quad f(x) = \log \left( \frac{x}{2+x} \right).$$

For  $x > 0$  er  $f$  differentiabel, herunder specielt for  $x = E(X_1)$ . Vi har  $f'(x) = \frac{2}{x(2+x)}$  og  $f'(E(X_1)) = \frac{(1-e^\theta)^2}{2e^\theta}$ . Vi benytter deltametoden, og får

$$\begin{aligned} \hat{\theta} = f \left( \frac{1}{n} \sum_{i=1}^n X_i \right) &\stackrel{as}{\sim} N \left( f \left( \frac{2e^\theta}{1-e^\theta} \right), \frac{1}{n} \left( \frac{(1-e^\theta)^2}{2e^\theta} \right)^2 \frac{2e^\theta}{(1-e^\theta)^2} \right) \\ &= N \left( \theta, \frac{1}{n} \frac{(1-e^\theta)^2}{2e^\theta} \right) \end{aligned}$$

som før.

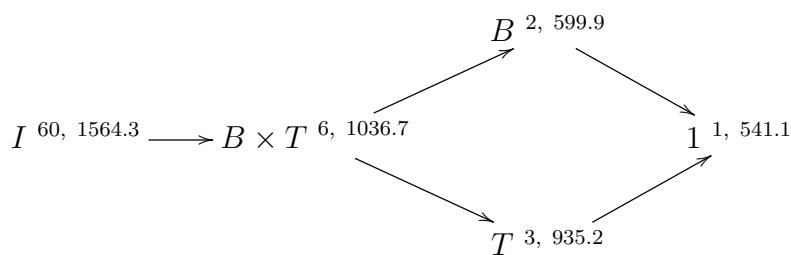
- (g) Gør rede for at  $\hat{p}$  er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling, parametreret ved  $p$ .

**Solution:**  $\hat{p} = g(\hat{\theta})$  hvor  $g(x) = e^x$  er målelig og differentiabel,  $g'(x) = e^x$ . Deltametoden giver

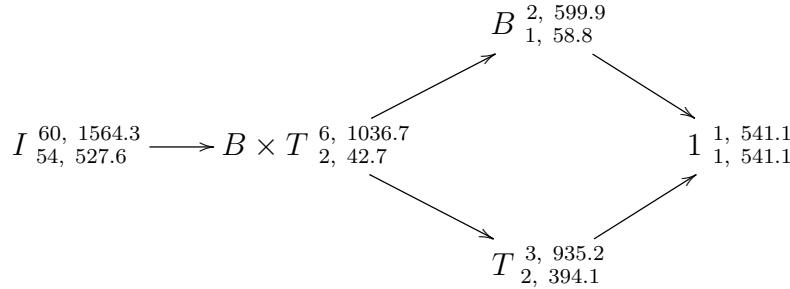
$$\hat{p} = g(\hat{\theta}) \stackrel{as}{\sim} N \left( e^\theta, \frac{1}{n} e^{2\theta} \frac{(1-e^\theta)^2}{2e^\theta} \right) = N \left( p, \frac{1}{n} \frac{p(1-p)^2}{2} \right)$$

## Opgave 2

2. Betragt de surjektive faktorer  $B$  og  $T$ , der antages at være usammenlignelige og geometrisk ortogonale, og deres tilhørende underrum  $L_B$  og  $L_T$ . Betragt det annoterede faktorstrukturdiagram:



- (a) Udfyld resten af det annoterede faktorstrukturdiagram i den ortogonale dekomposition.

**Solution:**

- (b) Test den additive hypotese op mod vekselvirkningsmodellen ved brug af det annoterede faktorstrukturdiagram.

**Solution:** Test af  $B + T$  mod  $B \times T$ :

$$\begin{aligned} F &= \frac{(\|P_{B+T}X\|^2 - \|P_{B+T}X\|^2) / (\dim \mathcal{L}_{B+T} - \dim \mathcal{L}_{B+T})}{(\|X\|^2 - \|P_{B+T}X\|^2) / (|I| - \dim \mathcal{L}_{B+T})} \\ &= \frac{\|Q_{B+T}X\|^2 / \dim(Q_{B+T})}{\|Q_I\|^2 / \dim(Q_I)} = \frac{42.7/2}{527.6/54} = 2.18 \end{aligned}$$

der skal evalueres i en F-fordeling med 2 og 54 frihedsgrader.  $p$ -værdien bliver 0.123, og vi accepterer den additive hypotese.

- (c) Test om der er en effekt af faktor B mod den additive hypotese ved brug af det annoterede faktorstrukturdiagram.

**Solution:** For at teste om der er effekt af faktor B testes  $T$  mod  $B + T$ :

$$\begin{aligned} F &= \frac{(\|P_{B+T}X\|^2 - \|P_TX\|^2) / (\dim \mathcal{L}_{B+T} - \dim \mathcal{L}_T)}{(\|X\|^2 - \|P_{B+T}X\|^2) / (|I| - \dim \mathcal{L}_{B+T})} \\ &= \frac{\|Q_B X\|^2 / \dim(Q_B)}{(\|Q_I\|^2 + \|Q_{B+T}\|^2) / \dim(Q_I + Q_{B+T})} = \frac{58.8/1}{(527.6 + 42.7)/(54 + 2)} = 5.78 \end{aligned}$$

der skal evalueres i en F-fordeling med 1 og 56 frihedsgrader.  $p$ -værdien er 0.0196, og vi afviser at B ingen effekt har.

- (d) Test om der er en effekt af faktor T mod den additive hypotese ved brug af det annoterede faktorstrukturdiagram.

**Solution:** Test af  $B$  mod  $B + T$ :

$$\begin{aligned} F &= \frac{(\|P_{B+T}X\|^2 - \|P_B X\|^2) / (\dim \mathcal{L}_{B+T} - \dim \mathcal{L}_B)}{(\|X\|^2 - \|P_{B+T}X\|^2) / (|I| - \dim \mathcal{L}_{B+T})} \\ &= \frac{\|Q_T X\|^2 / \dim(Q_T)}{(\|Q_I\|^2 + \|Q_{B \times T}\|^2) / \dim(Q_I + Q_{B \times T})} = \frac{394.1/2}{(527.6 + 42.7)/(54 + 2)} = 19.35 \end{aligned}$$

der skal evalueres i en F-fordeling med 2 og 56 frihedsgrader.  $p$ -værdien er  $< 0.0001$ , og vi afviser at  $T$  ingen effekt har.

(e) Er faktoren  $B \times T$  surjektiv?

**Solution:** Ja. Det kan ses ud fra det annoterede faktorstrukturdiagram, hvor dimensionen af  $B \times T$  ses at være  $6 = 2 \times 3 =$  produktet af dimensionerne af enkeltfaktorerne.

(f) Hvor mange observationer er der i datasættet?

**Solution:** Der er 60 observationer. Det kan ses ud fra det annoterede faktorstrukturdiagram, hvor dimensionen af  $I$  er 60.

(g) Hvor mange labels er der for faktoren B?

**Solution:** Der er 2 labels for faktoren B. Det kan ses ud fra det annoterede faktorstrukturdiagram, hvor dimensionen af  $B$  er 2.

(h) Er  $L_B$  og  $L_T$  ægte ortogonale?

**Solution:** Nej. Det kan blandt andet ses udfra at summen af deres dimensioner ( $=5$ ) ikke er lig dimensionen af sumunderrummet ( $=4$ ). Det kan også ses udfra at  $L_B \cap L_T = L_1 \neq \{0\}$ .

### Opgave 3

3. Ved en undersøgelse af virkningen af forskellige dæktyper på benzinförbruget af offentlige busser blev følgende forsøg gennemført: 2 busser,  $A$  og  $B$ , gennemkørte hver 30 gange samme rundstrækning på ca. 10 km. I hver kørsel brugte de en af tre forskellige dæktyper,  $K$ ,  $L$  eller  $M$ , således at hver kombination af bus og dæktype blev testet 10 gange, og benzinförbruget i milliliter blev målt. Der var 10 chauffører til at køre de ialt 60 ture.

Data er tilgængelige i filen `benzin.txt` og består af variablene `bus`, `dæk`, `cha` og `benzin`, hvor den sidste angiver benzinförbruget.

Delopgaverne (b), (e), (f) og (g) skal løses i R, og det er nok at angive værdier fundet i output fra analyserne.

- (a) Opstil en varianskomponentmodel med en fast effekt af `bus` og en fast effekt af `dæk`, således at de indgår additivt, og en tilfældig effekt af `cha`. Alle de forklarende variable skal indgå som faktorer.

**Solution:** Vi skriver  $B$ ,  $D$  og  $C$  for faktorerne `bus`, `dæk` og `cha`. Statistisk model for data:

Lad  $I = \{1, \dots, 60\}$  være indeksmængden for observationerne.  $X$  er regulært normalfordelt på  $\mathbb{R}^I$  med middelværdivektor  $\xi = (\xi_i)_{i \in I} \in L_{B+D}$  og kovariansmatrix  $\sigma^2 \Sigma = \sigma^2(I + \lambda BB^T)$ , hvor  $B$  er effektmatricen hørende til effektparret  $(C, 1)$ . Middelværdiunderrummet har dimension 4. Kovariansmatricen er givet ved

$$\sigma^2 \Sigma = \sigma^2 \begin{pmatrix} \Sigma_1 & \cdots & 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_1 & 0 & \cdots & \Sigma_2 \\ \Sigma_2 & \cdots & 0 & \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_2 & 0 & \cdots & \Sigma_1 \end{pmatrix}$$

hvor

$$\Sigma_1 = \begin{pmatrix} 1 + \lambda & \lambda & \lambda \\ \lambda & 1 + \lambda & \lambda \\ \lambda & \lambda & 1 + \lambda \end{pmatrix}; \quad \Sigma_2 = \begin{pmatrix} \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda \end{pmatrix}$$

Alternativt:

$$X = A\beta + Z + \varepsilon$$

hvor  $A$  er designmatrix for middelværdiunderrummet  $L_{B+D}$ ,  $\beta \in \mathbb{R}^4$  er middelværdiparametervektoren,  $Z = BY \sim \mathcal{N}(0, \nu^2 BB^T)$  er den tilfældige effekt, hvor  $B$  er effektmatricen hørende til effektparret  $(C, 1)$ . Derudover er  $Y \sim \mathcal{N}(0, \nu^2 I_{10})$  og  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{60})$ . Bemærk at  $\nu^2 = \lambda\sigma^2$ .

- (b) Estimer parametrene i modellen, både ved ML-princippet og ved REML-princippet.

**Solution:** Her benyttes R til at estimere parametrene  $(\beta, \nu^2, \sigma^2)$ .

Estimater med ML-princippet:

R-kode (behøves ikke at vedlægges besvarelseren):

```
require(lme4)
data <- read.table("benzin.txt", head=TRUE)
# Med intercept:
m1ml <- lmer(benzin ~ bus+daek+(1|cha), data=data, REML=FALSE)
summary(m1ml)
# Uden intercept:
m1ml <- lmer(benzin ~ bus+daek-1+(1|cha), data=data, REML=FALSE)
summary(m1ml)
```

Resultat:

$$\begin{aligned}\hat{\beta} &= (2772.6, 2.7, 330.1, -191.3) \quad (\text{parametriseret med intercept}) \\ \hat{\beta} &= (2772.6, 2775.2, 330.1, -191.3) \quad (\text{parametriseret uden intercept}) \\ \hat{\nu}^2 &= 88934; \quad \hat{\nu} = 298.2 \\ \hat{\sigma}^2 &= 10510; \quad \hat{\sigma} = 102.5\end{aligned}$$

Den studerende behøver kun at angive en af parametriseringerne, det gælder også i de følgende besvarelser.

Estimater med REML-princippet:

```
# Med intercept:
m1reml <- lmer(benzin ~ bus+daek+(1|cha), data=data)
summary(m1reml)
# Uden intercept:
m1reml <- lmer(benzin ~ bus+daek-1+(1|cha), data=data)
summary(m1reml)
```

$$\begin{aligned}\hat{\beta} &= (2772.6, 2.7, 330.1, -191.3) \quad (\text{parametriseret med intercept}) \\ \hat{\beta} &= (2772.6, 2775.2, 330.1, -191.3) \quad (\text{parametriseret uden intercept}) \\ \hat{\nu}^2 &= 98898; \quad \hat{\nu} = 314.5 \\ \hat{\sigma}^2 &= 11181; \quad \hat{\sigma} = 105.7\end{aligned}$$

- (c) Diskuter forskelle/ligheder i resultaterne mellem de to metoder.

**Solution:** Vi ser at middelværdiestimaterne er det samme for de to metoder (de er ikke altid nøjagtig det samme, men tæt på hinanden).

Vi ser at begge variansestimater er højere for REML end for ML. REML er netop en metode til at korrigere for at variansen underestimeres med MLE, især de tilfældige effekter underestimeres.

- (d) Når der testes hypoteser vedrørende faste effekter med et kvotienttest, skal man så bruge ML-princippet eller REML-princippet?

**Solution:** Man skal altid benytte ML-princippet, da resultatet ellers afhænger af den valgte parametrisering af middelværdiunderrummet.

- (e) Test om dæktypen påvirker benzinförbruget.

**Solution:** Der testes ved at fitte modellen kun med `bus` i middelværdien overfor den additive model og benytte anova i R.

R-kode:

```
m2ml <- lmer(benzin ~ bus+(1|cha), data=data, REML=FALSE)
anova(m1ml,m2ml)
```

Bemærk ovenfor at man ikke behøver at fitte med `REML=FALSE` da R automatisk refitter når man kører `anova` på modellerne. Her fås at  $-2 \log Q = 91.99$  med 2 frihedsgrader, hvilket giver en  $p$ -værdi på under 0.0001, og vi afviser derfor hypotesen om at der ikke er forskel på de forskellige dæktypers benzinförbrug. Vi fortsætter med den additive model.

- (f) Undersøg om der er en signifikant forskel på de to bussers benzinförbrug.

**Solution:** Der testes ved at fitte modellen kun med `dæk` i middelværdien overfor den additive model og benytte anova i R.

R-kode:

```
m3ml <- lmer(benzin ~ daek+(1|cha), data=data, REML=FALSE)
anova(m1ml,m3ml)
```

Her fås at  $-2 \log Q = 0.0101$  med en frihedsgrad, hvilket giver en  $p$ -værdi på 0.92, og vi accepterer derfor hypotesen om at der ikke er forskel på de to bussers benzinförbrug. Den endelige model bliver derfor at middelværdien kun afhænger af dæktypen.

- (g) I den endelige model skal du kun beholde de faste effekter, der var signifikante. Angiv estimererne i den endelige model. Hvilken bus og dæktype vil du anbefale?

**Solution:** Her kan man evt teste den konstante model overfor en model med kun dæk i middelværdien, men det er ikke noget krav. R-kode:

```
m4ml <- lmer(benzin ~ 1+(1|cha), data=data, REML=FALSE)
anova(m3ml,m4ml)
```

Hvis det gøres fås at  $-2 \log Q = 92$  med 2 frihedsgrader, hvilket giver en meget lille  $p$ -værdi under 0.0001, og vi afviser igen at dæk ikke har nogen effekt på benzinforsbruget. Den endelige model bliver derfor at middelværdien kun afhænger af dæktypen.

Estimerne i den endelige model findes i R (her med REML, da disse er estimererne der bør rapporteres).

R-kode:

```
m3reml <- lmer(benzin ~ daek+(1|cha), data=data)
summary(m3reml)
m3reml <- lmer(benzin ~ daek-1+(1|cha), data=data)
summary(m3reml)
```

$$\hat{\beta} = (2773.9, 330.1, -191.3) \quad (\text{parametrizeret med intercept})$$

$$\hat{\beta} = (2773.9, 3104.0, 2582.6) \quad (\text{parametrizeret uden intercept})$$

$$\hat{\nu}^2 = 98937; \quad \hat{\nu} = 314.5$$

$$\hat{\sigma}^2 = 10951; \quad \hat{\sigma} = 104.6$$

Det er således dæktype M der bruger mindst benzin. Anbefalingen er derfor at det er ligegyldigt hvilken bus, man vælger, men man bør vælge dæktype M.

# Eksamens i Statistik 2

23. juni 2016

Eksamens varer 4 timer. Alle hjælpemidler er tilladt under eksamen, også computer, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant.

Eksamenssættet består af tre opgaver med i alt 18 delspørgermål. De tre opgaver vægtes ens. Data til opgave 3 ligger i filen `bus.txt` på en USB-stick. Stickens skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den kan altså ikke indgå som en del af besvarelsen.

## Opgave 1

- For fast  $r$ , hvor  $r$  er et naturligt tal, betragt fordelingen med tæthed

$$f_p(x) = \binom{x+r-1}{x} p^x (1-p)^r \quad \text{for } x \in \mathbb{N}_0$$

mht tællemålet på  $\mathbb{N}_0$ . Fordelingen afhænger af parameteren  $p \in (0, 1)$ .

Du kan uden bevis benytte at  $f_p$  er en tæthed. Det kan ligeledes benyttes uden bevis at der for  $|a| < 1$  gælder

$$\sum_{k=1}^{\infty} k \cdot a^k = \frac{a}{(1-a)^2} \quad ; \quad \sum_{k=1}^{\infty} k^2 \cdot a^k = \frac{a(1+a)}{(1-a)^3}$$

Lad  $X_1, \dots, X_n$  være uafhængige og identisk fordelte stokastiske variable med tæthed  $f_p$ , med kendt  $r \in \mathbb{N}$  og ukendt  $p \in (0, 1)$ .

- Opskriv likelihoodfunktionen og loglikelihoodfunktionen.

**Solution:** Likelihoodfunktion:

$$\begin{aligned} L_X(p) = \prod_{i=1}^n f_p(X_i) &= \prod_{i=1}^n \left[ \binom{X_i+r-1}{X_i} p^{X_i} (1-p)^r \right] \\ &= p^{\sum_{i=1}^n X_i} (1-p)^{nr} \prod_{i=1}^n \binom{X_i+r-1}{X_i} \end{aligned}$$

(Minus) loglikelihoodfunktion:

$$l_X(p) = -\log L_X(p) = -\sum_{i=1}^n \log \left( \frac{X_i + r - 1}{X_i} \right) - \sum_{i=1}^n X_i \log p - nr \log(1-p)$$

- (b) Find scorefunktionen og informationsfunktionen. Find fortegnet på den forventede information.

**Solution:** Scorefunktion:

$$\frac{d}{dp} l_X(p) = -\frac{1}{p} \sum_{i=1}^n X_i + \frac{nr}{1-p}$$

Informationsfunktion:

$$\frac{d^2}{dp^2} l_X(p) = \frac{1}{p^2} \sum_{i=1}^n X_i + \frac{nr}{(1-p)^2}$$

Da  $X_i \geq 0$  for alle  $i = 1, \dots, n$  er informationsfunktionen åbenlyst positiv.

- (c) Gør rede for at der er en entydig maksimaliseringsestimator  $\hat{p}$  og skriv den op.

**Solution:** Likelihoodligningen findes ved at sætte scorefunktionen lig nul. Informationsfunktionen er skarpt positiv for alle  $0 < p < 1$ . En eventuel løsning til likelihoodligningen vil derfor være et globalt minimum for  $l_X(p)$ . Likelihoodligningen er

$$\frac{1}{p} \sum_{i=1}^n X_i = \frac{nr}{1-p}$$

hvis løsning giver maksimaliseringsestimatoren

$$\hat{p} = \frac{\frac{1}{n} \sum_{i=1}^n X_i}{r + \frac{1}{n} \sum_{i=1}^n X_i}$$

- (d) Sæt nu  $r = 1$ . Undersøg om  $\hat{p}$  er konsistent.

**Solution:** Lad  $r = 1$ . Da er

$$\hat{p} = \frac{\frac{1}{n} \sum_{i=1}^n X_i}{1 + \frac{1}{n} \sum_{i=1}^n X_i}$$

og  $f_p(x) = p^x(1-p)$  for  $x \in \mathbb{N}_0$ . Vi skal bruge  $E(X)$ :

$$E(X) = \sum_{x=0}^{\infty} xp^x(1-p) = \frac{p}{1-p}$$

der ses at eksistere (sum af positive led), så LLN giver at

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \frac{p}{1-p}.$$

Vi har at

$$\hat{p} = f\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \quad \text{hvor} \quad f(x) = \frac{x}{1+x}.$$

Da  $f(x)$  er kontinuert for alle  $x > 0$  (og specielt i  $f(p)$ ) kan (5.3) benyttes, og vi får at

$$\hat{p} \xrightarrow{P} f\left(\frac{p}{1-p}\right) = p.$$

Således er  $\hat{p}$  konsistent.

- (e) Sæt nu  $r = 1$ . Gør rede for at  $\hat{p}$  er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling.

**Solution:** Vi skal også bruge  $\text{Var}(X)$ . Først ses at andetmomentet eksisterer:

$$E(|X|) = \sum_{x=0}^{\infty} x^2 p^x(1-p) = \frac{p(1+p)}{(1-p)^2} < \infty$$

Vi får

$$\text{Var}(X) = \frac{p(1+p)}{(1-p)^2} - \frac{p^2}{(1-p)^2} = \frac{p}{(1-p)^2}$$

CLT giver da at

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{as} N\left(\frac{p}{1-p}, \frac{p}{n(1-p)^2}\right)$$

Vi benytter nu deltametoden da  $f$  er differentiabel for  $p \in (0, 1)$ :

$$\hat{p} = f\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \xrightarrow{as} N\left(f\left(\frac{p}{1-p}\right), \frac{1}{n} \left(f'\left(\frac{p}{1-p}\right)\right)^2 \frac{p}{(1-p)^2}\right)$$

således at

$$\hat{p} \xrightarrow{as} N\left(p, \frac{1}{n} p(1-p)^2\right)$$

## Opgave 2

2. Betragt de to faktorer:

$$\begin{aligned} F &: \{1, \dots, N\} \longrightarrow \{\text{F1, F2, F3, F4, F5}\} \\ G &: \{1, \dots, N\} \longrightarrow \{\text{G1, G2, G3}\} \end{aligned}$$

Faktoren  $F \times G$  antages at være surjektiv.

Betragt varianskomponentmodellen  $X \sim \mathbf{N}(A\beta, \sigma^2\Sigma)$ , hvor  $A\beta \in L \subset \mathbb{R}^N$ ,  $\dim(L) = k$ ,  $\sigma^2 > 0$  og  $\Sigma = I + \lambda BB^T$ . Her er  $I$  identitetsmatricen og  $\lambda \geq 0$ . Vi antager yderligere at  $A = A_G$  er designmatricen for faktorunderrummet for faktor  $G$  og matricen  $B$  er effektmatricen hørende til effektparret  $(F, 1)$ .

- (a) Er  $X_i$ 'erne uafhængige? (At svare ja eller nej er nok)

**Solution:** Nej hvis  $\lambda > 0$ . Ja hvis  $\lambda = 0$ .

- (b) Hvad er  $k$ ? (Her skal både angives i ord hvad det er og angives en numerisk værdi)

**Solution:**  $k = 3$  er dimensionen af middelværdiunderrummet.

- (c) Antag at  $\dim(F \times G) = N$  og at datasættet er ordnet efter faktor  $F$ , således at først kommer alle observationer med label  $F1$  i faktor  $F$ , dernæst alle observationer med label  $F2$ , osv. Opskriv kovariansmatricen.

**Solution:**  $\text{Var}(X) = \sigma^2\Sigma$  hvor

$$\sigma^2\Sigma = \sigma^2 \begin{pmatrix} \Sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \Sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \Sigma_3 & 0 & 0 \\ 0 & 0 & 0 & \Sigma_4 & 0 \\ 0 & 0 & 0 & 0 & \Sigma_5 \end{pmatrix}$$

og for  $j = 1, \dots, 5$  er

$$\Sigma_j = \begin{pmatrix} 1 + \lambda & \lambda & \lambda \\ \lambda & 1 + \lambda & \lambda \\ \lambda & \lambda & 1 + \lambda \end{pmatrix}$$

- (d) Opskriv likelihoodfunktionen.

**Solution:** Likelihoodfunktion:

$$L_X(\beta, \sigma^2, \lambda) = \left( \frac{1}{2\pi\sigma^2} \right)^{N/2} \frac{1}{\sqrt{|\Sigma|}} \exp\{-(X - A\beta)^T \Sigma^{-1} (X - A\beta)/2\sigma^2\}$$

- (e) Er der en anden estimator af parametrene i modellen end maksimaliseringsestimatoren, man kunne foretrække? Argumenter for dit svar.

**Solution:** Ja, MLE underestimerer variansparametrene, især  $\lambda$  underestimeres. I stedet benyttes REML.

Resten af spørgsmålene drejer sig ikke om varianskomponentmodellen ovenfor.

- (f) Betragt de surjektive faktorer  $B$  og  $T$ , der antages at være usammenlignelige. De er begge forskellige fra den konstante faktor 1. Betragt deres tilhørende underrum  $L_B$  og  $L_T$ . Angiv hvilke af følgende udsagn, der er henholdsvis korrekte, falske eller ikke kan afgøres uden at vide mere om faktorerne.

- A.  $L_B + L_T \subseteq L_{B \times T}$
- B.  $L_{B \times T} \subseteq L_B + L_T$
- C.  $L_{B \times T} \subseteq L_{B \wedge T}$
- D.  $L_{B \wedge T} \subseteq L_{B \times T}$
- E.  $L_B + L_T \subseteq L_{B \wedge T}$
- F.  $L_{B \wedge T} \subseteq L_B + L_T$
- G.  $L_B + L_T \subseteq L_1$
- H.  $L_1 \subseteq L_B + L_T$
- I.  $L_1 \subseteq L_{B \wedge T}$
- J.  $L_{B \wedge T} \subseteq L_1$

**Solution:** A. Sand

B. Kan ikke afgøres. Det er næsten altid falsk, men her er et eksempel hvor det er sandt: 2 kategorier for hver faktor, 3 observationer, med antalstabellen:

1	1
1	0

Her er  $B \wedge T = 1$  (en sammenhængskomponent i designgrafen) så  $\dim(L_B + L_T) = \dim(L_B) + \dim(L_T) - \dim(L_{B \wedge T}) = 2 + 2 - 1 = 3$  og fra antalstabellen ses at  $\dim(L_{B \times T}) = 3$ . Da de har samme dimension og  $L_B + L_T \subseteq L_{B \times T}$  må  $L_B + L_T = L_{B \times T}$

C. Falsk

- D. Sand
- E. Falsk (da de er usammenlignelige, og derfor ikke kan være ens)
- F. Sand
- G. Falsk
- H. Sand
- I. Sand
- J. Det kan man ikke afgøre, det afhænger af  $\dim(L_{B \wedge T})$

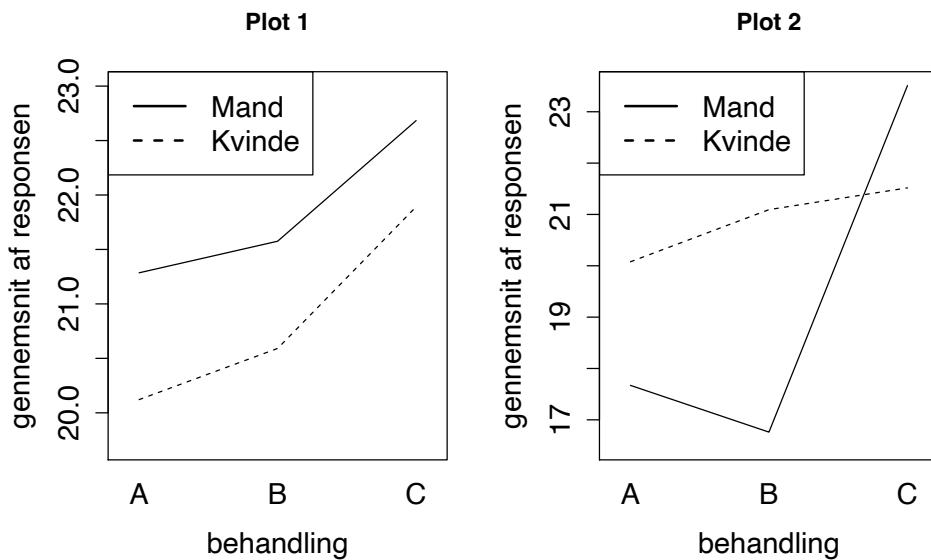
(g) Lad  $L_1$  og  $L_2$  være to underrum, begge forskellige fra  $\{0\}$ . Hvilke af følgende udsagn er korrekte?

- A. Hvis  $L_1 \perp_G L_2$  så er  $L_1 \subset L_2$
- B. Hvis  $L_1 \subset L_2$  så er  $L_1 \perp_G L_2$
- C. Hvis  $L_1 \subset L_2$  så er  $L_1 \perp L_2$
- D. Hvis  $L_1 \perp L_2$  så er  $L_1 \perp_G L_2$
- E. Hvis  $L_1 \perp_G L_2$  så er  $L_1 \perp L_2$

**Solution:** A. Falsk

- B. Sand
- C. Falsk
- D. Sand
- E. Falsk

(h) Betragt følgende interaktionsplots mellem de to faktorer Behandling og Køn, med henholdsvis 3 og 2 kategorier.



- Vurder for hvert af ovenstående to interaktionsplots om de bedst beskrives med en vekselvirkningsmodel eller med en additiv model.

**Solution:** Plot 1: Additiv. Plot 2: Vekselvirkning.

- Antag at responsen er lungekapacitet, og at man gerne vil have at den er stor. Hvilken behandling bør anbefales i hvert tilfælde?

**Solution:** Plot 1: Behandling C. Plot 2: Behandling C.

- Antag at responsen er blodtryk, og at man gerne vil have at den er lille. Hvilken behandling bør anbefales i hvert tilfælde?

**Solution:** Plot 1: Behandling A. Plot 2: Behandling A for kvinder, behandling B for mænd.

### Opgave 3

3. Ved en undersøgelse af virkningen af forskellige dæktyper på benzinförbruget af offentlige busser blev følgende forsøg gennemført: 3 busser,  $A$ ,  $B$  og  $C$  gennemkørte adskillige gange samme rundstrækning på ca. 10 km med 3 forskellige dæktyper  $K$ ,  $L$  og  $M$ , og benzinförbruget i milliliter blev målt.

Data er tilgængelige i filen `bus.txt` og består af variablene `bus`, `dæk` og `benzin`, hvor den sidste angiver benzinförbruget.

Vi antager i det følgende at de målte benzinförbrugstal kan ses som realisationer af uafhængige, normalfordelte stokastiske variable med samme varians  $\sigma^2$  og med en middelværdi der potentielt afhænger af bussen og dæktypen. Vi indicerer observationerne ved mængden  $I$ , og betragter to faktorer:

$$\begin{aligned} \text{Bus} &: I \longrightarrow \{A, B, C\} \\ \text{Dæk} &: I \longrightarrow \{K, L, M\} \end{aligned}$$

I spørgsmålene nedenfor bør angives relevante kvadrerede projektlængder, dimensioner, F-test størrelser og fordelinger, både teoretisk og med numeriske værdier.

- (a) Gør rede for at de to faktorer er geometrisk ortogonale og opstil en passende statistisk model for data.

**Solution:** Vi kalder faktorerne for B (Bus) og D (Dæk). Faktorerne er geometrisk ortogonale hvis de opfylder balanceequationen Sætning 14.8 - alternativt kan lemma 13.11 benyttes da designet for de to faktorer er sammenhængende.

Antalstabbel:

	K	L	M	
A	2	2	2	6
B	4	4	4	12
C	6	6	6	18
	12	12	12	36

Vi tjekker ligningerne i den første søjle, de øvrige er det samme:

$$2 = \frac{6 \cdot 12}{36}; \quad 4 = \frac{12 \cdot 12}{36}; \quad 6 = \frac{18 \cdot 12}{36}$$

Faktorerne er således geometrisk ortogonale.

Statistisk model for data:

$X$  er regulært normalfordelt på  $\mathbb{R}^I$  med middelværdi  $\xi \in L_{B \times D}$  og varians  $\sigma^2 I$ ,

hvor  $I$  er  $36 \times 36$  identitetsmatricen.

Alternativt:

$$X_i \sim \mathcal{N}(\xi_i, \sigma^2)$$

uafhængige, hvor middelværdivektoren  $\xi = (\xi_i)_{i \in I} \in L_{B \times D}$ .

- (b) Undersøg om der er en signifikant vekselvirkning mellem de to faktorer.

**Solution:** Hypotese:  $\xi \in L_{B+D}$ . Relevante størrelser (udregnet med formel (12.10)):

Faktor $F$	$\ P_F X\ ^2$	$\dim L_F$
$I$	329532176	36
$B \times D$	329139343	9
$B$	326914949	3
$D$	328516570	3
1	326308096	1

Dette giver yderligere følgende størrelser (formel (13.5), designet er ortogonalt og sammenhængende):

$$\|P_{B+D} X\|^2 = \|P_B X\|^2 + \|P_D X\|^2 - \|P_1 X\|^2 = 329123424$$

$$\dim P_{B+D} = \dim P_B + \dim P_D - \dim P_1 = 5$$

$F$ -test for vekselvirkning mellem de to faktorer:

$$\begin{aligned} F &= \frac{(\|P_{B \times D} X\|^2 - \|P_{B+D} X\|^2) / (\dim L_{B \times D} - \dim L_{B+D})}{(\|X\|^2 - \|P_{B \times D} X\|^2) / (\dim L_I - \dim L_{B \times D})} \\ &= \frac{(329139343 - 329123424) / (9 - 5)}{(329532176 - 329139343) / (36 - 9)} = 0.2735 \end{aligned}$$

der skal vurderes i en  $F(4, 27)$ -fordeling. Vi får  $p = 0.8925$  og accepterer derfor nulhypotesen om ingen vekselvirkning mellem de to faktorer.

- (c) Fortsæt med den additive model. Undersøg om der er en signifikant forskel på de tre bussers benzinforsbrug. Test om dæktypen påvirker benzinforsbruget.

**Solution:** For at undersøge om der er signifikant forskel på bussernes benzinforsbrug, opstilles hypotesen om ingen forskel:  $H_1 : \xi \in L_D$ .  $F$ -teststørrelsen

bliver:

$$\begin{aligned} F &= \frac{(||P_{B+D}X||^2 - ||P_D X||^2) / (\dim L_{B+D} - \dim L_D)}{(||X||^2 - ||P_{B+D}X||^2) / (\dim L_I - \dim L_{B+D})} \\ &= \frac{(329123424 - 328516570) / (5 - 3)}{(329532176 - 329123424) / (36 - 5)} = 23.01 \end{aligned}$$

der skal vurderes i en  $F(2, 31)$ -fordeling. Vi får  $p < 0.00001$  og afviser derfor nulhypotesen om ingen virkning af behandling.

For at undersøge om dæktypen påvirker benzinförbruget, opstilles nulhypotesen:  $H_2 : \xi \in L_B$ .  $F$ -teststørrelsen bliver:

$$\begin{aligned} F &= \frac{(||P_{B+D}X||^2 - ||P_B X||^2) / (\dim L_{B+D} - \dim L_B)}{(||X||^2 - ||P_{B+D}X||^2) / (\dim L_I - \dim L_{B+D})} \\ &= \frac{(329123424 - 326914949) / (5 - 3)}{(329532176 - 329123424) / (36 - 5)} = 83.75 \end{aligned}$$

der skal vurderes i en  $F(2, 31)$ -fordeling. Vi får  $p < 0.00001$  og afviser derfor nulhypotesen om ingen forskel mellem dæktyper.

- (d) Estimer parametrene i den additive model hvor begge faktorer indgår, og angiv deres simultane fordeling.

**Solution:** Vi benytter Korollar 10.21 og formel (10.27) til  $\sigma^2$ . Her er  $A$  designmatricen, der er  $36 \times 5$ . Vi får (parametrisering med intercept)

$$(A^T A)^{-1} = \begin{bmatrix} 36 & 12 & 18 & 12 & 12 \\ 12 & 12 & 0 & 4 & 4 \\ 18 & 0 & 18 & 6 & 6 \\ 12 & 4 & 6 & 12 & 0 \\ 12 & 4 & 6 & 0 & 12 \end{bmatrix}^{-1} = \begin{bmatrix} 0.22 & -0.17 & -0.17 & -0.08 & -0.08 \\ -0.17 & 0.25 & 0.17 & 0.00 & 0.00 \\ -0.17 & 0.17 & 0.22 & 0.00 & 0.00 \\ -0.08 & 0.00 & 0.00 & 0.17 & 0.08 \\ -0.08 & 0.00 & 0.00 & 0.08 & 0.17 \end{bmatrix}$$

og  $\hat{\beta} = (A^T A)^{-1} A^T X$ , hvilket giver

$$\hat{\beta} = (2985.6, -181.8, 108.50, -255.0, 349.3)^T \text{ hvor } \hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (A^T A)^{-1})$$

Her er  $\beta_1$  intercept-estimatet, der angiver middelværdien af benzinförbruget for bus A med dæktype K,  $\beta_2$  angiver det yderligere bidrag hvis det er bus B,  $\beta_3$  angiver det yderligere bidrag hvis det er bus C,  $\beta_4$  angiver det yderligere bidrag hvis det er dæktype L, og  $\beta_5$  angiver det yderligere bidrag hvis det er dæktype M.

Hvis den studerende vælger parametrisering uden intercept fås

$$(A^T A)^{-1} = \begin{bmatrix} 6 & 0 & 0 & 2 & 2 \\ 0 & 12 & 0 & 4 & 4 \\ 0 & 0 & 18 & 6 & 6 \\ 2 & 4 & 6 & 12 & 0 \\ 2 & 4 & 6 & 0 & 12 \end{bmatrix}^{-1} = \begin{bmatrix} 0.22 & 0.06 & 0.06 & -0.08 & -0.08 \\ 0.06 & 0.14 & 0.06 & -0.08 & -0.08 \\ 0.06 & 0.06 & 0.11 & -0.08 & -0.08 \\ -0.08 & -0.08 & -0.08 & 0.17 & 0.08 \\ -0.08 & -0.08 & -0.08 & 0.08 & 0.17 \end{bmatrix}$$

og  $\hat{\beta} = (A^T A)^{-1} A^T X$ , hvilket giver

$$\hat{\beta} = (2985.6, 2803.8, 3094.1, -255.0, 349.3)^T \text{ hvor } \hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (A^T A)^{-1})$$

I tabelform fås følgende estimerater for middelværdierne:

	Bus = A	Bus = B	Bus = C
Dæk = K	2985.6	2803.8	3094.1
Dæk = L	2730.6	2548.8	2839.1
Dæk = M	3334.8	3153.1	3443.3

Benzinforbruget estimeres derfor til at være mindst for bus B med dæktype L. Varianseestimatet er

$$\tilde{\sigma}^2 = \frac{\|X\|^2 - \|P_{B+D}X\|^2}{(\dim L_I - \dim L_{B+D})} = 13185.56; \quad \tilde{\sigma}^2 \sim \frac{\sigma^2}{31} \chi^2_{31}$$

Derudover er  $\hat{\beta}$  og  $\tilde{\sigma}^2$  uafhængige. Det er OK at angive estimaterne for  $\xi$  i stedet for  $\beta$ , men oversættelsen til noget fortolkeligt må da gerne følge med.

- (e) De to busser  $A$  og  $B$  er samme mærke bus, hvorimod bus  $C$  er af et andet mærke. Dermed kan det tænkes at busserne  $A$  og  $B$  virker ens. Opstil og test denne hypotese.

**Solution:** Vi definerer en ny faktor med to labels:

$$\text{B2 (Bus2)} : I \longrightarrow \{AB, C\}$$

Vi opstiller hypotesen  $H_3 : \xi \in L_{B2+D}$ .  $F$ -teststørrelsen bliver:

$$\begin{aligned} F &= \frac{(\|P_{B+D}X\|^2 - \|P_{B2+D}X\|^2) / (\dim L_{B+D} - \dim L_{B2+D})}{(\|X\|^2 - \|P_{B+D}X\|^2) / (\dim L_I - \dim L_{B+D})} \\ &= \frac{(329123424 - 328991292) / (5 - 4)}{(329532176 - 329123424) / (36 - 5)} = 10.021 \end{aligned}$$

der skal vurderes i en  $F(1, 31)$ -fordeling. Vi får  $p = 0.00346$  og afviser derfor nulhypotesen om at de to busser A og B har samme benzinförbrug.

## Eksamens i Statistik 1, vejledende besvarelse

### 6. april 2016

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `april17.R`.

#### Opgave 1

1. Likelihoodfunktionen er

$$L_{x,y}(\theta) = \prod_{i=1}^n \left( \frac{1}{\theta} e^{-x_i/\theta} \right) (\theta e^{-\theta y_i}) = e^{-\frac{1}{\theta} \sum x_i - \theta \sum y_i} = e^{-\frac{1}{\theta} S_x - \theta S_y}$$

hvor  $S_x = \sum_{i=1}^n x_i$  og  $S_y = \sum_{i=1}^n y_i$  som i opgaveteksten.

Vi får således

$$\begin{aligned}\ell_{x,y}(\theta) &= -\log L_{x,y}(\theta) = \frac{1}{\theta} S_x + \theta S_y \\ \ell'_{x,y}(\theta) &= -\frac{1}{\theta^2} S_x + S_y \\ I_{x,y}(\theta) &= S'_{x,y}(\theta) = \frac{2}{\theta^3} S_x \\ i(\theta) &= E_\theta I_{X,Y}(\theta) = \frac{2}{\theta^3} E_\theta S_x = \frac{2}{\theta^3} n\theta = \frac{2n}{\theta^2}\end{aligned}$$

hvor vi til sidst har benyttet at  $E_\theta X_i = \theta$ .

2. Vi løser først scoreligningen for en observation  $(x, y)$ :

$$S_{x,y}(\theta) = 0 \Leftrightarrow \frac{1}{\theta^2} S_x = S_y \Leftrightarrow \theta^2 = \frac{S_x}{S_y} \Leftrightarrow \theta = \sqrt{\frac{S_x}{S_y}}$$

Der er således et entydigt stationært punkt. Da der desuden gælder  $I_{x,y}(\theta) > 0$  for alle  $\theta > 0$ , giver det stationære punkt anledning til et minimum for  $\ell_{x,y}$ . Dette gælder for alle  $(x, y)$ , så vi får at ML estimatoren er

$$\hat{\theta} = \sqrt{\frac{S_x}{S_y}}$$

hvor  $S_X = \sum_{i=1}^n X_i$  og  $S_Y = \sum_{i=1}^n Y_i$  som i opgaveteksten.

Den asymptotiske fordeling af  $\hat{\theta}$  er

$$\hat{\theta} \stackrel{as}{\sim} N\left(\theta, i(\theta)^{-1}\right), \text{ dvs. } \hat{\theta} \stackrel{as}{\sim} N\left(\theta, \frac{\theta^2}{2n}\right)$$

3. Eftersom eksponentiaffordelinger er gammafordelinger med formparameter 1, får vi

$$\begin{aligned}S_X &\sim \Gamma(n, \theta) \text{ og dermed } \frac{1}{\theta} S_X \sim \Gamma(n, 1) \\ S_Y &\sim \Gamma(n, 1/\theta) \text{ og dermed } \theta S_X \sim \Gamma(n, 1)\end{aligned}$$

Alle  $X_i$ 'er  $Y_i$ 'er uafhængige, så  $\frac{1}{\theta}S_X$  og  $\theta S_Y$  er uafhængige. Vi får derfor fra vinket at

$$Z = \theta^2 \frac{S_Y}{S_X} = \frac{\theta S_Y}{\frac{1}{\theta}S_X} \sim F(2n, 2n).$$

Hvis  $f_1$  og  $f_2$  er 2.5% og 97.5% fraktilerne i  $F(2n, 2n)$  fordelingen gælder der altså for alle  $\theta$  at

$$0.95 = P(f_1 < Z < f_2) = P\left(f_1 < \theta^2 \frac{S_Y}{S_X} < f_2\right) = P\left(\sqrt{\frac{f_1 S_X}{S_Y}} < \theta < \sqrt{\frac{f_2 S_X}{S_Y}}\right),$$

således at

$$\left(\sqrt{\frac{f_1 S_X}{S_Y}}, \sqrt{\frac{f_2 S_X}{S_Y}}\right) = (\hat{\theta}\sqrt{f_1}, \hat{\theta}\sqrt{f_2})$$

er et eksakt 95% konfidensinterval for  $\theta$ .

4. For  $n = 7$  er  $f_1 = 0.3357$  og  $f_2 = 2.9786$ . For de givne data er  $S_x = 21.81$  og  $S_y = 4.45$ . Indsættes værdierne i formlen for konfidensintervaller får vi det eksakte 95% konfidensinterval  $(1.283, 3.821)$ .

MLE for de givne data er  $\hat{\theta} = 2.214$  og den estimerede Fisherinformation er  $i(\hat{\theta}) = 2.856$ . Det approksimative 95% konfidensinterval baseret på den falske Waldteststørrelse bliver således

$$\hat{\theta} \pm 1.96 \frac{1}{\sqrt{i(\hat{\theta})}} = 2.214 \pm 1.160 = (1.054, 3.374).$$

Wald KI er mindre end det eksakte, men vi kender ikke dets præcise dækningsgrad, så vi kan ikke umiddelbart sige at vi foretrækker det.

Udfra simulationer viser det sig at dækningsgraden for Wald KI faktisk er snublende tæt på 95% selv for  $n$  så lille som 7, og at den gennemsnitlige længde faktisk *er* kortere for Wald KI end for det eksakte KI, således at Wald faktisk er at foretrække. Men dette er ikke en del af besvarelsen...

## Opgave 2

1.  $\Sigma$  er en lovlig variansmatrix hvis og kun hvis den er positiv semidefinit. Pga. „blokstrukturen“ i  $\Sigma$  er det ensbetydende med at følgende to ting gælder:

- $\Sigma_{33} \geq 0$ , dvs.  $\varphi \geq 0$
- Den øvre  $2 \times 2$  matrix er positiv semidefinit. Da  $\Sigma_{11} > 0$  gælder dette hvis og kun hvis  $4 - \varphi^2 \geq 0$ , dvs.  $\varphi^2 \leq 4$  eller  $-2 \leq \varphi \leq 2$ .

Altså er  $\Sigma$  en lovlig variansmatrix hvis og kun hvis  $0 \leq \varphi \leq 2$ .

Fordelingen er regulær hvis og kun hvis  $\Sigma$  er invertibel. Determinanten

$$\det(\Sigma) = (4 - \varphi^2)\varphi$$

er ikke-negativ for alle de lovlige værdier af  $\varphi$  og positiv hvis ydermere  $\varphi \notin \{0, 2\}$ . Fordelingen er altså regulær for  $\varphi \in (0, 2)$  og singulær for  $\varphi \in \{0, 2\}$ .

2. Hvis vi definerer

$$C = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

og bruger transformationssætningen for normalfordelingen, får vi at

$$\begin{pmatrix} X_1 + X_2 \\ X_3 \end{pmatrix} = CX \sim N(0, C\Sigma C^T).$$

Variansmatricen viser sig at være

$$C\Sigma C^T = \begin{pmatrix} 5+2\varphi & 0 \\ 0 & \varphi \end{pmatrix}.$$

3. Fra spørgsmål 2 ses at  $X_1 + X_2 \sim N(0, 5 + 2\varphi)$ . Det følger at

$$\begin{aligned} E\tilde{\varphi} &= \frac{E(X_1 + X_2)^2 - 5}{2} = \frac{5 + 2\varphi - 5}{2} = \varphi \\ \text{Var}(\tilde{\varphi}) &= \frac{1}{4}\text{Var}(X_1 + X_2)^2 = \frac{1}{4}2(5 + 2\varphi)^2 = \frac{1}{2}(5 + 2\varphi)^2 \end{aligned}$$

hvor vi har benyttet vinket om at hvis  $Z \sim N(0, \sigma^2)$ , så er  $\text{Var}(Z^2) = 2\sigma^2$ .

Desuden er  $X_3 \sim N(0, \varphi)$ , så

$$E\hat{\varphi} = EX_3^2 = \varphi, \quad \text{Var}(\hat{\varphi}) = \text{Var}(X_3^2) = 2\varphi^2$$

Både  $\tilde{\varphi}$  og  $\hat{\varphi}$  er altså centrale estimatorer for  $\varphi$ .

Bemærk evt. at

$$\text{Var}(\tilde{\varphi}) = \frac{1}{2}(5 + 2\varphi)^2 > \frac{1}{2}(2\varphi)^2 = 2\varphi^2 = \text{Var}(\hat{\varphi}),$$

så  $\hat{\varphi}$  har mindst varians. (Dette er ikke så overraskende:  $\varphi$  bestemmer korrelationen mellem  $X_1$  og  $X_2$  som er endnu sværere at estimere præcist end en varians.)

Antag til sidst at  $\varphi = 1$ . Så er  $X_1 + X_2 \sim N(0, 7)$  og

$$\begin{aligned} P(0 < \tilde{\varphi} < 2) &= P(5 < (X_1 + X_2)^2 < 9) \\ &= P(-3 < X_1 + X_2 < \sqrt{5}) + P(\sqrt{5} < X_3 < 3) \\ &= 0.141 \end{aligned}$$

og  $X_3 \sim N(0, 1)$ , så

$$P(0 < \hat{\varphi} < 2) = P(0 < X_3^2 < 2) = P(-\sqrt{2} < X_3 < \sqrt{2}) = 0.843.$$

Estimatoren  $\hat{\varphi}$  rammer altså det lovlige område meget oftere end  $\tilde{\varphi}$  (i hvert fald når den sande værdi er  $\varphi = 1$ ) og har desuden mindst varians. Vi foretrækker derfor  $\hat{\varphi}$  fremfor  $\tilde{\varphi}$ .

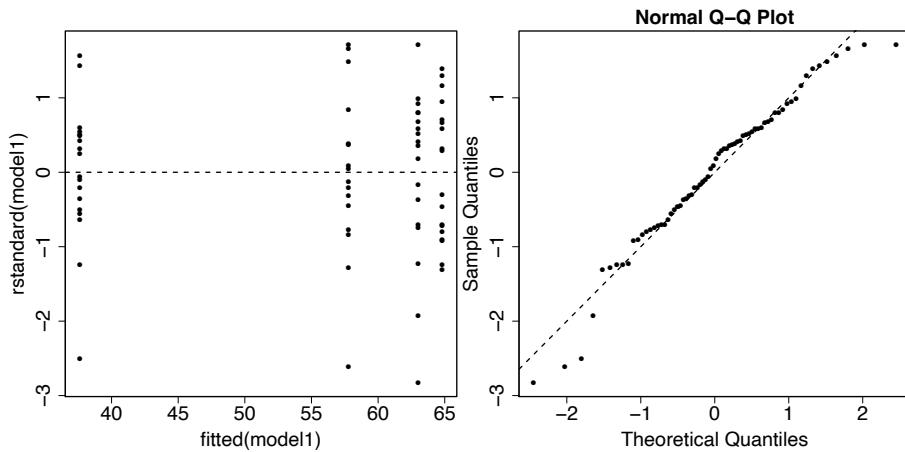
### Opgave 3

1. Data skal analyseres med en ensidet variansanalyse. Vi antager altså at vektoren af skudhøjder  $x = (x_1, \dots, x_{70})$  er udfald af en en stokastisk variabel  $X \sim N(\xi, \sigma^2 I)$  hvor  $\xi \in L_G$  og  $\sigma^2 > 0$  er de ukendte parametre og  $L_G$  er faktorunderrummet hørende til de fire kombinationer af eksperiment og svampebehandling.

Modellen fittes fx med kommandoen

```
model1 <- lm(sh ~ group, data=shData)
```

Residualplot og QQ-plot for de standardiserede residualer er vist nedenfor:



Begge plots ser ganske fornuftige ud! I residualplottet ligger værdierne cirka symmetrisk om nul (sådan vil det faktisk nødvendigvis være i en ensidet variansanalyse) og med cirka samme spredning for de fire grupper af data. I QQ-plottet er der fire observationer der stikker lidt ud, men ellers ligger punkterne omkring 0-1 linjen. (Hvis man simulerer 70 standardfordelte observationer får man ofte noget hvor afvigelserne fra den rette linie er af samme størrelsesorden.)

2. Hypotesen er at der ikke er forskel på middelværdierne i de fire grupper, altså  $H : \xi \in L_1$ . Hypotesen testes med et  $F$ -test, fx med følgende kode:

```
model2 <- lm(sh ~ 1, data=shData)
anova(model2, model1)
```

Den observerede værdi af teststørrelsen er  $f = 19.49$ . Vurderet i  $F$ -fordelingen med  $(3, 66)$  frihedsgrader giver dette  $p$ -værdien  $p = 3.7 \cdot 10^{-9}$ . Der er altså stærk evidens i data for forskel mellem de fire grupper.

3. Parameteren  $\delta_1$  er en af parametrene i `model1`, så estimat og 95% konfidensinterval aflæses direkte ved brug af `summary` og `confint` (dog skal fortegnet skiftes):

$$\hat{\delta}_1 = -5.23, \quad 95\% \text{ KI: } (-13.45, 3.00)$$

Estimat og KI for  $\delta_2$  kan fx findes ved at fitte modellen med gruppe `control2` som referencegruppe. Så fås

$$\hat{\delta}_2 = -27.20, \quad 95\% \text{ KI: } (-35.18, -19.22)$$

4. Hvis de fire gruppemiddelværdier betegnes  $\alpha_{1c}, \alpha_{1f}, \alpha_{2c}$  og  $\alpha_{2f}$ , så er

$$\delta_1 = \alpha_{1f} - \alpha_{1c}, \quad \delta_2 = \alpha_{2f} - \alpha_{2c}.$$

Altså er

$$\bar{\delta} = \frac{1}{2}(\delta_1 + \delta_2) = \frac{1}{2}(\alpha_{1f} - \alpha_{1c} + \alpha_{2f} - \alpha_{2c})$$

Hvis vi fitter modellen uden referencegruppe („uden intercept“), så optræder  $\alpha$ -parameterene i rækkefølgen  $\alpha_{1c}, \alpha_{2c}, \alpha_{1f}, \alpha_{2f}$ , således at

$$\bar{\delta} = \left( -\frac{1}{2} \quad -\frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \right) \alpha.$$

Vi kan derfor benytte eksempel 10.30 med  $\psi^T = \left( -\frac{1}{2} \quad -\frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \right)$ . Vi får

$$\hat{\delta} = \psi^T \hat{\alpha}, \quad \text{Var}(\hat{\delta}) = \psi^T \text{Var}(\hat{\alpha}) \psi$$

Den ønskede parametrisering opnås med kommandoen

```
model4 <- lm(sh ~ group-1, data=shData)
```

og estimat  $\hat{\alpha}$  og variansmatrix  $\text{Var}(\hat{\alpha})$  fås fx. med `coef` og `vcov`. Vi får  $\hat{\delta} = -16.213$  med estimeret spredning  $\text{SE}(\hat{\delta}) = 2.870$ . Bemærk at estimatet (naturligvis) er gennemsnittet af  $\hat{\delta}_1$  og  $\hat{\delta}_2$ .

Det tilhørende 95% konfidensinterval er

$$-16.213 \pm 1.997 \cdot 2.870 = -16.213 \pm 5.731 = (-21.944, -10.481)$$

hvor vi har benyttet at 97.5% fraktilen i  $t_{66}$  fordelingen er 1.997.

Konfidensintervallet indeholder kun negative værdier, så det tyder på at svampebehandlingen påvirker plantevæksten negativt — det man gerne ville undgå.

## Opgave 4

1. Likelihoodfunktionen er (på nær en multiplikativ konstant)

$$L_x(\sigma^2) = (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} SS_x},$$

og det følger på sædvanlig vis, fx via lemma 4.18, at  $L_x$  har maksimum for  $\sigma^2 = \frac{1}{n} SS_x$ . Det er også fint at argumentere direkte via sætning 10.19 med den modifikation at der ikke er en middelværdi der skal estimeres.

MLE er altså  $\hat{\sigma}^2 = \frac{1}{n} SS_x$ .

Log-likelihooden er (på nær en additiv konstant)

$$\ell_x(\sigma^2) = -\log L_x(\sigma^2) = \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} SS_x.$$

Specielt er

$$\ell_x(\hat{\sigma}^2) = \frac{n}{2} \log\left(\frac{SS_x}{n}\right) + \frac{n}{2}, \quad \ell_x(1) = \frac{1}{2} SS_x$$

Likelihood ratio teststørrelsen for hypotesen  $H : \sigma^2 = 1$  er derfor

$$LR(x) = 2(\ell_x(1) - \ell_x(\hat{\sigma}^2)) = SS_x - n \log(SS_x) + n \log(n) - n$$

2. For det givne  $x$  er  $SS_x = 25.1867$ , så  $\hat{\sigma}^2 = 2.519$ .

Endvidere fås  $LR(x) = 5.95$ . Det sædvanlige asymptotiske resultat siger at  $LR(X) \xrightarrow{as} \chi_1^2$  da dimensionalfaldet ved hypotesen er 1. Derfor fås  $p$ -værdien

$$p = p(x) = P(W \geq 5.95) = 0.015.$$

hvor  $W \sim \chi^2_1$ . Hypotesen afvises: Der er evidens i data for at variansen ikke er lig 1.

For  $n = 10$  er de relevante fraktiler for det eksakte test  $z_1 = 3.247$  (2.5% fraktilen) og  $z_2 = 20.483$  (97.5% fraktilen). Da  $SS_x$  ligger udenfor intervallet  $(z_1, z_2)$ , afvises hypotesen, dvs.  $x \in \mathcal{H}$ .

For de observerede data er de to testmetoder er altså enige om at hypotesen skal afvises.

3. Jeg har kørt 10000 simulationer og fået følgende relative hyppigheder:

Sand værdi af $\sigma^2$	Relativ hyppighed hvormed hypotesen forkastes	
	LR test	Alternativt test
1	0.0519	0.0480
0.5	0.2926	0.2328
1.5	0.1652	0.1952

For  $\sigma^2 = 1$  er hypotesen sand, og vi ser at det faktiske niveau er tæt på 5% for begge test. Det vidste vi godt for det alternative test, men ikke for LR-testet. For  $\sigma^2 = 0.5$  er LR-testet det stærkeste, mens det modsatte er tilfældet for  $\sigma^2 = 1.5$ . Man kan altså ikke sige at det ene test er uniformt sterkere/bedre end det andet.

## Eksamens i Statistik 1, vejledende besvarelse

### 29. juni 2017

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `august17.R`.

#### Opgave 1

- Likelihoodfunktionen er (proportional med)

$$L_x(\theta) = \prod_{i=1}^n \theta t_i x_i^{\theta t_i - 1} \propto \theta^n \prod_{i=1}^n x_i^{\theta t_i}$$

Vi får således (på nær en additiv) konstant

$$\ell_x(\theta) = -\log L_x(\theta) = -n \log \theta - \theta \sum_{i=1}^n t_i \log x_i$$

og dermed

$$\begin{aligned} S_x(\theta) &= \ell'_x(\theta) = -\frac{n}{\theta} - \sum_{i=1}^n t_i \log x_i \\ I_x(\theta) &= S'_x(\theta) = \frac{n}{\theta^2} \\ i(\theta) &= E_\theta I_x(\theta) = \frac{n}{\theta^2} \end{aligned}$$

- Vi løser først scoreligningen for en observation  $x$ :

$$S_{x,y}(\theta) = 0 \Leftrightarrow \frac{n}{\theta} = -\sum_{i=1}^n t_i \log x_i \Leftrightarrow \theta = -\frac{n}{\sum_{i=1}^n t_i \log x_i}$$

Der er således et entydigt stationært punkt. Da der desuden gælder  $I_x(\theta) > 0$  for alle  $\theta > 0$ , giver det stationære punkt anledning til et minimum for  $\ell_{x,y}$ . Bemærk at løsningen til scoreligningen er positiv da alle  $x_i \in (0, 1)$ , så løsningen ligger i parametermængden.

Ovenstående gælder for alle  $x \in (0, 1)^n$ , så vi får at ML estimatoren er

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^n t_i \log X_i}.$$

Den asymptotiske fordeling af  $\hat{\theta}$  er

$$\hat{\theta} \stackrel{as}{\sim} N\left(\theta, i(\theta)^{-1}\right), \text{ dvs. } \hat{\theta} \stackrel{as}{\sim} N\left(\theta, \frac{\theta^2}{n}\right)$$

3. Lad  $t > 0$  være givet og lad  $X$  have tæthed  $f(x) = \theta t \cdot x^{\theta t - 1}$  for  $x > 0$ . Definer desuden funktionen  $h : (0, 1) \rightarrow (0, \infty)$  ved  $h(x) = -t \log x$  og  $Y = h(X)$ . Funktionen  $h$  er strengt aftagende, og dermed bijektiv, samt kontinuert differentielabel. Vi får følgende:

$$h^{-1}(y) = e^{-y/t}, \quad Dh^{-1}(y) = -\frac{1}{t}e^{-y/t}, \quad y > 0,$$

og det følger fra den endimensionale transformationssætning af tætheden for  $Y$  er givet ved  $g(y) = 0$  for  $y \leq 0$  og

$$g(y) = f(h^{-1}(y)) |Dh^{-1}(y)| = \theta t (e^{-y/t})^{\theta y - 1} \frac{1}{t} e^{-y/t} = \theta e^{-\theta y}$$

for  $y > 0$ . Således er  $Y$  eksponentiafordelt med middelværdi  $1/\theta$  eller gammafordelt med formparameter 1 og skalaparameter  $1/\theta$ :  $Y_i \sim \Gamma(1, 1/\theta)$ .

Altså er  $Y_1, \dots, Y_n$  uafhængige (fordi  $X_i$ 'erne er det) og alle  $Y_i \sim \Gamma(1, 1/\theta)$ . Pga. foldningsegenskaben for gammafordelingen, får vi så  $S_Y \sim \Gamma(n, 1/\theta)$  og endelig  $\theta S_Y \sim \Gamma(n, 1)$ .

Således er  $\theta S_Y$  en pivot, og hvis  $g_1$  og  $g_2$  er 2.5% og 97.5% fraktilerne i  $\Gamma(n, 1)$  fordelingen, så er

$$0.95 = P(g_1 < \theta S_Y < g_2) = P\left(\frac{g_1}{S_Y} < \theta < \frac{g_2}{S_Y}\right)$$

og

$$\left(\frac{g_1}{S_Y}, \frac{g_2}{S_Y}\right)$$

er et eksakt 95% konfidensinterval for  $\theta$ .

4. Vi har fra tidligere at

$$\log L_x(\theta) = n \log \theta + \theta \sum_{i=1}^n t_i \log x_i = n \log \theta - n \theta \bar{y},$$

og at  $\hat{\theta} = 1/\bar{y}$ . Derfor er

$$\begin{aligned} LR(\theta, x) &= 2 \left( \log L_x(\hat{\theta}) - \log L_x(\theta) \right) \\ &= 2 \left( -n \log \bar{y} - n - n \log \theta + n \theta \bar{y} \right) \\ &= 2n \left( -\log \bar{y} - 1 - \log \theta + \theta \bar{y} \right). \end{aligned}$$

For hypotesen  $H : \theta = 4$  fås  $LR(4, x) = 1.83$ . Hvis vi benytter  $\chi^2$  approksimationen til fordelingen af  $LR(\theta, X)$  under hypotesen, fås  $p$ -værdien  $P(LR(X, 4) \geq 1.83) = 0.18$ , så vi kan ikke afvise hypotesen: Der er altså ikke evidens i data for at  $\theta$  er forskellig fra 4.

5. For det givne datasæt og  $n = 10$  er

$$S_y = 1.578, \quad \bar{y} = 0.1578, \quad \hat{\theta} = 6.337, \quad g_1 = 4.795, \quad g_2 = 17.08$$

Det eksakte 95% konfidensinterval er således

$$\left(\frac{g_1}{S_y}, \frac{g_2}{S_y}\right) = (3.039, 10.827).$$

Konfidensintervallet baseret på  $LR(\theta, X)$  er defineret ved

$$\{\theta > 0 \mid LX(\theta, x) < q_{0.95}\}$$

hvor  $q_{0.95} = 3.84$  er 95% fraktilen i  $\chi^2$  fordelingen med en frihedsgad. Eftersom  $\theta \rightarrow LR(\theta, x)$  er konveks og  $LR(\hat{\theta}, x) = 0$ , er endepunkterne i konfidensintervallet løsningerne til ligningen  $LR(\theta, x) = 3.84$ . Hvis vi indsætter  $\theta = 3.1754$  og  $\theta = 11.1151$  i udtrykket for  $LR(\theta, x)$  får vi netop 3.84 (på nær afrundingsfejl).

6. Den asymptotiske fordeling af  $\hat{\theta}$  er  $N(\theta, \theta^2/n)$ , specielt er spredning for  $\hat{\theta}$  approximativt  $\theta/\sqrt{n}$ . Jeg fik følgende skema med 5000 simulationer:

n	$\theta$	Simulation		Asymptotisk fordeling	
		middelværdi	spredning	gennemsnit	spredning
10	5	5.59	1.99	5	1.58
25	5	5.21	1.08	5	1
250	5	5.02	0.32	5	0.32

Vi ser at middelværdi og spredning i den asymptotiske fordeling først er fornuftige for  $n = 250$ . For  $n = 10$  og  $n = 25$  er den faktiske middelværdi og den faktiske varians begge større end den asymptotiske fordeling tilsliger.

Specielt er  $\hat{\theta}$  ikke en central estimator, thi så skulle den have den korrekte middelværdi også for små værdier af  $n$ . (Derimod er  $1/\hat{\theta}$  faktisk central for  $1/\theta$ , hvorfaf det i øvrigt via Jensens ulighed følger at  $\hat{\theta}$  ikke er central for  $\theta$ .)

## Opgave 2

1. Hvis vi definerer

$$C = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix},$$

så er  $Y = CX$ , og transformationssætningen for normalfordelingen giver

$$Y = CX \sim N(0, C\Sigma C^T).$$

Variansmatricen viser sig at være

$$VY = C\Sigma C^T = \begin{pmatrix} 13 & 19 \\ 19 & 30 \end{pmatrix}.$$

Variansmatricen har determinant 29 og er dermed regulær, så  $Y$  er regulært normalfordelt på  $\mathbb{R}^2$ .

2. Vi har at  $\det(\Sigma) = 0$ , så  $X$  er singulært normalfordelt på  $\mathbb{R}^3$ .

Sæt  $D = (1 \ -1 \ -1)$ . Så er  $Z = X_1 - X_2 - X_3 = DX \sim N(0, D\Sigma D^T) = N(0, 0)$ . Altså er  $Z = 0$ , eller  $X_3 = X_1 - X_2$ , med sandsynlighed 1, dvs.  $P(X \in V) = 1$ . Det er desuden klart at  $X$  ikke kan være koncentreret på en mængde af lavere dimension eftersom fx  $(X_1, X_2)^T$  er regulært normalfordelt på  $\mathbb{R}^2$ .

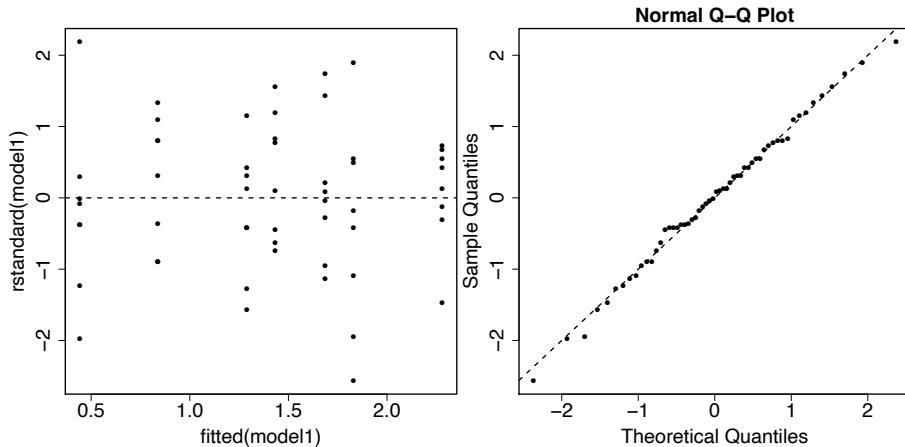
Man kan også argumentere udfra resultatet i spørgsmål 1: Fordelingen af  $Y$  er singulær, dvs.  $Y_2 = a + bY_1$  med sandsynlighed 1 for passende værdier  $a$  og  $b$ . Udfra middelværdierne ser vi at  $a = 0$ , udfra varianserne at  $b = 2$ . Altså er  $Y_2 = 2Y_1$ , eller  $X_3 = X_1 - X_2$ .

## Opgave 3

1. Modellen er en multipel regressionsmodel og fittes med kommandoen

```
model1 <- lm(styrke ~ A + B + C, data=limData)
```

Residualplot og QQ-plot for de standardiserede residualer er vist nedenfor:



Begge plots ser yderst fornuftige ud! I residualplottet ligger værdierne cirka symmetrisk om nul og med cirka samme spredning henover  $x$ -aksen. I QQ-plottet ligger punkterne nydeligt omkring 0/1 linien.

*Bemærkning:* Alle tre prædiktorer har kun to mulige værdier (0/1), og den multiple regressionsmodel er derfor sammenfaldende med den additive tresidede variansanalysemodel fra Stat2.

2. Estimaterne er følgende:

$$\hat{\alpha} = 0.4425, \hat{\beta}_1 = 0.9906, \hat{\beta}_2 = 0.8469, \hat{\beta}_3 = 0.3963, \hat{\sigma}^2 = 0.1705^2 = 0.0291.$$

Parameteren  $\alpha$  er den forventede styrke uden tilslætning af nogen af de tre komponenter. Parametrene  $\beta_1, \beta_2, \beta_3$  er den forventede ændring i styrke når  $A, B$  hhv.  $C$  tilslættes. Endelig er  $\sigma$  spredningen i fordelingen, dvs. udtryk for den "typiske" afvigelse fra middelværdien.

3. Den prædikterede værdi er

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_3 = 1.829.$$

Prædiktionsintervallet er givet i eksempel 10.31, og kan beregnes i R vha. funktionen predict:

```
newData <- data.frame(A=1, B=0, C=1)
predict(model1, newData, interval="p")
```

Vi får intervallet (1.474, 2.185). En ny observation med  $A$  og  $C$ , men ikke  $B$  tilslat vil med 95% sandsynlighed havne i dette interval.

4. Den interessante parameterfunktion er  $\delta = \beta_1 - \beta_2$ . Denne kan skrives som

$$\delta = (0 \ 1 \ -1 \ 0) \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \psi^T \gamma$$

hvor definitionen af  $\psi$  og  $\gamma$  fremgår af opskrivningen. Vi kan nu bruge eksempel 10.30 til at bestemme de relevante størrelser:

$$\hat{\delta} = \psi^T \hat{\gamma} = 0.1438$$

$$\text{Var}(\hat{\delta}) = \psi^T \text{Var}(\hat{\gamma}) \psi = 0.003632$$

$$\text{SE}(\hat{\delta}) = 0.0603.$$

Vi kan bruge `vcov` til at finde variansmatricen i R. Det tilhørende 95% konfidensinterval er

$$0.1438 \pm 2.007 \cdot 0.0603 = (0.0228, 0.2647)$$

hvor vi har benyttet at 97.5% fraktilen i  $t_{52}$  fordelingen er 2.007. Konfidensintervallet indeholder ikke nul, så der er tegn på at komponent  $A$  virker bedre end komponent  $B$ .

Alternativt kan vi teste hypotesen  $H : \beta_1 = \beta_2$ . Vi får

$$t = \frac{\hat{\delta}}{\text{SE}(\hat{\delta})} = 2.39$$

der skal vurderes in  $t_{52}$  fordelingen. Dette giver  $p$ -værdien 0.021, så hypotesen forkastes og konklusionen er (naturligvis) som følger.

5. Synergieffekten svarer til at middelværdien har formen

$$\text{E}Y_i = \alpha + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i + \varphi A_i B_i, \quad i = 1, \dots, 56$$

hvor det sidste led jo netop er 1 hvis både  $A$  og  $B$  er tilsat. Modellen fittes fx som følger, hvor vi først laver produktvariablen  $AB$ :

```
limData <- transform(limData, AB=A*B)
model2 <- lm(styrke ~ A + B + C + AB, data=limData)
```

Synergiparameteren estimeres til  $\hat{\varphi} = 0.0333$  med 95% konfidensinterval  $(-0.1660, 0.2326)$ . Hypotesen  $H : \varphi = 0$  kan testes med et  $t$ -test hvor man får  $t = 0.336$  og  $p = 0.74$ . Der er altså ikke evidens for synergি.

# Reeksamen i Statistik 2, 24. august 2017

Vejledende besvarelse

## Opgave 1

1. Variablene  $X_i$  er uafhængige og identisk Bernouilli-fordelte med  $P(X_i = 1) = (1/2)^\alpha$ . Det følger specialet af Den Centrale Grænseværdidisætning ... og EH eksempel ... , at  $\hat{\theta}_n \xrightarrow{as} \mathcal{N}((1/2)^\alpha, \frac{(1/2)^\alpha(1-(1/2)^\alpha)}{n})$ .
2. Benyttes Deltametoden med  $f(y) = -\frac{\log(y)}{\log(2)}$  fås, at  $\tilde{\alpha}_n = f(\hat{\theta}_n)$  er asymptotisk normalfordelt med middelværdi  $f((1/2)^\alpha) = \alpha$  og asymptotisk varians  $\frac{(1/2)^\alpha(1-(1/2)^\alpha)}{(-(1/2)^\alpha \log(2))^2}$ .
3. Tætheden for fordelingen af  $Y_1$  kan skrives på formen

$$f_\alpha(y) = \alpha y^{\alpha-1} = \frac{1}{1/\alpha} \exp(\alpha \cdot \log(y)) \cdot \frac{1}{y},$$

hvoraf det ses, at tætheden er en eksponentiel familie. Den kanoniske stikprøvefunktion er  $\theta = \frac{1}{\alpha}$  og normeringskonstanten er  $c(\theta) = \frac{1}{\theta}$ .

4. Likelihoodfunktion

$$\begin{aligned} L_{Y_1, \dots, Y_n}(\alpha) &= \prod_{i=1}^n \alpha Y_i^{\alpha-1} \\ &= \alpha^n \left( \prod_{i=1}^n Y_i^{\alpha-1} \right) \end{aligned}$$

(Minus) loglikelihoodfunktion

$$l_{Y_1, \dots, Y_n}(\alpha) = c - n \log(\alpha) - \alpha \cdot \sum_i \log(Y_i)$$

Scorefunktion

$$l'_{Y_1, \dots, Y_n}(\alpha) = -\frac{n}{\alpha} - \sum_i \log(Y_i)$$

5. Den observerede information

$$l''_{Y_1, \dots, Y_n}(\alpha) = \frac{n}{\alpha^2}$$

er strengt positiv, hvorfor en eventuel løsning til likelihoodligningen vil være et globalt minimum for  $l_{Y_1, \dots, Y_n}(\alpha)$ . Løses likelihoodligningen fås følgende udtryk for maksimaliseringsestimatoren  $\hat{\alpha}_n = -\frac{n}{\sum_i \log(Y_i)}$ .

Den asymptotiske fordeling af MLE kan bestemmes enten ved at kombinere Den Centrale Grænseværdidisætning (anvendt på  $\frac{\sum_i \log(Y_i)}{n}$ ) med Deltametoden eller vha. Cramér's sætning (EH: Sætning 5.23). Vi konkluderer, at  $\hat{\alpha}_n \xrightarrow{as} \mathcal{N}(\alpha, \frac{\alpha^2}{n})$ .

## Opgave 2

1. Den eneste model i R udskriften, der indeholder den relevante vekselvirkning beskriver  $X = (X_i)_{i \in I}$  som regulært normalfordelt på  $\mathbb{R}^I$  med middelværdi  $\xi \in L_T + L_{G \times V}$  og varians  $\sigma^2 I$ . Den additive hypotese,  $H_0 : \xi \in L_T + L_G + L_V$  kan testes ved  $F$ -teststørrelsen  $F = 0.7249$ , der under  $H_0$  følger en  $F$ -fordeling med (2,9)-frihedsgrader.  $P$ -værdien ses at være 0.5106, hvorfor vi accepterer nulhypotesen. Det er ikke et krav at model og hypotese opskrives, men angivelse af teststørrelse,  $P$ -værdi og konklusion anses som minimum for en fuldstændig besvarelse. Resultaterne er aflæst i R-udskriften efter `anova(mod2, mod1)`.
2. MLE for parametrene i middelværdistrukturen er givet ved formlen  $\hat{\beta} = (A^T A)^{-1} A^T X$ , hvor  $A$  er designmatricen for den valgte parametrisering af den additive model med alle tre faktorer (`mod2`). Det følger af EH korollar 10.21, at  $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (A^T A)^{-1})$ , hvor  $A^T A$  og dens inverse er anført i R-udskriften.

Der er i R-udskriften angivet 7 parameterestimater svarende til middelværdistrukturen. Det første estimat (=522.39) svarer til referencegruppen `tomat = 1`,  $G = G1$ ,  $V = V1$ , mens det øvrige 6 estimater angiver forskellen på det forventede udbytte i forhold til referencegruppen, hvis man ændrer en af de tre faktor `tomat`, `G` eller `V`.

3. Fordelingen af MLE for variansen er ifølge EH korollar 10.21 givet ved  $\hat{\sigma}^2 \sim \chi^2$  – fordelt med  $N - k = 18 - 7 = 11$  frihedsgrader og skalaparameter  $\sigma^2/N$ .
4. En designgraf viser, at der er ikke-trivielt minimum mellem faktorerne `gødning` (`G`) og `vanding` (`V`), som har to niveauer. Faktoren angiver, om plantekassen har modtaget en behandling eller ej. Ved at krydstabellere datasættet mht. de to faktorer ses, at hvor af de fremkomne diagonalblokke opfylder balanceingen. Dermed er de to faktorer geometrisk ortogonale.
5. Den simple løsning består i at anvende formel (13.3) i EH, hvor man bemærker, at der er to sammenhængskomponenter i designgrafen for de to faktorer.

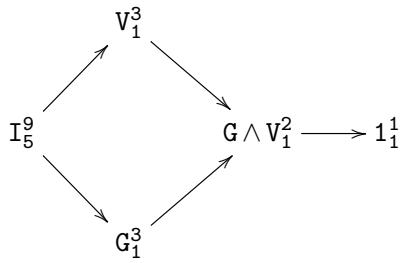
Alternativt kan bemærkes, at mængden

$$\mathbb{G} = \{V, G, V \wedge G, 1\}$$

udgør et geometrisk ortogonal design, som er afsluttet over for dannelse af minimum. Dimensionerne af  $V_G$ -rummene,  $G \in \mathbb{G}$ , fra sætningen om den ortogonale dekomposition (EH: Sætning 14.21) kan let beregnes, hvorefter vi finder, at

$$\begin{aligned} \dim(L_V + L_G) &= \dim(V_V) + \dim(V_G) + \dim(V_{V \wedge G}) + \dim(V_1) \\ &= 1 + 1 + 1 + 1 = 4. \end{aligned}$$

Det kan her være nyttigt at støtte sig op ad et faktorstrukturdiagram, for at holde styr på ordningen af faktorerne



## Opgave 3

1. Et faktorstrukturdiagram ser ud som følger

$$I_{56}^{70} \longrightarrow \text{person}_{12}^{14} \longrightarrow \text{morgen}_1^2 \longrightarrow 1_1^1$$

2. Modellen kan udtrykkes ved at  $X = (X_i)_{i \in I}$  er normalfordelt på  $\mathbb{R}^{70}$  med

$$\xi_i = EX_i = \alpha + \beta \cdot \text{puls}_i$$

og  $VX = \sigma^2 I + v^2 BB^T$ . Her er  $B$  effektmatricen hørende til effektparret (**person**, 1).

Da der i opgaveformuleringen blot lægges op til, at man skal modellere sammenhængen mellem **tid** og **puls**, så er det helt ok, hvis man ved besvarelsen af delopgave 2.-4. i stedet benytter **puls** som responsvariabel og **tid** som forklarende variabel. Bemærk dog, at det i delopgave 5. implicit fremgår, at det gennem hele opgaven er tanken, at **tid** skal benyttes som responsvariabel.

3. Parametrene i middelværdistrukturen fremgår af

```
data <- read.table("Stat2aug2017opg3.txt", header = T)
library(lme4)
```

```
head(data) ## NB: variablen 'lap' skal ikke benyttes!
```

```
##   morgen lap puls tid person
## 1     ja   1  143 445     P1
## 2     ja   2  156 431     P1
## 3     ja   3  156 428     P1
## 4     ja   4  165 383     P1
## 5     ja   5  163 401     P1
## 6     ja   1  154 429     P2

m1 <- lmer(tid ~ puls + (1|person), data = data)
summary(m1)$coefficients

##                   Estimate Std. Error t value
## (Intercept) 902.387649 26.5355840 34.00670
## puls        -3.058166  0.1592837 -19.19949
```

og disse estimeres til  $\hat{\alpha} = 902.4$  og  $\hat{\beta} = -3.058$ .

Variansparametrene estimeres til  $\hat{\sigma}^2 = 9.4572^2 = 89.4$  og  $\hat{V}^2 = 14.7736^2 = 218.3$  hvilket fremgår af udskriften

```
VarCorr(m1)
```

```
## Groups     Name      Std.Dev.
## person    (Intercept) 14.7736
## Residual           9.4572
```

4. Regressionsmodellen kan udvides ved at ændre middelværdistrukturen således at skæring og hældning tillades at afhænge af, om personen har løbet sine ture om morgenen eller om eftermiddagen svarende til

$$\xi_i = EX_i = \alpha(\text{morgen}_i) + \beta(\text{morgen}_i) \cdot \text{puls}_i.$$

Modellen kan testes direkte imod modellen fra delopgave 2., hvorved vi finder at

```
m3 <- lmer(tid ~ morgen + morgen:puls + (1|person) - 1, data = data)
anova(m1, m3)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## m1: tid ~ puls + (1 | person)
## m3: tid ~ morgen + morgen:puls + (1 | person) - 1
##   Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m1  4 555.27 564.27 -273.64    547.27
## m3  6 558.52 572.01 -273.26    546.52  0.7542      2      0.6858
```

Likelihoodratioteststørrelsen bliver  $LRT = 0.7542$  der ved et opslag i en tabel over  $\chi^2$ -fordelingen med 2 frihedsgrader giver et P-værdi på 0.6858. Det lader således ikke til, at der er forskel på sammenhængen mellem `tid` og `puls` for personer der løber om morgenen og senere på dagen.

**Testet kan alternativt** udføres i to trin, hvor man først tester om hældning og dernæst om skæringen kan antages at være ens for de to niveauer af faktoren `morgen`. R kode og resultater fremgår nedenfor (den overordnede konklusion ændres ikke!)

```
m2 <- lmer(tid ~ puls + (1|person) + morgen, data = data)
anova(m3, m2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
```

```

## m2: tid ~ puls + (1 | person) + morgen
## m3: tid ~ morgen + morgen:puls + (1 | person) - 1
##   Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m2 5 556.68 567.92 -273.34    546.68
## m3 6 558.52 572.01 -273.26    546.52 0.1586      1     0.6905

anova(m1, m2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## m1: tid ~ puls + (1 | person)
## m2: tid ~ puls + (1 | person) + morgen
##   Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m1 4 555.27 564.27 -273.64    547.27
## m2 5 556.68 567.92 -273.34    546.68 0.5956      1     0.4402

```

5. Kovariansmatricen for de 5 målinger af omgangstiden fra person P1 er givet ved

$$\begin{pmatrix} \sigma^2 + v^2 & v^2 & v^2 & v^2 & v^2 \\ v^2 & \sigma^2 + v^2 & v^2 & v^2 & v^2 \\ v^2 & v^2 & \sigma^2 + v^2 & v^2 & v^2 \\ v^2 & v^2 & v^2 & \sigma^2 + v^2 & v^2 \\ v^2 & v^2 & v^2 & v^2 & \sigma^2 + v^2 \end{pmatrix}$$

## Supplerende R kode til løsning af opgave 3

Følgende R kode kan benyttes til analyserne, hvis `puls` benyttes som responsvariablen i delopgave 2.-4.

```

l1 <- lmer(puls ~ tid + (1|person), data = data)
summary(l1)$coefficients

##                   Estimate Std. Error t value
## (Intercept) 274.8088601 5.85128704 46.96554
## tid         -0.2762089 0.01431843 -19.29045

VarCorr(l1)

## Groups   Name       Std.Dev.
## person   (Intercept) 4.5444
## Residual             2.8268

l3 <- lmer(puls ~ morgen + morgen:tid + (1|person) - 1, data = data)
anova(l1, l3)

```

```

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## l1: puls ~ tid + (1 | person)
## l3: puls ~ morgen + morgen:tid + (1 | person) - 1
##   Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## l1  4 386.95 395.94 -189.47   378.95
## l3  6 389.37 402.87 -188.69   377.37 1.5761      2     0.4547

l2 <- lmer(puls ~ tid + (1|person) + morgen, data = data)
anova(l3, l2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## l2: puls ~ tid + (1 | person) + morgen
## l3: puls ~ morgen + morgen:tid + (1 | person) - 1
##   Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## l2  5 387.92 399.16 -188.96   377.92
## l3  6 389.37 402.87 -188.69   377.37 0.5436      1     0.4609

anova(l1, l2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## l1: puls ~ tid + (1 | person)
## l2: puls ~ tid + (1 | person) + morgen
##   Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## l1  4 386.95 395.94 -189.47   378.95
## l2  5 387.92 399.16 -188.96   377.92 1.0325      1     0.3096

```

# Eksamens i Statistik 2, 22. juni 2017

Vejledende besvarelse

## Opgave 1

- Da  $E[1(Y_1 = 2)] = P(Y_1 = 2) = p_2$  følger af Store Tals Lov, at  $\hat{\theta}_n \xrightarrow{P} p_2$  for  $n \rightarrow \infty$ . Specielt er estimatoren konsistent for  $p_2$ . Da  $V[1(Y_1 = 2)] = p_2(1 - p_2)$  giver Den Centrale Grænseværdidisætning, at  $\hat{\theta}_n \xrightarrow{as} \mathcal{N}(p_2, \frac{p_2(1-p_2)}{n})$ .
- Fordelingen af  $Y$  har tæthed (mht. tællemålet på  $\{0, 1, 2\}$ ) givet ved

$$\begin{aligned} f_{\theta}(y) &= p_0^{1(y=0)} \cdot p_1^{1(y=1)} \cdot p_2^{1(y=2)} \\ &= (1 - p_1 - p_2)^{1-1(y=1)-1(y=2)} \cdot p_1^{1(y=1)} \cdot p_2^{1(y=2)} \\ &= \exp \left( \log \left( \frac{p_1}{1 - p_1 - p_2} \right) \cdot 1(y=1) + \log \left( \frac{p_2}{1 - p_1 - p_2} \right) \cdot 1(y=2) \right) \\ &\quad \times (1 - p_1 - p_2), \end{aligned}$$

Dette er en eksponentiel familie med kanonisk stikprøvefunktion  $t(y) = (1_{(y=1)}, 1_{(y=2)})^T$ , parameter  $\theta = (\theta_1, \theta_2)^T$  (se opgaveformuleringen) og normeringskonstant  $c(\theta) = \frac{1}{1-p_1-p_2} = \frac{1}{1+\exp(\theta_1)+\exp(\theta_2)}$ .

- Fra delopgave 1 ved vi, at  $\hat{\theta}_n$  er konsistent for  $p_2 = p^2$ . Det følger af EH s. 174 formel (5.3) og *Deltametoden* (EH: Sætning 5.15) med  $f(x) = \sqrt{x}$ , at  $f(\sqrt{\hat{\theta}_n})$  er konsistent (for  $p$ ) og asymptotisk normalfordelt med (asymptotisk) middelværdi  $p = f(p^2)$  og asymptotisk varians  $\frac{1}{n}f'(p^2)p^2(1-p^2)f'(p^2)$ . Den asymptotiske varians kan omskrives til  $\frac{1-p^2}{4n}$ .
- Likelihoodfunktion

$$\begin{aligned} L_{Y_1, \dots, Y_n}(p) &= \prod_{i=1}^n \binom{2}{Y_i} p^{Y_i} (1-p)^{2-Y_i} \\ &= \left( \prod_{i=1}^n \binom{2}{Y_i} \right) \cdot p^{\sum_i Y_i} (1-p)^{2n - \sum_i Y_i} \end{aligned}$$

(Minus) loglikelihoodfunktion

$$l_{Y_1, \dots, Y_n}(p) = c - \sum_i Y_i \cdot \log(p) - \log(1-p) \cdot (2n - \sum_i Y_i)$$

Scorefunktion

$$l'_{Y_1, \dots, Y_n}(p) = -\frac{\sum_i Y_i}{p} + \frac{2n - \sum_i Y_i}{1-p}$$

## 5. Den observerede information

$$l''_{Y_1, \dots, Y_n}(p) = \frac{\sum_i Y_i}{p^2} + \frac{2n - \sum_i Y_i}{(1-p)^2}$$

er strengt positiv, hvorfor en eventuel løsning til likelihoodligningen vil være et globalt minimum for  $l_{Y_1, \dots, Y_n}(p)$ . Løses likelihoodligningen fås følgende udtryk for maksimaliseringsestimatoren  $\hat{p}_n = \frac{\sum_i Y_i}{2n}$ .

**Bemærk:** Med de praktiske antagelser omkring problemstillingen er vi kun interesserede i parameterværdier i det åbne interval  $0 < p < 1$ . For  $\sum_i Y_i = 0$  eller  $\sum_i Y_i = 2n$  eksisterer MLE således ikke, men dette indtræffer med ssh 0 i grænsen, så MLE er asymptotisk veldefineret. Hvis man vælger at betragte likelihoodfunktionen over hele  $[0, 1]$  så eksisterer MLE altid, men ligger på randen for de to typer af *panikobservationer* indikeret ovenfor. Der gives fuldt point, selvom man ikke forholder sig til disse specialtilfælde.

Den asymptotiske fordeling af MLE kan bestemmes enten vha. Den Centrale Grænseværdisætning eller vha. Cramér's sætning (EH: Sætning 5.23). Da  $Y$  følger en binomialfordeling med antalsparameter 2 og sandsynlighedsparameter  $p$ , så er  $EY = 2p$  og den forventede information (for een observation!) kan fx. findes ved at indsætte i udtrykket fra 4. (svarende til  $n = 1$ )

$$E_p[l''_Y(p)] = \frac{2p}{p^2} + \frac{2-2p}{(1-p)^2} = 2\frac{1-p+p}{p(1-p)}.$$

Vi konkluderer, at  $\hat{p}_n \xrightarrow{as} \mathcal{N}(p, \frac{p(1-p)}{2n})$ .

**Kommentar:** Den asymptotiske varians på estimatoren fra delspørgsmål 3. er lig med  $\frac{1+p}{2p}$  gange den asymptotiske varians for MLE. Det ses, at denne faktor altid er  $\geq 1$ . Dette er en konsekvens af den asymptotiske optimalitet af MLE som diskuteret i EH kapitel 5.4.

## 6. Reparameteriseringsafbildningen fra $p$ til $B$ er givet ved afbildningen

$$B = \phi(p) = \frac{D}{p}$$

MLE i den alternative parametrisering bliver blot  $\hat{B}_n = \phi(\hat{p}_n)$  (se f.x. EH eksempel 15.19), så den asymptotiske fordeling kan bestemmes vha. Deltametoden. For at udtrykke den asymptotiske varians i  $B$ -parametriseringen er det nyttigt at bemærke, at  $p = \phi^{-1}(B) = \frac{D}{B}$ . Opgaven løses nu ved at omregne følgende udtryk til  $B$ -parametriseringen

$$D\phi(p) \frac{p(1-p)}{2n} D\phi(p).$$

**Kommentar:** Den asymptotiske varians er en aftagende funktion af diameteren  $D$ . Det giver mening at en papskive med kendt diameter meget tæt på  $B$  gør det muligt at bestemme bredden af plankerne med stor præcision. En naiv estimator (med varians 0!) ville være  $D$ , men denne er ikke central og bliver udkonkurreret af MLE i det lange løb. Der gælder et tilsvarende resultat, hvis man benytter sig af papbrikker med andre former blot sandsynligheden for at berøre to planker vokser lineært med diameteren af den mindste cirkulære skive, som kan indeholde papbrikken. Et udartet tilfælde fås ved at betragte uendelig tynde pinde (=tændstikformede papbrikker), hvor  $p = \frac{D}{B}(1/2 + 1/\pi)$ . Varianter af denne situation omtales som *Buffons nåleproblem*.

## Opgave 2

- Udgangsmodellen (=vekselvirkningsmodellen) udtrykker, at  $X = (X_i)_{i \in I}$  er regulært normalfordelt på  $\mathbb{R}^I$  med middelværdi  $\xi \in L_{G \times V}$  og varians  $\sigma^2 I$ . Den additive hypotese,  $H_0 : \xi \in L_G + L_V$  kan testes ved  $F$ -teststørrelsen  $F = 0.8124$ , der under  $H_0$  følger en  $F$ -fordeling med  $(1, 11)$ -frihedsgrader.  $P$ -værdien ses at være 0.3867, hvorfor vi accepterer nulhypotesen. Det er ikke et krav at model og hypotese opskrives, men angivelse af teststørrelse,  $P$ -værdi og konklusion anses som minimum for en fuldstændig besvarelse. Resultaterne er aflæst i R-udskriften efter `anova(mod2, mod1)`.
- Dimensionen af det additive underrum er her 3. Det er fint, blot at argumentere med, at der optræder 3 estimerer for middelværdistrukturen, når man laver et summary af modellen `mod2`. Det demonstrerer dog et større overblik, hvis man benytter den generelle formel

$$\dim(L_G + L_V) = \dim(L_G) + \dim(L_V) - \dim(L_{G \wedge V}) = 2 + 2 - 1,$$

hvor det benyttes at minimumsfaktoren  $G \wedge V$  er identisk med den konstante faktor. Den størrelse der ønskes beregnet i opgaven er  $F$ -teststørrelsen for test af den additive hypotese i tosidet variansanalyse med *geometrisk ortogonale faktorer*. Det konstateres, at teststørrelsen bliver 1.046 som *ikke* er lig med teststørrelsen fra R-udskriften. Dette skyldes, at faktorerne  $G$  og  $V$  i dette tilfælde ikke er geometrisk ortogonale, netop fordi observationen fra et af forsøgsplottene mangler.

- MLE for parametrene i middelværdistrukturen er givet ved formlen  $\hat{\beta} = (A^T A)^{-1} A^T X$ , hvor  $A$  er designmatricen for den valgte parametrisering af den additive model. Det følger af EH korollar 10.21, at  $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (A^T A)^{-1})$ , hvor  $A^T A$  og dens inverse er anført i R-udskriften. Vi har, at

$$\hat{\beta} = (9.777, -0.322, -2.280)^T.$$

Her angiver  $\beta_1$  middelværdien for kombinationen  $G = \text{høj}$ ,  $V = \text{II}$ , mens  $\beta_2$  angiver forskellen mellem  $V = \text{I}$  og  $V = \text{II}$ , og  $\beta_3$  angiver forskellen mellem  $G = \text{lav}$  og  $G = \text{høj}$ .

- Da hver kombination af godtning ( $G$ ) og vanding ( $V$ ) optræder i forsøgsplanen, og da sort ( $S$ ) optræder for hver vandret række (gødning) og for hver lodret søjle (vanding) konkluderes udmiddelbart, at

$$G \wedge V = G \wedge S = V \wedge S = 1.$$

Da hver sort optræder netop en gang inden for hver jordtype haves desuden  $J \wedge S = 1$ . Derimod viser antalstabellerne nedenfor, at  $G \wedge J$  og  $V \wedge J$  er (to forskellige) faktorer med to niveauer.

```
table(g, j)

##      j
##  g    I II III IV
##  G1  2  2   0   0
##  G2  2  2   0   0
##  G3  0  0   2   2
##  G4  0  0   2   2
```

```



```

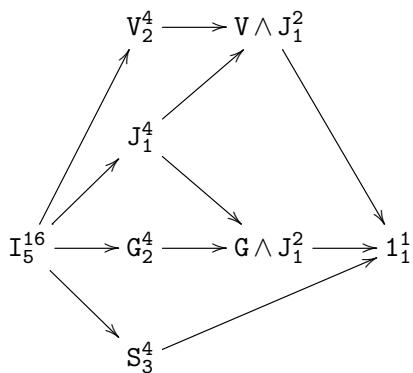
5. Tilføjes de to ikke-trivielle minima fra delspørøgsmål 4.konstateres, at mængden

$$\mathbb{G} = \{V, G, S, J, V \wedge J, G \wedge J, 1\}$$

udgør et geometrisk ortogonalt design, som er afsluttet over for dannelse af minimum. Dimensionerne af  $V_G$ -rummene,  $G \in \mathbb{G}$ , fra sætningen om den ortogonale dekomposition (EH: Sætning 14.21) kan let beregnes, hvorefter vi finder, at

$$\begin{aligned}
 \dim(L_V + L_G + L_J + L_S) &= \dim(V_V) + \dim(V_G) + \dim(V_J) \\
 &\quad + \dim(V_S) + \dim(V_{G \wedge J}) + \dim(V_{V \wedge J}) + \dim(V_1) \\
 &= 2 + 2 + 1 + 3 + 1 + 1 + 1 = 11.
 \end{aligned}$$

Det kan her være nyttigt at støtte sig op ad et faktorstrukturdiagram, for at holde styr på ordningen af faktorerne



En **alternativ løsningsmetode** består i at indtaste et fiktivt datasæt (fx. med simulerede målinger) svarende til det angivne forsøgsdesign. Fittes den additive model til dette datasæt i R, så vil antallet af parameterestimater i den additive model angive dimensionen af det ønskede additive underrum.

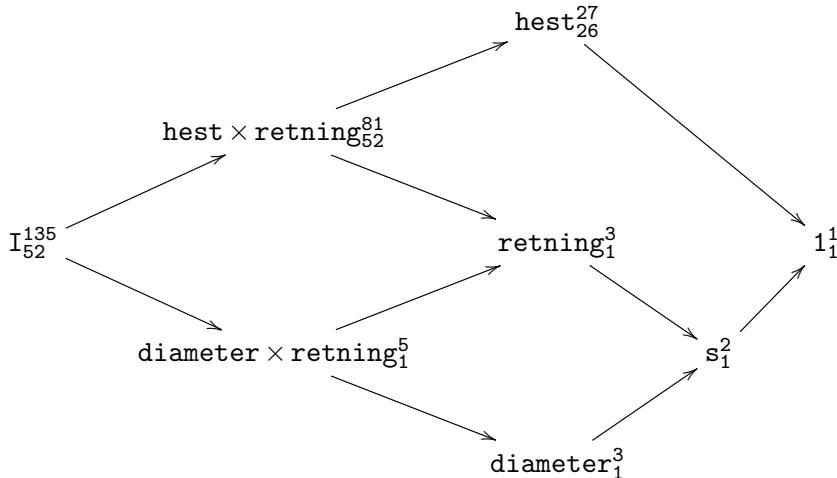
## Opgave 3

1. Antalstabellen for faktorerne **diameter** og **retning** viser, at minimum  $s = \text{diameter} \wedge \text{retning}$  er en faktor med 2 labels, hvor hver "blok" opfylder balancealigningen (EH: Sætning 14.8).

```
##      diameter
## retning 0 8 16
##       M 27 0 0
##       H 0 27 27
##       V 0 27 27
```

Dermed er disse to faktorer geometrisk ortogonale. Minimumsfaktoren angiver om symmetrimålingen stammer fra en måleserie, hvor hesten går lige ud eller bevæger sig i cirkler.

Produktfaktoren mellem **diameter** og **retning** er reelt kun en faktor med 5 labels. Da datasættet består af netop een måling fra hver hest for hvert niveau af produktfaktoren, så vil alle antalstabeller konstrueret ud fra faktorer i designet opfynde balancealigningen. Specielt er der tale om et geometrisk ortogonalt design og dimensionerne af underrummene som indgår i den ortogonale dekomposition fra EH sætning 14.21 kan beregnes rekursivt ved *håndkraft*.



2. Formålet med forsøget vedrører ikke de konkrete 27 heste der indgår i eksperimentet, så alle faktorer indeholdende **hest** bør indgå som tilfældige effekter. Der bliver således to oplagte tilfældige effekter svarende til effektparrene  $(\text{hest}, 1)$  og  $(\text{hest} \times \text{retning}, 1)$ . Lader vi  $B_1$  og  $B_2$  betegne effektmatricerne hørende til de to effektpar, så kan modellen udtrykkes ved at  $X = (X_i)_{i \in I}$  er normalfordelt på  $\mathbb{R}^{135}$  med  $\xi = EX \in L_{\text{diameter} \times \text{retning}}$  og  $VX = \sigma^2 I + v_1^2 B_1 B_1^T + v_2^2 B_2 B_2^T$ .

Der gives et fradrag, hvis man (uden nogen form for argumentation) vælger *ikke* at indrage en tilfældig effekt af **hest**  $\times$  **retning** i modellen.

3. Parameterestimatorerne for variansparametrene (baseret på REML-estimation) bliver

$$\hat{\sigma}^2 = 0.3632^2 \quad \hat{v}_1^2 = 0.2739^2 \quad \hat{v}_2^2 = 0.5166^2.$$

Den totale varians på en symmetriscore bliver  $\sigma^2 + v_1^2 + v_2^2$  og kovariansmatricen for de 5 målinger på en given hest kan udtrykkes som

$$\begin{pmatrix} \sigma^2 + v_1^2 + v_2^2 & v_1^2 & v_1^2 & v_1^2 & v_1^2 \\ v_1^2 & \sigma^2 + v_1^2 + v_2^2 & v_1^2 + v_2^2 & v_1^2 & v_1^2 \\ v_1^2 & v_1^2 + v_2^2 & \sigma^2 + v_1^2 + v_2^2 & v_1^2 & v_1^2 \\ v_1^2 & v_1^2 & v_1^2 & \sigma^2 + v_1^2 + v_2^2 & v_1^2 + v_2^2 \\ v_1^2 & v_1^2 & v_1^2 & v_1^2 + v_2^2 & \sigma^2 + v_1^2 + v_2^2 \end{pmatrix}$$

(her er målingerne organiseret inden for hver hest som anført i datasættet).

4. Test for reduktion i middelværdistrukturen foretages ved brug af likelihoodratioteststørrelsen. Vi benytter her en  $\chi^2$ -fordeling som approksimation ved beregning af p-værdien.

Vekselsvirkningen  $\text{diameter} \times \text{retning}$  ( $H_0 : \xi \in L_{\text{diameter}} + L_{\text{retning}}$ ,  $LRT = 0.9237$ ,  $p = 0.3365$ ) og hovedeffekten af  $\text{retning}$  ( $H_0 : \xi \in L_{\text{diameter}}$ ,  $LRT = 0.8605$ ,  $p = 0.3536$ ) lader ikke til at bidrage væsentligt til at forklare variationen i symmetriscorerne.

Yderligere reduktion af modellen svarende til hypotesen  $H_0 : \xi \in L_1$  forkastes ( $LRT = 75.266$ ,  $p < 0.0001$ ), hvilket indikerer at diameteren har betydning for symmetriscoren.

Baseret på faktorstrukturdiagrammet kan man vælge at indskyde hypotesen  $H_0 : \xi \in L_s$  om, at det kun betyder noget for symmetriscoren, om hesten løber ligeud eller i cirkler. Denne hypotese forkastes dog ( $LRT = 33.433$ ,  $p < 0.0001$ ).

5. Vi tilføjer en ny variable  $\text{invdiam}$  til datasættet, og tester hypotesen om at middelværdistrukturen er givet ved  $\xi_i = EX_i = \alpha + \beta \cdot \text{invdiam}_i$  mod modellen  $\xi \in L_{\text{diameter}}$ . Et likelihoodratiotest for denne hypotese giver teststørrelsen  $LRT = 5.166$  der ved opslag i en  $\chi^2$ -fordeling med 1 frihedsgrad svarer til en p-værdi på 0.023. Der er således (svag) evidens imod hypotesen om, at symmetriscoren afhænger lineært af den reciproke diameter.

## Eksempel på R-kode som kunne være brugt til løsning af opgave 3

```
acc_data <- read.table(file = "stat2juni2017opg3.txt", header = T)
head(acc_data)

##   hest retning diameter      S
## 1 G01     M      0 -6.257128
## 2 G01     H      8 -4.723692
## 3 G01     H     16 -5.000378
## 4 G01     V      8 -4.297942
## 5 G01     V     16 -4.589719
## 6 G02     M      0 -5.348065

table(acc_data$retning, acc_data$diameter)

##
##      0  8 16
## H  0 27 27
## M 27  0  0
## V  0 27 27

### tilføj minimum af retning og diameter
acc_data$cirkel <- acc_data$diameter != 0

library(lme4)
modelfull <- lmer(S ~ factor(diameter) * retning +
                      (1 | hest) + (1 | hest : retning)
                      , data = acc_data, REML = TRUE)
modelfull

## Linear mixed model fit by REML ['lmerMod']
## Formula: S ~ factor(diameter) * retning + (1 | hest) + (1 | hest:retning)
## Data: acc_data
## REML criterion at convergence: 247.835
## Random effects:
## Groups      Name      Std.Dev.
## hest:retning (Intercept) 0.5166
## hest        (Intercept) 0.2739
## Residual          0.3632
## Number of obs: 135, groups: hest:retning, 81; hest, 27
## Fixed Effects:
##             (Intercept)    factor(diameter)8
##                 -6.09429           1.24956
##             factor(diameter)16      retningV
##                 0.84203           0.07725
## factor(diameter)8:retningV
##                 0.13242
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 4 columns / coefficients
```

```

model0 <- lmer(S ~ factor(diameter) * retning +
                 (1 | hest) + (1 | hest : retning)
                 , data = acc_data, REML = FALSE)
model1 <- lmer(S ~ factor(diameter) + retning +
                 (1 | hest) + (1 | hest : retning)
                 , data = acc_data, REML = FALSE)
anova(model1, model0)

## Data: acc_data
## Models:
## model1: S ~ factor(diameter) + retning + (1 | hest) + (1 | hest:retning)
## model0: S ~ factor(diameter) * retning + (1 | hest) + (1 | hest:retning)
##          Df     AIC     BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model1  7 250.08 270.42 -118.04    236.08
## model0  8 251.16 274.40 -117.58    235.16 0.9237      1      0.3365

model2a <- lmer(S ~ retning + (1 | hest) + (1 | hest : retning)
                  , data = acc_data, REML = FALSE)
model2b <- lmer(S ~ factor(diameter) + (1 | hest) + (1 | hest : retning)
                  , data = acc_data, REML = FALSE)
anova(model2a, model1)

## Data: acc_data
## Models:
## model2a: S ~ retning + (1 | hest) + (1 | hest:retning)
## model1: S ~ factor(diameter) + retning + (1 | hest) + (1 | hest:retning)
##          Df     AIC     BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model2a  6 281.45 298.88 -134.72    269.45
## model1   7 250.08 270.42 -118.04    236.08 33.364      1 7.643e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(model2b, model1)

## Data: acc_data
## Models:
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
## model1: S ~ factor(diameter) + retning + (1 | hest) + (1 | hest:retning)
##          Df     AIC     BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model2b  6 248.94 266.38 -118.47    236.94
## model1   7 250.08 270.42 -118.04    236.08 0.8605      1      0.3536

```

```

model3 <- lmer(S ~ 1 + (1 | hest) + (1 | hest : retning)
                 , data = acc_data, REML = FALSE)
anova(model3, model2b)

## Data: acc_data
## Models:
## model3: S ~ 1 + (1 | hest) + (1 | hest:retning)
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
##          Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model3   4 320.21 331.83 -156.10    312.21
## model2b  6 248.94 266.38 -118.47    236.94 75.266      2 < 2.2e-16 ***
## ...
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model3a <- lmer(S ~ cirkel + (1 | hest) + (1 | hest : retning)
                  , data = acc_data, REML = FALSE)
anova(model3a, model2b)

## Data: acc_data
## Models:
## model3a: S ~ cirkel + (1 | hest) + (1 | hest:retning)
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
##          Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model3a  5 280.38 294.90 -135.19    270.38
## model2b  6 248.94 266.38 -118.47    236.94 33.433      1 7.375e-09 ***
## ...
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### genfit model med REML-estimation
model2bfinal <- lmer(S ~ factor(diameter) + (1 | hest : retning) +(1 | hest)
                      , data = acc_data, REML = TRUE)
VarCorr(model2bfinal)

## Groups       Name     Std.Dev.
## hest:retning (Intercept) 0.51557
## hest         (Intercept) 0.27462
## Residual           0.36286

confint(model2bfinal)

## Computing profile confidence intervals ...

##                   2.5 %     97.5 %
## .sig01            0.3851479  0.6615867
## .sig02            0.0000000  0.4706491
## .sigma            0.2999166  0.4413148
## (Intercept)      -6.3530665 -5.8355101
## factor(diameter)8  1.0615006  1.6472870
## factor(diameter)16 0.5877646  1.1735511

```

```

acc_data$invdiam <- 1/acc_data$diameter
acc_data$invdiam[acc_data$diameter == 0] <- 0
head(acc_data)

##   hest retning diameter      S cirkel invdiam
## 1 G01      M      0 -6.257128 FALSE  0.0000
## 2 G01      H      8 -4.723692 TRUE   0.1250
## 3 G01      H     16 -5.000378 TRUE   0.0625
## 4 G01      V      8 -4.297942 TRUE   0.1250
## 5 G01      V     16 -4.589719 TRUE   0.0625
## 6 G02      M      0 -5.348065 FALSE  0.0000

model3b <- lmer(S ~ invdiam + (1 | hest) + (1 | hest : retning)
                 , data = acc_data, REML = FALSE)
anova(model3b, model2b)

## Data: acc_data
## Models:
## model3b: S ~ invdiam + (1 | hest) + (1 | hest:retning)
## model2b: S ~ factor(diameter) + (1 | hest) + (1 | hest:retning)
##           Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## model3b  5 252.11 266.63 -121.06    242.11
## model2b  6 248.94 266.38 -118.47    236.94 5.1656      1  0.02304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

## Eksamens i Statistik 1, April 2018

Vejledende besvarelse udarbejdet af Steffen Lauritzen og Helle Sørensen

### Opgave 1

Ved opsendelsen af rumfærgen Challenger den 28. januar 1986 omkom syv astronauter i forbindelse med en ekspllosion. En undersøgelseskommission blev nedsat og fastslog at årsagen til ulykken var en såkaldt O-ring, som gik i stykker i forbindelse med opsendelsen, sandsynligvis forårsaget af ekstremt koldt vejr. Filen `challenger.txt` indeholder data fra 23 tidligere opsendelser af rumfærgen. Kolonnen `temp` angiver temperaturen i Fahrenheit ved opsendelsen og kolonnen `critical` angiver om man i forbindelse med opsendelsen har observeret en kritisk tilstand for en af rumfærgens seks O-ringe, idet 1 angiver at man har observeret et problem og 0 at man ikke har.

1. Opstil en passende generaliseret lineær model til beskrivelse af afhængigheden mellem opsendelsestemperaturen og en kritisk begivenhed og angiv maximum likelihood estimatet for de indgående parametre.

Lad  $Y_1, \dots, Y_{23}$  være indbyrdes uafhængige og Bernoullifordelte med parametre  $\mu_i$ , hvor

$$\eta_i = \text{logit}(\mu_i) = \log \frac{\mu_i}{1 - \mu_i} = \alpha + \beta t_i$$

hvor  $t_i$  angiver temperaturen ved den  $i$ -te opsendelse og  $Y_i$  indikerer om der har været en kritisk hændelse i forbindelse med samme opsendelse.

Dette er en generaliseret lineær model med binomial fejlfordeling, kanonisk link, og opsendelsestemperaturen som kovariat. Den specificeres som følger:

```
> logreg <- glm(critical ~ temp, family="binomial", data=challenger)
```

og estimatorer for parametrerne  $(\alpha, \beta)$  fås via kommandoen

```
> summary(logreg)
```

```
...
```

Coefficients:

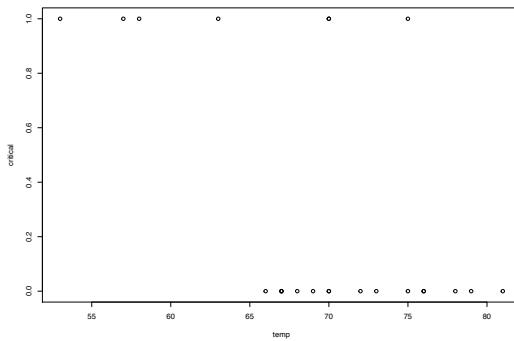
	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	15.0429	7.3786	2.039	0.0415 *
temp	-0.2322	0.1082	-2.145	0.0320 *

hvilket giver estimatorne  $(\hat{\alpha}, \hat{\beta}) = (15.0439, -0.2322)$ .

- Brug modellen og de angivne data til at afgøre om opsendelsestemperaturen har betydning for sandsynligheden for en kritisk hændelse.

Et asymptotisk test for temperaturafhængigheden giver iflg. ovenstående output en  $p$ -værdi på 0.032 hvilket er signifikant på 5% niveau; så man må konkludere at opsendelsestemperaturen har en betydning for forekomst af kritiske hændelser.

Et plot af data giver samme konklusion idet der aldrig er set kritiske hændelser ved temperaturer over 75 grader og altid kritiske hændelser ved temperaturer under 65 grader.



- Da rumfærgen blev sendt op var temperaturen 31 grader Fahrenheit. Estimer sandsynligheden for en kritisk hændelse med en O-ring ved denne temperatur.

Her fås

$$\hat{\mu}(31) = \frac{e^{\hat{\alpha}+31\hat{\beta}}}{1+e^{\hat{\alpha}+31\hat{\beta}}} = 0.9996$$

så hvis denne voldsomme ekstrapolation ellers står til troende, ville man med stor sikkerhed forvente en kritisk hændelse.

- Find et approksimativt 95% konfidensinterval for den samme sandsynlighed som estimeret under punkt 3. *Vink:* Brug for eksempel deltametoden på funktionen  $f(\alpha, \beta) = \alpha + 31\beta$ .

Den approximative kovariansmatrix for estimaterne kan fås ved kommandoen

```
> vcov(logreg)
            (Intercept)      temp
(Intercept) 54.4441826 -0.79638547
temp        -0.7963855  0.01171512
```

Ved deltametoden fås dernæst

$$\mathbf{V}\{f(\hat{\alpha}, \hat{\beta})\} \approx \mathbf{V}(\hat{\alpha}) + 31^2 \mathbf{V}(\hat{\beta}) + 62 \mathbf{V}(\hat{\alpha}, \hat{\beta}) = 16.3265$$

og videre fås nu et approksimativt 95% konfidensinterval for den lineære prediktor  $f(\alpha, \beta)$

$$\hat{\alpha} + 31\hat{\beta} \pm 1.96\sqrt{16.3265} = (-0.07, 15.8)$$

og dermed et approximativt 95% konfidensinterval for den ønskede sandsynlighed til

$$\left( \frac{e^{-0.07}}{1+e^{-0.07}}, \frac{e^{15.8}}{1+e^{15.8}} \right) = (0.482, 1).$$

Bemærk, at usikkerheden på denne interpolation er ganske stor, så selvom man estimerer sandsynligheden til at være tæt på 100%, kunne den også være så lille som 48%. Men selv den nedre grænse er stor nok til, at man klart vil fraråde opsendelse.

## Opgave 2

Paretofordelingen anvendes for eksempel til at beskrive fordelingen af formuer over en givet tærskelværdi  $c$  og den har tæthedsfunktion

$$f_\theta^c(x) = \frac{\theta c^\theta}{x^{\theta+1}}, \quad \text{for } x > c,$$

hvor  $c > 0$  er fast og kendt mens  $\theta > 0$  er en ukendt parameter som kaldes fordelingens *index*.

1. Gør rede for, at familien af Paretofordelinger med fast tærskel  $c$  udgør en eksponentiel familie; angiv familiens grundmål.

Vi omskriver tæthedsfunktionen således

$$f_\theta(x) = \frac{\theta c^\theta}{x^{\theta+1}} = \theta c^\theta e^{-\theta \log x} \frac{1}{x}, \quad x > c$$

hvorefter vi genkender den eksponentielle form med grundmål  $\mu = \mathbf{1}_{(c,\infty)}(x)/x \cdot \lambda$ , hvor  $\lambda$  er Lebesguemålet.

2. Angiv familiens dimension, den kanoniske parameter, den kanoniske stikprøvefunktion, og kumulantfunktionen.

Familien har dimension 1, den kanoniske parameter er  $\theta$  og den kanoniske stikprøvefunktion er  $t(x) = -\log x$ ; kumulantfunktionen er

$$\psi(\theta) = -\log(\theta c^\theta) = -\log \theta - \theta \log c.$$

Alternativt kan man skrive

$$f_\theta(x) = \theta e^{-\theta \log(x/c)} \frac{1}{x}, \quad x > c$$

hvor nu den kanoniske stikprøvefunktion er  $-\log(x/c)$  og kumulantfunktionen bliver  $\psi^*(\theta) = -\log \theta$ .

Endnu et alternativ er at skrive

$$f_\theta(x) = \frac{\theta c^\theta}{x^{\theta+1}} = \theta c^\theta e^{-(\theta+1)\log x}, \quad x > c$$

med kanonisk parameter  $\theta + 1$  og grundmål Lebesguemålet på den positive akse.

3. Lad nu  $X_1, \dots, X_n$  være uafhængige og Paretofordelte med kendt tærskel  $c$  og ukendt index  $\theta$ . Find maximum likelihood estimatoren for  $\theta$  og angiv dens asymptotiske fordeling.

Vi finder først middelværdifunktionen

$$\tau(\theta) = \mathbf{E}_\theta(-\log X) = \psi'(\theta) = -1/\theta - \log c$$

hvilket fører til likelihood ligningen

$$\sum_{i=1}^n \log x_i = \frac{n}{\theta} + n \log c$$

som har den entydige løsning

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n \{\log(x_i) - \log c\}}.$$

$\hat{\theta}$  er asymptotisk normalfordelt med den reciprokke Fisher information som asymptotisk varians. Vi får

$$\kappa(\theta) = i(\theta) = \psi''(\theta) = \tau'(\theta) = \theta^{-2}$$

og dermed er

$$\hat{\theta} \stackrel{\text{as}}{\sim} N(\theta, \theta^2/n).$$

Tidsskriftet *Forbes magazine* angiver hvert år en liste over verdens største personlige formuer. Nedenfor ses for 2018 størrelsen af alle personlige formuer over 50 milliarder US dollars som angivet af Forbes magazine.

År	Formue i milliarder US dollars								
2018	112 90 84 72 71 70 67 60 58.5								

4. Under antagelse af at disse observationer følger en Paretofordeling med tærskel  $c = 50$  og ukendt indexparameter, ønskes værdien af maximum likelihood estimatoren  $\hat{\theta}$  samt et approksimativt 95% konfidensinterval for indexparameteren  $\theta$ .

Maximum likelihood estimatet beregnes her til  $\hat{\theta} = 2.502$  og den tilsvarende approximative standardafvigelse til  $\sqrt{\hat{\theta}^2/9} = \hat{\theta}/3 = 0.834$ .

Baseret på den asymptotiske fordeling får vi så konfidensintervallet

$$\hat{\theta} \pm 1.96 \cdot \hat{\theta}/3 = (0.867, 4.137).$$

### Opgave 3

Lad  $X_1, \dots, X_n$  være uafhængige og identisk gammafordelte med med samme skala- og formparameter, d.v.s. deres fordeling har tæthed

$$f_\theta(x) = \frac{x^{\theta-1} e^{-x/\theta}}{\Gamma(\theta)\theta^\theta}, \quad x > 0$$

hvor  $\theta > 0$  er ukendt.

I det følgende indgår *digamma*funktionen  $\psi$  og *trigamma*funktionen  $\psi'$ , hvor

$$\psi(y) = D \log \Gamma(y) = \frac{\Gamma'(y)}{\Gamma(y)}, \quad \psi'(y) = D^2 \log \Gamma(y).$$

Begge funktioner er implementeret som standard i R og kaldes som `digamma()` og `trigamma()`.

Antag nu, at der foreligger en observation  $x = (x_1, \dots, x_n)$ .

1. Bestem log-likelihoodfunktionen og scorefunktionen.

Vi får, idet vi ignorerer led som kun afhænger af observationerne

$$\ell_x(\theta) = -\log L_x(\theta) = n \log \Gamma(\theta) + n\theta \log \theta - \theta \sum_i \log x_i + \sum_i x_i / \theta$$

og for scorefunktionen ved differentiation

$$S(x, \theta) = n\psi(\theta) + n \log \theta + n - \sum_i \log x_i - \sum_i x_i / \theta^2.$$

2. Bestem informationsfunktionen og Fisherinformationen.

Ved yderligere differentiation fås

$$I(x, \theta) = n\psi'(\theta) + n/\theta + 2 \sum_i x_i / \theta^3. \quad (1)$$

I en gammafordeling  $\Gamma(\alpha, \beta)$  hvor  $\beta$  er skalaparameter, er middelværdien  $\mathbf{E}(X) = \alpha\beta$  så i vores tilfælde er middelværdien  $\mathbf{E}_\theta(X) = \theta^2$ . Heraf følger at informationsfunktionen i tilfældet  $n = 1$  er givet som

$$i(\theta) = \psi'(\theta) + 1/\theta + 2/\theta = \psi'(\theta) + 3/\theta.$$

3. Scoreligningen kan ikke løses explicit. Vis, at scoreligningen har en entydig løsning og at denne løsning  $\hat{\theta}$  er maximum likelihood estimator for  $\theta$ ; det kan uden bevis benyttes, at

$$\psi'(y) = \sum_{k=0}^{\infty} \frac{1}{(y+k)^2}.$$

Idet  $\psi'(\theta) > 0$  og  $x_i > 0$  giver (??) at scorefunktionen er strengt voksende; da vi for alle  $x$  har at

$$\lim_{\theta \rightarrow 0} S(x, \theta) = -\infty, \quad \lim_{\theta \rightarrow \infty} S(x, \theta) = \infty$$

har scoreligningen præcis en løsning.

4. Angiv maximum likelihood estimatorens asymptotiske fordeling.

Maksimum likelihood estimatoren er asymptotisk normalfordelt med den reciproke information som varians, altså er

$$\hat{\theta} \stackrel{\text{as}}{\sim} N\left(\theta, \frac{1}{n(\psi'(\theta) + 3/\theta)}\right).$$

5. En alternativ estimator for  $\theta$  er givet som

$$\tilde{\theta} = \sqrt{\frac{\sum_i x_i}{n}} = \sqrt{\bar{x}}.$$

Gør rede for, at denne estimator er en momentestimator; er estimatoren central?

Estimatoren er en momentestimator i og med at

$$\mathbf{E}_\theta \left( \frac{\sum_i X_i}{n} \right) = \theta^2$$

så estimatoren  $\tilde{\theta}$  ovenfor er bestemt ved at løse denne ligning m.h.t.  $\theta$ ; estimatoren er ikke central idet

$$\mathbf{E}_\theta(\tilde{\theta}) = \mathbf{E}_\theta(\sqrt{\bar{X}}) < \sqrt{\mathbf{E}_\theta(\bar{X})} = \theta.$$

ifølge Jensens ulighed.

Nedenfor er angivet et eksempel på en stikprøve som anført ovenfor med  $n = 10$ :

```
> x
[1] 0.69 3.64 0.96 1.28 1.79 0.12 4.82 0.68 0.55 0.52
```

Værdien af maksimum likelihood estimatoren for  $\theta$  baseret på den angivne stikprøve kan beregnes til  $\hat{\theta} = 1.2278$ .

6. Betragt nu hypotesen  $H_0 : \theta = 1$ , svarende til, at observationerne stammer fra en eksponentialfordeling; beregn en approksimativ  $p$ -værdi for likelihood ratio testet for den pågældende hypotese. Kunne observationerne være eksponentialfordelte?

Log-likelihoodfunktion i minimumspunktet beregnes til 13.96 og i punktet 1 fås værdien 15.05, så kvotientteststørrelsen bliver

$$LR = 2(15.05 - 13.96) = 2.187$$

og baseret på  $\chi^2$  fordelingen med 1 frihedsgrad fås en  $p$ -værdi på 0.14; man kan altså ikke afvise at tallene stammer fra en eksponentialfordeling.

## Opgave 4

Lad  $(X, Y)$  være normalfordelt på  $\mathbb{R}^2$  med middelværdi 0 og varians  $\Sigma$ , altså  $(X, Y) \sim N(0, \Sigma)$ , hvor

$$\Sigma = \begin{pmatrix} \alpha & \frac{\alpha}{2} \\ \frac{\alpha}{2} & \alpha \end{pmatrix}.$$

Her er  $\alpha$  en konstant der opfylder visse betingelser, se spørgsmål 1.

- For hvilke værdier af  $\alpha$  er  $\Sigma$  en lovlig variansmatrix? For hvilke værdier af  $\alpha$  er fordelingen af  $X$  en regulær hhv. singulær normalfordeling?

$\Sigma$  er en lovlig variansmatrix hvis og kun hvis den er positiv semidefinit, dvs. hvis og kun hvis

- $\Sigma_{11} \geq 0$ , dvs.  $\alpha \geq 0$
- $\det(\Sigma) \geq 0$  dvs.  $\frac{3}{4}\alpha^2 \geq 0$ .

Tilsammen får vi altså at  $\Sigma$  en lovlig variansmatrix hvis og kun hvis  $\alpha \geq 0$ .

Fordelingen er regulær hvis og kun hvis  $\Sigma$  er invertibel. Da  $\det(\Sigma) = \frac{3}{4}\alpha^2$ , har vi altså at fordelingen er regulær for  $\alpha > 0$  og singulær for  $\alpha = 0$ .

- Bestem fordelingen af  $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix}$ , og vis derved at  $X+Y$  og  $X-Y$  er uafhængige,  $X+Y \sim N(0, 3\alpha)$  og  $X-Y \sim N(0, \alpha)$ .

Hvis vi definerer

$$C = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

og bruger transformationssætningen for normalfordelingen, får vi at

$$\begin{pmatrix} X+Y \\ X-Y \end{pmatrix} = C \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, C\Sigma C^T).$$

Variansmatricen viser sig at være

$$C\Sigma C^T = \begin{pmatrix} 3\alpha & 0 \\ 0 & \alpha \end{pmatrix}.$$

Det følger umiddelbart at  $X+Y$  og  $X-Y$  er uafhængige da elementet udenfor diagonalen er nul, og marginalfordelingerne aflæses også direkte:  $X+Y \sim N(0, 3\alpha)$  henholdsvis  $X-Y \sim N(0, \alpha)$ .

I det følgende kan du uden bevis benytte omskrivningen

$$XY = \frac{1}{4}(X+Y)^2 - \frac{1}{4}(X-Y)^2$$

samt at  $E(Z^4) = 3\sigma^4$  hvis  $Z \sim N(0, \sigma^2)$ .

- Definer estimatoren

$$\hat{\alpha} = 2XY$$

Vis at  $\hat{\alpha}$  er en central estimator for  $\alpha$  og bestem variansen  $V(\hat{\alpha})$ . Gør desuden rede for at  $P(\hat{\alpha} \geq 0) < 1$  hvis  $\alpha > 0$ .

Vi definerer estimatoren  $\hat{\alpha} = 2XY$ , der er central fordi

$$\mathbf{E}\hat{\alpha} = 2\mathbf{E}XY = 2\text{Cov}(X, Y) = 2\frac{\alpha}{2} = \alpha$$

Ved hjælp af det første vink of uafhængigheden af  $X + Y$  og  $X - Y$ , får vi variansen:

$$\mathbf{V}(XY) = \frac{1}{16}\mathbf{V}((X+Y)^2) + \frac{1}{16}\mathbf{V}((X-Y)^2)$$

Det andet vink giver at  $\mathbf{V}(Z) = 3\sigma^4 - \sigma^4 = 2\sigma^4$  hvis  $Z \sim N(0, \sigma^2)$ , så

$$\mathbf{V}(XY) = \frac{1}{16}2(3\alpha)^2 + \frac{1}{16}2\alpha^2 = \frac{5}{4}\alpha^2$$

og dermed er

$$\mathbf{V}(\hat{\alpha}) = 4\mathbf{V}(XY) = 5\alpha^2.$$

Hvis  $\alpha > 0$ , så er fordelingen af  $(X, Y)$  regulær på  $\mathbb{R}^2$ , således at  $(X, Y)$  „lever på“ hele  $\mathbb{R}^2$ , og der er derfor positiv sandsynlighed for at havne i anden eller fjerde kvadrant hvor  $XY < 0$ . Altså er  $P(\hat{\alpha} > 0) < 1$ , og vi kan altså risikere at få et estimat udenfor parametermængden.

#### 4. Definer estimatoren

$$\tilde{\alpha} = \frac{1}{2}(X^2 + Y^2)$$

Vis at  $\tilde{\alpha}$  er en central estimator for  $\alpha$  og at variansen er  $\mathbf{V}(\tilde{\alpha}) = 5\alpha^2/4$ . Diskuter kortfattet om du foretrækker  $\tilde{\alpha}$  eller  $\hat{\alpha}$  som estimator for  $\alpha$ .

Vi definerer en ny estimator,  $\tilde{\alpha} = \frac{1}{2}(X^2 + Y^2)$ . Estimatoren er central da

$$\mathbf{E}\tilde{\alpha} = \frac{1}{2}(\alpha + \alpha) = \alpha$$

Andetmomentet er

$$\begin{aligned}\mathbf{E}\tilde{\alpha}^2 &= \frac{1}{4}\mathbf{E}(X^4 + Y^4 + 2X^2Y^2) \\ &= \frac{1}{4}(3\alpha^2 + 3\alpha^2 + 2\mathbf{E}(X^2Y^2)) \\ &= \frac{9}{4}\alpha^2\end{aligned}$$

hvor det til sidst er benyttet at

$$\mathbf{E}(X^2Y^2) = \mathbf{V}(XY) + (\mathbf{E}XY)^2 = \frac{5}{4}\alpha^2 + \frac{1}{4}\alpha^2 = \frac{3}{2}\alpha^2$$

Således er, som ønsket,

$$\mathbf{V}\tilde{\alpha} = \frac{9}{4}\alpha^2 - \alpha^2 = \frac{5}{4}\alpha^2.$$

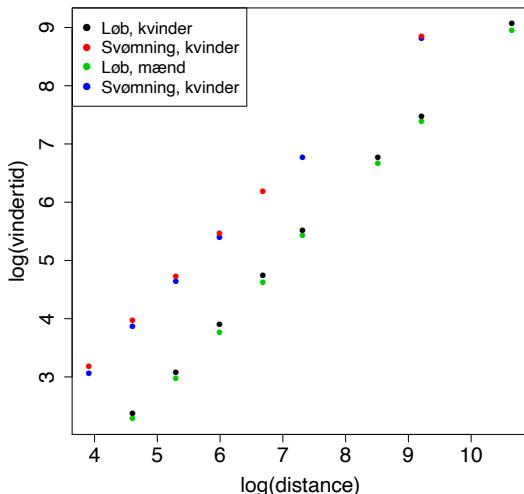
Estimatorerne  $\hat{\alpha}$  og  $\tilde{\alpha}$  er begge centrale, og  $\tilde{\alpha}$  har mindre varians end  $\hat{\alpha}$  (pånær i det uinteressante tilfælde hvor  $\alpha = 0$ ). Desuden er  $P(\tilde{\alpha} \geq 0) = 1$ , så  $\tilde{\alpha}$  rammer altid parameterområdet hvilket ikke gælder for  $\hat{\alpha}$ . Estimatoren  $\tilde{\alpha}$  er af disse grunde klart at foretrække.

## Opgave 5

Data til denne opgave består af vindertiderne i løbe- og svømmedisiplinerne ved OL i Rio de Janeiro, 2016. Data er tilgængelige i filen `rio2016.txt` på den vedlagte USB-nøgle. Der er følgende variable:

- køen: Køn, med værdien 0 for mænd og 1 for kvinder
- type: Typen af disciplin, med værdien 0 for løb og 1 for svømning
- distance: Distancen for disciplinen
- vindertid: Vindertiden i den pågældende disciplin, målt i sekunder

Figuren nedenfor viser data, hvor både distance og vindertid er log-transformeret og punkterne er farvet efter kombinationen af køn og typen af disciplin.



Du skal først betragte den multiple regressionsmodel hvor `log(vindertid)` benyttes som responsvariabel og `log(distance)`, type og køen benyttes som forklarende variable. Hvis data er indlæst i R som `rio2016`, så kan modellen fittes med kommandoen

```
reg <- lm(log(vindertid) ~ type + køen + log(distance), data=rio2016)
```

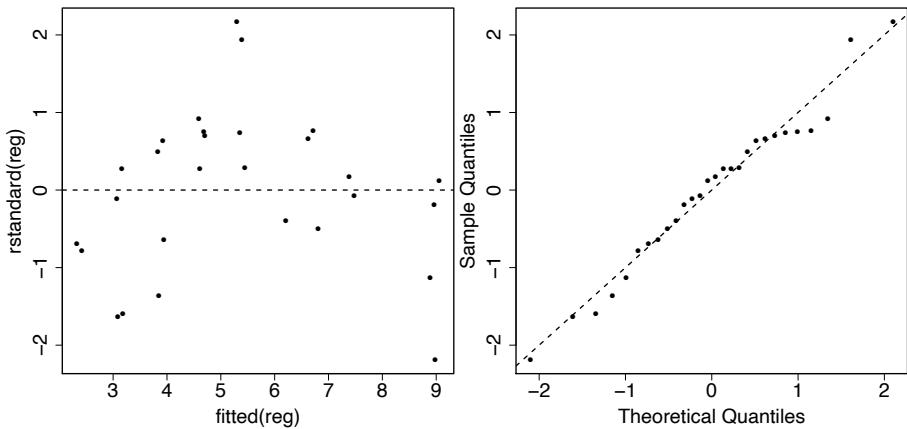
1. Opskriv den statistiske model svarende til `reg` (med papir og blyant). Fit modellen i R, og udfør modelkontrol. Svaret vedrørende modelkontrol skal indeholde skitser af relevante figurer og kommentarer til figurerne.

Den statistiske model antager at observationen  $x = (x_1, \dots, x_{28})$  af log-vindertiderne er udfald af en stokastisk variabel  $X \sim N(\xi, \sigma^2 I)$  hvor  $\xi \in L$  og  $\sigma^2 > 0$  er ukendte parametre  $\xi$  har formen

$$\xi_i = \beta_0 + \beta_1 \cdot \mathbf{1}_{\text{type}=\text{svømning}} + \beta_2 \cdot \mathbf{1}_{\text{køen}=\text{kvinde}} + \beta_3 \cdot \log(\text{distance})$$

hvor  $\mathbf{1}$  er indikatorfunktioner.

Residualplot og QQ-plot for de standardiserede residualer er vist nedenfor:



QQ-plottet ser fint ud eftersom punkterne ligger omkring 0/1-linien. Residualplottet ser derimod ikke helt godt ud: Der er en tendens til at punkterne udgør en „sur“ parabel svarende til at modellen overvurderer tiden på langsomme og hurtige discipliner og undervurderer tiden mellemhurtige discipliner.

Hvis man kigger nøjere efter, viser det sig at de to datapunkter der ligger lidt for sig selv, med store residualer, viser sig at svare til 1500 m løb for mænd og kvinder. Her er de observerede vindertider altså væsentligt større end forventet udfra modellen, hvilket indikerer at disse distancer er hårdere end de øvrige.

Uanset hvad du konkluderede vedrørende modelkontrol i spørgsmål 1, så skal du fortsætte med modellen i spørgsmål 2–4.

Husk desuden at både `distance` og `vindertid` indgår log-transformeret i modellen.

2. Kvinderne svømmer 800 m og mændene svømmer 1500 m, men ikke omvendt. Brug modellen til at bestemme et estimat for vindertiden for kvinder på 1500 m og et estimat for vindertiden for mænd på 800 m (hvis disse discipliner fandtes).

Middelværdien af log-vindertiden for 1500 m svømning for kvinder er

$$\xi' = \beta_0 + \beta_1 + \beta_2 + \beta_3 \cdot \log(1500).$$

Hvis vi indsætter parameterestimaterne, så fås estimatet for  $\xi'$ :

$$\hat{\xi}' = -2.732 + 1.506 + 0.094 + 1.098 \cdot \log(1500) = 6.899$$

Dette er på log-skala, så et estimat for vindertiden er  $\exp(6.899) = 991.2$  sekunder. Dette svarer til 16 min, 31 sek.

Middelværdien af log-vindertiden for 800 m svømning for mænd er

$$\xi'' = \beta_0 + \beta_1 + \beta_3 \cdot \log(800).$$

der estimeres til  $\hat{\xi}'' = 6.114756$ . Dette giver estimatet 452.5 sekunder, eller 7 min, 32 sek.

3. Betragt to distancer hvor den ene er dobbelt så lang som den anden. Gør rede for at modellen antager at den forventede forskel i  $\log(\text{vindertid})$  er den samme for alle fire kombinationer af køn og disciplintype, og bestem et estimat for den fælles forventede

forskel. Bestem derefter et estimat for den faktor, som vindertiden forøges med, når distancen fordobles (uanset køn og disciplintype).

Betruger distancerne  $d$  og  $2d$ . Uanset køn og disciplintype er forskellen i forventet log-vindertid mellem de to distancer lig

$$\delta = \beta_3 \log(2d) - \beta_3 \log(d) = \beta_3 \log(2).$$

Dette skyldes at de øvrige led i modellen enten optræder eller ikke optræder for begge distancer, og således udgår når der vi betragter forskellen.

Forskellen i forventet log-vindertid estimeres derfor til

$$\hat{\delta} = \hat{\beta}_3 \log(2) = 1.098 \cdot \log(2) = 0.761$$

Dette svarer til at vindertiden (ikke log-transformeret) øges med en faktor  $2^{\hat{\beta}_2} = \exp(0.761) = 2.14$  når distancen fordobles.

4. Gør rede for at modellen antager at den forventede forskel i `log(vindertid)` mellem svømning og løb er den samme for alle distancer og begge køn, og bestem et estimat for den fælles forventede forskel. Bestem derefter et estimat for den faktor, som vindertiden er længere ved svømning end ved løb (uanset køn og distance).

Forskellen i forventet log-vindertid mellem svømning og løb er, for alle distancer og begge køn, lig parameteren  $\beta_1$ . Dette skyldes at de øvrige led indgår i begge forventede værdier og dermed går ud når der tages differens.

Forskellen i forventet log-vindertid estimeres derfor til  $\hat{\beta}_1 = 1.506$ , hvilket svarer til at vindertiden (ikke log-transformeret) er en faktor  $\exp(\hat{\beta}_1) = \exp(1.506) = 4.51$  større for svømning end for løb.

## Eksamens i Statistik 1

### 28. juni 2018

Eksamens varer 4 timer. Alle hjælpemidler er tilladt under hele eksamen, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant. Eksamenssættet består af tre opgaver med i alt 14 delspørgsmål; alle delspørgsmål vægtes ens i bedømmelsen. Data til Opgave 3 ligger på en USB-nøgle. Nøglen skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den skal altså ikke indgå som del af besvarelsen.

## Opgave 1

Betræt  $X$  og  $Y$  uafhængige og eksponentiaffordelte med  $\mathbf{E}(X) = \lambda$  og  $\mathbf{E}(Y) = 3\lambda$ . Betragt følgende to estimatorer af  $\lambda$ :

$$\hat{\lambda} = (3X + Y)/6, \quad \tilde{\lambda} = \sqrt{XY/3}.$$

I det følgende kan det benyttes uden bevis at  $\Gamma(1.5) = \Gamma(0.5)/2 = \sqrt{\pi}/2$ , hvor

$$\Gamma(y) = \int_0^\infty u^{y-1} e^{-u} du$$

er gammafunktionen.

1. Hvilke af disse estimatorer er centrale for  $\lambda$ ?

Den første er central for  $\lambda$  mens vi for  $\tilde{\lambda}$  har

$$\mathbf{E}(\tilde{\lambda}) = \mathbf{E}(\sqrt{X})\mathbf{E}(\sqrt{Y})/\sqrt{3}$$

og

$$\mathbf{E}(\sqrt{X}) = \frac{1}{\lambda} \int_0^\infty \sqrt{x} e^{-x/\lambda} dx = \sqrt{\lambda} \Gamma(1.5) = \sqrt{\lambda} \frac{\sqrt{\pi}}{2}$$

og tilsvarende for  $Y$

$$\mathbf{E}(\sqrt{Y}) = \sqrt{3\lambda} \frac{\sqrt{\pi}}{2}$$

hvoraf

$$\mathbf{E}(\tilde{\lambda}) = \lambda \frac{\pi}{4} < \lambda$$

så  $\tilde{\lambda}$  er ikke central.

2. Beregn variansen for begge estimatorer.

Idet variansen i en eksponentiaffordeling med middelværdi  $\lambda$  er  $\lambda^2$  har vi

$$\mathbf{V}(\hat{\lambda}) = \lambda^2(9+9)/36 = \frac{1}{2}\lambda^2.$$

For at finde variansen på  $\tilde{\lambda}$  fås

$$\mathbf{E}(\tilde{\lambda}^2) = \mathbf{E}(X)\mathbf{E}(Y)/3 = \lambda^2$$

så

$$\mathbf{V}(\tilde{\lambda}) = \lambda^2 - \left(\frac{\pi}{4}\lambda\right)^2 = \frac{16-\pi^2}{16}\lambda^2.$$

3. Sammenlign varianserne med Cramér–Raos nedre grænse og kommenter resultatet.

Vi finder log-likelihoodfunktionen pånær irrelevante konstantled

$$\ell_{X,Y}(\lambda) = \frac{X}{\lambda} + \frac{Y}{3\lambda} + 2\log\lambda$$

og videre score- og information

$$S_{X,Y}(\lambda) = \frac{\partial \ell_{X,Y}}{\partial \lambda} = -\frac{X}{\lambda^2} - \frac{Y}{3\lambda} + \frac{2}{\lambda}$$

$$I_{X,Y}(\lambda) = \frac{\partial^2 \ell_{X,Y}}{\partial \lambda^2} = 2\frac{X}{\lambda^3} + 2\frac{Y}{3\lambda^3} - \frac{2}{\lambda^2}$$

som giver Fisherinformationen

$$i(\lambda) = \mathbf{E}\{I_{X,Y}(\lambda)\} = \frac{2}{\lambda^2}$$

og dermed er Cramér–Rao grænsen for en central estimator  $\lambda^2/2$ .

Vi ser, at  $\mathbf{V}(\hat{\lambda})$  netop har denne mindste varians.

Den nedre grænse for den ikke-centrale estimator  $\tilde{\lambda}$  er bestemt som

$$\mathbf{V}(\tilde{\lambda}) \geq \frac{\{\mathbf{E}'(\tilde{\lambda})\}^2}{i(\lambda)} = \frac{(\pi/4)^2}{2/\lambda^2} = \frac{\pi^2\lambda^2}{32} = v_{\min}.$$

Da

$$\mathbf{V}(\tilde{\lambda}) - v_{\min} = \frac{32 - 3\pi^2}{32} > 0$$

idet  $3\pi^2 = 29.60881\dots$  antager  $\mathbf{V}(\tilde{\lambda})$  ikke den nedre grænse.

4. Hvilken estimator har mindst kvadratisk middelfejl (mean square error)?

For den centrale estimator er den kvadratiske middelfejl lig med variansen. For den anden estimator fås

$$\text{MSE}(\tilde{\lambda}) = \mathbf{V}(\tilde{\lambda}) + \{\mathbf{E}(\tilde{\lambda}) - \lambda\}^2 = \left\{ \frac{16-\pi^2}{16} + \frac{(4-\pi)^2}{16} \right\} \lambda^2 = \frac{4-\pi}{2} \lambda^2 < \lambda^2/2.$$

Så  $\tilde{\lambda}$  har den mindste kvadratiske middelfejl.

## Opgave 2

Den inverse normalfordeling anvendes til at beskrive fordelingen af visse typer ventetider. I det speciale tilfælde, hvor middelværdi og varians er ens siges fordelingen at være *standardiseret* og i så fald har den tæthedsfunktion

$$f_\mu(x) = \frac{\mu}{\sqrt{2\pi x^3}} e^{-\frac{(x-\mu)^2}{2x}}, \quad x > 0,$$

Det kan uden bevis benyttes at  $\int_0^\infty f_\mu(x) dx = 1$  for alle  $\mu > 0$  og at  $\mathbf{E}(X) = \mathbf{V}(X) = \mu > 0$ .

Lad nu  $X_1, \dots, X_n$  være uafhængige og standardiseret invers normalfordelte med ukendt  $\mu > 0$  som ovenfor.

- Bestem scorefunktionen  $S(x, \mu)$ , informationsfunktionen  $I(x, \mu)$ , og Fisherinformationen  $i(\mu)$ . *Vink:* Benyt at scorefunktionen har middelværdi 0.

Scorefunktionen for en enkelt observation er

$$S(x, \mu) = \frac{\partial \ell_x(\mu)}{\partial \mu} = -\frac{1}{\mu} - \frac{(x-\mu)}{X} = -\frac{1}{\mu} - 1 + \frac{\mu}{x}$$

og informationsfunktionen

$$I(x, \mu) = \frac{\partial^2 \ell_x(\mu)}{\partial \mu^2} = \frac{\partial S(x, \mu)}{\partial \mu} = \frac{1}{\mu^2} + \frac{1}{x}.$$

Idet scorefunktionen har middelværdi 0, fås

$$\mathbf{E}(1/X) = \frac{1}{\mu} + \frac{1}{\mu^2} \tag{1}$$

og dermed

$$i(\mu) = \mathbf{E}\{I(X, \mu)\} = \frac{1}{\mu} + \frac{2}{\mu^2} = \frac{\mu+2}{\mu^2}.$$

Fisherinformation for  $n$  observationer fås ved at gange med antallet af observationer. Score og informationsfunktionerne fås ved at lægge størrelserne sammen

$$S_n(\mu) = -\frac{n}{\mu} - n + \sum_i \frac{\mu}{x_i}, \quad I_n(\mu) = \frac{n}{\mu^2} + \sum_i \frac{1}{x_i}, \quad i_n(\mu) = ni(\mu).$$

- Gør rede for, at familien af standardiserede inverse normalfordelinger med ukendt middelværdi  $\mu > 0$  udgør en eksponentiel familie og angiv familiens grundmål.

Vi omskriver tætheden ved at udvikle kvadratet:  $(x - \mu)^2 / (2x) = x/2 + \mu^2/(2x) - \mu$  så

$$f_\mu(x) = e^{\frac{\mu^2}{2}(-x^{-1})+(\mu+\log\mu)} \frac{1}{\sqrt{2\pi x^3}} e^{-\frac{x}{2}} \mathbf{1}_{(0,\infty)}(x).$$

Herat ser vi, at der er tale om en en-dimensional eksponentiel familie med grundmål

$$v(dx) = \frac{1}{\sqrt{2\pi x^3}} e^{-\frac{x}{2}} \mathbf{1}_{(0,\infty)}(x) \cdot dx.$$

3. Angiv den kanoniske parameter, den kanoniske stikprøvefunktion, samt kumulantfunktionen.

Den kanoniske parameter er  $\theta = \mu^2/2$ , den kanoniske stikprøvefunktion  $t(x) = -1/x$ , og idet  $\mu = \sqrt{2\theta}$  er kumulantfunktionen

$$\psi(\theta) = -\mu - \log \mu = -\sqrt{2\theta} - \frac{1}{2} \log \theta - \frac{\log 2}{2}.$$

Identiteten (1) kan naturligvis også fås ved at beregne  $\tau(\theta) = \psi'(\theta) = \mathbf{E}_\theta(1/X)$ .

4. Vis at maximum likelihood estimatoren  $\hat{\mu}$  for  $\mu$  er givet som  $\hat{\mu} = (1 + \sqrt{1 + 4\bar{T}})/(2\bar{T})$ , hvor  $\bar{T} = \frac{1}{n} \sum_{i=1}^n \frac{1}{X_i}$ .

I en eksponentiel familie er maximum likelihood estimatoren bestemt ved at sætte den kanoniske stikprøvefunktion lig sin middelværdi, altså

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{X_i} = \mathbf{E}_\mu(1/X) = \frac{1}{\mu} + \frac{1}{\mu^2}.$$

Sættes  $\bar{T} = \frac{1}{n} \sum_{i=1}^n \frac{1}{X_i}$  fører det til andengrads ligningen

$$\bar{T}\mu^2 - \mu - 1 = 0$$

som har præcis en positiv rod

$$\hat{\mu} = \frac{1 + \sqrt{1 + 4\bar{T}}}{2\bar{T}}.$$

5. Gør rede for, at  $\hat{\mu}$  er asymptotisk normalfordelt med parametre

$$\hat{\mu} \stackrel{\text{as}}{\sim} N\left(\mu, \frac{\mu^2}{n(\mu+2)}\right).$$

At estimatoren er asymptotisk normalfordelt følger af, at der er tale om en eksponentiel familie og  $\mu = \sqrt{2\theta}$  er en differentiabel funktion af  $\theta$ . Den asymptotiske varians er

$$\frac{1}{ni(\mu)} = \frac{\mu^2}{n(\mu+2)},$$

idet Fisherinformationen  $i(\mu)$  blev fundet ovenfor.

Nedenfor er angivet et eksempel på en stikprøve som anført ovenfor med  $n = 10$ :

```
> x
[1] 0.60 0.80 2.65 0.78 0.27 0.96 1.59 1.28 1.62 1.08
```

6. Under antagelse af at disse observationer følger en standard invers normalfordeling ønskes et approximativt 95% konfidensinterval for middelværdien  $\mu$ .

Stikprøvefunktionen  $\bar{t}$  beregnes til

```
> tbar = mean(1/x)
> tbar
[1] 1.227484
```

og dermed fås

```
> hatmu=(1+sqrt(1+4*tbar))/(2*tbar)
> hatmu
[1] 1.397589
```

med den asymptotiske varians og standardafvigelse

```
> asvar= hatmu^2/(length(x)*(2+hatmu))
> asvar
[1] 0.05748945
> sqrt(asvar)
[1] 0.2397696
```

hvilket giver et konfidensinterval (0.93, 1.87);

7. Er observationerne i overensstemmelse med hypotesen  $H_0 : \mu = 1$ ?

Da 1 ligger i konfidensintervallet er observationerne i overensstemmelse med hypotesen.

### Opgave 3

I Ugeskrift for Læger kunne man i 1974 læse, at der var mistanke om et særligt højt antal tilfælde af lungekraeft i byen Fredericia, sammenlignet med det observerede antal i naboyerne. For eksempel var der 64 tilfælde af lungekraeft blandt mænd i Fredericia i perioden 1968–1971, mens der i Vejle kun var 41 tilfælde.

Filen cancer.txt indeholder data som omhandler antal tilfælde af lungekraeft (Freq) i perioden 1968–1971 hos mænd i Vejle og Fredericia i forskellige aldersgrupper samt det omtrentlige antal mænd (Population) i de samme aldersgrupper, bosiddende i disse byer. Aldersgrupperne er kodet som følger

	A	B	C	D	E	F
Alder	40–54	55–59	60–64	65–69	70–74	>74

For at undersøge om der kunne være tale om tilfældigheder, kunne man betragte antallet af lungekraefttilfælde  $X_{ab}$  i en given aldersgruppe  $a$  og en given by  $b$  som uafhængige og Poisson-fordelte med en middelværdi af formen

$$\mathbf{E}(X_{ab}) = \alpha_a \beta_b N_{ab}$$

hvor  $N_{ab}$  angiver antallet af mandlige personer i aldersgruppe  $a$  i byen  $b$  og betragtes som fast og kendt, mens  $\alpha_a$  og  $\beta_b$  er ukendte parametre, som beskriver variationen i hyppighed over aldersgrupper og lokaliteter. Modellen kan for eksempel specificeres som følger

```
glm(Freq ~ Age + Town, offset=log(Population), family="poisson", data=cancer)
```

1. Undersøg om en model af den angivne form kan beskrive data og angiv estimererne for modellens parametre.

Data indlæses

```
> cancer <- read.table("data/cancer.txt", header=TRUE)
> cancer
   Age      Town Population Freq
1   A Fredericia     3059    11
2   A      Vejle     2520     5
3   B Fredericia     800    11
4   B      Vejle     878     7
5   C Fredericia    710    11
6   C      Vejle     839    10
7   D Fredericia    581    11
8   D      Vejle     631    14
9   E Fredericia    509    11
10  E      Vejle     539     8
11  F Fredericia    605    10
12  F      Vejle     619     7
```

Og modellen specificeres som følger

```
> p1<-glm(Freq~Age+Town, offset=log(Population),family="poisson",data=cancer)
```

Hvilket giver flg. output

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-5.7328	0.2600	-22.051	< 2e-16 ***
AgeB	1.3398	0.3438	3.897	9.75e-05 ***
AgeC	1.5794	0.3323	4.754	2.00e-06 ***
AgeD	1.9929	0.3204	6.220	4.97e-10 ***
AgeE	1.8620	0.3395	5.485	4.14e-08 ***
AgeF	1.5931	0.3485	4.572	4.83e-06 ***
TownVejle	-0.2918	0.1873	-1.558	0.119

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 63.1074 on 11 degrees of freedom
Residual deviance: 1.9361 on 5 degrees of freedom
```

Med en residualdevians på 1.9361 på 5 frihedsgrader giver modellen en fin tilpasning.

2. Brug modellen og analysen til at afgøre, om hyppigheden af lungekræft er forskellig i de to byer udover, hvad der kan forklares af en forskellig aldersfordeling.

Koefficienten som angiver hvor forskellig Vejle er fra Fredericia, er estimeret til -0.2918 med en standardafvigelse på 0.1873, så den er ikke signifikant forskellig fra 0 på noget rimeligt signifikansniveau. Med andre ord er der ingen grund til at antage at hyppigheden er forskellig i de to byer.

3. Angiv estimatorer for morbiditetsraterne  $\alpha_a$  i aldersgrupperne 40–54 og 65–69 under antagelse af at disse er ens i de to byer, altså  $\beta_b \equiv 1$ .

Vi specificerer en model, hvor byen ikke indgår, som følger.

```
> p2<- glm(Freq ~ Age, offset=log(Population), family="poisson", data=cancer)
> summary(p2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.8542    0.2500 -23.417 < 2e-16 ***
AgeB         1.3192    0.3436   3.839 0.000123 ***
AgeC         1.5533    0.3318   4.681 2.86e-06 ***
AgeD         1.9730    0.3202   6.163 7.15e-10 ***
AgeE         1.8440    0.3393   5.434 5.50e-08 ***
AgeF         1.5775    0.3483   4.529 5.93e-06 ***
---
Residual deviance: 4.3831 on 6 degrees of freedom
```

Nu regnes effekten om ved at tage eksponentialfunktionen af koefficienterne

```
> eff<-exp(p2$coefficients)
> round(eff,4)
(Intercept)      AgeB      AgeC      AgeD      AgeE      AgeF
 0.0029     3.7404    4.7272    7.1924    6.3216    4.8429
```

Altså er antallet af lungekræfttilfælde 2.9 pr tusind indbyggere i aldersgruppen 40–54; de øvrige faktorer angiver den faktor, som raten bliver forøget med i de andre aldersgrupper. For eksempel er der omrent 21 tilfælde pr tusind indbyggere i aldersgruppen 65–69. Den samlede effekt i de øvrige aldersgrupper kan for eksempel beregnes således

```
> toteff<- eff[1]*eff[2:6]
> round(toteff,4)
AgeB  AgeC  AgeD  AgeE  AgeF
0.0107 0.0136 0.0206 0.0181 0.0139
```

# Reeksamen i Statistik 2, 23. august 2018

Vejledende besvarelse

## Opgave 1

1. Ifølge EH Korollar 9.43 er  $X$  regulært normalfordelt hvis og kun hvis variansen  $\Sigma$  er invertibel. Det ses at determinanten af  $\Sigma$  er lig

$$\frac{1}{2} \{1 \cdot 2 \cdot 1 + 1 \cdot (-1) \cdot 0 + 0 \cdot 1 \cdot (-1) - 1 \cdot (-1) \cdot (-1) - 1 \cdot 1 \cdot 1 - 0 \cdot 2 \cdot 0\} = 0,$$

hvorfor  $\Sigma$  ikke er invertibel. Alternativt kan bemærkes at række 1 i  $\Sigma$  er lig med summen af række 2 og 3, hvorfor  $\Sigma$  ikke er invertibel. Det konkluderes at  $X$  følger en singulær normalfordeling.

2. Det følger af EH Lemma 9.47 at  $Y$  er normalfordelt med varians  $\Sigma_Y = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ . M.  
Der er tale om en regulær normalfordeling og af EH Sætning 9.42 konkluderes, at den tilhørende præcision er givet ved

$$\langle x, y \rangle = x^T \Sigma_Y^{-1} y = x^T \cdot \underbrace{2 \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}}_{:= \Sigma_Y^{-1}} \cdot y.$$

Tæthedens for  $Y$  i  $y = (1, 1)$  bliver ifølge EH Sætning 9.20

$$f(y) = \frac{(det \Sigma_Y^{-1})^{1/2}}{(2\pi)^{2/2}} \exp\left(-\frac{1}{2} y^T \Sigma_Y^{-1} y\right) = \frac{4^{1/2}}{2\pi} \exp\left(-\frac{1}{2} 4\right) \approx 0.0431.$$

3. Det følger EH Sætning 9.47 og 9.48 at  $X_1$  og  $X_3$  er uafhængige og at begge variable er normalfordelte med middelværdi 0 og varians  $1/2$ . Dermed er  $\tilde{X}_1 = \sqrt{2}X_1$  og  $\tilde{X}_3 = \sqrt{2}X_3$  uafhængige og standardnormalfordelte. Dermed er  $\tilde{X}_1^2 + \tilde{X}_3^2 = 2X_1^2 + 2X_3^2$  per definition  $\chi^2$ -fordelte med 2 frihedsgrader. Det er netop dette udtryk som fremkommer, når man udregner matrixproduktet fra opgaveformuleringen.

En alternativ løsning består i at bemærke, at  $Z = (X_1, X_3)^T$  er regulært normalfordelt med middelværdi  $(0, 0)^T$  og varians  $\Sigma_Z = \frac{1}{2}I_2$ . Dermed gælder ifølge EH Sætning 9.29 at  $\|Z\|_{\Sigma_Z^{-1}}^2 = Z^T (\frac{1}{2}I_2)^{-1} Z$   $\chi^2$ -fordelt med 2 frihedsgrader. Opgaven løses nu ved at indse, at  $\|Z\|_{\Sigma_Z^{-1}}^2$  er identisk med matrixproduktet i opgaveformuleringen.

## Opgave 2

1. Med  $\beta = (\beta_1, \beta_2)^T$  bliver designmatricen

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix}.$$

Maksimaliseringsestimateren i den lineære normale model findes ved brug af EH Korollar 10.21. Vi finder, at

$$\hat{\beta} = (A^T A)^{-1} A^T W = (-0.7333, 2.1142)^T$$

og

$$\hat{\sigma}^2 = \frac{\|W - A\hat{\beta}\|^2}{6} = 0.184127.$$

2. Tilsvarende giver EH Korollar 10.21 at  $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(A^T A)^{-1})$  og at  $\hat{\sigma}^2$  er  $\chi^2$ -fordelt med  $6 - 2 = 4$  frihedsgrader og skalaparamater  $\sigma^2/6$ . Spredningen på estimatet for  $\beta_2$  kan estimeres ved at indsætte et estimat for  $\sigma$  i udtrykket  $\sigma^2(A^T A)^{-1}$  for variansen. Benyttes den centrale estimator  $\tilde{\sigma}^2 = \frac{6}{4}\hat{\sigma}^2$  så estimeres standard error for estimatet på  $\beta_2$  til

$$se(\hat{\beta}_2) = \sqrt{\frac{6}{4} \cdot 0.184127 \cdot 0.05714286} = 0.1256.$$

Grænserne for et 95 % - konfidensinterval kan beregnes som  $2.1142 \pm 2.7764 \cdot 0.1256$ , hvor 2.7764 angiver 97.5 % - fraktilen i en  $t$ -fordeling med  $6 - 2 = 4$  frihedsgrader. Konfidensintervallet bliver  $[1.765 - 2.463]$ .

Følgende R kode er benyttet til den vejledende besvarelse af opgave 2, men alle beregninger kan ret let foretages i hånden.

Først beregnes maksimaliseringsestimaterne

```
W <- matrix(ncol = 1, data = c(1,4,6,7,10,12))
A <- cbind(1, 1:6)
bhat <- solve(t(A) %*% A) %*% t(A) %*% W
bhat # estimat for beta

##           [,1]
## [1,] -0.7333333
## [2,]  2.1142857

shat2 <- sum((W - A %*% bhat)^2)/6
shat2 # estimat for sigma^2

## [1] 0.184127
```

Dernæst bestemmes (det estimeres værdier) for parametrene i fordelingen af maksimaliseringsestimatorerne.

```
bhatvar <- shat2*6/4 * solve(t(A)%*%A)
bhatvar

##           [,1]      [,2]
## [1,]  0.2393651 -0.05523810
## [2,] -0.0552381  0.01578231

bhatse <- sqrt(diag(bhatvar))
bhatse

## [1] 0.4892495 0.1256277

bhat2ci <- bhat[2] + c(-1, 1) * bhatse[2] * qt(0.975, 6 - 2)
bhat2ci

## [1] 1.765487 2.463084
```

Det er også muligt at lade `lm()`-funktionen foretage beregningerne.

```
summary(lm(W ~ A - 1))

##
## Call:
## lm(formula = W ~ A - 1)
##
## Residuals:
##       1        2        3        4        5        6
## -0.38095  0.50476  0.39048 -0.72381  0.16190  0.04762
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## A1     -0.7333    0.4892  -1.499   0.208
## A2      2.1143    0.1256  16.830 7.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 0.5255 on 4 degrees of freedom
## Multiple R-squared:  0.9968, Adjusted R-squared:  0.9952
## F-statistic: 624.4 on 2 and 4 DF,  p-value: 1.019e-05

confint(lm(W ~ A - 1))

##
##           2.5 %    97.5 %
## A1 -2.091708  0.6250411
## A2  1.765487  2.4630841
```

## Opgave 3

1. Det fremgår af tabellen over forsøgsdesignet i opgaveformuleringen, at designet er sammenhængende og at faktorerne  $M$  og  $D$  opfylder *balancealigningen* fra EH Lemma 13.11. Dermed er faktorerne geometrisk ortogonale. For et sammenhængende design er minimum af faktorerne blot den konstante faktor. Dermed bliver dimensionen af det additive underrum

$$\dim(L_M + L_D) = \dim L_M + \dim L_D - \dim L_1 = 2 + 2 - 1 = 3.$$

2.  $F$  teststørrelsen for test af den additive hypotese kan ifølge EH formel (10.31) udtrykkes som

$$F = \frac{(||P_{M \times D}||^2 - ||P_{M+D}X||^2)/1}{(||X||^2 - ||P_{M \times D}X||^2)/(24 - 4)}.$$

For tosiden variansanalyse med geometrisk ortogonale faktorer kan vi benytte EH formel (13.5) til at beregne

$$||P_{M+D}X||^2 = ||P_M X||^2 + ||P_D X||^2 - ||P_1 X||^2,$$

hvor alle tre størrelser på højresiden fremgår af faktorstrukturdiagrammet i opgaveformuleringen. Vi finder at  $||P_{M+D}X||^2 = 429918.342$  og dermed, at

$$F = \frac{(429921.699 - 429918.342)/1}{(431533.199 - 429921.699)/(24 - 4)} = 0.0417.$$

Under hypotesen om at der ikke er nogen vekselvirkning vil  $F$  teststørrelsen følge en  $F$ -fordeling med  $(1, 20)$  frihedsgrader. Vi finder således den tilsvarende  $P$ -værdi = 0.840. Vi kan således ikke forkaste hypotesen om, at der ikke er vekselvirkning mellem medikament og dosis.

3. De estimerede middelværdier bliver

M	D	E[X]
A	lav	99.507
A	hoj	134.814
B	lav	135.419
B	hoj	168.894
0	0	97.733

4. Data fra USB-nøglen indlæses i R. Teststørrelsen bliver  $F = 0.0434$  med tilhørende  $P$ -værdi = 0.8361. Der lader således ikke til at være en vekselvirkning mellem dosis og behandling.
5. Det er helt legalt blot at kigge på antallet af estimater for middelværdistrukturen i R udskriften, når man skal argumentere for, at dimensionen af den additive model er 4.

Ønsker man at regne mere formelt på tingene kan benyttes, at

$$\dim(L_M + L_D) = \dim L_M + \dim L_D - \dim L_{M \wedge D}.$$

Udfordringen ligger i at minimum  $M \wedge D$  her er en faktor på 2 niveauer, som blot holder styr på om målingen stammer fra en person, som har modtaget et medikament (dvs.  $M = A$  eller  $M = B$ ) eller ej (dvs.  $M = 0$ ). Minimum kan umiddelbart aflæses ud fra antalstabellen til beskrivelse af det fulde forsøgsdesign som findes i opgaveformuleringen. Indsættes i ovenstående formel fås nu, at

$$\dim(L_M + L_D) = \dim L_M + \dim L_D - \dim L_{M \wedge D} = 3 + 3 - 2 = 4.$$

Det virker ikke rimeligt at reducere modellen yderligere ved at se bort fra effekten af  $M$  ( $F = 89.143, p < 0.0001$ ) eller faktoren  $D$  ( $F = 69.519, p < 0.0001$ ).

6. Som altid er der ikke entydighed omkring valget af designmatrix ved parametrisering af det additive middelværdiunderrum. Nedenfor angives estimaterne fra en et parametrisering som benytter kontrolgruppen som reference (estimat: 97.733, 95 %-KI: [93.3 – 102.1]). Gives den lave dosis af  $A$  øges estimatet med 2.079 (95 % - KI: [-5.8-10.0]), mens den tilsvarende effekt for medikament  $B$  estimeres til 37.534 (95 % - KI: [30.979048 44.088322]). Gives i stedet den høje dosis øges estimatet med 34.086 (95 % - KI: [25.8-42.4]) uanset medikament (da vi betragter en additiv model!).

På baggrund af konfidensintervallerne for estimaterne i den valgte parametrisering kan vi umiddelbart konkludere: i) at medikament  $A$  ikke har effekt i den lave dosis, ii) at medikament  $B$  *har* effekt i den lave dosis, iii) der er effekt af at bruge høj dosis i stedet for lav dosis.

Afhængigt af den valgte parametrisering kan andre aspekter af effekten af de forskellige behandlingskombinationer undersøges. For at få fuldt point for delopgaven er det nok, at der udtrækkes relevante konklusioner fra estimater og konfidensintervaller fra mindst en fornuftig parametrisering af modellen.

Følgende R-kode er benyttet i forbindelse med den vejledende besvarelse

```
data2 <- read.table("stat2_2018_aug_opg3.txt", header = T)
```

```
head(data2)

##   M   D      x
## 1 A  lav  95.77016
## 2 A  lav  84.50122
## 3 A  lav  99.35571
## 4 A  lav 102.70881
## 5 A  lav 117.35284
## 6 A  lav  97.35289
```

```

mod0 <- lm(x ~ M:D - 1, data = data2)

### Delopgave 4: test af vekselvirkning

mod1 <- lm(x ~ M + D, data = data2)
anova(mod1, mod0) # F = 0.0434, P = 0.8361

## Analysis of Variance Table
##
## Model 1: x ~ M + D
## Model 2: x ~ M:D - 1
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     36 2707.4
## 2     35 2704.1  1    3.3564 0.0434 0.8361

### Delopgave 5: yderligere modelreduktion

mod2a <- lm(x ~ M, data = data2)
anova(mod2a, mod1) # F = 69.519, p < 0.0001

## Analysis of Variance Table
##
## Model 1: x ~ M
## Model 2: x ~ M + D
##   Res.Df   RSS Df Sum of Sq      F      Pr(>F)
## 1     37 7935.7
## 2     36 2707.4  1    5228.3 69.519 6.297e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mod2b <- lm(x ~ D, data = data2)
anova(mod2b, mod1) # F = 89.143, p < 0.0001

## Analysis of Variance Table
##
## Model 1: x ~ D
## Model 2: x ~ M + D
##   Res.Df   RSS Df Sum of Sq      F      Pr(>F)
## 1     37 9411.5
## 2     36 2707.4  1    6704.1 89.143 2.819e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

### Delopgave 6: estimerer fra additive model

summary(mod1)

##
## Call:
## lm(formula = x ~ M + D, data = data2)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -15.4196 -5.6774 -0.4192  6.5403 17.5406
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.733     2.168   45.079 < 2e-16 ***
## MA          2.079     3.892   0.534    0.596
## MB          37.534     3.232  11.613 9.87e-14 ***
## Dhoj        34.086     4.088   8.338 6.30e-10 ***
## Dlav         NA        NA      NA      NA
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.672 on 36 degrees of freedom
## Multiple R-squared:  0.8954, Adjusted R-squared:  0.8867
## F-statistic: 102.8 on 3 and 36 DF,  p-value: < 2.2e-16

confint(mod1)

##             2.5 %    97.5 %
## (Intercept) 93.336023 102.129992
## MA          -5.813555  9.972101
## MB          30.979048  44.088322
## Dhoj        25.794768  42.376833
## Dlav         NA        NA

# Konklusion:
# A, lav: ingen effekt (2.079)
# B, lav: signifikant effekt (37.534)
# Forskel p<U+00E5> høj og lav: signifikant (34.086)

mod1alt <- lm(x ~ D + M , data = data2)
summary(mod1alt)

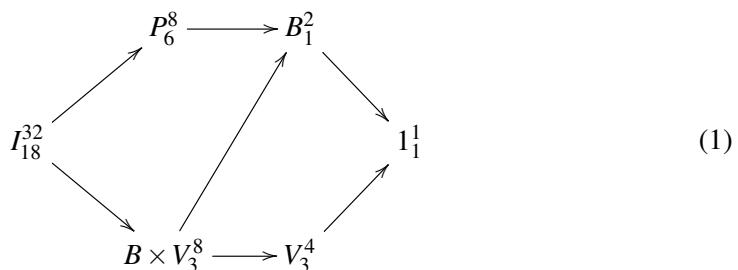
##
## Call:
## lm(formula = x ~ D + M, data = data2)

```

```
##  
## Residuals:  
##      Min       1Q   Median      3Q     Max  
## -15.4196  -5.6774  -0.4192   6.5403  17.5406  
##  
## Coefficients: (1 not defined because of singularities)  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  97.733     2.168  45.079 < 2e-16 ***  
## Dhoj        71.619     4.336  16.517 < 2e-16 ***  
## Dlav        37.534     3.232  11.613 9.87e-14 ***  
## MA         -35.454     3.755 -9.442 2.82e-11 ***  
## MB            NA        NA       NA       NA  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 8.672 on 36 degrees of freedom  
## Multiple R-squared:  0.8954, Adjusted R-squared:  0.8867  
## F-statistic: 102.8 on 3 and 36 DF,  p-value: < 2.2e-16
```

## Opgave 4

- Alle kombinationer af  $P$  og  $V$  er afprøvet netop een gang i forsøget. Derfor er disse to faktorer geometriske ortogonale med trivielt minimum (= den konstante faktor). Da  $B$  er grovere end  $P$  følger det umiddelbart af regneregler for minimum, at alle øvrige faktorer i designet er geometrisk ortogonale, og at minimumskonstruktion ikke fører til nye faktorer.
- Den væsentlige udfordring er at lade faktorstrukturdiagrammet afspejle, at  $B$  er grovere end  $P$ . Det udfyldte diagram med dimensioner ser ud som følger



- Modellen kan udtrykkes ved at vektoren  $X = (X_i)_{i \in I}$  bestående af udbyttet på de enkelte parceller er normalfordelt på  $\mathbb{R}^{32}$  med  $\xi = EX \in L_{B \times V}$  og varians  $VX = \sigma^2 + v_1^2 B_1$ , hvor  $B_1$  er effektmatricen hørende til parret ( $P, 1$ ).

Den totale varians på udbyttet bliver  $\sigma^2 + v_1^2$  og kovariansmatricen for de 4 målinger på samme plot kan udtrykkes som

$$\begin{pmatrix} \sigma^2 + v_1^2 & v_1^2 & v_1^2 & v_1^2 \\ v_1^2 & \sigma^2 + v_1^2 & v_1^2 & v_1^2 \\ v_1^2 & v_1^2 & \sigma^2 + v_1^2 & v_1^2 \\ v_1^2 & v_1^2 & v_1^2 & \sigma^2 + v_1^2 \end{pmatrix}$$

- Følgende R kode kan benyttes til at estimere parametrene i den ønskede varianskompontimentmodel i R

```
data4 <- read.table("stat2_2018_aug_opg4.txt", header = T)
```

```
head(data4)

##   P   B     V udbytte
## 1 1 Ja  Lami 52.3836
## 2 1 Ja  Lofa 49.6232
## 3 1 Ja Salka 49.6232
## 4 1 Ja  Zita 49.7334
## 5 2 Ja  Lami 55.4953
## 6 2 Ja  Lofa 52.7372
```

```

library(lme4)

## Loading required package: Matrix

m0 <- lmer(udbytte ~ V * B + (1|P), data = data4)
VarCorr(m0)

## Groups   Name        Std.Dev.
## P         (Intercept) 1.1926
## Residual             1.8243

```

Varianseestimaterne bliver  $\hat{\nu}_1 = 1.1926$  og  $\hat{\sigma} = 1.8243$ .

5. Middelværdiestimaterne hørende til de ønskede kombinationer af  $B$  og  $V$  fremgår af følgende R-udskrift, hvor de ønskede grupper optræder i linjerne 6, 1, 5.

```

m0alt <- lmer(udbytte ~ V : B - 1 + (1|P), data = data4)
coef(summary(m0alt))

##                   Estimate Std. Error t value
## VLami:BJa    55.64332  1.089786 51.05893
## VLofa:BJa    49.89000  1.089786 45.77962
## VSalka:BJa   54.11810  1.089786 49.65937
## VZita:BJa    52.82257  1.089786 48.47058
## VLami:BNej   61.21607  1.089786 56.17255
## VLofa:BNej   54.89210  1.089786 50.36960
## VSalka:BNej  57.70762  1.089786 52.95316
## VZita:BNej   58.52837  1.089786 53.70629

```

6. Konfidensintervaller for middelværdierne i de otte grupper givet ved  $B \times V$  kan beregnes i R ved bruge af `confint()`-funktionen.

```

confint(m0alt)

## Computing profile confidence intervals ...

##                   2.5 %     97.5 %
## .sig01      0.000000  2.197756
## .sigma       1.219952  2.159005
## VLami:BJa   53.716425 57.570225
## VLofa:BJa   47.963100 51.816900
## VSalka:BJa  52.191200 56.045000
## VZita:BJa   50.895675 54.749475
## VLami:BNej  59.289175 63.142975
## VLofa:BNej  52.965200 56.819000
## VSalka:BNej 55.780725 59.634525
## VZita:BNej  56.601475 60.455275

```

For at vi kan udtale os om effekten af bayleton for sorten **Lami** benyttes en ny parametrering af modellen, hvorfaf forskellene mellem grupperne  $\{Ja, Lami\}$  og  $\{Nej, Lami\}$  direkte kan aflæses.

```
m0altny <- lmer(udbytte ~ V + B:V - 1 + (1|P), data = data4)
coef(summary(m0altny))

##           Estimate Std. Error   t value
## VLami      55.643325  1.089786 51.058933
## VLofa      49.890000  1.089786 45.779618
## VSalka     54.118100  1.089786 49.659369
## VZita      52.822575  1.089786 48.470581
## VLami:BNej 5.572750   1.541191  3.615873
## VLofa:BNej 5.002100   1.541191  3.245608
## VSalka:BNej 3.589525   1.541191  2.329060
## VZita:BNej 5.705800   1.541191  3.702203

confint(m0altny)

## Computing profile confidence intervals ...

##           2.5 %    97.5 %
## .sig01    0.0000000 2.197756
## .sigma    1.2199519 2.159005
## VLami     53.7164246 57.570225
## VLofa     47.9630996 51.816900
## VSalka    52.1911996 56.045000
## VZita     50.8956746 54.749475
## VLami:BNej 2.8477013 8.297799
## VLofa:BNej 2.2770513 7.727149
## VSalka:BNej 0.8644763 6.314574
## VZita:BNej 2.9807513 8.430849
```

Forskellen mellem grupperne aflæses til  $5.572750$  [ $95 - \%KI : 2.847701 - 8.297799$ ]. Da konfidensintervallet *ikke* indholder værdien 0, så lader det til at sprøjtning har nogen effekt på udbyttet for sorten **Lami**. Effekten af at sprøjte med bayleton kan i øvrigt genfindes for alle fire sorten, hvilken kunne uddybes ved en mere systematisk analyse af datasættet.

# Eksamens i Statistik 2

21. juni 2018

Eksamens varer 4 timer. Alle hjælpemidler er tilladt under hele eksamen, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant.

Eksamenssættet består af fire opgaver og dækker både pensum fra 2017 og 2018. Hvis man ønsker at blive bedømt i forhold til pensum fra 2018 afleverer man besvarelser af opgave 1, 3, og 4. Hvis man ønsker at blive bedømt i forhold til gammelt pensum (2017) afleverer man besvarelser af opgaverne 2, 3, og 4.

*Hvis man ønsker at blive bedømt efter gammelt pensum skal dette angives tydeligt på første side i besvarelsen.*

Hvis intet er angivet, vil besvarelsen blive bedømt i forhold til pensum fra 2018 og kun besvarelser af opgave 1, 3, og 4 vil indgå i bedømmelsen. Angives, at besvarelsen ønskes bedømt i forhold til gammelt pensum, indgår kun besvarelser af opgaverne 2, 3, og 4 i bedømmelsen.

I begge tilfælde vil der være i alt 17 delspørgsmål, som skal bedømmes; alle delspørgsmål vægtes ens i bedømmelsen.

Data til Opgave 3 og 4 ligger på en USB-nøgle i filerne **moral.txt** og **koed.txt**. Nøglen skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den skal altså ikke indgå som del af besvarelsen.

# Opgave 1

Bemærk: besvarelse af denne opgave bedømmes kun i forhold til 2018 (nyt) pensum.

Lad  $X = (X_1, X_2, X_3, X_4)^T \in \mathbb{R}^4$  være normalfordelt  $\mathcal{N}(\xi, \Sigma)$ , hvor

$$\xi = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \frac{1}{2} \begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

1. Er  $X$  regulært eller singulært normalfordelt? Begrund dit svar.

### Vejledende besvarelse

Da  $\Sigma$  er singulær er  $X$  singulært normalfordelt, jvf. Korollar 9.43.

2. Lad  $X = (X_1, X_2, X_3, X_4)^T$  være som ovenfor. Angiv om  $Y = (X_1, X_2)^T$ , hhv.  $Z = (X_3, X_4)^T$  er regulært eller singulært normalt fordelte.

### Vejledende besvarelse

Vi ser at de marginale fordelinger (jvf. Sætning 9.47) er  $Y \sim \mathcal{N}(0, \Sigma_Y)$  og  $Z \sim \mathcal{N}(0, \Sigma_Z)$ , med

$$\Sigma_Y = \frac{1}{2} \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} \quad \text{og} \quad \Sigma_Z = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

hvoraf  $\Sigma_Y^{-1}$  eksisterer, mens  $\Sigma_Z$  er singulær. Kovariansmatricerne viser derfor at  $Y$  er regulært normalfordelt, mens  $Z$  er singulært normalfordelt, jvf. Korollar 9.43.

3. Angiv fordelingen af  $(Y+Z)^T \Sigma_{Y+Z}^{-1} (Y+Z)$ , hvor  $\Sigma_{Y+Z}$  er variansmatricen for  $Y+Z$ .

### Vejledende besvarelse

Bemærk først at  $Y \perp\!\!\!\perp Z$  da  $\text{Cov}(Y, Z) = 0$ , jvf. Sætning 9.48. Derfor er  $Y+Z$  en sum af uafh. normalfordelte variable hvilket giver at den selv er 2-dimensionalt normal fordelt, med middelværdi  $\mathbb{E}Y + \mathbb{E}Z = 0$  og varians

$$\Sigma_{Y+Z} = \Sigma_Y + \Sigma_Z = \frac{1}{4} \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$$

Da  $\Sigma_{Y+Z}^{-1}$  eksisterer er den regulært normalfordelt.

Da  $Y+Z$  er 2-dimensionelt regulært normalfordelt med præcision  $\langle \cdot, \cdot \rangle$  givet ved  $\Sigma_{Y+Z}^{-1}$  vil  $\|Y+Z\|_{\Sigma_{Y+Z}}^2 = (Y+Z)^T \Sigma_{Y+Z}^{-1} (Y+Z)$  være  $\chi^2$ -fordelt med 2 frihedsgrader, jvf. Sætning 9.29.

Bemærk at de næste delopgaver er uden sammenhæng med de foregående.

Lad nu  $W \sim \mathcal{N}(\xi, \sigma^2 I_4)$  være regulært normal fordelte.

4. Opskriv design matricen  $A$  for følgende specifikation af  $\xi$

$$\xi = A\beta = \begin{pmatrix} \beta_1 + \beta_3 \\ \beta_1 \\ \beta_2 + \beta_3 \\ \beta_2 \end{pmatrix}$$

#### Vejledende besvarelse

Idet der indgår tre parametre  $(\beta_1, \beta_2, \beta_3)$  i en 4-dimensional fordeling kan vi se at  $A$  skal være  $4 \times 3$ . Udfra kombinationerne af  $\beta$ 'er i  $\xi$  ser vi at

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Med  $\beta = (\beta_1, \beta_2, \beta_3)$  ser vi at  $A\beta$  giver  $\xi$  som beskrevet.

5. Betragt en observation  $w = (2, 5, 3, 1)^T$  fra  $W$  defineret som i spørgsmål 4).

Find maksimum likelihood estimatorerne for  $\beta$  og  $\sigma^2$  med designmatricen fra spørgsmål 4) og angiv standard errors for  $\hat{\beta}_1, \hat{\beta}_2$  og  $\hat{\beta}_3$ .

Du må her gerne benytte maksimalisering-estimatoren for  $\hat{\sigma}^2$ , fremfor den centrale estimator til at beregne standard errors for  $\hat{\beta}$ .

#### Vejledende besvarelse

Fra Korollar 10.21 har vi at

$$\hat{\beta} = (A^T A)^{-1} A^T W$$

$$\hat{\sigma}^2 = \frac{\|W - A\hat{\beta}\|^2}{N}$$

Med  $A$  som i forrige spørgsmål er

$$A^T A = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

hvilket giver

$$(A^T A)^{-1} = \frac{1}{4} \begin{pmatrix} 3 & 1 & -2 \\ 1 & 3 & -2 \\ -2 & -2 & 4 \end{pmatrix}$$

Dvs. når vi indsætter  $w$  får vi

$$\hat{\beta} = (A^T A)^{-1} A^T w = \frac{1}{4} \begin{pmatrix} 1 & 3 & -1 & 1 \\ -1 & 1 & 1 & 3 \\ 2 & -2 & 2 & -2 \end{pmatrix} \begin{pmatrix} 2 \\ 5 \\ 3 \\ 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 15 \\ 9 \\ -2 \end{pmatrix}$$

Indsætter vi  $A\hat{\beta}$  i varianseestimatoren får vi

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\|W - A\hat{\beta}\|^2}{N} = \frac{W^T W - (A\hat{\beta})^T A\hat{\beta}}{N} \\ &= \frac{39 - 32.75}{4} = 1.5625 = \frac{25}{16} \end{aligned}$$

Endeligt bliver de estimerede standard errors for  $\hat{\beta}$   $\sqrt{\text{-roden af diagonalen i matricen}}$

$$\hat{\sigma}^2 (A^T A)^{-1} = \frac{25}{64} \begin{pmatrix} 3 & 1 & -2 \\ 1 & 3 & -2 \\ -2 & -2 & 4 \end{pmatrix} = \begin{pmatrix} 1.17 & 0.39 & -0.78 \\ 0.39 & 1.17 & -0.78 \\ -0.78 & -0.78 & 1.56 \end{pmatrix}$$

dvs.

$$\text{SE}(\hat{\beta}) = (\sqrt{1.17}, \sqrt{1.17}, \sqrt{1.56})^T = (1.08, 1.08, 1.25)^T = \frac{5}{8}(\sqrt{3}, \sqrt{3}, 2)^T$$

Teknisk set bør vi bruge den centrale estimator  $\tilde{\sigma}^2$  til standard errors for  $\hat{\beta}$ , men det undlader vi her, som efterspurgt. Hvis  $\tilde{\sigma}^2$  er benyttet, er tælleren i estimatoren  $N - k = 4 - 3 = 1$ , dvs.  $\tilde{\sigma}^2 = 4\hat{\sigma}^2$  og derfor bliver  $\tilde{\sigma} = 2\hat{\sigma}$ . Det giver så standard errors for  $\hat{\beta}$ 'erne der er dobbelt så store som angivet ovenfor

$$\text{SE}(\hat{\beta}) = (2.17, 2.17, 2.50)^T = \frac{5}{4}(\sqrt{3}, \sqrt{3}, 2)^T$$

## Opgave 2

Bemærk: besvarelse af denne opgave bedømmes kun i forhold til 2017 (gammelt) pensum.

Betrægt fordelingen med tæthed

$$f_\alpha(x) = \left( \frac{1}{\sqrt{2\pi}} \right) x^{-3/2} \exp \left\{ -\frac{(x-\alpha)^2}{2\alpha^2 x} \right\} \quad \text{for } x > 0$$

med hensyn til Lebesgue-målet. Fordelingen afhænger af parameteren  $\alpha > 0$ .

Du kan uden bevis benytte at  $f_\alpha$  er en tæthed.

1. Reparametriser  $f_\alpha$ -fordelingen ved  $\theta = -\frac{1}{2\alpha^2}$  så det fremgår at det er en eksponentiel familie med kanonisk stikprøvefunktion  $t(x) = x$ .

### Vejledende besvarelse

Vi får

$$\begin{aligned} f_\alpha(x) &= \frac{1}{\sqrt{2\pi}} x^{-3/2} \exp \left\{ -\frac{(x-\alpha)^2}{2\alpha^2 x} \right\} \\ &= \frac{1}{\sqrt{2\pi}} x^{-3/2} \exp \left\{ -\frac{x^2 - 2x\alpha + \alpha^2}{2\alpha^2 x} \right\} \\ &= \frac{1}{\sqrt{2\pi}} x^{-3/2} \exp \left\{ -\frac{1}{2\alpha^2} x + \frac{1}{\alpha} - \frac{1}{2x} \right\} \\ &= \frac{1}{\sqrt{2\pi}} x^{-3/2} \exp \left\{ -\frac{1}{2x} \right\} \exp \left\{ \frac{1}{\alpha} \right\} \exp \left\{ -\frac{1}{2\alpha^2} x \right\} \\ &= \frac{1}{\sqrt{2\pi}} x^{-3/2} \exp \left\{ -\frac{1}{2x} \right\} \exp \left\{ \sqrt{-2\theta} \right\} \exp \left\{ \theta x \right\} \end{aligned}$$

hvor  $\theta = -\frac{1}{2\alpha^2}$ . Vi får således (se def. 2.13 i lærebogen)

$$f_\alpha(x) = \underbrace{\frac{1}{\sqrt{2\pi}} x^{-3/2} e^{-\frac{1}{2x}}}_{\text{funktion af } x} \underbrace{e^{\sqrt{-2\theta}}}_{\text{funktion af } \theta} \underbrace{e^{\theta x}}_{t(x) = x}$$

2. Identifier grundmålet. Identifier normeringskonstanten  $c(\theta)$ .

### Vejledende besvarelse

Lad  $\mu$  have tæthed

$$\frac{1}{\sqrt{2\pi}} x^{-3/2} e^{-\frac{1}{2x}} \quad \text{for } x > 0$$

med hensyn til Lebesgue målet. Udtrykt ved  $\theta = -\frac{1}{2\alpha^2}$  har  $X$  tæthed

$$e^{\sqrt{-2\theta}} e^{\theta x} \quad \text{for } x \in \mathbb{R}^+$$

med hensyn til  $\mu$ . Der er altså tale om en en-dimensional eksponential familie på  $\mathbb{R}^+$  med kanonisk stikprøvefunktion  $t(x) = x$ , grundmål  $\mu$  og normeringskonstant

$$c(\theta) = e^{-\sqrt{-2\theta}}.$$

3. Lad  $X$  have tæthed  $f_\alpha$ . Argumenter for at  $X$  har momenter af enhver orden. Find middelværdi og varians af  $X$ , både som funktion af  $\theta$  og som funktion af  $\alpha$ .

#### Vejledende besvarelse

Ifølge Lemma 2.19 har  $t(X) = X$  momenter af enhver orden. Bemærk at Lemma 2.19 udtaler sig om momenter af  $t(X)$ , ikke om momenter af  $X$ . Det er derfor vigtigt at den kanoniske stikprøvefunktion er identitetsfunktionen.

Ifølge (2.15) (eller formlerne lige over, eller Lemma 2.20) er

$$E(t(X)) = E(X) = \frac{d}{d\theta} \log c(\theta) = \frac{1}{\sqrt{-2\theta}} = \alpha$$

og

$$\text{Var}(t(X)) = \text{Var}(X) = \frac{d^2}{d\theta^2} \log c(\theta) = \frac{1}{(\sqrt{-2\theta})^3} = \alpha^3$$

4. Opskriv likelihoodligningen for  $\theta$ . Find maksimaliseringestimatoren for  $\theta$ . Find maksimaliseringestimatoren for  $\alpha$ .

#### Vejledende besvarelse

Likelihoodligningen for  $\theta$  er givet i (5.6), det følger derfor fra resultaterne i (c) at likelihoodligningen er

$$\frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{\sqrt{-2\theta}}$$

Løses denne fås maksimaliseringestimatoren for  $\theta$ :

$$\hat{\theta} = -\frac{1}{2 \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2}$$

Maksimaliseringestimatoren for  $\alpha$ :

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^n X_i$$

5. Gør rede for at  $\hat{\theta}$  er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling, parametreret ved  $\theta$ .

**Vejledende besvarelse**

Da  $\text{Var}(X_1) > 0$  benytter vi direkte nederste formel på side 187, og får at

$$\hat{\theta} \stackrel{as}{\sim} N\left(\theta, \frac{1}{n}(\sqrt{-2\theta})^3\right)$$

Man kan også gå en lille omvej, og bruge at ifølge CLT, Sætning 5.11, er

$$\frac{1}{n} \sum_{i=1}^n X_i \stackrel{as}{\sim} N\left(\frac{1}{\sqrt{-2\theta}}, \frac{1}{n(\sqrt{-2\theta})^3}\right)$$

Vi har at

$$\hat{\theta} = f\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \quad \text{hvor} \quad f(x) = -\frac{1}{2x^2}.$$

For  $x > 0$  er  $f$  differentiel, herunder specielt for  $x = E(X_1)$ . Vi har  $f'(x) = \frac{1}{x^3}$  og  $f'(E(X_1)) = (\sqrt{-2\theta})^3$ . Vi benytter deltametoden, og får

$$\begin{aligned} \hat{\theta} = f\left(\frac{1}{n} \sum_{i=1}^n X_i\right) &\stackrel{as}{\sim} N\left(f\left(\frac{1}{\sqrt{-2\theta}}\right), \frac{1}{n} \left((\sqrt{-2\theta})^3\right)^2 \frac{1}{(\sqrt{-2\theta})^3}\right) \\ &= N\left(\theta, \frac{1}{n}(\sqrt{-2\theta})^3\right) \end{aligned}$$

som før.

## Opgave 3

I et studie<sup>1</sup> undersøgte man om der var forskel på mænd og kvinders moral i den amerikanske kystvagt. Specifikt noterede man en score for hver person udfra den såkaldte *Rest's Defining Issues Test (DIT)*, hvor en højere score indikerer en højere moral.

Studiet involverede 225 personer fordelt på 4 grupper efter køn og rang. Tabel 1 viser

		Rang	
		Officer	Menig
Køn	Mand	60	120
	Kvinde	15	30

Tabel 1: Antalstabell for studiet omkring mænd og kvinders moral i US Coast Guard.

antallet af personer i hver af de fire grupper. Data til besvarelse af opgaven findes på USB nøglen med filnavnet **moral.txt**.

Betragt faktorerne  $K = \{\text{mand, kvinde}\}$  og  $R = \{\text{officer, menig}\}$  med hver to niveauer.

1. Gør rede for at de to faktorer  $K$  og  $R$  er geometrisk ortogonale og angiv hvorvidt designet

$$\mathbb{G} = \{K, R, K \times R\}$$

er ortogonalt og minimums-stabilt.

### Vejledende besvarelse

Tabel 1 er præcis antals-tabellen for de to faktorer  $K$  og  $R$ . Indsætter vi række- og søjle-sum har vi

		Rang		
		Officer	Menig	
Køn	Mand	60	120	180
	Kvinde	15	30	45
		75	150	225

<sup>1</sup>Kilde: R.D. White, Jr. (1999). "Are Women More Ethical? Recent Findings on the Effects of Gender Under Moral Development," Journal of Public Administration Research and Theory, Vol. 9, #3, pp.459-471

og da

$$\begin{aligned}\frac{180 \cdot 75}{225} &= 60 \\ \frac{180 \cdot 150}{225} &= 120 \\ \frac{45 \cdot 75}{225} &= 15 \\ \frac{45 \cdot 150}{225} &= 30\end{aligned}$$

ser vi at balance-ligningen er opfyldt (her kan henvises til enten Lemma 13.11 da  $K \wedge R = 1$ , eller den mere generelle Sætning 14.8). Dermed er  $K \perp_R G$ . Bemærk også at  $R, K \leq K \times R$ , dvs.  $K, R \perp_G K \times R$ .

Designet  $\mathbb{G}$  er derfor ortogonalt (de indbyrdes faktorer er geometrisk ortogonale), men det er *ikke* minimums-stabilt. Der mangler en minimumsfaktor:  $K \wedge R = 1$ , dvs.

$$\mathbb{G}' = \{1, K, R, K \times R\}$$

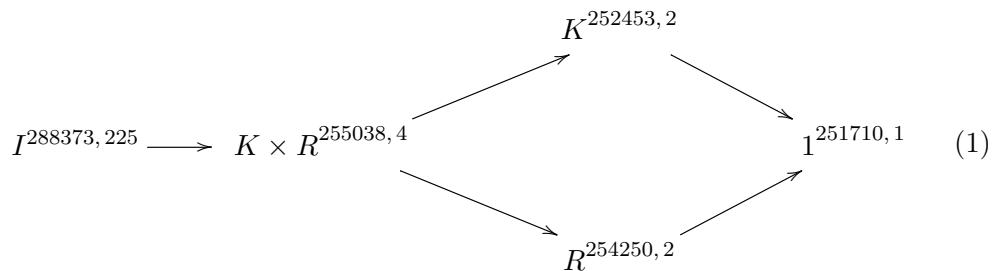
er både  $\wedge$ -stabilt (og ortogonalt).

2. Find  $\dim(V_G)$  og  $\|Q_G X\|^2$  for hver faktor  $G$  i faktorstrukturdiagrammet hvor  $V_G$  er den ortogonale dekomposition

$$V_G = L_G \ominus \sum_{G' \in \mathbb{G}, G' < G} L_{G'}, \quad \text{for hvert } G \in \mathbb{G}$$

med tilhørende projektion  $Q_G$ , og angiv  $\dim L_K + L_R$ .

Værdierne i diagrammet (1) kan benyttes.



### Vejledende besvarelse

Sætning 14.21 (ortogonal dekomposition) benyttes til at finde  $V_G$  og  $\dim(V_G)$ , hvor

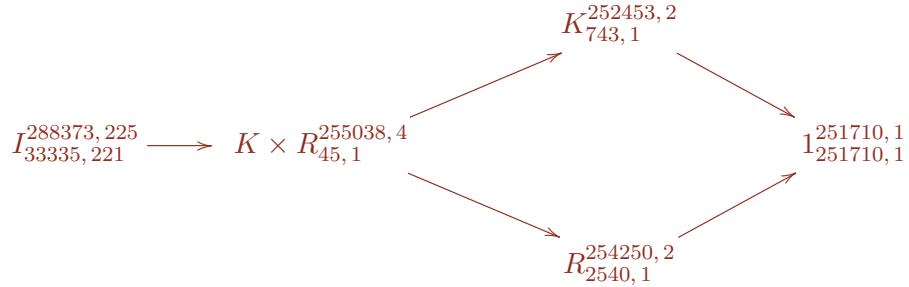
der startes bagfra:

$$\begin{aligned}
 \dim(V_1) &= \dim(L_1) \\
 \dim(V_K) &= \dim(L_K) - \dim(V_1) \\
 \dim(V_R) &= \dim(L_R) - \dim(V_1) \\
 \dim(V_{K \times R}) &= \dim(L_{K \times R}) - \dim(V_K) - \dim(V_R) - \dim(V_1) \\
 \dim(V_I) &= \dim(L_I) - \dim(V_{K \times R}) - \dim(V_K) - \dim(V_R) - \dim(V_1)
 \end{aligned}$$

tilsvarende for projektionerne  $\|Q_G X\|^2$ :

$$\begin{aligned}
 \|Q_1 X\|^2 &= \|X\|^2 \\
 \|Q_K X\|^2 &= \|P_K X\|^2 - \|Q_1 X\|^2 \\
 \|Q_R X\|^2 &= \|P_R X\|^2 - \|Q_1 X\|^2 \\
 \|Q_{K \times R} X\|^2 &= \|P_{K \times R} X\|^2 - \|Q_K X\|^2 - \|Q_R X\|^2 - \|Q_1 X\|^2 \\
 \|Q_I X\|^2 &= \|P_I X\|^2 - \|Q_{G \times R} X\|^2 - \|Q_G X\|^2 - \|Q_R X\|^2 - \|Q_1 X\|^2
 \end{aligned}$$

heraf får vi



Derudover ser vi at  $\dim L_{K+R} = \dim L_R + \dim L_R - \dim L_{K \wedge R} = 2 + 2 - 1 = 3$ .

3. Idet vi antager en lineær normal model

$$X \sim \mathcal{N}(\xi, \sigma^2 I_{225})$$

hvor  $I_{225}$  betegner den 225-dimensionale identitetsmatrix, udfør  $F$ -testet der sammenligner modellerne  $\xi \in L_{K \times R}$  og  $\xi \in L_K + L_R$ , udfra projektionerne  $Q_G$  fra spørgsmål 2).

**Vejledende besvarelse**

Vi stiller  $F$ -testet op som i formel (10.31) og får

$$\begin{aligned} F_{\text{test}} &= \frac{\|P_{G \times R} - P_{G+R}\|^2 / (\dim L_{G \times R} - \dim L_{G+R})}{(\|x\|^2 - \|P_{G \times R} - x\|^2) / (N - \dim L_{G \times R})} \\ &= \frac{\|Q_{K \times R}X\|^2 / (\dim V_{K \times R})}{\|Q_I X\|^2 / \dim(V_I)} \\ &= \frac{45/1}{33335/221} \sim 0.298 \sim F_{1,221} \end{aligned}$$

Det giver en  $p$ -værdi på  $1 - \text{pf}(0.298, 1, 221) = 0.586$ .

4. Angiv de estimerede middelværdier for de fire grupper, estimeret ud fra modellen

$$X \sim \mathcal{N}(\xi, \sigma^2 I_{225}),$$

hvor  $\xi \in L_{K \times R}$

og undersøg om residualerne kan antages at følge en normal-fordeling med konstant varians.

### Vejledende besvarelse

Data kan indlæses i R med kommandoen

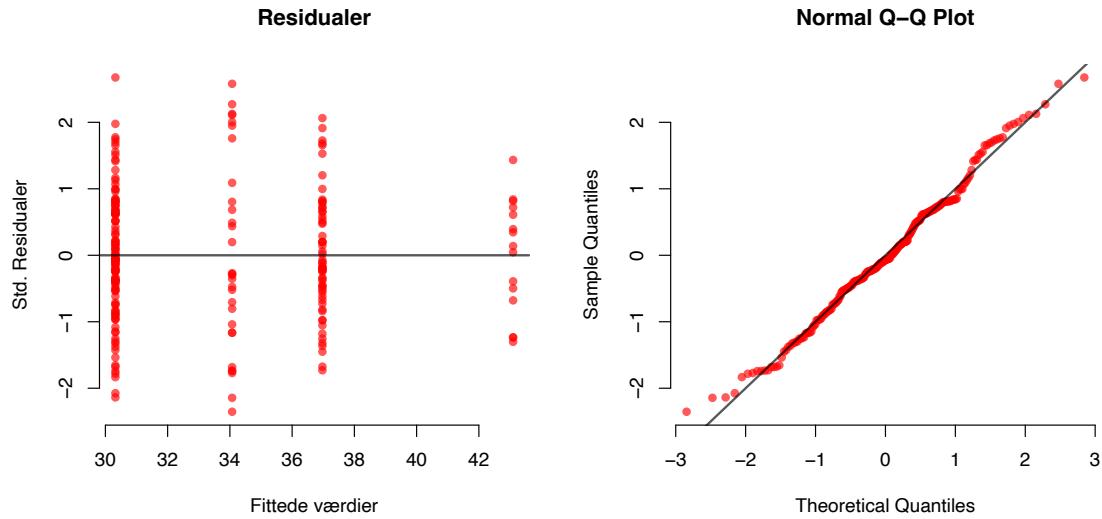
```
read.delim("moral.txt", header=TRUE)
```

Herefter kan modellen fittes i R med kommandoen

```
m0 = lm(score~koen*rang, data=moral)
```

og `summary(m0)` giver nu parametrene, mm.

Nedenfor ses de standardiserede residualer plottet mod de fittede værdier,  $\hat{\xi}$ , og et fraktil-plot (mod standard normal-fordelingen). Det giver ingen anledning til bekymring, dog ses at variansen er en smule mindre for kvindelige officerer, men det er ikke noget vi vil bekymre os mere om da der her også er færre observationer til at estimere variationen. Fraktil-plottet er overbevisende påt i forhold til identitetslinien. Alt efter hvordan modellen er parametreret og hvilke grupper er valgt som reference, vil sammensætningen af outputtet variere. Her benyttes referencerne 'kvinde' og 'menig' og modellen er parametreret med kontraster, dvs. inklusiv (`intercept`)-parameteren. `summary`-kommandoen giver flg. estimater for modellen med interaktion ( $K \times R$ ):



Parameter	Estimat
(Intercept)	34.073
koenmand	-3.751
rangofficer	9.027
koenmand:rangofficer	-2.376

Heraf fås (med de angivne referencegrupper) de estimerede gruppe-middelværdier:

$$\begin{aligned}
 \alpha_{kvinde,menig} &= 34.073 & = 34.073 \\
 \alpha_{kvinde,officer} &= 34.073 + 9.027 & = 43.100 \\
 \alpha_{mand,menig} &= 34.073 - 3.751 & = 30.322 \\
 \alpha_{mand,officer} &= 34.073 - 3.751 + 9.027 - 2.376 & = 36.973
 \end{aligned}$$

5. Test om det kan antages at der er vekselvirkning mellem faktorerne  $K$  og  $R$  med et signifikans niveau på 5% og forklar i ord hvad resultatet fortæller.

### Vejledende besvarelse

Der kan benyttes en af 3 fremgangsmetoder:

- Fra `summary(m0)` aflæses  $t$ -testet for at fjerne interaktionen til -0.547 med tilhørende  $p$ -værdi: 0.585.
- Alternativt fittes modellen under  $H_0$  i R med kommandoen `m1 = lm(score~koen+rang,`

`data=moral`) hvorved `anova(m1,m0)` kan bruges til at finde

$$F_{\text{test}} = 0.2994 \sim F_{1,221}$$

hvilket giver en  $p$ -værdi på 0.585 (tilsvarende  $t$ -testet ovenfor).

- (c) Endeligt kan man benytte  $F$ -testet fra spørgsmål 3) hvor man evt. mangler at beregne  $p$ -værdien:  $1 - \text{pf}(0.298, 1, 221) = 0.586$ . Der er en mindre forskel til `anova(...)` og  $t$ -testet pga. afrunding.

Vi ser at  $H_0$  ej forkastes og vi kan derfor antage at der ikke er interaktion mellem  $K$  og  $R$ . Altså indeholder modellen effekter af både køn og rang, men effekterne kan antages uafhængige af hinanden.

6. Antag nu modellen  $\xi \in L_K + L_R$ , dvs. modellen under  $H_0$  i forrige spørgsmål, uanset din konklusion for  $H_0$ . Undersøg hvorvidt denne model kan reduceres yderligere.

#### Vejledende besvarelse

Ved at benytte `summary(m1)`, hvor `m1` er modellen estimeret som ovenfor, ses at alle parametrene er signifikante. Dog er  $p$ -værdien for at fjerne  $K$  lig med 0.027, dvs. parameteren er signifikant på et 5% signifikans niveau, men havde man valgt et signifikans-niveau på 1% ville man have kunnet fjerne den. Det er altså tale om en borderline-case.

7. Angiv konfidensintervaller for parametrene i modellen hvor  $\xi \in L_K + L_R$  og beskriv hvad den fortæller om moralen for mænd og kvinder, samt menige og officerer i den amerikanske kystvagt.

### Vejledende besvarelse

Vi ender i en model med additiv virkning af faktoerne  $K$  og  $R$ . Denne model kan ikke reduceres yderligere på et 5% konfidens niveau. Vi ender derfor med 3 parametre: **(intercept)** (for referencegruppen: {kvinde, menig}) samt to kontrast parametre: en for køn en for rang.

Parameter	Fit med intercept			Fit uden intercept		
	Estimat	2.5 %	97.5 %	Estimat	2.5 %	97.5 %
(Intercept)	34.71	30.93	38.48	34.71	30.93	38.48
koenmand	-4.54	-8.57	-0.52	30.16	28.03	32.29
rangofficer	7.13	3.71	10.54	7.13	3.71	10.54

Modellen fortæller os at den gennemsnitlige score for en menig kvinde er 34.71 ([30.93; 38.48]), svarende til **(intercept)**-parameteren. Derudover ser vi at officerer scorer noget højere, gennemsnitligt 7.13 ([3.71; 10.54]) point mere end menige. Endeligt ser vi at mænd generelt scorer en forskel på -4.54 ([-8.57; -0.52]) point i forhold til kvinder. Vi kan derfor konkludere på baggrund af modellen at der rent faktisk *er* en signifikant (på 5%) forskel i mænd og kvinders moral i den amerikanske kystvagt.

## Opgave 4

I et eksperiment med grisekød har man undersøgt hvordan kødets farve (intensitet af rød) falmer over tid som følge af to forskellige opbevaringer:  $B = \{\text{lyst, mørkt}\}$ . Da farven på kød indikerer hvor attraktivt kødet vurderes af forbrugerne, er man interesseret i hvordan de to behandlinger påvirker kødets farve, og specielt hvorvidt der er forskel på de to behandlinger.

I eksperimentet målte man på 10 slagtede grise. Fra hver gris tog man 6 stykker kød, hvoraf 3 blev lagret lyst i hhv. 1,4 og 6 dage og 3 blev lagret mørkt i hhv. 1,4 og 6 dage. Dette giver i alt  $10 \times 3 \times 2 = 60$  observationer. Tabel 2 viser designet for hver gris.

Opbevaring	1 dag	4 dage	6 dage
Lyst	Stykke 1	Stykke 2	Stykke 3
Mørkt	Stykke 4	Stykke 5	Stykke 6

Tabel 2: Fordeling af kødstykker fra hver gris.

Figur 1 viser et plot af observationerne fordelt på de to opbevaringsmetoder.

Som det fremgår af Tabel 3 er der i alt 3 faktorer i spil.

Faktor	Betydning	Niveauer
$G$	Gris	$\{1, \dots, 10\}$
$T$	Tid	$\{1, 4, 6\}$
$B$	Opbevaring	$\{\text{lyst, mørkt}\}$

Tabel 3: Faktorer i eksperimentet med grisekød.

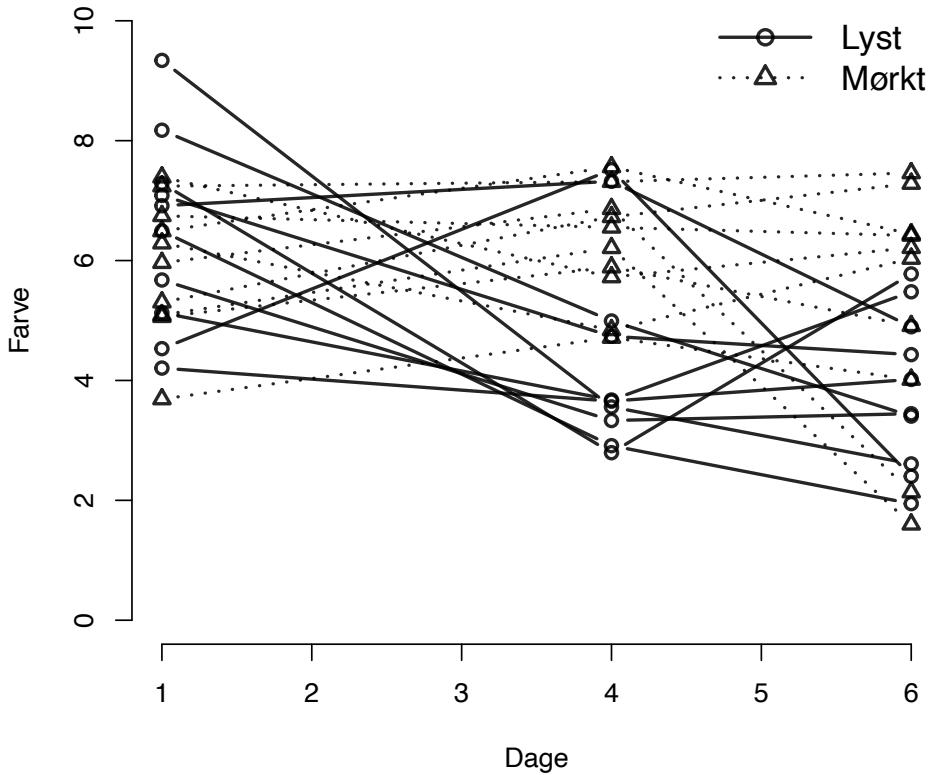
Betræt  $G$  som tilfældig effekt og  $T, B$  som faste effekter. Betræt derudover  $T$  (tid) som faktor og ej numerisk.

Data til besvarelse af opgaven findes på USB nøglen med filnavnet `koed.txt`.

Lad  $i = 1, 2, j = 1, 2, 3, k = 1, \dots, 10$  angive  $i$ 'te behandling (opbevaring) til  $j$ 'te tidspunkt for den  $k$ 'te gris.

- Opskriv en varianskomponentmodel for rødheden af kødet, med  $T \times B$  (vekselvirkning mellem  $T$  og  $B$ ) som fast effekt og  $G$  som tilfældigt intercept. Angiv variansmatricen for en vilkårlig gris udtrykt ved hjælp af de teoretiske parametre i modellen.

### Vejledende besvarelse



Figur 1: Effekt af 2 forskellige opbevaringer af grisekød til 3 forskellige tidspunkter.

Modellen har interaktion mellem  $T \times B$  der har 6 niveauer. Vi indekserer som flg.

Opbevaring:  $i = 1, 2$

Tid:  $j = 1, 2, 3$

Gris:  $k = 1, \dots, 10$

og skriver da

$$\begin{aligned} X_{ijk} &= \alpha_{ij} + Y_k + \varepsilon_{ij} \\ Y_k &\sim \mathcal{N}(0, \nu^2) \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

hvor  $\alpha_{ij}$  betegner kombinationen af den  $i$ 'te behandling og  $j$ 'te tidspunkt.  $Y_k$  angiver det tilfældige intercept for den  $k$ 'te gris.

Idet vi lader  $1_n$  og  $1_{m \times n}$  betegne hhv. en  $n$ -dimensional vektor at 1-taller og en  $m \times n$  matrix at 1-taller, kan vi skrive effektmatricen tilhørende effektparret  $(G, 1)$ , dvs. den

tilfældige effekt på interceptet, som

$$B = \begin{pmatrix} 1_6 & 0 & \dots & 0 \\ 0 & 1_6 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1_6 \end{pmatrix}$$

hvilket giver at

$$BB^T = \begin{pmatrix} 1_{6 \times 6} & 0 & \dots & 0 \\ 0 & 1_{6 \times 6} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1_{6 \times 6} \end{pmatrix}$$

Vi ser derfor at for en vilkårlig gris bliver kovariansstrukturen en  $6 \times 6$  matrix på formen

$$\text{Cov}(X_{ijk}, X_{i'j'k'}) = \sigma^2 I_6 + \nu^2 1_{6 \times 6} = \begin{pmatrix} \sigma^2 + \nu^2 & \nu^2 & \dots & \nu^2 \\ \nu^2 & \sigma^2 + \nu^2 & \dots & \nu^2 \\ \vdots & \vdots & \ddots & \vdots \\ \nu^2 & \nu^2 & \dots & \sigma^2 + \nu^2 \end{pmatrix}, \quad \text{for } k = k'$$

- Angiv estimaterne for variansparametrene i modellen fra spørgsmål 1). Husk at variablerne **gris** og **tid** skal betragtes som faktorer og ej numeriske!

#### Vejledende besvarelse

Data indlæses med `read.delim("data/grise.txt")`. Husk at lave variablene **gris** og **tid** om til faktorer

```
koed$gris = as.factor(koed$gris)
koed$tid = as.factor(koed$tid)
```

herefter kan modellen fittes ved

```
m0 = lmer(farve ~ opbevaring * tid + (1 | gris), data=koed)
```

`summary(m0)` angiver parametrene, heraf fås

$$\begin{aligned} \sigma^2 &= 1.9502 \\ \nu^2 &= 0.2727 \end{aligned}$$

3. Angiv de estimerede middelværdier for rødheden af kød for de tre  $B \times T$ -niveauer:

$\{\text{lyst, 1 dag}\}$   
 $\{\text{mørkt, 1 dag}\}$   
 $\{\text{mørkt, 6 dage}\}$

udfra estimaterne fra modellen fra spørgsmål 1). Bemærk at der spørges til estimerede middelværdier for niveauerne i produktfaktoren, snarere end parameterestimater. Sidstnævnte afhænger af den parametrisering du har valgt, men det gør middelværdierne ikke!

#### Vejledende besvarelse

Som før angiver `summary(m0)` parametrene (husk  $ij$  betegner den  $i$ 'te behandling til  $j$ 'te tidspunkt. Her er reference grupperne for de to faktorer  $B$  og  $T$  hhv.  $\{\text{lyst}\}$  og  $\{1\}$ , dvs. (`intercept`) svarer til  $B \times T$  gruppen  $\{\text{lyst}, 1\}$ . Heraf ses flg. niveauer

$$\begin{array}{ll}
 \alpha_{11} = 6.5885 & = 6.5885 \\
 \alpha_{12} = 6.5885 - 2.1508 & = 4.4377 \\
 \alpha_{13} = 6.5885 - 2.7813 & = 3.8072 \\
 \alpha_{21} = 6.5885 - 0.6933 & = 5.8952 \\
 \alpha_{22} = 6.5885 - 0.6933 - 2.1508 + 2.4769 & = 6.2213 \\
 \alpha_{23} = 6.5885 - 0.6933 - 2.7813 + 2.1438 & = 5.2577
 \end{array}$$

De tre der spørges til er hhv.  $\alpha_{11} = 6.5885$ ,  $\alpha_{21} = 5.8952$  og  $\alpha_{23} = 5.2577$ .

4. Opskriv konfidensintervaller for middelværdierne tilhørende de faste effekter i modellen med vekselvirkning, dvs. hvor  $\xi \in L_{T \times B}$ . Kommenter på effekten af de to behandlinger.

#### Vejledende besvarelse

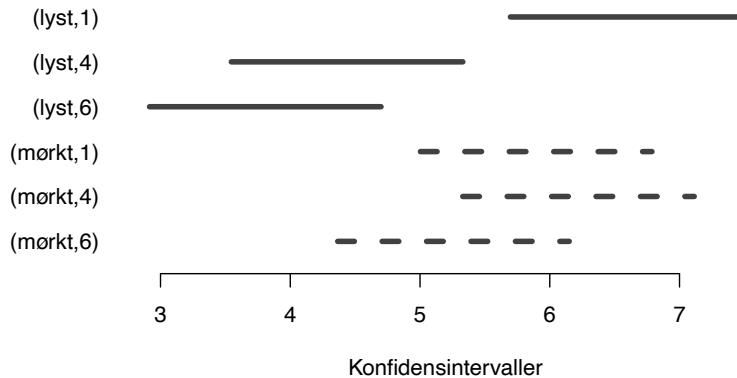
Her benyttes `confint(...)` til at estimere konfidensintervallerne. Vi omparametriserer modellen så de faste effekter svarer til niveauerne i de 6 grupper af  $B \times T$  faktoren:

```
m0.1 = lmer(farve ~ opbevaring : tid - 1 + (1 | gris), data = koed)
```

Det giver flg. konfidensintervaller for de faste effekter (med `confint(m0.1)`):

	2.5 %	97.5 %
opbevaringlyst:tid1	5.70	7.48
opbevaringlyst:tid4	3.55	5.33
opbevaringlyst:tid6	2.92	4.70
opbevaringmoerkt:tid1	5.00	6.79
opbevaringmoerkt:tid4	5.33	7.11
opbevaringmoerkt:tid6	4.37	6.15

Det ses at intervallerne for  $B \times T$  grupperne  $\{\text{mørkt}, i\}$ ,  $i = 1, 2, 3$  overlapper hinanden, hvorimode  $\{\text{lyst}, i\}$ ,  $i = 1, 2, 3$  ikke gør: de to intervaller for grupperne til tid 4 og 6 er signifikant lavere end tid 1. Plotter vi konfidens intervallerne er tendensen endnu tydeligere



Det konkluderes derfor at kød der opbevares lyst ser ud til at forringes i farve kvalitet over de tre tidsperioder, mens kød der opbevares mørkt ikke viser samme trend.

5. Angiv et forslag til at teste hvorvidt tid har nogen effekt når kødet er opbevaret mørkt. Du behøver ikke at udføre testet.

#### Vejledende besvarelse

Vi kan definere en ny faktor der er hhv. 0 for kød der er opbevaret mørkt og 1 for kød der er opbevaret lyst. Benytter vi faktoren som numerisk (dummy-variabel), kan vi derfor lave interaktionen med `tid`, således at den kun har effekt for gruppen af kød der er opbevaret lyst.

Vi kan lave dummy variablen ved følgende kommando i R:

```
koed$opbevaring2 = 1*(koed$opbevaring == "lyst")
```

og kan nu fitte en modellen med den numeriske dummy `opbevaring2` som

```
m2 = lmer(farve~tid:opbevaring2+(1|gris), data=koed)
```

Denne model har interaktion mellem `tid` og `opbevaring2`, men da `opbevaring2` er en dummy variabel, og  $\{\text{mørkt}\}$  gruppen bliver reference i  $B$  faktoren vil (`intercept`) parameteren svare til niveauet for  $\{\text{mørkt}\}$  gruppen. Interaktionen gør at kontrasterne til  $\{\text{lyst}\}$  gruppen er forskellige til hvert tidspunkt. Modellen har derfor 4 parametre for de faste effekter:

Parameter	Gruppe	Estimat
(intercept)	{mørkt, $i\}$ , $i = 1, 2, 3$	5.7914
tid1:opbevaring2	{lyst, 1}	0.7971
tid4:opbevaring2	{lyst, 2}	-1.3537
tid6:opbevaring2	{lyst, 3}	-1.9842

Bemærk at parametrene til {lyst} gruppen er kontraster, dvs. (intercept) parameteren skal lægges til hver af de tre {lyst} grupper, dvs. {mørkt} gruppen har et konstant niveau på 5.79, mens {lyst} gruppens niveauer aftager over tid.

Et `anova(m2,m0)` test (dvs. i forhold til startmodellen) giver en  $p$ -værdi på 0.2638. Ser vi på konfidensintervallerne for de faste parametre, ses det at det kan antages at kød der opbevares 1 dag kan antages at have samme niveau for de to opbevaringsmetoder, eftersom `tid1:opbevaring2` ikke signifikant forskellig fra 0, der ligger i intervallet.

	2.5 %	97.5 %
(Intercept)	5.19	6.39
tid1:opbevaring2	-0.20	1.79
tid4:opbevaring2	-2.35	-0.36
tid6:opbevaring2	-2.98	-0.99

Man kunne evt. undersøge nærmere om der er forskel på kød der har ligget lyst i hhv. 4 og 6 dage. I det tilfælde må det antages at farven aftager ikke-lineært over tid, men med kun 3 tidspunkter kan det være svært at afgøre en mere konkret effekt af kombinationen lyst opbevaret kød og tid.

## Eksamens i Statistik 1 — Vejledende besvarelse

### 11. april 2019

Eksamens varer 4 timer. Alle hjælpemidler er tilladt under hele eksamen, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant. Eksamenssættet består af tre opgaver med i alt 13 delspørger; alle delspørger vægtes ens i bedømmelsen.

### Opgave 1

Betræt den negative binomialfordeling med antalsparameter 2 og sandsynlighedsparameter  $p \in (0, 1)$ , dvs. fordelingen der har tæthed

$$f_p(x) = (x+1)p^2(1-p)^x, \quad x \in \mathbb{N}_0$$

med hensyn til tællemålet på  $\mathbb{N}_0$ .

Det kan uden bevis benyttes at  $f_p$  faktisk er en tæthed for alle  $p \in (0, 1)$ , og at en stokastisk variabel  $X$  med tæthed  $f_p$  har middelværdi og varians givet ved

$$\mathbb{E}_p X = \frac{2(1-p)}{p}, \quad \mathbb{V}_p X = \frac{2(1-p)}{p^2}.$$

Lad  $X_1, \dots, X_n$  være uafhængige stokastiske variable der alle har fordeling givet ved tæthed  $f_p$  med ukendt  $p \in (0, 1)$ . Definer desuden  $X_+ = \sum_{i=1}^n X_i$ .

Q1: Opskriv log-likelihoodfunktionen, scorefunktionen og den observerede informationsfunktion. Bestem derefter Fisherinformationen.

For  $x \in \mathbb{N}_0^n$  og  $x_+ = \sum x_i$  er de ønskede funktioner givet ved (unødvendige multiplikative eller additive konstanter er udeladt):

$$\begin{aligned} L_x(p) &= p^{2n}(1-p)^{x_+} \\ l_x(p) &= -2n \log p - x_+ \log(1-p) \\ Dl_x(p) &= -\frac{2n}{p} + \frac{x_+}{1-p} \\ D^2l_x(p) &= \frac{2n}{p^2} + \frac{x_+}{(1-p)^2} \end{aligned}$$

Derefter fås Fisherinformationen

$$i(p) = \mathbb{E}_p D^2 l_X(p) = \frac{2n}{p^2} + \frac{\mathbb{E} X_+}{(1-p)^2} = \frac{2n}{p^2} + \frac{2n(1-p)}{p(1-p)^2} = \frac{2n}{p^2(1-p)}$$

Q2: Gør rede for at hvis  $X_+ > 0$ , så er

$$\hat{p} = \frac{2n}{X.+2n} \quad (1)$$

en entydig maximum likelihood estimator for  $p$ . Gør desuden rede for at formlen for  $\hat{p}$  også giver mening når  $X_+ = 0$ .

*Vink til sidste del:* Hvilken fordeling svarer værdien  $p = 1$  til (når  $0^0$  defineres til 1)?

Vi løser scoreligningen:

$$Dl_x(p) = 0 \Leftrightarrow \frac{2n}{p} = \frac{x.}{1-p} \Leftrightarrow p = \frac{2n}{x.+2n}$$

Når  $x. > 0$  ligger dette  $p$  i mængden  $(0, 1)$ . Da  $D^2l_x(p) > 0$  for alle  $p \in (0, 1)$  og  $(0, 1)$  er en åben mængde, får vi derfor at  $l_x$  har entydigt minimum for dette  $p$  når  $x. > 0$ . Altså: ML estimatoren er

$$p = \frac{2n}{X.+2n}$$

når  $X. > 0$ .

Hvis  $x. = 0$  giver formlen at  $p = 1$  der strengt taget ligger udenfor parameterområdet. Men det giver god mening fordi fordelingen svarende til  $p = 1$  er et punktmålet (den udartede fordeling) i 0: Hvis alle  $x_i$  er 0, er det et fornuftigt bud at fordelingen er denne udartede fordeling, altså et godt bud at  $p = 1$ . Faktisk maksimerer  $p = 1$  likelihoodfunktionen over intervallet  $(0, 1]$  i tilfældet  $x. = 0$ , thi så er  $L_x(p) = p^2$ .

I det følgende er  $\hat{p}$  defineret ved (1) uanset om  $X_+ > 0$  eller  $X_+ = 0$ .

Q3: Vis at  $1/\hat{p}$  er central for  $1/p$ , men at  $\hat{p}$  ikke er central for  $p$ .

*Vink:* Benyt Jensens ulighed.

Vi har  $1/\hat{p} = (X.+2n)/(2n)$  så

$$E_p\left(\frac{1}{\hat{p}}\right) = \frac{E_p X.+2n}{2n} = \frac{E_p X.}{2n} + 1 = \frac{2n(1-p)}{2np} + 1 = \frac{1}{p}$$

for alle  $p \in (0, 1)$ , så  $1/\hat{p}$  er central for  $1/p$ .

Funktionen  $t : x \rightarrow 1/x$  er strengt konveks på  $(0, \infty)$  thi  $t''(x) = 2x^{-3} > 0$  for alle  $x > 0$ . Endvidere har  $1/\hat{p}$  middelværdi (se ovenfor) og  $\hat{p}$  har middelværdi da den ligger i et begrænset interval. Jensens ulighed giver derfor at

$$E_p \hat{p} = E_p t(1/\hat{p}) \geq t(E_p(1/\hat{p})) = t(1/p) = p$$

med lighedstegn hvis og kun hvis fordelingen af  $1/\hat{p}$  er udartet — men det er den ikke når  $p \in (0, 1)$ . Således er  $\hat{p}$  ikke central for  $p$ .

Q4: Benyt “den falske Wald-teststørrelse” til at bestemme et approksimativt 95% konfidensinterval for  $p$ . Du behøver ikke at gøre rede for forudsætningerne for den asymptotiske fordeling af Wald-teststørrelsen.

Den falske Wald-teststørrelse er givet ved

$$W = (\hat{p} - p) i(\hat{p})(\hat{p} - p) = (\hat{p} - p)^2 \frac{2n}{\hat{p}^2(1-\hat{p})}$$

og er approksimativt  $\chi_1^2$  fordelt når  $n$  er stor.

Et approksimativt 95% konfidensinterval er derfor givet ved

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}^2(1-\hat{p})}{2n}}.$$

- Q5: En studerende lægger kabale hver aften i en uge. Hver aften lægger hun kabalen indtil den er gået op to gange, og skriver ned hvor mange gange kabalen *ikke* er gået op.

Hun får følgende observationer:

$$0 \quad 12 \quad 8 \quad 3 \quad 7 \quad 9 \quad 14 \tag{2}$$

Observationen 0 svarer altså til at kabelen gik op i de to første forsøg, mens observationen 12 svarer til at hun måtte lægge kabalen i alt 14 gange den pågældende aften.

Gør rede for at situationen naturligt kan beskrives med den statistiske model fra denne opgave, hvor  $p$  er sandsynligheden for at kabelen går op når den studerende lægger den en enkelt gang. Beregn derefter et estimat og et approksimativt 95% konfidensinterval for  $p$  baseret på observationerne i (2).

Lad  $X_1, \dots, X_7$  være de stokastiske variable svarende til de syv dage. Hvis den studerende ikke lærer med tiden, kan de antages at være uafhængige og identisk fordelte.

Hver dag antages de enkelte kabaler at gå op uafhængigt af hinanden og med samme sandsynlighed,  $p$ . At  $X_i = x$  betyder at den studerende har måttet lægge kabalen  $x + 2$  gange. Heraf gik kabelen op sidste gang, og desuden netop en ud af de første  $x + 1$  gange. Der er  $x + 1$  forskellige muligheder for dette, som alle har sandsynlighed  $p^2(1-p)^x$ . Vi får derfor  $P(X_i = x) = (x+1)p^2(1-p)^x$  som ønsket.

Vi indsætter således blot i formlerne fra spørgsmål 2 og 4, og får:  $x. = 53$ ,  $\hat{p} = 0.209$  og det approksimative 95% konfidensinterval  $(0.112, 0.306)$ .

## Opgave 2

Betrægt fordelingen  $P_\lambda$  der har tæthed

$$f_\lambda(x) = \frac{x}{\lambda} e^{-x^2/(2\lambda)} \cdot 1_{(0,\infty)}(x)$$

med hensyn til Lebesguemålet på  $\mathbb{R}$ . Fordelingen er bestemt af parameteren  $\lambda > 0$ .

Lad  $X_1, \dots, X_n$  være uafhængige reelle stokastiske variable, der alle har fordeling  $P_\lambda$  med ukendt  $\lambda > 0$ .

- Q1: Gør rede for at familien  $\mathcal{P} = \{P_\lambda, \lambda > 0\}$  er en eksponentiel familie og bestem familiens grundmål, kanoniske stikprøvefunktion, og kanoniske parameter.

Vi sætter  $\theta = 1/\lambda$  og får

$$f_\theta(x) = \theta x e^{-\theta x^2/2} 1_{(0,\infty)}(x) = \theta e^{-\theta x^2/2} x 1_{(0,\infty)}(x)$$

som identificerer familien som en eksponentiel familie med grundmål  $\mu(dx)x \cdot \eta(dx)$  hvor  $\eta$  er Lebesguemål på  $(0, \infty)$ , kanonisk parameter  $\theta = 1/\lambda$ , kanonisk stikprøvefunktion  $t(x) = -x^2/2$ , og kumulantfunktion  $\psi(\theta) = -\log \theta$ .

Q2: Bestem maksimum likelihood estimatoren for  $\lambda$ .

I en eksponentiel familie bestemmes maksimum likelihood estimatoren ved at sætte den kanoniske stikprøvefunktion lig med sin middelværdi. Vi får for middelværdien

$$E_\lambda(-X^2/2) = \psi'(\theta) = -1/\theta = -\lambda$$

og derfor er

$$\hat{\lambda} = \frac{1}{2n} \sum_i X_i^2.$$

Q3: Bestem fordelingen af maksimaliseringestimatoren, og vis at den er en central estimator for  $\lambda$ .

Vink: Find først fordelingen af  $X_i^2$ .

Lad  $Y_i = X_i^2$ , således at

$$\hat{\lambda} = \frac{1}{2n} \sum_{i=1}^n Y_i = \frac{1}{2} \bar{Y}$$

Transformationssætningen for endimensionale transformationer giver at  $Y_i$  har tæthed

$$g_\lambda(y) = \frac{1}{2\lambda} e^{-y/(2\lambda)} \cdot 1_{(0,\infty)}(y)$$

mht. Lebesguemålet på  $\mathbb{R}$ .

Vi får derfor:

- $Y_1, \dots, Y_n$  er uafhængige og identisk exponentialfordelte med skalaparameter  $2\lambda$ , dvs. gammafordelte med formparameter 1 og skalaparameter  $2\lambda$ .
- $\sum_{i=1}^n Y_i$  er gammafordelt med formparameter  $n$  og skalaparameter  $2\lambda$  jf. foldningsegenskaben for gammafordelingen
- $\hat{\lambda}$  gammafordelt med formparameter  $n$  og skalaparameter  $\lambda/n$ .

Specielt er

$$E_\lambda \hat{\lambda} = n \frac{\lambda}{n} = \lambda$$

så  $\hat{\lambda}$  er en central estimator for  $\lambda$ .

Q4: Betrag hypotesen  $H : \lambda = \lambda_0$  for en given værdi  $\lambda_0 > 0$ . Opskriv kvotientteststørrelsen,  $q(x_1, \dots, x_n)$ , for hypotesen. Vis derefter at fordelingen af  $Q = q(X_1, \dots, X_n)$  under hypotesen ikke afhænger af den specifikke værdi af  $\lambda_0$ .

Kvitientteststørrelsen er givet ved

$$\begin{aligned} q(x) &= \frac{L_x(\lambda_0)}{L_x(\hat{\lambda})} = \frac{\hat{\lambda}^n \exp\left(-\frac{1}{2\lambda_0} \sum x_i^2\right)}{\lambda_0^n \exp\left(-\frac{1}{2\hat{\lambda}} \sum x_i^2\right)} \\ &= \left(\frac{\hat{\lambda}}{\lambda_0}\right)^n \exp\left(-\frac{1}{2\lambda_0} \sum x_i^2 + n\right) \\ &= \left(\frac{\sum x_i^2}{2n\lambda_0}\right)^n \exp\left(-\frac{1}{2\lambda_0} \sum x_i^2 + n\right) \end{aligned}$$

Den stokastiske version er

$$Q = q(X) = \left( \frac{\sum X_i^2}{2n\lambda_0} \right)^n \exp \left( -\frac{1}{2\lambda_0} \sum X_i^2 + n \right)$$

og viser at  $Q$  kun afhænger af  $(X, \lambda_0)$  gennem  $\frac{1}{2\lambda_0} \sum_{i=1}^n X_i^2$ .

Under hypotesen er  $\sum_{i=1}^n X_i^2$  gammafordelt med formparameter  $n$  og skalaparameter  $2\lambda_0$ , så  $\frac{1}{2\lambda_0} \sum_{i=1}^n X_i^2$  er gammafordelt med formparameter  $n$  og skalaparameter 1. Altså afhænger fordelingen af  $Q$  ikke af den specifikke værdi af  $\lambda_0$ .

### Opgave 3

En sygeplejerske er mistænkt for at have forgiftet et antal patienter og for at belyse dette spørgsmål laves en opgørelse over antallet af dødsfald på den afdeling, hvor hun var tjenstgørende. Opgørelsen er delt op efter om dødsfaldene var sket på en vagt lige før hun var mødt ind, under hendes vagt, eller på vagten umiddelbart efter at hun var gået hjem. Opgørelsen er fordelt på to perioder, hvor hun var ansat på den pågældende afdeling. Resultatet af opgørelsen er angivet nedenfor.

	Inden	Under	Efter
Periode A	12	32	12
Periode B	6	18	7

Under antagelse af, at antallet af dødsfald i de forskellige vagtperioder er uafhængige og Poissonfordelte ønskes det blyst, om der er særligt mange dødsfald under den pågældendes vagt, sammenlignet med andre vagtperioder.

Betrægt derfor den multiplikative Poissonmodel, altså hvor det antages, at

$$E(X_{pv}) = \alpha_p \beta_v, \quad p = A, B; \quad v = \text{Inden, Under, Efter}.$$

hvor  $X_{pv}$  er antal dødsfald på vagten  $v$  i perioden  $p$  og  $\alpha_v, \beta_p \in \mathbb{R}_+$ .

Q1: Gør rede for, at ovennævnte model er en generaliseret lineær model og angiv den tilhørende linkfunktion.

Poissonfordelingen er en velkendt dispersionsfamilie og modellen er anvender det kano-niske link  $g(\mu) = \log \mu$  hvorefter log-middelværdien specificeres til at tilhøre det lineære underrum

$$\log \mu_{pv} = \log \alpha_p + \log \beta_v = \eta_p + \gamma_v.$$

Q2: Angiv maksimum likelihood estimatoren for  $E(X) = \{E(X_{vp})\}$  under antagelse af ovennævnte model.

Data kan eventuelt indlæses i R ved at lave en tekstfil `nurse.txt` med følgende indhold

Periode	Vagt	Antal
A	Inden	12
A	Under	32
A	Efter	12
B	Inden	6
B	Under	18
B	Efter	7

og derefter køre kommandoen

```
nurse <- read.table("nurse.txt", header=TRUE)
```

Etter kommandoen

```
m<- glm(Antal~Vagt+Periode, family="poisson", data=nurse)
```

fås følgende fittede værdier

```
m$fitted.values
```

1	2	3	4	5	6
11.586207	32.183908	12.229885	6.413793	17.816092	6.770115

hvilket i tabelform svarer til

	Inden	Under	Efter
Periode A	11.59	32.18	12.23
Periode B	6.41	17.82	6.77

Q3: Kan det antages, at  $\beta_v$  ikke afhænger af vagten v?

Kommandoen

```
summary(m)
```

giver blandt andet følgende output

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )							
(Intercept)	2.50388	0.24289	10.309	< 2e-16 ***							
VagtInden	-0.05407	0.32892	-0.164	0.86943							
VagtUnder	0.96758	0.26950	3.590	0.00033 ***							
PeriodeB	-0.59136	0.22386	-2.642	0.00825 **							
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 28.577991 on 5 degrees of freedom
Residual deviance: 0.056908 on 2 degrees of freedom
```

Den meget lille residualdevians giver anledning til at have tillid til den multiplikative model. Det ses at koefficienten til "VagtUnder" er stærkt signifikant, så man kan ikke antage, at dødeligheden er ens på de tre vagttyper.

Q4: Hvad siger ovennævnte undersøgelse om den oprindelige problemstilling?

Der er konstateret en klart forøget mortalitet under den pågældendes vagter og dette skyldes ikke tilfældigheder, men må have en anden forklaring.

# Eksamens i Matematisk Statistik, 20. juni 2019

Steffen Lauritzen og Anders Tolver

Vejledende besvarelse, 8. juli 2019

## Opgave 1

1. Vi omskriver tætheden som

$$\begin{aligned} f_{(1,\lambda)}(x) &= \exp \left\{ \lambda \frac{-(x-1)^2}{2x} + \frac{1}{2} \log \lambda \right\} \frac{1}{\sqrt{2\pi x^3}} \\ &= \exp \left\{ -\lambda(x+1/x)/2 + \lambda + \frac{1}{2} \log \lambda \right\} \frac{1}{\sqrt{2\pi x^3}} \end{aligned}$$

2. Man kan fx direkte skrive log-likelihood funktionen som

$$\ell(\lambda) = -n\lambda - \frac{n}{2} \log \lambda + \frac{\lambda}{2} \sum_{i=1}^n (X_i + 1/X_i)$$

og differentiere, hvilket giver

$$S(\lambda) = -n - \frac{n}{2\lambda} + \frac{n}{2} (\bar{X}_n + \bar{Y}_n)$$

hvoraf resultatet følger.

3. Vis at informationen  $i_0(\lambda)$  for  $\lambda$  i delfamilien  $\mathcal{P}_0$  baseret på en enkelt observation er bestemt som

$$i_0(\lambda) = \frac{1}{2\lambda^2}$$

og angiv den asymptotiske fordeling af  $\hat{\lambda}_n$ .

I en eksponentiel familie som er kanonisk parametreret, er informationen lig med den anden afledede af kumulantfunktionen

$$i_0(\lambda) = \psi''(\lambda) = I(\lambda) = S'(\lambda) = \frac{1}{2\lambda^2}.$$

Den asymptotiske fordeling af  $\hat{\lambda}_n$  er  $N(\lambda, \lambda^2/2n)$ .

4. Vi omskriver eksponenten som

$$-\frac{\lambda(x-\mu)^2}{2\mu^2 x} = -\frac{\lambda}{2\mu^2} x - \frac{\lambda}{2x} + \frac{\lambda}{\mu} = \theta_1 \frac{-x}{2} + \theta_2 \frac{-1}{2x} + \sqrt{\theta_1 \theta_2}$$

og tilsvarende faktoren inden eksponentialfunktionen som

$$\sqrt{\frac{\lambda}{2\pi x^3}} = e^{\frac{1}{2}\log \theta_2} \frac{1}{2\pi x^3}$$

hvorfor hele tætheden kan skrives som

$$f_{\mu,\lambda}(x) = \exp\{\theta^T t(x) - \psi(\theta)\} \cdot \frac{1}{2\pi x^3}$$

så grundmålet er

$$d\mu(x) = \frac{1}{2\pi x^3} \mathbf{1}_{(0,\infty)} dx$$

hvor  $dx$  betegner standard Lebesgue mål.

5. Vi differentierer kumulantfunktionen og finder

$$\mathbf{E}(-X/2) = \frac{\partial}{\partial \theta_1} \psi(\theta) = -\frac{1}{2} \sqrt{\frac{\theta_2}{\theta_1}} = -\frac{\mu}{2}$$

og videre

$$\mathbf{V}(-X/2) = \frac{\partial^2}{\partial \theta_1^2} \psi(\theta) = \frac{1}{4\theta_1} \sqrt{\frac{\theta_2}{\theta_1}} = \frac{1}{4} \frac{\mu^3}{\lambda}$$

hvoraf resultatet følger.

6. I en regulær eksponentiel familie finder vi MLE ved at sætte de kanoniske stikprøvefunktioner lig med deres middelværdi. Vi har allerede fundet  $\mathbf{E}_{\mu,\lambda}(-X/2) = -\mu/2$ , hvoraf  $\tilde{\mu}_n = \bar{X}_n$ . Vi skal bruge middelværdien af den anden komponent:

$$\mathbf{E}(-1/(2X)) = \frac{\partial}{\partial \theta_2} \psi(\theta) = -\frac{1}{2} \sqrt{\frac{\theta_1}{\theta_2}} - \frac{1}{\theta_2}$$

hvoraf

$$\mathbf{E}(Y) = \mathbf{E}\left(\frac{1}{X}\right) = \frac{1}{\mu} + \frac{1}{\lambda}.$$

Indsættes  $\tilde{\mu}_n = \bar{X}_n$  fås derfor

$$\bar{Y}_n = \frac{1}{\bar{X}_n} + \frac{1}{\tilde{\lambda}_n}$$

og resultatet fremkommer nu ved at løse ligningen. Ligningen har en entydig løsning hvis og kun hvis  $\bar{X}_n \bar{Y}_n > 1$ , hvilket sker med sandsynlighed 1 hvis  $n \geq 2$  idet ikke to værdier af  $X_i$  så er ens.

7. Delfamilien svarende til  $H_0$  har dimension 1 og er en krum delfamilie af  $\mathcal{P}$  med parameteringen  $(\theta_1, \theta_2) = (\beta, \beta)$  idet  $\mu = 1 \iff \theta_1 = \theta_2$ . Derfor vil  $\Lambda_n$  være asymptotisk  $\chi^2$ -fordelt med  $2 - 1 = 1$  frihedsgrad.

## Opgave 2

1. Ifølge EH Korollar 9.43 er  $X$  regulært normalfordelt hvis og kun hvis variansen  $\Sigma$  er invertibel. Det ses at determinanten af  $\Sigma$  er lig med

$$2 \cdot 1 \cdot 2 + 1 \cdot 1 \cdot 1 + 1 \cdot 1 \cdot 1 - 2 \cdot 1 \cdot 1 - 1 \cdot 1 \cdot 2 - 1 \cdot 1 \cdot 1 = 1,$$

hvorfor  $\Sigma$  er invertibel. Vi konkluderer at  $X$  er regulært normalfordelt.

2. Det følger af EH Lemma 9.47 at  $Y = (Y_1, Y_2)^T$  er normalfordelt med middelværdi

$$EY = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

og varians

$$\Sigma_Y = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \Sigma \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Dette viser, at  $Y_1$  og  $Y_2$  er uafhængige og standardnormalfordelte. Kvadratet  $Y_2^2$  på en standardnormalfordelt variabel er per definition  $\chi^2$ -fordelt med 1 frihedsgrad.

Da  $Y_1 \sim N(0, 1)$  og uafhængig af  $Y_2^2$  som er  $\chi^2_1$ -fordelt, så er forholdet  $\frac{Y_1}{\sqrt{Y_2^2}}$  t-fordelt med 1 frihedsgrad. Denne konstruktion har fx. været benyttet i forbindelse med udledning af konfidensintervaller og prædiktionsintervaller i den lineære normale model, eller mere direkte i HS opgave 9. Resultatet er også kendt fra MI eksempel 20.27.

## Opgave 3

1. Vi ser at faktorerne  $G$  og  $\text{time}$  opfylder EH Sætning 14.8, hvorfor de er geometrisk ortogonale. I praksis er dette en konsekvens af, at der er foretaget to målinger for *alle* børn.

```
data <- read.table(file = "MatStatJuni2019.txt", header = T)
table(data$G, data$time)

##
##           after before
## B/C        41    41
## Control    36    36
## N          42    42
## P/S         39    39
## S/J         30    30
```

Faktoren  $\text{subj}$  er finere end  $G$ , hvorfor  $\text{subj} \wedge G = \text{subj}$ .

2. Modellen kan udtrykkes ved at vektoren  $X = (X_i)_{i \in I}$  bestående af målinger af børnenes  $\text{liking}$  er normalfordelt på  $\mathbb{R}^{376}$  med  $\xi = EX \in L_{G \times \text{time}}$  og varians  $VX = \sigma^2 + v_1^2 B_1$ , hvor  $B_1$  er effektmatricen hørende til parret ( $\text{subj}, 1$ ).

Kovariansmatricen for de 2 målinger på samme barn kan udtrykkes som

$$\begin{pmatrix} \sigma^2 + v_1^2 & v_1^2 \\ v_1^2 & \sigma^2 + v_1^2 \end{pmatrix}.$$

3. Modellen fitteres i R med følgende kode (angivelse af kode ikke påkrævet)

```
library(lme4)

## Error in library(lme4): there is no package called 'lme4'

mod1 <- lmer(liking ~ G * time + (1|subj), data = data)

## Error in lmer(liking ~ G * time + (1 | subj), data = data): could not
## find function "lmer"

mod1

## Error in eval(expr, envir, enclos): object 'mod1' not found
```

Estimaterne (REML!) for variansparametrene er  $\hat{\sigma} = 1.4249$  og  $\hat{v}_1 = 0.9154$ . Den forventede  $\text{liking}$  for børn i gruppen  $G = \text{B/C}$  bliver  $4.9268$  (after) og  $4.9268 - 1.0244$  (before).

4. Modellen fra foregående delspørgsmål genfitteres nu med en parametrisering, hvor man direkte kan aflæse ændring i  $\text{liking}$  (before til after) for hver af de 5 eksponeringsgrupper. På baggrund af estimer og konfidensintervaller for tilvæksterne kan man nu konkludere på effekten af interventionen (dvs. eksponering til pågældende snackbar) på ændring i  $\text{liking}$ .

```

mod1refit <- lmer(liking ~ G + G : time - 1 + (1|subj), data = data)

## Error in lmer(liking ~ G + G:time - 1 + (1 | subj), data = data): could
not find function "lmer"

confint(mod1refit)

## Error in confint(mod1refit): object 'mod1refit' not found

```

Vi ser, at der i eksponeringsgrupperne B/C, P/S og S/J sker signifikante stigninger i `liking` (baseret på et 5 % signifikansniveau).

Det er også muligt at konstruere et overordnet likelihoodtest for hypotesen om, at  $\xi \in L_6$  svarende til at der ikke sker ændringer i `liking` for nogle af eksponeringsgrupperne.

```

mod2 <- lmer(liking ~ G + (1|subj), data = data)

## Error in lmer(liking ~ G + (1 | subj), data = data): could not find function
## "lmer"

anova(mod2, mod1)

## Error in anova(mod2, mod1): object 'mod2' not found

```

Hypotesen forkastes med et brag.  $P$ -værdi  $< 0.0001$  baseret på et opslag i en  $\chi^2$ -fordeling med 5 frihedsgrader.

## Opgave 4

1. Modellen er en ensidet variansanalysemodel, så middelværdiunderrummet er  $L_{race}$  og har dimension 5 (=antal forskellige racer i datasættet). I R udskriften er benyttet en parametrering med middelværdien for `race = Border_Terrier` og forskellene i forhold til denne gruppe. Estimaterne for parametrene til beskrivelse af middelværdivektoren bliver

$$\begin{aligned}\hat{\alpha}_{\text{Border\_Terrier}} &= 5.0812 \\ \hat{\alpha}_{\text{Grand_Danios}} &= 5.0812 + 30.3866 \\ \hat{\alpha}_{\text{Labrador}} &= 5.0812 + 13.8429 \\ \hat{\alpha}_{\text{Petit_Basset}} &= 5.0812 + 7.0097 \\ \hat{\alpha}_{\text{Whippet}} &= 5.0812 + 5.4599\end{aligned}$$

og (det centrale) variansestimat er  $\tilde{\sigma} = 4.304$ . MLE for variansen er  $\hat{\sigma}^2 = \frac{92}{97}\tilde{\sigma}^2$ . Lader vi  $\xi = A\beta$ , så giver EH Korollar 10.21 at  $\hat{\beta} \sim N(\beta, \sigma^2(A^T A)^{-1})$  og at  $\hat{\sigma}^2$  er  $\chi^2$ -fordelt med 92 frihedsgrader og skalaparameter  $\sigma^2/97$ .

2. Ved at betragte residualplottet for `modA` ses, at variansen for de standardiserede residualer ser ud til at vokse med størrelsen på de fittede værdier. Det lader derfor ikke til at antagelsen om, at residualerne er normalfordelte med samme varians er opfyldt, når man laver en regressionsanalyse af `maxLA` på `wgt`.

På residualplottet for `modB` ses et mønster fx. med hensyn til residualernes placering i forhold til 0. Residualerne hørende til observationer med små fittede værdier er oftest negative, mens residualer knyttet til lidt større fittede værdier har en tendens til at være positive. Dette strider imod en antagelse om, at residualerne skal have samme middelværdi.

Residualplottet for `modC` giver ikke umiddelbart anledning til at anfægte antagelserne om, at residualerne er normalfordelte med samme varians.

QQ-plottet af de standardiserede residualer ligger for alle modeller rimelig pænt omkring en ret linje med hældning 1 og skæring 0.

Regressionsmodellen `modC` udtrykker, at middelværdien af  $\log(\text{maxLA}_i)$  (for den  $i$ -te hund i datasættet) er givet ved  $\alpha + \beta \cdot \log(\text{wgt}_i)$ .

3. Tager vi udgangspunkt i formel (10.42) fra EH Eksempel 10.31 med  $\phi = (1, \log(25))^T$  så kan vi beregne et 95 % - prædiktionsinterval for  $\log(\text{maxLA})$  for en hund på 25 kg på baggrund af R-udskriften i opgaven.

```
beta_hat <- c(-0.11931, 0.89163)
sigma_hat <- 0.2344
ATAinv <- matrix(nrow = 2, ncol = 2, byrow = T
                  , data = c(0.170702, -0.054205, -0.054205, 0.018319))
phi <- matrix(nrow = 2, ncol = 1, data = c(1, log(25)))
z_alpha <- qt(1 - 0.05/2, 97 - 2)
pred_low <- t(phi) %*% beta_hat - z_alpha *
             sqrt(sigma_hat^2 * (1 + t(phi) %*% ATAinv %*% phi))
pred_up <- t(phi) %*% beta_hat + z_alpha *
```

```
sqrt(sigma_hat^2 * (1 + t(phi) %*% ATAinv %*% phi))  
c(pred_low, pred_up)  
  
## [1] 2.282714 3.218759
```

Dette kan omregnes til et 95 % - prædiktionsinterval for maxLA (målt i mL) ved tilbagetransformation med eksponentialfunktionen

```
exp(c(pred_low, pred_up) )  
  
## [1] 9.803249 24.997072
```

# Eksamens i Matematisk Statistik, 22. august 2019

Vejledende besvarelse

## Opgave 1

- Da både  $X_1$  og  $X_2$  er standardnormalfordelte (og dermed specielt regulært normalfordelte), så følger det af EH Sætning 9.26, at  $X = (X_1, X_2)^T$  er regulært normalfordelt på  $\mathbb{R}^2$  med middelværdi  $\xi = (0, 0)^T$  og varians  $\Sigma$ . Da  $X_1$  of  $X_2$  er uafhængige, så følger specielt (fx. af EH Sætning 9.48), at kovarianserne  $\Sigma_{12} = \Sigma_{21} = 0$  og vi har at  $\Sigma_{11} = \Sigma_{22} = 1$ . Dermed er argumenteret for, at  $X$  er standardnormalfordelt på  $\mathbb{R}^2$ . Det er også muligt at gennemføre dette argument ved at trækkes på EH Korollar 9.39.

Det følger af EH Korollar 9.46 at  $Y = (Y_1, Y_2)^T$  er normalfordelt med middelværdi

$$EY = \begin{pmatrix} 1 & 1 \\ a & b \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

og varians

$$\Sigma_Y = \begin{pmatrix} 1 & 1 \\ a & b \end{pmatrix} \Sigma \begin{pmatrix} 1 & a \\ 1 & b \end{pmatrix} = \begin{pmatrix} 2 & a+b \\ a+b & a^2+b^2 \end{pmatrix}.$$

Ifølge EH Korollar 9.43 er  $Y$  regulært normalfordelt hvis og kun hvis variansen  $\Sigma$  er invertibel. Det ses at determinanten af  $\Sigma$  er lig med

$$2 \cdot (a^2 + b^2) - (a+b)^2 = a^2 + b^2 - 2ab = (a-b)^2.$$

Vi konkluderer at  $Y$  er regulært normalfordelt netop hvis  $a \neq b$ .

- Fra delopgave 1. ved vi, at covariansen mellem  $Y_1$  og  $Y_2$  er  $a+b$ , hvorfor det specielt følger af EH Sætning 9.48, at de er uafhængige for  $a=1$  og  $b=-1$ .

For dette valg af konstanter har vi desuden, at  $Y_1$  og  $Y_2$  begge er normalfordelte med middelværdi 0 og varians 2. Dermed er  $\frac{1}{\sqrt{2}}Y_1 \sim N(0, 1)$  og  $(\frac{1}{\sqrt{2}}Y_1)^2 = \frac{1}{2}Y_1^2$  er  $\chi_1^2$ -fordelt.

Tilsvarende er  $\frac{1}{\sqrt{2}}Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2) \sim N(0, 1)$  og uafhængig af  $\frac{1}{2}Y_1^2$ . Det følger at forholdet  $\frac{\frac{1}{\sqrt{2}}Y_2}{\sqrt{\frac{1}{2}Y_1^2}} = \frac{X_1 - X_2}{|X_1 + X_2|}$  er  $t$ -fordelt med 1 frihedsgrad. Denne konstruktion har fx. været benyttet i forbindelse med udledning af konfidensintervaller og prædiktionsintervaller i den lineære normale model, eller mere direkte i HS opgave 9. Resultatet er også kendt fra MI eksempel 20.27.

## Opgave 2

Lad  $X_1, \dots, X_n, Y_1, \dots, Y_n$  være uafhængige stokastiske variable, hvor alle  $X_i$  er poissonfordelte med middelværdi  $\lambda$ , og alle  $Y_i$  er poissonfordelte med middelværdi  $\lambda^2$ . Her er  $\lambda > 0$  en ukendt parameter.

- Opskriv log-likelihoodfunktionen og vis at scorefunktionen er givet som

$$S_n(\lambda; x, y) = Dl_{x,y}(\lambda) = -\frac{S_x + 2S_y}{\lambda} + n + 2n\lambda$$

For givne observationer  $x = (x_1, \dots, x_n) \in \mathbb{N}_0^n$  og  $y = (y_1, \dots, y_n) \in \mathbb{N}_0^n$  får vi (på nær en multiplikativ konstant) likelihoodfunktionen

$$L_{x,y}(\lambda) = \prod_{i=1}^n (\lambda^{x_i} e^{-\lambda}) \prod_{i=1}^n (\lambda^{2y_i} e^{-\lambda^2}) = \lambda^{S_x + 2S_y} e^{-n\lambda} e^{-n\lambda^2}, \quad \lambda > 0$$

hvor  $S_x = \sum_{i=1}^n x_i$  og  $S_y = \sum_{i=1}^n y_i$ . Derefter fås

$$\begin{aligned} l_{x,y}(\lambda) &= -\log L_{x,y}(\lambda) = -(S_x + 2S_y) \log \lambda + n\lambda + n\lambda^2 \\ Dl_{x,y}(\lambda) &= -\frac{S_x + 2S_y}{\lambda} + n + 2n\lambda. \end{aligned}$$

hvor jeg har brugt notationen  $S_X = \sum_{i=1}^n X_i$  og  $S_Y = \sum_{i=1}^n Y_i$ .

- Vis at maksimaliseringestimatoren (MLE) for  $\lambda$  er asymptotisk veldefineret og entydigt givet ved

$$\hat{\lambda} = \frac{-n + \sqrt{n^2 + 8n(S_X + 2S_Y)}}{4n}$$

hvor  $S_X = \sum_{i=1}^n X_i$  og  $S_Y = \sum_{i=1}^n Y_i$ .

Scoreligningen er

$$Dl_{x,y}(\lambda) = 0 \Leftrightarrow 2n\lambda^2 + n\lambda - S_x - 2S_y = 0,$$

dvs. en andengrads ligning i  $\lambda$ . Diskriminantens er  $n^2 + 8n(S_X + 2S_Y) > 0$ , så der er to reelle løsninger:

$$\frac{-n \pm \sqrt{n^2 + 8n(S_X + 2S_Y)}}{4n}$$

Løsningen med “+” er altid ikke-negativ, mens løsningen med “-” altid er ikke-positiv, så scoreligningen har netop en løsning i parameterområdet hvis og kun hvis  $S_X + S_Y > 0$  hvilket sker med en sandsynlighed gående mod 1, så estimatoren er asymptotisk veldefineret. Da vi endvidere har at  $D^2l_{x,y}(\lambda) > 0$  for alle  $\lambda > 0$ , er løsningen et minimum for  $l_{x,y}$  på  $(0, 1)$ . Samlet set får vi at

$$\hat{\lambda} = \frac{-n + \sqrt{n^2 + 8n(S_X + 2S_Y)}}{4n}$$

er en entydig MLE for  $\lambda$ .

3. Vis at familien  $\mathcal{P} = \{P_\lambda, \lambda > 0\}$  angivet ovenfor kan repræsenteres som en regulær eksponentiel familie af dimension 1. Angiv familiens kanoniske parameter, kanoniske stikprøvefunktion, samt grundmål.

Vi lader  $\theta = \log \lambda$  og skriver tæthedens som

$$p(x_1, \dots, x_n, y_1, \dots, y_n; \theta) = \prod_i \frac{1}{x_i!} \prod_i \frac{1}{y_i!} e^{\theta(S_X + 2S_Y) - n(e^\theta + e^{2\theta})}$$

hvoraf det fremgår at grundmålet er  $\prod_i \frac{1}{x_i!} \prod_i \frac{1}{y_i!} \cdot m$  hvor  $m$  er tællemål, den kanoniske stikprøvefunktion er  $(S_X + 2S_Y)$ , osv.

4. Vis at Fisherinformationen for  $\lambda$  er givet som

$$i_n(\lambda) = \frac{n}{\lambda} + 4n$$

og angiv den asymptotiske fordeling af maksimaliseringsestimatoren  $\hat{\lambda}_n$ .

Vi får Fisherinformationen

$$i(\lambda) = ED^2 l_{X,Y} = E\left(\frac{S_X + 2S_Y}{\lambda^2} + 2n\right) = \frac{n\lambda + 2n\lambda^2}{\lambda^2} + 2n = \frac{n}{\lambda} + 4n.$$

I en regulær eksponentiel familie er MLE asymptotisk normalfordelt med den inverse Fisherinformation som varians:

$$\hat{\lambda}_n \stackrel{\text{as}}{\sim} N\left(\lambda, \frac{\lambda}{n(1+4\lambda)}\right).$$

5. Betrag nu den alternative estimator  $\tilde{\lambda}_n$  hvor

$$\tilde{\lambda}_n = \frac{S_X/n + \sqrt{S_Y/n}}{2},$$

og vis, at den asymptotiske fordeling er

$$\tilde{\lambda}_n \stackrel{\text{as}}{\sim} N\left(\lambda, \frac{4\lambda+1}{16n}\right).$$

Vi har at  $S_X/n \stackrel{\text{as}}{\sim} N(\lambda, \lambda/n)$  og  $S_Y/n \stackrel{\text{as}}{\sim} N(\lambda^2, \lambda^2/n)$ . Deltametoden giver at

$$\sqrt{S_Y/n} \stackrel{\text{as}}{\sim} N\left(\sqrt{\lambda^2}, \frac{1}{(2\sqrt{\lambda^2})^2} \frac{\lambda^2}{n}\right) = N\left(\lambda, \frac{1}{4n}\right).$$

Alt i alt får vi at

$$\tilde{\lambda}_n \stackrel{\text{as}}{\sim} N\left(\frac{\lambda+\lambda}{2}, \frac{1}{4}\left(\frac{\lambda}{n} + \frac{1}{4n}\right)\right) = N\left(\lambda, \frac{4\lambda+1}{16n}\right)$$

6. Sammenlign de to estimatorer  $\hat{\lambda}_n$  og  $\tilde{\lambda}_n$ .

Begge estimatorer er asymptotisk konsistente, men den alternative estimator har større asymptotisk varians end MLE. Forholdet mellem de asymptotiske varianser er

$$\frac{V_{as}(\tilde{\lambda}_n)}{V_{as}(\hat{\lambda}_n)} = \frac{4\lambda+1}{16} \frac{1+4\lambda}{\lambda} = \frac{(1+4\lambda)^2}{16\lambda} > 1.$$

Det er særlig slemt for store værdier af  $\lambda$ , hvor forholdet går mod uendelig.

## Opgave 3

1. Ved at lave en krydstabel over antallet af observationer for de forskellige kombinationer af faktorerne fås

G	V = -	V = +	G	K = 0	K = lav	K = høj
A	20	20	A	0	20	20
B	20	20	B	0	20	20
C	20	20	C	40	0	0

Vi ser at faktorerne G og V opfylder EH Sætning 14.8, hvorfor de er geometrisk ortogonale. Tilsvarende ses, at minimum af G og K er en faktor med to niveauer, som angiver om en jordlod er blevet kunstgødet (med G = A eller G = B) eller ej (svarende til G = C).

2. Det følger af EH (13.1) og Lemma 14.6 at

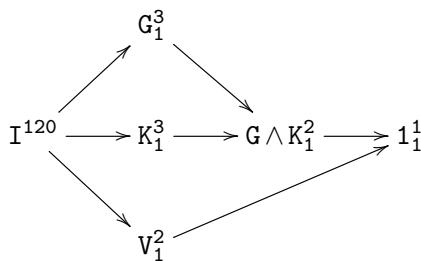
$$\dim(L_G + L_V) = \dim(L_G) + \dim(L_V) - \dim(L_{G \wedge V}) = 3 + 2 - 1 = 4,$$

hvor det er benyttet, at minimum af faktorerne G og V er den konstante faktor 1.

Designet  $\mathbb{G} = \{G, V, G \wedge K, K, 1\}$  er geometrisk ortonalt og afsluttet over for dannelsen af minimum. Dimensionerne af  $V_G$ -rummene,  $G \in \mathbb{G}$ , fra sætningen om den ortogonale decompositon (EH: Sætning 14.21) kan let beregnes, hvorefter vi finder, at

$$\begin{aligned}\dim(L_G + L_V + L_K) &= \dim(V_G) + \dim(V_V) + \dim(V_{G \wedge K}) + \dim(V_K) + \dim(V_1) \\ &= 1 + 1 + 1 + 1 + 1 = 5.\end{aligned}$$

Det kan her være nyttigt at støtte sig op ad et faktorstrukturdigram, for at holde styr på ordningen af faktorerne



3. Der er tale om en lineær normal model, hvor middelværdiunderrummet er parametriseret ved designmatricen  $A$ . Det følger derfor at EH Korollar 10.21 at

$$\hat{\beta} = (A^T A)^{-1} A^T X \sim N(\beta, \sigma^2 (A^T A)^{-1})$$

og at

$$\hat{\sigma}^2 = \frac{\|X - PX\|^2}{120} = \frac{\|X - A\hat{\beta}\|^2}{120} \sim \frac{\sigma^2}{120} \chi^2_{120-k} - \text{fordelt.}$$

## Opgave 4

1. Modellen kan udtrykkes ved at vektoren  $X = (X_i)_{i \in I}$  bestående af målinger af den maksimale bøjningsvinkel  $v$  er normalfordelt på  $\mathbb{R}^{64}$  med  $\xi = EX \in L_{ben \times fase}$  og varians  $VX = \sigma^2 + v_1^2 B_1 B_1^T$ , hvor  $B_1$  er effektmatrixen hørende til parret  $(subj, 1)$ . Vi er ikke specifikt interesserede i lige præcis de 16 personer, som indgår i eksperimentet, hvorfor vi lader  $subj$  indgå med en tilfældig effekt.

Kovariansmatrixen for de 4 målinger på samme person kan udtrykkes som

$$\begin{pmatrix} \sigma^2 + v_1^2 & v_1^2 & v_1^2 & v_1^2 \\ v_1^2 & \sigma^2 + v_1^2 & v_1^2 & v_1^2 \\ v_1^2 & v_1^2 & \sigma^2 + v_1^2 & v_1^2 \\ v_1^2 & v_1^2 & v_1^2 & \sigma^2 + v_1^2 \end{pmatrix}.$$

En mere avanceret løsning består i at inddrage vekselvirkningerne  $subj \times ben$  og  $subj \times fase$  i modellen. Alle faktorer der involverer  $subj$  bør indgå med tilfældig effekt, således at der bliver op til tre tilfældige effekter svarende til hvert af effektparrene  $(subj, 1)$ ,  $(subj \times ben, 1)$  og  $(subj \times fase, 1)$ . Lader vi  $B_1$ ,  $B_2$  og  $B_3$  betegne effektmatrixerne hørende til de tre effektpar, så kan modellen udtrykkes ved at  $X = (X_i)_{i \in I}$  er normalfordelt på  $\mathbb{R}^{64}$  med  $\xi = EX \in L_{ben \times fase}$  og  $VX = \sigma^2 I + v_1^2 B_1 B_1^T + v_2^2 B_2 B_2^T + v_3^2 B_3 B_3^T$ .

Organiseres de 4 målinger for samme person, så vi først har de to målinger for  $ben = D$  og dernæst de to målinger for  $ben = N$ , så bliver kovariansmatrixen

$$\begin{pmatrix} \sigma^2 + v_1^2 + v_2^2 + v_3^2 & v_1^2 + v_2^2 & v_1^2 + v_3^2 & v_1^2 \\ v_1^2 + v_2^2 & \sigma^2 + v_1^2 + v_2^2 + v_3^2 & v_1^2 & v_1^2 + v_3^2 \\ v_1^2 + v_3^2 & v_1^2 & \sigma^2 + v_1^2 + v_2^2 + v_3^2 & v_1^2 + v_2^2 \\ v_1^2 & v_1^2 + v_3^2 & v_1^2 + v_2^2 & \sigma^2 + v_1^2 + v_2^2 + v_3^2 \end{pmatrix}.$$

Det betragtes som en tilstrækkelig besvarelse, selvom man vælger ikke at inddrage vekselvirkninger med  $subj$  som tilfældige effekter i modellen.

2. Modellen fitteres i R med følgende kode (angivelse af kode ikke påkrævet)

```
library(lme4)
knee <- read.table(file = "MatStatAug2019.txt", header = T)
mod1 <- lmer(v ~ ben * fase + (1|subj), data = knee)
mod1

## Linear mixed model fit by REML ['lmerMod']
## Formula: v ~ ben * fase + (1 | subj)
##   Data: knee
## REML criterion at convergence: 406.8466
## Random effects:
##   Groups   Name        Std.Dev.
##   subj     (Intercept) 3.989
##   Residual           5.720
## Number of obs: 64, groups: subj, 16
## Fixed Effects:
##   (Intercept)      benN      fasesving  benN:fasesving
##   12.6222       -0.7094      41.8129      -1.6462
```

Estimaterne (REML!) for variansparametrene er  $\hat{\sigma} = 5.720$  og  $\hat{v}_1 = 3.989$ . Den forventede maksimale bøjningsvinkel for det dominante ben (ben = D) bliver 12.6222 grader (fase = kontakt) og  $12.6222 + 41.8129$  grader (fase = sving).

- Modellen fra foregående delspørgsmål genfittes nu med en parametrisering, hvor man direkte kan aflæse forskellen i maksimal bøjningsvinkel (ben = N fratrukket ben = D) for hver af de faser. På baggrund af estimerer og konfidensintervaller for tilvæksterne kan man nu konkludere på de ønskede forskelle.

```
mod1refit <- lmer(v ~ fase + ben : fase - 1 + (1|subj), data = knee)
mod1refit

## Linear mixed model fit by REML ['lmerMod']
## Formula: v ~ fase + ben:fase - 1 + (1 | subj)
##   Data: knee
## REML criterion at convergence: 406.8466
## Random effects:
##   Groups     Name        Std.Dev.
##   subj       (Intercept) 3.989
##   Residual             5.720
## Number of obs: 64, groups:  subj, 16
## Fixed Effects:
##           fasiekontakt      fasesving  fasiekontakt:benN      fasesving:benN
##                 12.6222          54.4351         -0.7094          -2.3555

confint(mod1refit)

## Computing profile confidence intervals ...

##                  2.5 %    97.5 %
## .sig01          1.910588  6.469507
## .sigma          4.591540  6.862495
## fasiekontakt   9.245042 15.999395
## fasesving      51.057917 57.812270
## fasiekontakt:benN -4.625514  3.206736
## fasesving:benN  -6.271671  1.560580
```

Vi ser at der hverken er signifikante forskelle i bøjningsvinklen for ben = N og ben = D når man betragter den maksimale vinkel under kontakt-fasen (-0.71 [-4.63-3.21]) eller under spring-fasen (forskel: -2.36 [-6.27-1.56]).

Det er også muligt at lave et (simultan) test for, om maksimal bøjningsvinkel varierer mellem ben = N og ben = D.

```
mod2 <- lmer(v ~ fase + (1|subj), data = knee)
anova(mod2, mod1)

## refitting model(s) with ML (instead of REML)

## Data: knee
## Models:
```

```

## mod2: v ~ fase + (1 | subj)
## mod1: v ~ ben * fase + (1 | subj)
##          Df      AIC      BIC    logLik deviance Chisq Chi Df Pr(>Chisq)
## mod2   4 427.56 436.20 -209.78     419.56
## mod1   6 430.01 442.96 -209.00     418.01 1.5528      2     0.4601

```

Her fås en likelihoodratio-teststørrelsen  $LRT = 1.5528$  som i en tabel over  $\chi^2$ -fordelingen med 2 frihedsgrader oversættes til en approksimativ P-værdi på 0.4601. Vi kan altså ikke afvise en hypotese om, at der ikke er forskel på maksimal bøjningsvinkel for  $ben = N$  og  $ben = D$ .

# Matematisk Statistik: Vejledende besvarelse af eksamen

Steffen Lauritzen og Niels Richard Hansen

18. juni, 2020

## Spørgsmål 1.1

Vi har tæthedten

$$f_{(\beta, \gamma)}(x, y) = \frac{1}{\beta} e^{-x/\beta} \frac{1}{\gamma} e^{-y/\gamma}$$

som vi omskriver på eksponentiel familie form ved at lade  $\theta_1 = 1/\beta$  og  $\theta_2 = 1/\gamma$  og derfor med  $t(x, y) = -(x, y)^\top$

$$f_\theta(x, y) = \theta_1 \theta_2 e^{\theta^\top t(x, y)} = e^{\theta^\top t(x, y) - \log \theta_1 - \log \theta_2}.$$

Idet  $\lambda_1 x + \lambda_2 y = c$  for næsten alle  $(x, y)$  medfører  $\lambda_1 = \lambda_2 = 0$ , fremgår det, at den specificerede familie uden restriktioner på  $\theta$  er en regulære og minimalt repræsenteret eksponentiel med dimension 2. Under hypotesen defineres en krum familie idet

$$\phi(\beta) = \begin{pmatrix} 1/\beta \\ 1/\beta^2 \end{pmatrix}$$

er en glat homeomorfi med Jacobi matrix

$$D\phi(\beta) = (-1/\beta, -2/\beta^2)$$

som har fuld rang 1.

## Spørgsmål 1.2

Vi får likelihoodfunktionen

$$L_n(\beta) = \prod_{i=1}^n \frac{e^{-x_i/\beta} e^{-y_i/\beta}}{\beta^3}.$$

Lad  $S_x = \sum_{i=1}^n X_i$ ,  $S_y = \sum_{i=1}^n Y_i$ , så får vi log-likelihoodfunktionen

$$\ell_n(\beta) = 3n \log \beta + \frac{S_x}{\beta} + \frac{S_y}{\beta^2}$$

og videre scorefunktion ved differentiation

$$S_n(\beta) = \frac{3n}{\beta} - \frac{S_x}{\beta^2} - \frac{2S_y}{\beta^3}$$

og en gang til for at få informationsfunktionen

$$I_n(\beta) = -\frac{3n}{\beta^2} + \frac{2S_x}{\beta^3} + \frac{6S_y}{\beta^4}.$$

## Spørgsmål 1.3

Scoreligningen  $S_n(\beta) = 0$ :

$$\frac{3n}{\beta} - \frac{S_x}{\beta^2} - \frac{2S_y}{\beta^3} = 0$$

omskrives til idet  $\bar{x}_n = S_x/n$ ,  $\bar{y}_n = S_y/n$

$$3\beta^2 - \bar{x}_n\beta - 2\bar{y}_n = 0$$

Med entydig løsning i området  $\beta > 0$

$$\hat{\beta} = \frac{\bar{x}_n + \sqrt{\bar{x}_n^2 + 24\bar{y}_n}}{6}.$$

Idet

$$\ell_n(\beta) = 3n \log \beta + \frac{S_x}{\beta} + \frac{S_y}{\beta^2}$$

ser vi at  $\ell_n(\beta) \rightarrow \infty$  for  $\beta \rightarrow 0$  og  $\beta \rightarrow \infty$ , så det må være et minimum.

## Spørgsmål 1.4

Vi har Fisherinformationen

$$\begin{aligned} i_n(\beta) &= \mathbf{E}_\beta \{I_n(\beta)\} \\ &= -\frac{3n}{\beta^2} + \frac{2\mathbf{E}_\beta \{S_x\}}{\beta^3} + \frac{6\mathbf{E}_\beta \{S_y\}}{\beta^4} \\ &= -\frac{3n}{\beta^2} + \frac{2n\beta}{\beta^3} + \frac{6n\beta^2}{\beta^4} = \frac{5n}{\beta^2} \end{aligned}$$

hvoraf vi slutter at MLE er asymptotisk normalfordelt

$$\hat{\beta}_n \sim N\left(\beta, \frac{\beta^2}{5n}\right).$$

## Spørgsmål 1.5

(a): Vi at bruge et likelihood ratio test. Vi får

$$\begin{aligned} 2\ell(\hat{\beta}_{10}) - 2\ell(\hat{\theta}_{10}) &= 60\log(\hat{\beta}_{10}) + 20\frac{\bar{x}_{10}}{\hat{\beta}_{10}} + 20\frac{\bar{y}_{10}}{\hat{\beta}_{10}^2} \\ &\quad - 20\log\bar{x}_{10} - 20\log\bar{y}_{10} - 20\frac{\bar{x}_{10}}{\bar{x}_{10}} - 20\frac{\bar{y}_{10}}{\bar{y}_{10}} \\ &= 60\log(\hat{\beta}_{10}) + 20\frac{\bar{x}_{10}}{\hat{\beta}_{10}} + 20\frac{\bar{y}_{10}}{\hat{\beta}_{10}^2} - 20\log\bar{x}_{10} - 20\log\bar{y}_{10} - 40 \\ &= 0.04. \end{aligned}$$

Denne skal vurderes i en  $\chi^2$ -fordeling med  $2 - 1 = 1$  frihedsgrader, hvilket giver en  $p$ -værdi på  $p = 0.837$ , så der er absolut ingen grund til at forkaste  $H_0$ .

(b): Denne gang vælger vi at bruge den ægte Wald størrelse for den simple hypotese. Idet  $\hat{\beta}_{10} = 2.28$  fås

$$W_n = 5n(\hat{\beta}_n - 1)^2 = 50 \times 1.28^2 = 81.97$$

som skal vurderes i en  $\chi^2$  med 1 frihedsgrad, hvilket giver en  $p$ -værdi tæt på 0, så hypotesen kan ikke opretholdes.

Man kunne naturligvis også have brugt et LR test

$$\Lambda = 2(\ell(1) - \ell(\hat{\beta}_{10})) = 61.13$$

med samme resultat.

```
# data

xbar = 2.15
ybar = 5.349

# mle

hatbeta=(2.15+sqrt(xbar^2+24*ybar))/6

# LR sammensat hypotese

ell_0 =30*log(hatbeta)+10*xbar/hatbeta+10*ybar/(hatbeta^2)
ell_1 = 10*log(xbar) +10*log(ybar) +20
Lambda=2*(ell_0-ell_1)

# p værdi
1-pchisq(Lambda,1)

[1] 0.8375556

Lambda

[1] 0.04203367

# Wald
w=50*(hatbeta-1)^2
w

[1] 81.97329

# p værdi
1-pchisq(w,1)

[1] 0

# LR for simpel hypotese

ell_2=10*xbar+10*ybar
Lambda2= 2*(ell_2-ell_1)
Lambda2

[1] 61.13245

# p værdi
1-pchisq(Lambda2,1)

[1] 5.329071e-15
```

## Spørgsmål 2.1

Da  $\mathbf{E}_\theta(Z) = 0$  og  $X$  og  $Y$  er uafhængige, er

$$m(\theta) = \mathbf{E}_\theta(Z^2) = \mathbf{V}_\theta(X - Y) = \mathbf{V}_\theta(X) + \mathbf{V}_\theta(Y).$$

Da  $X$  og  $Y$  begge er poissonfordelte med middelværdi, og dermed varians,  $e^\theta$  ser vi, at

$$m(\theta) = 2e^\theta.$$

Momentestimatoren er dermed givet som løsning til

$$2e^\theta = \frac{1}{n} \sum_{i=1}^n Z_i^2,$$

dvs.

$$\tilde{\theta}_n = \log \left( \frac{1}{2n} \sum_{i=1}^n Z_i^2 \right),$$

som er veldefineret, hvis ikke alle  $Z_i$ -erne er 0.

## Spørgsmål 2.2

Vi efterviser betingelserne for BMS, sætning 2.17. Da poissonfordelte variable har momenter af enhver orden, har  $t(Z)$  specielt endelig varians, endvidere er momentfunktionen glat og injektiv, og

$$m'(\theta) = 2e^\theta \neq 0.$$

Heraf følger, at  $\tilde{\theta}_n$  er konsistent og asymptotisk normalfordelt.

Vi finder nu

$$\mathbf{V}_\theta(t(Z)) = \mathbf{V}_\theta(Z^4) = \mathbf{E}_\theta(Z^4) - (\mathbf{E}_\theta(Z^2))^2 = 2e^\theta + 12e^{2\theta} - 4e^{2\theta} = 2e^\theta + 8e^{2\theta}.$$

BMS, sætning 2.17, giver den asymptotiske varians

$$\sigma^2(\theta) = \mathbf{V}(t(Z))/m'(\theta)^2 = \frac{1}{4}e^{-2\theta}(2e^\theta + 8e^{2\theta}) = 2 + \frac{1}{2}e^{-\theta}.$$

Med andre ord er

$$\tilde{\theta}_n \xrightarrow{\text{as}} N(\theta, (2 + e^{-\theta}/2)/n).$$

## Spørgsmål 2.3

Det er mest interessant at undersøge den asymptotiske fordeling for negative værdier af  $\theta$  og/eller små værdier af  $n$ , og eventuelt sammenholde med større værdier.

Her præsenteres resultaterne for fire kombinationer. De asymptotiske standardafvigelser (standard errors) beregnes endvidere for alle fire kombinationer.

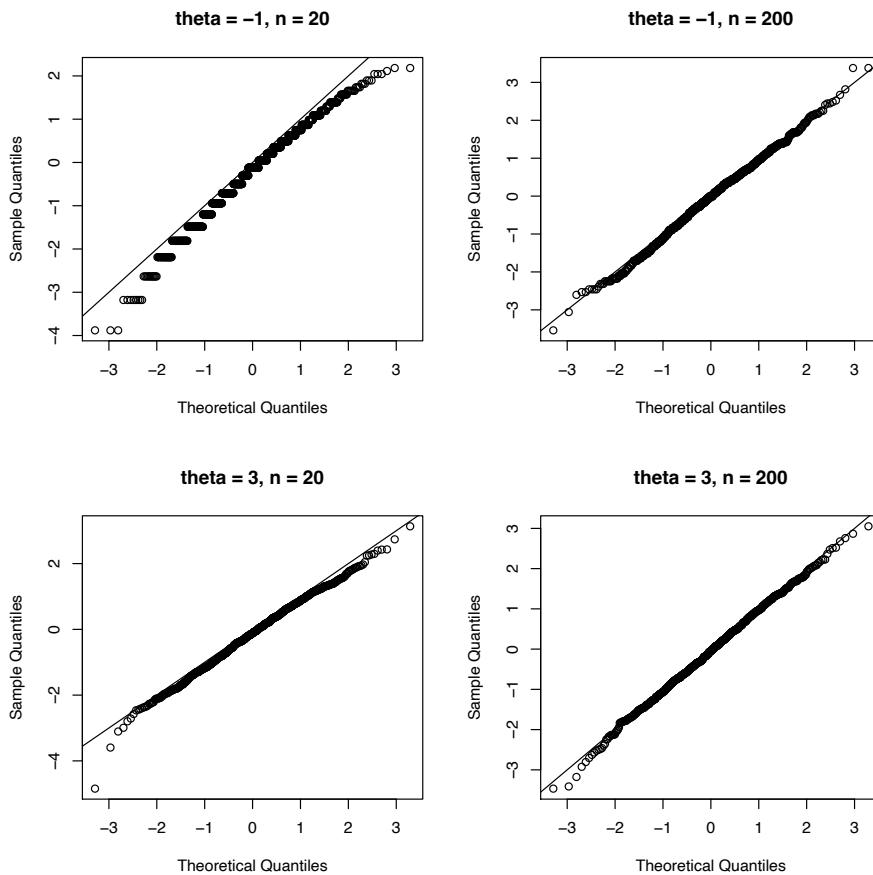
```
set.seed(11)
B <- 1000
theta1 <- -1 ## Middelværdi exp(-1) = 0.37
n <- 20
se1 <- sqrt((2 + exp(-theta1) / 2) / n)
theta_tilde1 <- replicate(B, {
  x <- rpois(n, exp(theta1))
  y <- rpois(n, exp(theta1))
  log(mean((x - y)^2)/2)
})
theta2 <- 3 ## Middelværdi exp(3) = 20
se2 <- sqrt((2 + exp(-theta2) / 2) / n)
```

```

theta_tilde2 <- replicate(B, {
  x <- rpois(n, exp(theta2))
  y <- rpois(n, exp(theta2))
  log(mean((x - y)^2)/2)
})
n <- 200
se3 <- sqrt((2 + exp(-theta1) / 2) / n)
theta_tilde3 <- replicate(B, {
  x <- rpois(n, exp(theta1))
  y <- rpois(n, exp(theta1))
  log(mean((x - y)^2)/2)
})
se4 <- sqrt((2 + exp(-theta2) / 2) / n)
theta_tilde4 <- replicate(B, {
  x <- rpois(n, exp(theta2))
  y <- rpois(n, exp(theta2))
  log(mean((x - y)^2)/2)
})

```

Vi sammenligner nu med den asymptotiske normalfordeling via qqplot.



Det er klart, at for  $\theta = -1$  er estimatoren ikke normalfordelt for  $n = 20$ , mens normalfordelingen er en OK approksimation for  $\theta = 3$ , selv for  $n = 20$ . For  $n = 200$  er normalfordelingen en god approksimation til fordelingen af estimatoren, selv for  $\theta = -1$

Vi kan også sammenligne de asymptotiske standardafvigelser med de empiriske.

```
tibble(theta = c(theta1, theta2, theta1, theta2), n = c(20, 20, 200, 200),
       teo_se = c(se1, se2, se3, se4),
       emp_sd = c(sd(theta_tilde1), sd(theta_tilde2), sd(theta_tilde3), sd(theta_tilde4)))
```

	theta	n	teo_se	emp_sd
1	-1	20	0.410	0.418
2	3	20	0.318	0.319
3	-1	200	0.130	0.133
4	3	200	0.101	0.101

Her ser vi en ret god overensstemmelse, selv for  $\theta = -1$  og  $n = 20$ .

## Spørgsmål 3.1

Vi indlæser data til opgaven.

```
gener <- read_csv("MatStat2020Juni_opg3.txt")
```

Opgavens første del løses ved krydstabulering af C og S. Sålænge det er gjort korrekt, og argumenterne er rigtige, spiller det ingen rolle for bedømmelsen, hvordan det præcist er implementeret. Nedenfor følger en måde at løse opgaven på.

```
count(gener, C, S)
```

```
# A tibble: 38 x 3
  C           S             n
  <chr> <chr> <int>
1 bird  Aquila_chrysaetos     9
2 bird  Cariama_cristata      6
3 bird  Charadrius_vociferus   9
4 bird  Eurypyga_helias        6
5 bird  Gallus_gallus        10
6 bird  Haliaeetus_leucocephalus 10
7 bird  Pygoscelis_adeliae      9
8 bird  Serinus_canaria        10
9 fish  Ictalurus_punctatus      9
10 fish Lepisosteus_oculatus     9
# ... with 28 more rows
```

I opgaven oplyses det, at der er 38 arter (hvilket i øvrigt også kan tjekkes ved tabulering), og da der ligeledes er 38 rækker i denne tabel følger det, at  $C \leq S$ . Enhver værdi af S må jo så forekomme netop en gang i tabellen, og bestemmer således værdien af C.

Det følger også af (den fulde version af) tabellen ovenfor, at C optræder på tre niveauer (bird, fish, mammal), så  $\dim(L_C) = 3$ .

Da  $C \leq S$  er endvidere  $C \times G \leq S \times G$ , og  $G \leq C \times G$ , så vi får, jf. også det tilsvarende design i eksempel 14.16 i EH, faktorstrukturdigrammet

$$\begin{array}{ccccc}
 S \times G & \longrightarrow & C \times G & \longrightarrow & G^{10} \\
 \downarrow & & \downarrow & & \downarrow \\
 S^{38} & \longrightarrow & C^3 & \longrightarrow & 1^1
 \end{array}$$

Diagrammet er ovenfor annoteret med dimensioner vi kender på nuværende tidspunkt.

### Spørgsmål 3.2

Da  $S \wedge C \times G = C$  og  $C \leq S$  er der kun to ikke-trivielle minima, der skal undersøges, nemlig  $C \wedge G$  og  $S \wedge G$ , jf. også det tilsvarende design i eksempel 14.20 i EH.

Det gøres ligeledes ved krydstabulering, hvor det her nok er lettest bare at bruge `table`.

```
table(gener$C, gener$G)
```

	1CQ7M	1CS4Z	1CSGF	1CV66	1CW69	1CXQA	1D1QW	1D229	1DD42	1DFZX
bird	8	3	8	8	6	8	6	6	8	8
fish	7	7	7	7	6	7	7	5	0	7
mammal	23	22	21	23	23	22	18	17	23	21

Af tabellen ovenfor fremgår det, at der kun er en enkelt kombination af  $C$  og  $G$  (fish og 1DD42), som ikke forekommer. Designgrafen indeholder derfor alle kanter pånær denne ene, og er oplagt sammenhængende, hvorfor  $C \wedge G = 1$ .

Bemærk i øvrigt at tabellen viser, at produktfaktoren  $C \times G$  forekommer på 29 niveauer, så  $\dim(L_{C \times G}) = 29$ .

```
table(gener$S, gener$G)
```

	1CQ7M	1CS4Z	1CSGF	1CV66	1CW69	1CXQA	1D1QW	1D229	1DD42	1DFZX
Acinonyx_jubatus	1	0	1	1	1	1	1	1	1	1
Aotus_nancymaae	1	1	1	1	1	1	0	1	1	1
Aquila_chrysaetos	1	0	1	1	1	1	1	1	1	1
Callithrix_jacchus	1	1	1	1	1	1	1	0	1	1
Cariama_cristata	1	0	1	1	0	1	0	0	1	1
Cebus_capucinus	1	1	1	1	1	1	1	0	1	1
Ceratotherium_simum	1	1	1	1	1	1	1	1	1	1
Charadrius_vociferus	1	0	1	1	1	1	1	1	1	1
Chinchilla_lanigera	1	1	1	1	1	1	1	1	1	1
Chrysocloris_asiatica	1	1	1	1	1	1	0	1	1	1
Elephantulus_edwardii	1	1	1	1	1	1	0	1	1	1
Eptesicus_fuscus	1	1	1	1	1	1	1	1	1	1
Eurypyga_helias	1	0	1	1	0	1	0	0	1	1
Gallus_gallus	1	1	1	1	1	1	1	1	1	1
Haliaeetus_leucocephalus	1	1	1	1	1	1	1	1	1	1
Ictalurus_punctatus	1	1	1	1	1	1	1	1	0	1
Lepisosteus_oculatus	1	1	1	1	1	1	1	1	0	1
Leptonychotes_weddellii	1	1	1	1	1	1	1	0	1	0
Lipotes_vexillifer	1	1	1	1	1	1	1	0	1	1
Loxodonta_africana	1	1	1	1	1	1	1	1	1	1

Manis_javanica	1	1	1	1	1	1	0	0	1	1
Myotis_lucifugus	1	1	1	1	1	1	1	1	1	0
Nannospalax_galili	1	1	1	1	1	1	1	1	1	1
Neolamprologus_brichardi	1	1	1	1	1	1	1	0	0	1
Octodon_degus	1	1	1	1	1	1	1	1	1	1
Orycteropus_afer	1	1	1	1	1	1	0	1	1	1
Pan_troglodytes	1	1	0	1	1	1	1	1	1	1
Papio_anubis	1	1	0	1	1	1	1	1	1	1
Poecilia_reticulata	1	1	1	1	0	1	1	1	0	1
Pteropus_alecto	1	1	1	1	1	1	1	1	1	1
Pygoscelis_adeliae	1	0	1	1	1	1	1	1	1	1
Saimiri_boliviensis	1	1	1	1	1	1	1	1	1	1
Sarcophilus_harrisii	1	1	1	1	1	0	1	0	1	1
Scleropages_formosus	1	1	1	1	1	1	1	1	0	1
Serinus_canaria	1	1	1	1	1	1	1	1	1	1
Sinocyclocheilus_grahami	1	1	1	1	1	1	1	1	0	1
Sorex_araneus	1	1	1	1	1	1	1	1	1	1
Xiphophorus_maculatus	1	1	1	1	1	1	1	0	0	1

Tabellen ovenfor viser, at der er visse art-gen kombinationer, der ikke forekommer, men f.eks. forekommer genet 1CQ7M sammen med alle arter, og da alle gener forekommer i kombination med mindst en art er der altid en vej i designgrafen mellem to knuder via 1CQ7M. Designgrafen er således sammenhængende, og  $S \wedge G = 1$ .

Designet er **ikke** ortogonal, f.eks. fordi tabellen for  $C \times G$  indeholder et 0, jf. lemma 13.11 i EH.

Vi kan nu endelig finde dimensionerne af sum-rummene ved at bruge formel (13.1) i EH sammen med lemma 14.6:

Da  $L_{S \wedge G} = L_1$  er

$$\dim(L_S + L_G) = \dim(L_S) + \dim(L_G) - \dim(L_{S \wedge G}) = 38 + 10 - 1 = 47$$

og da  $L_{S \wedge C \times G} = L_C$  er

$$\dim(L_S + L_{C \times G}) = \dim(L_S) + \dim(L_{C \times G}) - \dim(L_C) = 38 + 29 - 3 = 64$$

Bemærk at da designet ikke er ortogonal, kan vi principielt ikke benytte teknikken baseret på sætning 14.21 til at finde dimensionerne af sum-rummene ovenfor, selvom det vil give de rigtige dimensioner i dette tilfælde.

## Bonus: Minimum af $S$ og $C \times G$

Man kunne lave et tilsvarende argument som ovenfor via tabulering for at vise at  $S \wedge C \times G = C$ , men tabellen bliver uoverskuelig. I stedet kan vi finde minimum ved at finde sammenhængskomponenterne på følgende måde (som gennemgået ved forelæsningerne).

```
library(igraph)
tab <- count(gener, S, C, G)
g <- mutate(tab, CG = paste(C, G, sep = " "))
  select(S, CG) %>%
  as.matrix() %>%
  graph_from_edgelist(directed = FALSE)
tab <- mutate(tab, min.S.CG = components(g)$membership[as.character(S)])
count(tab, C, min.S.CG)

# A tibble: 3 x 3
  C      min.S.CG     n
  <dbl>        <dbl> <dbl>
1 0             0       1
2 1             1       1
3 2             2       1
```

```

<chr>      <dbl> <int>
1 bird        2     69
2 fish        3     60
3 mammal     1    213

```

Denne tabel viser, at minimum er identisk med C.

### Spørgsmål 3.3

Vi fitter de to modeller i R.

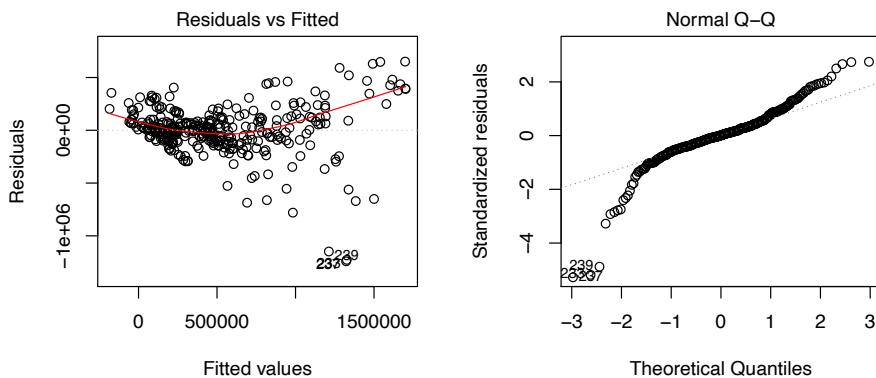
```

gen_lm <- lm(L ~ S + C * G, data = gener)
root_gen_lm <- lm(L^(1/3) ~ S + C * G, data = gener)

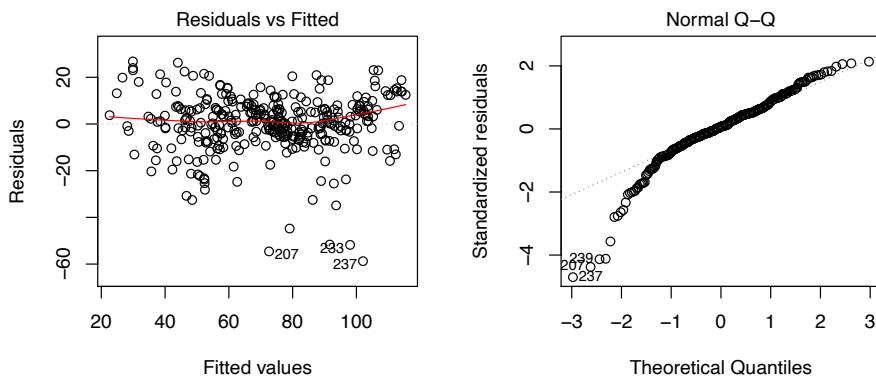
```

Dernæst ser vi på residualplot og qqplot for residualerne for de to modeller.

```
plot(gen_lm, 1:2)
```



```
plot(root_gen_lm, 1:2)
```



Det er klart, at modellen for L ikke fitter data særligt godt. Den krumme trægtform af residualplottet viser, at middelværdien ser misspecifieret ud, og at variansen ikke er konstant.

Modellen for  $\sqrt[3]{L}$  fitter bedre. Der er hverken nogen åbenlys misspecifikation af middelværdien, og variansen ser nogenlunde konstant ud på residualplottet. QQplottet viser dog, at residualerne ikke helt følger en normalfordeling, specielt ikke i den venstre hale. Og vi kan også identificere nogle ekstreme observationer.

## Spørgsmål 3.4

Da  $G \leq C \times G$  er  $L_G \subseteq L_{C \times G}$ , og deraf følger, at

$$L_S + L_G \subseteq L_S + L_{C \times G}.$$

Den additive hypotese  $S + G$  er således en hypotese i modellen specificeret ved  $S + C \times G$ .

Baseret på resultaterne i spørgsmål 3.4 vælger vi at teste den additive hypotese i modellen for  $\sqrt[3]{L}$  ved hjælp af et F-test. Bemærk at de 17 frihedsgrader i testet netop er dimensionsfaldet på  $64 - 47 = 17$ , som kan beregnes på basis af spørgsmål 3.2.

```
add_root_gen_lm <- lm(L^(1/3) ~ S + G, data = gener)
anova(add_root_gen_lm, root_gen_lm)
```

Analysis of Variance Table

	Model 1: $L^{(1/3)} \sim S + G$	Model 2: $L^{(1/3)} \sim S + C * G$
	Res.Df RSS Df Sum of Sq F Pr(>F)	Res.Df RSS Df Sum of Sq F Pr(>F)
1	295 72922	278 51747 17 21175 6.6916 2.227e-13 ***
2		---
		Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Her ser vi at  $p$ -værdien er meget lille ( $2.2 \times 10^{-13}$ ), så vi afviser den additive hypotese.

Havde vi udført testet i modellen for  $L$ , havde vi fået et tilsvarende resultat.

Analysis of Variance Table

	Model 1: $L \sim S + G$	Model 2: $L \sim S + C * G$
	Res.Df RSS Df Sum of Sq F Pr(>F)	Res.Df RSS Df Sum of Sq F Pr(>F)
1	295 2.5793e+13	278 1.8340e+13 17 7.4533e+12 6.6457 2.837e-13 ***
2		---
		Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Spørgsmål 4.1

Modellen er som i Eksempel 5.10 men middelværdi 0 men uden restriktioner på variansmatricen, og derfor er  $\hat{\Sigma} = \frac{1}{n}S$ . Hypotesen  $H_0$  er hypotesen om at  $\Sigma$  er diagonal, og det er den netop hvis  $\Sigma^{-1}$  er diagonal. Mængden af diagonalmatricer,  $M_0 \subseteq \text{Sym}_3$  udgør et underrum af dimension 3, så hypotesen er en lineær hypotese i den kanoniske parameter. Orthogonalprojektionen,  $q_0$ , på  $M_0$  består i at sætte ikke-diagonalindgangene til 0. Det følger igen af Eksempel 5.10 at likelihoodligningen er

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} = q_0(\Sigma) = \frac{1}{n}q_0(S) = \frac{1}{n} \begin{pmatrix} S_{11} & 0 & 0 \\ 0 & S_{22} & 0 \\ 0 & 0 & S_{33} \end{pmatrix},$$

hvorfaf det følger, at  $\hat{\sigma}_1^2 = \frac{1}{n}S_{11}$ ,  $\hat{\sigma}_2^2 = \frac{1}{n}S_{22}$ ,  $\hat{\sigma}_3^2 = \frac{1}{n}S_{33}$ .

Fra korollar 5.9, formel (14), følger det, at  $\log Q = \frac{n}{2}(\log \det(S) - \log \det(n\hat{\Sigma}_0)) = \frac{n}{2}(\log \det(S) - \log(S_{11}S_{22}S_{33}))$ , eller

$$Q = \left( \frac{\det(S)}{S_{11}S_{22}S_{33}} \right)^{n/2}.$$

## Spørgsmål 4.2

Da  $\text{Sym}_3$  har dimension 6 og hypotesen er en lineær hypotese af dimension 3 følger det af Wilks sætning (BMS, sætning 4.5) at  $-2 \log Q$  er asymptotisk  $\chi^2$ -fordelt med  $6 - 3 = 3$  frihedsgrader. Vi beregner  $p$ -værdien

```
z <- 30 * (sum(log(diag(S))) - determinant(S)$modulus)
c(test = z, pvalue = pchisq(z, 3, lower.tail = FALSE))
```

```
test      pvalue
15.17512045 0.00167295
```

Da  $p$ -værdien er relativt lille afvises hypotesen  $H_0$  om at variansmatricen er diagonal.

# Matematisk Statistik: Vejledende besvarelse af reeksamen

Steffen Lauritzen og Niels Richard Hansen

20. august, 2020

## Spørgsmål 1.1

$X$  har nu samme fordeling som  $\exp(\xi + \sigma Z)$  hvor  $Z \sim N(0, 1)$ . Vi har derfor

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} e^{\xi + \sigma z} \varphi(z) dz = e^{\xi} \int_{-\infty}^{\infty} e^{\sigma z} \varphi(z) dz = e^{\xi} e^{\sigma^2/2} = \exp(\xi + \sigma^2/2)$$

som ønsket.

## Spørgsmål 1.2

Vi bemærker først, at  $\xi = \log \mu - \sigma^2/2$ . Vi har så for log-likelihood funktionen, idet vi ignorerer irrelevante konstantledd ( $\sigma^2$  antages jo kendt)

$$\ell_n(\mu) = \sum_{i=1}^n \frac{(\log x_i - \log \mu + \sigma^2/2)^2}{2\sigma^2}$$

og videre

$$S_n(\mu) = \sum_{i=1}^n \frac{(\log \mu - \log x_i - \sigma^2/2)}{\mu \sigma^2}$$

samt

$$\begin{aligned} I_n(\mu) &= \sum_{i=1}^n \frac{\mu \sigma^2 / \mu - (\log \mu - \log x_i - \sigma^2/2) \sigma^2}{\mu^2 \sigma^4} \\ &= \frac{n}{\mu^2 \sigma^2} - \sum_{i=1}^n \frac{(\log \mu - \log x_i - \sigma^2/2)}{\mu^2 \sigma^2} = \frac{n}{\mu^2 \sigma^2} - \frac{1}{\mu} S_n(\mu). \end{aligned}$$

## Spørgsmål 1.3

Dette kan udledes vha deltametoden: Vi har at

$$\hat{\xi}_n = \frac{\sum_i \log x_i}{n}$$

og da MLE er ækvivariant har vi så

$$\hat{\mu}_n = \exp(\hat{\xi}_n + \sigma^2/2).$$

Idet vi ved, at  $\hat{\xi}_n \sim N(\xi, \sigma^2/n)$  anvender vi deltametoden for funktionen

$$f(t) = \exp(t + \sigma^2/2)$$

og får  $f'(t) = \exp(t + \sigma^2/2)$  og derfor er

$$\hat{\mu}_n \stackrel{\text{as}}{\sim} N\left(\mu \exp(2\xi + \sigma^2), \frac{\sigma^2}{n}\right) = N\left(\mu, \frac{\mu^2 \sigma^2}{n}\right).$$

Alternativt kan vi bruge resultatet fra det forrige spm. Da  $\mathbf{E}S_n(\mu) = 0$  er Fisher informationen

$$i_n(\mu) = \mathbf{E}(I_n(\mu)) = \frac{n}{\mu^2 \sigma^2} - \mathbf{E}(S_n(\mu)) = \frac{n}{\mu^2 \sigma^2} - 0 = \frac{n}{\mu^2 \sigma^2}$$

og derfor

$$\hat{\mu}_n \stackrel{\text{as}}{\sim} N\left(\mu, \frac{\mu^2 \sigma^2}{n}\right).$$

### Spørgsmål 1.4

Idet likelihood ratio teststørrelsen er ækvivariant, kan vi omformulere hypotesen til parameteren  $\xi$  som  $H_0 : \xi = \log 6 - 1/8$ . LR testet bliver derfor et simpelt Z-test og forkaster for numerisk store værdier af

$$Z = \sqrt{10} \frac{\sum_i \log x_i / 10 - \log 6 + 1/8}{\sigma} = 0.948$$

svarende til en  $p$ -værdi på 0.343, så observationerne understøtter hypotesen fint.

Man kan naturligvis også direkte beregne likelihood ratio størrelsen og bruge de asymptotiske resultater. Det giver samme konklusion men lidt andre talværdier.

### Spørgsmål 1.5

Idet  $m(\mu) = \mathbf{E}(X) = \mu$  er

$$\tilde{\mu}_n = \frac{\sum X_i}{n}.$$

### Spørgsmål 1.6

Vi skal finde  $\mathbf{V}(X)$  og har derfor brug for andet moment i fordelingen:

$$\mathbf{E}(X^2) = \int_{-\infty}^{\infty} e^{2\xi+2\sigma z} \varphi(z) dz = e^{2\xi} e^{2\sigma^2} = \exp(2\xi + 2\sigma^2)$$

og videre

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \exp(2\xi + 2\sigma^2) - \exp(2\xi + \sigma^2) = \mu^2(e^{\sigma^2} - 1).$$

Vi har tidligere beregnet Fisher information til  $i_n(\mu) = n/(\mu^2 \sigma^2)$ . Idet  $\tilde{\mu}_n$  per definition er en central estimator af  $\mu$  giver Cramer–Raos ulighed at

$$\mathbf{V}(\tilde{\mu}_n) = \frac{\mu^2}{n}(e^{\sigma^2} - 1) \geq \frac{\mu^2 \sigma^2}{n}.$$

Idet vi kan rækkeudvikle eksponentialfunktionen har vi

$$e^{\sigma^2} - 1 = \sigma^2 + \sum_{k=2}^{\infty} \frac{(\sigma^2)^k}{k!}$$

så forskellen er især stor for store værdier af  $\sigma^2$ , så MLE er klart at foretrække i det tilfælde, når  $n$  er stor, idet den nedre grænse er lig med MLEs asymptotiske varians.

### Spørgsmål 1.7

Simulation til sammenligning af estimatorer. Der arbejdes med tre forskellige værdier af  $\sigma$ .

```
M <- 5000
n <- 10
truexi <- 0
truesigma <- 0.1
truemean <- exp(truexi+truesigma^2/2)
```

Klargøring af arrays til resultaterne

```

muhat <- rep(0, M)
mutilde <- rep(0, M)
xihat <- rep(0,M)

```

Selve simulationerne

```

for (i in 1:M)
{
  simx <- rnorm(n)  # simulerede standardnormalfordelte
  simy <- exp(truexi+truesigma*simx)

  xihat[i] = mean(log(simy))
  muhat[i] <- exp(xihat[i]+truesigma^2/2)      # mle

  mutilde[i]<-mean(simy)                         # alternativ estimator

}

```

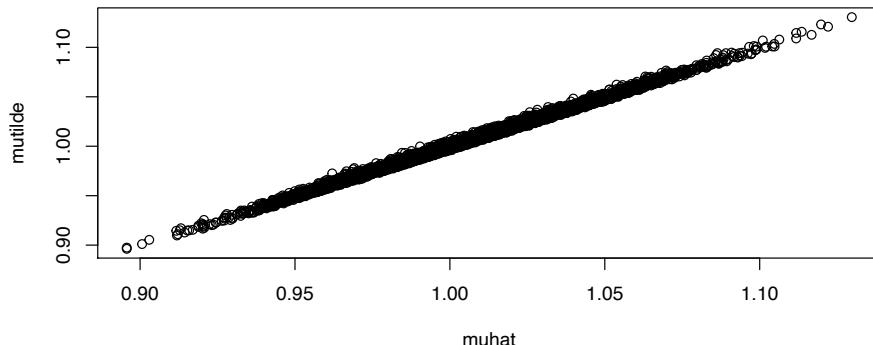
Scatterplot.

Der er næsten ingen forskel paa de to estimatorer, naar sigma er lille.

```

simData <- data.frame(muhat,mutilde)
plot(simData)

```



Resultaterne

```

truemean
[1] 1.005013
mean(mutilde)
[1] 1.005311
mean(muhat)
[1] 1.005832
sd(muhat)
[1] 0.03240237
sd(mutilde)
[1] 0.03245558

```

MLE har lidt mindre varians. Vi ser i stedet paa mean square error: MLE har lige akkurat den mindste MSE

```

msehat <- (mean(muhat)-truemean)^2+sd(muhat)^2
msehat
[1] 0.001050585

msetilde <- (mean(mutilde)-truemean)^2+sd(mutilde)^2
msetilde
[1] 0.001053454

Ønsker fælles akser
myRange <- range(c(muhat, mutilde))

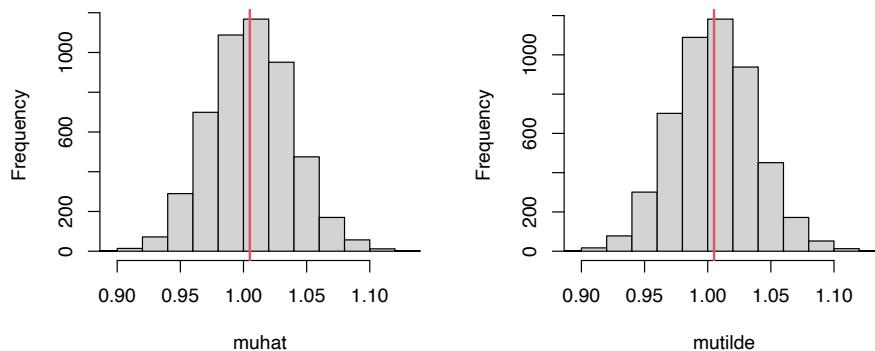
```

Histogrammer med sande værdi som lodret linie. De er næsten helt ens.

```

par(mfrow=c(1,2))
hist(muhat, xlim=myRange, main="")
abline(v=truemean, col=2, lwd=2)
hist(mutilde, xlim=myRange, main="")
abline(v=truemean, col=2, lwd=2)

```

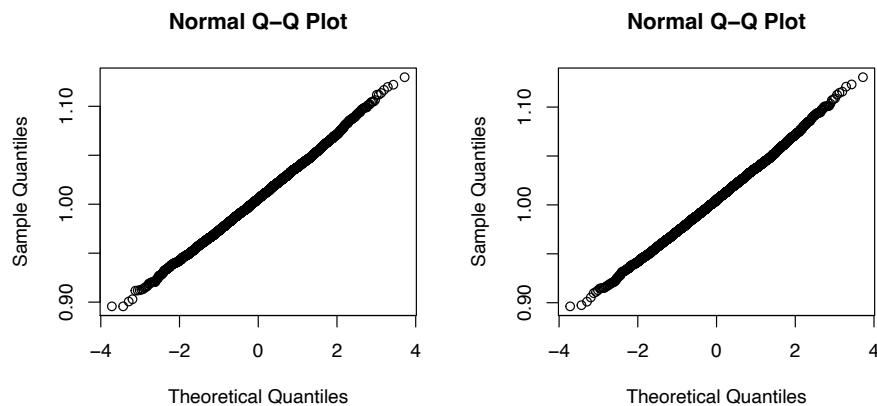


Og de er begge ret fint normalfordelte

```

par(mfrow=c(1,2))
qqnorm(muhat)
qqnorm(mutilde)

```



Gentag for større værdi af truesigma.

```

truesigma <- 0.5
truemean <- exp(truexi+truesigma^2/2)

```

Klargøring af arrays til resultaterne

```

muhat <- rep(0, M)
mutilde <- rep(0, M)
xihat <- rep(0,M)

```

Selve simulationerne

```

for (i in 1:M)
{
  simx <- rnorm(n)  # simulerede standardnormalfordelte
  simy <- exp(truexi+truesigma*simx)

  xihat[i] = mean(log(simy))
  muhat[i] <- exp(xihat[i]+truesigma^2/2)      # mle

  mutilde[i]<-mean(simy)                         # alternativ estimator
}

```

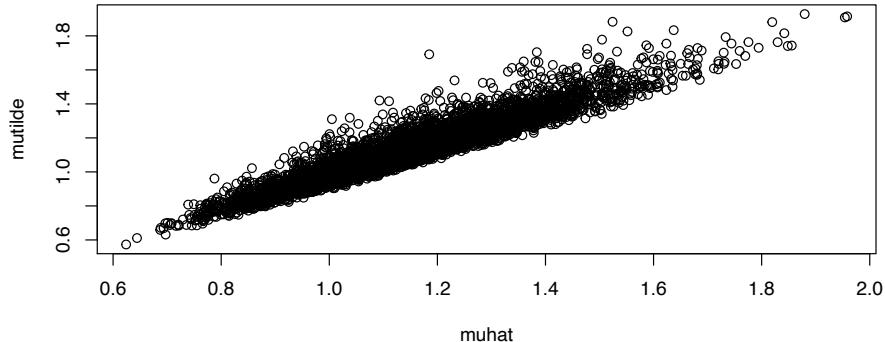
Scatterplot.

Nu er forskellen lidt større.

```

simData <- data.frame(muhat,mutilde)
plot(simData)

```



Resultaterne

```

truemean
[1] 1.133148
mean(mutilde)
[1] 1.132086
mean(muhat)
[1] 1.147048
sd(muhat)
[1] 0.1824333
sd(mutilde)
[1] 0.1897178

```

MLE har lidt mindre varians, men overvurderer middelvaerdien. Vi ser i stedet på mean square error: MLE har lig akkurat den mindste MSE

```

msehat <- (mean(muhat)-truemean)^2+sd(muhat)^2
msehat
[1] 0.0334751
msetilde <- (mean(mutilde)-truemean)^2+sd(mutilde)^2
msetilde
[1] 0.03599396

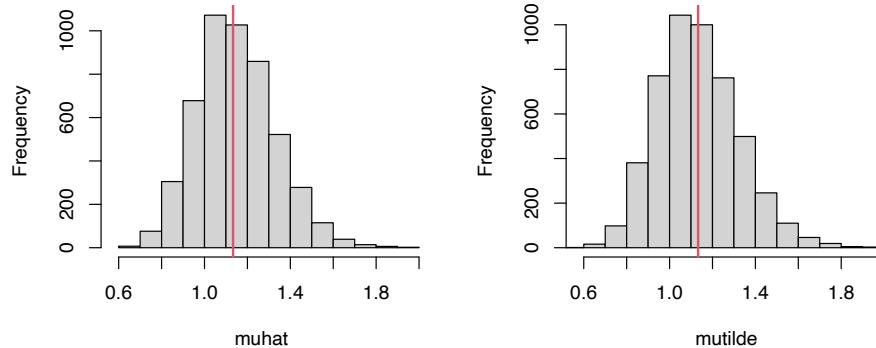
```

Histogrammer med sande værdi som lodret linie. De er ikke helt ens længere.

```

myRange <- range(c(muhat, mutilde))
par(mfrow=c(1,2))
hist(muhat, xlim=myRange, main="")
abline(v=truemean, col=2, lwd=2)
hist(mutilde, xlim=myRange, main="")
abline(v=truemean, col=2, lwd=2)

```

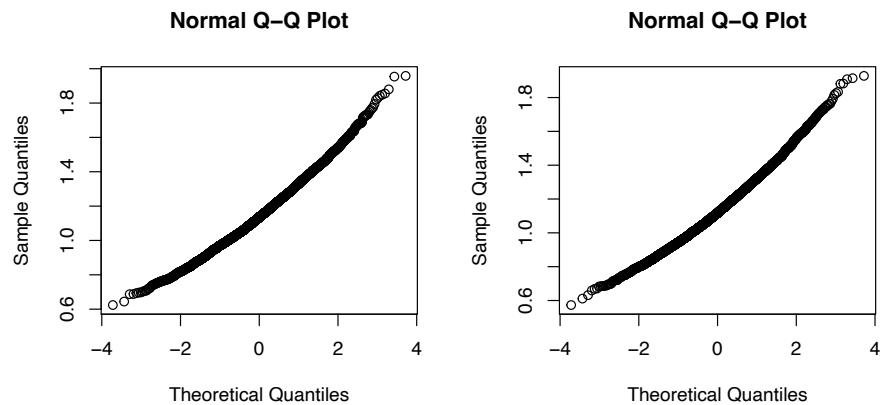


Og de er ikke længere påt normalfordelte.

```

par(mfrow=c(1,2))
qqnorm(muhat)
qqnorm(mutilde)

```



Gentag for meget større værdi af truesigma.

```

truesigma <- 1
truemean <- exp(truexi+truesigma^2/2)

```

Klargøring af arrays til resultaterne

```

muhat <- rep(0, M)
mutilde <- rep(0, M)

```

```
xihat <- rep(0,M)
```

Selve simulationerne

```
for (i in 1:M)
{
  simx <- rnorm(n) # simulerede standardnormalfordelte
  simy <- exp(truexi+truesigma*simx)

  xihat[i] = mean(log(simy))
  muhat[i] <- exp(xihat[i]+truesigma^2/2) # mle

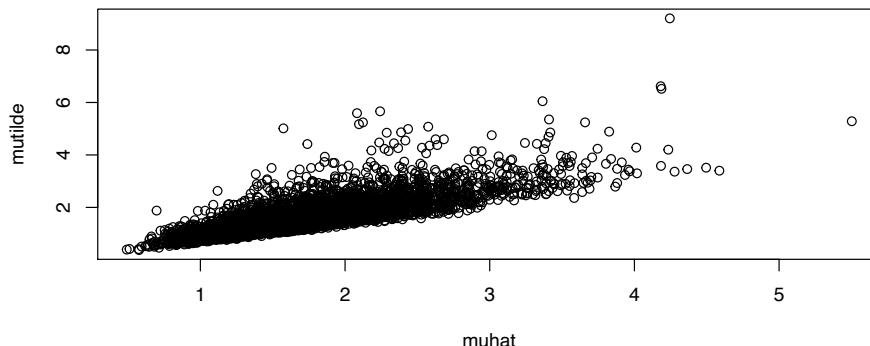
  mutilde[i]<-mean(simy) # alternativ estimator

}
```

Scatterplot.

Nu er de slet ikke ens.

```
simData <- data.frame(muhat,mutilde)
plot(simData)
```



Resultaterne

```
truemean
```

```
[1] 1.648721
```

```
mean(mutilde)
```

```
[1] 1.647151
```

```
mean(muhat)
```

```
[1] 1.737378
```

```
sd(muhat)
```

```
[1] 0.5697215
```

```
sd(mutilde)
```

```
[1] 0.6792565
```

MLE har nu klart mindre varians, men overvurderer stadig middelværdien. Vi ser i stedet på mean square error: MLE har klart den mindste MSE.

```
msehat <- (mean(muhat)-truemean)^2+sd(muhat)^2
```

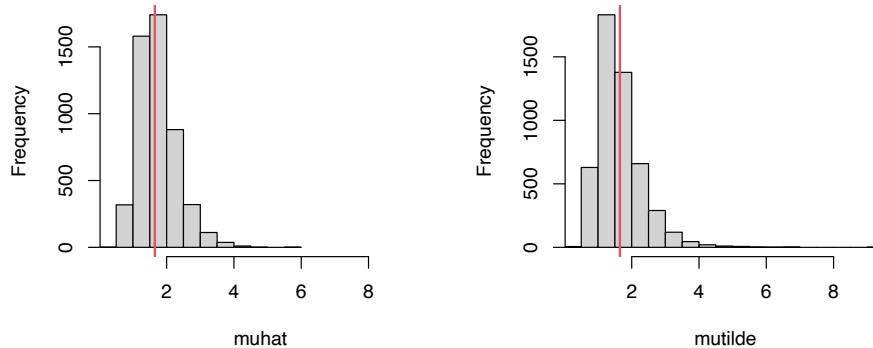
```
msehat
```

```
[1] 0.3324427
msetilde <- (mean(mutilde)-truemean)^2+sd(mutilde)^2
msetilde
```

```
[1] 0.4613919
```

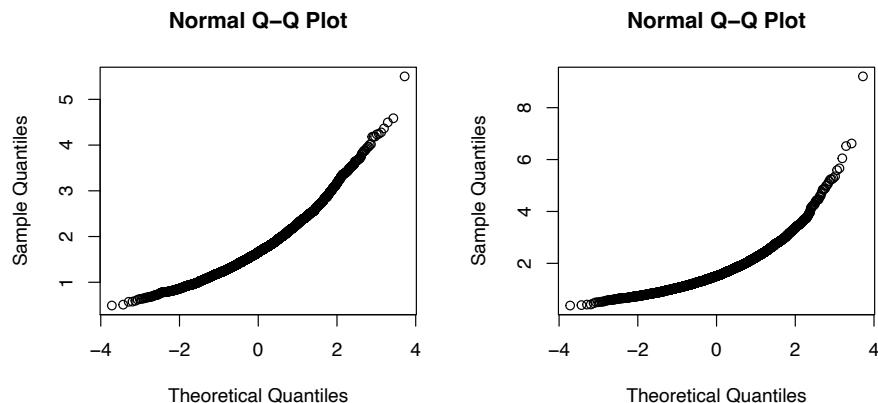
Histogrammer med sande værdi som lodret linie. De er nu meget forskellige og  $\hat{\mu}$  er sommetider meget for stor.

```
myRange <- range(c(muhat, mutilde))
par(mfrow=c(1,2))
hist(muhat, xlim=myRange, main="")
abline(v=truemean, col=2, lwd=2)
hist(mutilde, xlim=myRange, main="")
abline(v=truemean, col=2, lwd=2)
```



Og de er nu meget langt fra at være normalfordelte. Så  $n = 10$  er ikke nok til at den asymptotiske normalfordeling er en god approximation.

```
par(mfrow=c(1,2))
qqnorm(muhat)
qqnorm(mutilde)
```



## Spørgsmål 2.1

Vi indlæser data til opgaven.

```
restaurant <- read_csv("restaurant.txt",
                       col_types = cols(D = col_factor(
                           levels = c("m", "ti", "o", "to", "f", "l", "s"))))
```

Først tabuleres  $T \times B$ .

```
table(restaurant$T, restaurant$B)
```

	br	fr	mi
fa	8	20	8
ta	8	20	8

Vi ser, at alle seks kombinationer af  $T$  og  $B$  forekommer, så  $\dim(L_{T \times B}) = 6$ . Bemærk at alle indgange er positive og rækkerne er identiske, så  $T \wedge B = 1$  og faktorerne opfylder balancealigningen og er således geometrisk ortogonale.

Dernæst tabuleres  $B \times D$ .

```
table(restaurant$B, restaurant$D)
```

	m	ti	o	to	f	l	s
br	0	0	0	0	8	8	
fr	8	8	8	8	0	0	
mi	0	0	0	0	8	8	

Heraf fremgår det, at designgrafen har to sammenhængskomponenter. En svarende til frokost på hverdage, og en svarende til serveringer lørdage og søndage (altså i weekenden). Så minimum har to niveauer, og kan opfattes som en indikator for om det er en weekendservering. Bemærk iøvrigt også at indenfor hver sammenhængskomponent er designet fuldstændigt balanceret, så også  $B$  og  $D$  er geometrisk ortogonale.

Der er tre ikke-trivuelle minima tilbage at undersøge. Det er  $T \times B \wedge D$ ,  $D \wedge T$  og  $T \wedge W$ . Først ser vi at  $1 = B \wedge T \geq W \wedge T \geq 1$ , så  $W \wedge T = 1$ . Tabellen

```
table(interaction(restaurant$T, restaurant$B), restaurant$D)
```

	m	ti	o	to	f	l	s
fa.br	0	0	0	0	4	4	
ta.br	0	0	0	0	4	4	
fa.fr	4	4	4	4	0	0	
ta.fr	4	4	4	4	0	0	
fa.mi	0	0	0	0	4	4	
ta.mi	0	0	0	0	4	4	

viser at  $T \times B \wedge D = W$  og tabellen

```
table(restaurant$T, restaurant$D)
```

	m	ti	o	to	f	l	s
fa	4	4	4	4	8	8	
ta	4	4	4	4	8	8	

viser at  $D \wedge T = 1$ . Fakturstrukturdigrammet (med identitetsfaktoren  $I$  tilføjet) er som følger.

$$\begin{array}{ccccc}
 I^{72} & \longrightarrow & T \times B^6 & \longrightarrow & T^2 \\
 \downarrow & & \downarrow & & \downarrow \\
 D^7 & \longrightarrow & W^2 & \longrightarrow & 1^1
 \end{array}$$

Diagrammet er ovenfor annoteret med dimensioner vi kender på nuværende tidspunkt.

### Spørgsmål 2.2

Tabellerne ovenfor viser at balanceequationen (sætning 14.8) er opfyldt for  $T \times B$  og  $D$ ,  $D \times B$ ,  $D \times T$ , og  $B \times T$ , og disse faktorer er derfor geometrisk ortogonale. De øvrige faktorer i designet pånær  $W$  og  $T$  opfylder en ordningsrelation og er således geometrisk ortogonale ifølge lemma 14.11. Da  $B \geq W \geq B \wedge T = 1$  følger det af lemma 14.12 at også  $W$  og  $T$  er geometrisk ortogonale.

Alternativt kunne man indføre  $W$  i data som en faktor og lave tabellen derfra, men det er lidt mere bøvlet, eller finde  $T \times W$ -tabellen fra  $T \times B$ -tabellen

	weekend	hverdag
fa	16	20
ta	16	20

og indse at den ligeledes opfylder balanceequationen.

Konklusionen er, at designet er geometrisk ortogonalt, og da det ligeledes er  $\wedge$ -stabil kan vi bruge sætning 14.21 til at beregne  $V_G$ -dimensioner. Diagrammet nedenfor er annoteret med disse dimensioner

$$\begin{array}{ccccc}
 I_{61}^{72} & \longrightarrow & T \times B_2^6 & \longrightarrow & T_1^2 \\
 \downarrow & & \downarrow & & \downarrow \\
 D_5^7 & \longrightarrow & W_1^2 & \longrightarrow & 1_1^1
 \end{array}$$

Så

$$\dim(L_D + L_{T \times B}) = 1 + 1 + 5 + 1 + 1 + 2 = 11$$

(eller alternativt  $\dim(L_D + L_{T \times B}) = \dim(L_D) + \dim(L_{T \times B}) - \dim(L_W) = 7 + 6 - 2 = 11$ ) og

$$\dim(L_D + L_T + L_B) = 1 + 1 + 5 + 1 + 1 = 9.$$

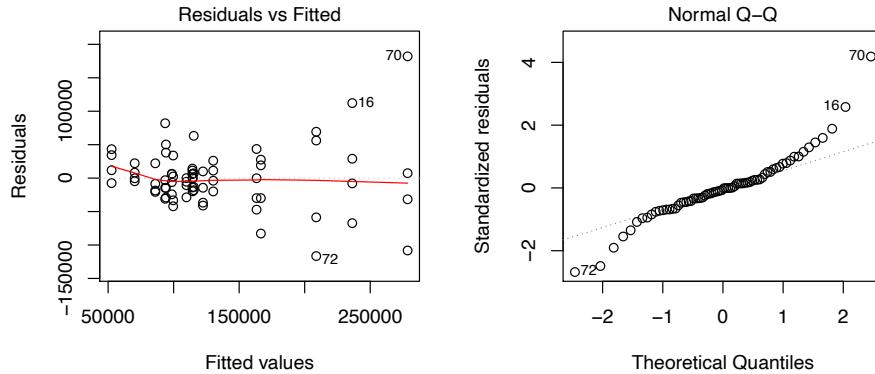
### Spørgsmål 2.3

Vi fitter de to modeller i R.

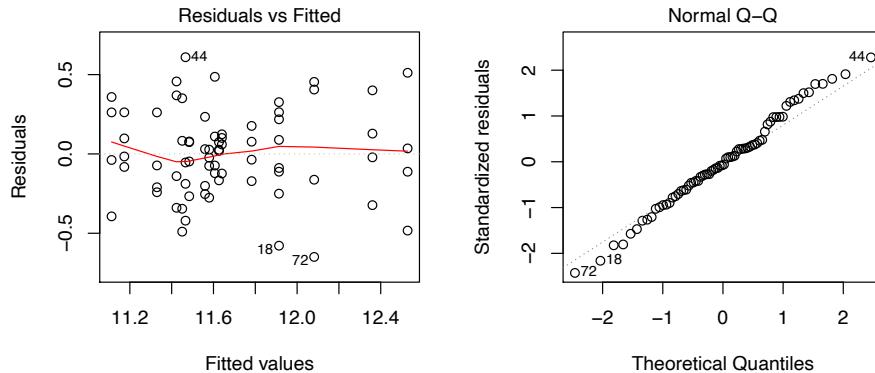
```
rest_lm <- lm(X ~ D + B * T, data = restaurant)
log_rest_lm <- lm(log(X) ~ D + B * T, data = restaurant)
```

Dernæst ser vi på residualplot og qqplot for residualerne for de to modeller.

```
plot(rest_lm, 1:2)
```



```
plot(log_rest_lm, 1:2)
```



Modellen for  $X$  fitter ikke data særligt godt. Der er en tydelig tragtform af residualplottet som viser, at variansen ikke er konstant. QQplottet viser også at residualerne ikke er normalfordelt, men har tungere haler end normalfordelingen.

Modellen for  $\log(X)$  fitter data meget bedre. Der er ingen åbenlyse systematiske afvigelser i residualplottet, og QQplottet viser at residualerne ser pænt normalfordelte ud.

### Spørgsmål 2.4

Det er klart at  $L_D + L_B \subseteq L_D + L_T + L_B$  da vi her blot lægger  $L_T$  til  $L_D + L_B$ . Endvidere er  $L_T + L_B \subseteq L_{T \times B}$ , så når vi lægger  $L_D$  til på begge sider fås

$$L_D + L_T + L_B \subseteq L_D + L_{T \times B}.$$

Vi benytter nedenfor modellen for  $\log(X)$  som udgangspunkt, da den fitter data bedst. Vi kan vælge at teste en effekt af reklameplatformen ( $T$ ) på omsætningen på flere måder. Her vælger vi først at teste den additive model mod modellen med vekselvirkningen, og dernæst at teste den additive effekt af  $T$ .

```
log_rest_lm %>% anova()

Analysis of Variance Table

Response: log(X)
          Df Sum Sq Mean Sq F value    Pr(>F)
D          6 4.7309  0.7885  9.2982 3.121e-07 ***
B          1 4.0118  4.0118 47.3091 3.751e-09 ***
T          1 0.7398  0.7398  8.7237  0.004456 **
B:T        2 0.1175  0.0587  0.6927  0.504121
Residuals 61 5.1727  0.0848
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi aflæser her en  $p$ -værdi på omkring 0.5 for  $F$ -testen for vekselvirkningen mellem  $T$  og  $B$ , så vi kan ikke afvise den additive hypotesen. Det efterfølgende test for at der ikke er en additiv effekt af  $T$  har en  $p$ -værdi omkring 0.0045, som er relativt lille, og vi afviser hypotesen om, at der ikke er en effekt af reklameformen på omsætningen. Vi konkluderer altså på basis af analysen at der er en additiv effekt af reklameplatformen på omsætningen på en log-skala.

Bemærk at anova udregner sekventielle test på en lidt anden måde, end hvis vi udførte dem et ad gangen, men konklusionen er den samme.

```
lm(log(X) ~ D + B + T, data = restaurant) %>% anova()
```

Analysis of Variance Table

```
Response: log(X)
          Df Sum Sq Mean Sq F value    Pr(>F)
D          6 4.7309  0.7885  9.3898 2.368e-07 ***
B          1 4.0118  4.0118 47.7751 2.845e-09 ***
T          1 0.7398  0.7398  8.8097  0.004233 **
Residuals 63 5.2902  0.0840
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Endelig kunne vi også have testet modellen  $D + B$  direkte

```
anova(lm(log(X) ~ D + B, data = restaurant), log_rest_lm)
```

Analysis of Variance Table

```
Model 1: log(X) ~ D + B
Model 2: log(X) ~ D + B * T
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1      64 6.0300
2      61 5.1727  3   0.85724 3.3697 0.0241 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

der giver en  $p$ -værdi på omkring 0.024, hvilket er en lille  $p$ -værdi, men den efterlader os ikke med en klar konklusion. Endvidere belyser dette ene test ikke hvorvidt der er en vekselvirkning eller ej.

## Spørgsmål 2.5

Vi tager udgangspunkt i den additive model fra spørgsmålet ovenfor af log-omsætningen, hvor effekten af reklameformen på omsætningen udtrykkes ved parameteren tilhørende faktoren  $T$ .

```
summary(lm(log(X) ~ D + B + T, data = restaurant)) %>% coef()
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.15019519	0.10799406	103.2482268	5.318629e-72
Dti	0.24979254	0.14488924	1.7240241	8.960772e-02
Do	0.27693412	0.14488924	1.9113506	6.051101e-02
Dto	0.45288933	0.14488924	3.1257624	2.682046e-03
Df	0.31059686	0.14488924	2.1436848	3.592506e-02
Dl	0.48484573	0.13553147	3.5773664	6.738595e-04
Ds	0.03787857	0.13553147	0.2794817	7.807907e-01
Bmi	0.70814493	0.10245216	6.9119570	2.844753e-09
Tta	0.20272641	0.06830144	2.9681132	4.232502e-03

Vi ser fra ovenstående summary at forskellen på de to reklameformer er estimeret til 0.2027 (parameteren hørende til Tta), som er positiv. Estimatet udtrykker at ta platformen på giver en forøgelse af omsætningen estimeret med en faktor  $\exp(0.2027) = 1.225$  i forhold til fa platformen. Vi kan udregne et (standard) 95% konfidensinterval på log-skalaen, men det er mere informativt at transformere det til den oprindelige skala.

```
confint(lm(log(X) ~ D + B + T, data = restaurant), "Tta")
```

```
2.5 %    97.5 %
Tta 0.06623687 0.3392159
exp(confint(lm(log(X) ~ D + B + T, data = restaurant), "Tta"))
```

```
2.5 %    97.5 %
Tta 1.06848 1.403846
```

Intervallet udtrykker altså at ta platformen forøger salget med en faktor mellem 1.068 og 1.404 i forhold til fa platformen.

## Spørgsmål 3.1

Vi indlæser først data.

```
bodyfat <- read_csv("bodyfat.txt")
n <- nrow(bodyfat)
n ## Der er 240 observationer
```

```
[1] 240
```

Fra sætning 5.5 (se også s. 29 i de supplerende noter) har vi, at MLE for  $\mu$  er gennemsnittet

```
colMeans(bodyfat)
```

```
Weight    BodyFat
178.04917 19.12167
```

og MLE for  $\Sigma$  er  $\frac{1}{n}S$ . Den simpleste måde at beregne MLE på er nok ved brug af cov og en reskalering (da den funktion beregner  $\frac{1}{n-1}S$ .)

```
Sigma_hat <- (n - 1) / n * cov(bodyfat)
Sigma_hat
```

```
Weight    BodyFat
Weight  695.1475 132.66356
BodyFat 132.6636  67.30078
```

Alternativt kan MLE beregnes på den her måde

```
as.matrix(bodyfat) %>%
  scale(scale = FALSE) %>%
  crossprod() %>%
  "/"(n)
```

	Weight	BodyFat
Weight	695.1475	132.66356
BodyFat	132.6636	67.30078

Hypotesen  $H_0$  er et specialtilfælde af  $H$  givet ved (16) i de supplerende noter (og identisk med hypotesen  $H$  givet ved (18) i eksempel 6.5), så korollar 6.3 giver at MLE af  $\mu$  fortsat er gennemsnittet som udregnet ovenfor, og MLE af  $\Sigma$  er diagonalmatricen

```
Sigma_hat %>%
  diag() %>%
  diag()
```

	[,1]	[,2]
[1,]	695.1475	0.00000
[2,]	0.0000	67.30078

Vi beregner først korrelationen.

```
rho <- Sigma_hat[1, 2] / sqrt(Sigma_hat[1, 1] * Sigma_hat[2, 2])
rho
[1] 0.6133426
cor(bodyfat$Weight, bodyfat$BodyFat) ## Alternativ beregning af korrelation
```

[1] 0.6133426

Vi tester hypotesen med et kvotienttest, og eksempel 6.5 giver at  $-2\log Q$  er

```
logQ <- - n * log (1 - rho^2)
logQ
```

[1] 113.2579

Sætning 6.4 giver at under  $H_0$  er  $-2\log Q$  asymptotisk  $\chi^2_1$ -fordelt, så  $p$ -værdien er

```
pchisq(logQ, df = 1, lower.tail = FALSE)
```

[1] 1.894542e-26

som er ekstremt lille, og vi afviser derfor hypotesen  $H_0$  om uafhængighed.

## Spørgsmål 3.2

Vi udregner først  $\Sigma^{-1}$  under hypotesen  $H_1$ . Idet

$$\det(\Sigma) = \sigma_1^2 \sigma_2^2 (1 - 1/4) = \frac{3}{4} \sigma_1^2 \sigma_2^2$$

har vi at

$$\Sigma^{-1} = \frac{4}{3\sigma_1^2 \sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\frac{1}{2}\sigma_1\sigma_2 \\ -\frac{1}{2}\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} = \frac{2}{3} \begin{pmatrix} \frac{2}{\sigma_1^2} & -\frac{1}{\sigma_1\sigma_2} \\ -\frac{1}{\sigma_1\sigma_2} & \frac{2}{\sigma_2^2} \end{pmatrix}.$$

Som opgaven er formuleret, er det selvfølgelig også godt nok at verificere direkte, at  $\Sigma^{-1}\Sigma = I$ .

Det følger af sætning 3.2 at modellen er en minimal og regulær eksponentiel familie med kanonisk parameter

$$(\eta, \Omega) = (\Sigma^{-1}\mu, \Sigma^{-1}) \in \mathbb{R}^2 \times \text{PD}_2,$$

hvor dimensionen af det kanoniske parameterrum er  $2 + 3 = 5$ . Hypotesen  $H_1$  giver ingen restriktioner på  $\eta$ , der fortsat frit kan variere i  $\mathbb{R}^2$ , mens  $\Omega = \Sigma^{-1} = \phi(\sigma_1, \sigma_2)$  for

$$(\sigma_1, \sigma_2) \in (0, \infty)^2.$$

Vi udregner de partielt afledte

$$\partial_{\sigma_1} \phi(\sigma_1, \sigma_2) = \frac{2}{3} \begin{pmatrix} -\frac{4}{\sigma_1^3} & \frac{1}{\sigma_1^2 \sigma_2} \\ \frac{1}{\sigma_1^2 \sigma_2} & 0 \end{pmatrix}$$

og

$$\partial_{\sigma_2} \phi(\sigma_1, \sigma_2) = \frac{2}{3} \begin{pmatrix} 0 & \frac{1}{\sigma_1 \sigma_2^2} \\ \frac{1}{\sigma_1 \sigma_2^2} & -\frac{4}{\sigma_2^3} \end{pmatrix}.$$

Det er klart at disse to matricer er lineært uafhængige da f.eks.  $\partial_{\sigma_1} \phi(\sigma_1, \sigma_2)_{11} \neq 0$  mens  $\partial_{\sigma_2} \phi(\sigma_1, \sigma_2)_{11} = 0$ . Jacobianen har derfor altid fuld rang. Endelig ser vi at  $\phi$  er en homeomorfi idet den har en invers på sit billede, der er restriktionen af den globalt kontinuerte afbildning

$$\text{PD}_2 \ni \Omega = \begin{pmatrix} \omega_1 & \omega_{12} \\ \omega_{12} & \omega_2 \end{pmatrix} \mapsto (\sqrt{4/(3\omega_1)}, \sqrt{4/(3\omega_2)}).$$

Vi har hermed vist at betingelserne i definition 2.25 i BMS er opfyldt med  $k = 5$  dimensionen af det kanoniske parameterrum for den eksponentielle familie, og med  $B = \mathbb{R}^2 \times (0, \infty)^2$ , som er en åben delmængde af  $\mathbb{R}^4$ , hvorfor  $m = 4$ . Dermed specificerer  $H_1$  en krum eksponentiel familie af dimension 4 og orden 5.

# Matematisk Statistik: Vejledende besvarelse af eksamen

Steffen Lauritzen og Niels Richard Hansen

24. juni, 2021

## Spørgsmål 1.1

Vi omskriver tæthedens idet vi har  $\beta = 1/\theta_2$  og  $\lambda = e^{\theta_1}/\theta_2$ :

$$f_{\theta}(x,y) = \exp \left\{ \theta_1 x - y\theta_2 - \frac{e^{\theta_1}}{\theta_2} + \log \theta_2 \right\} \frac{y^x}{(x!)^2}$$

hvoraf vi får den kanoniske stikprøvefunktion og kumulantfunktionen til

$$t(x,y) = \begin{pmatrix} x \\ -y \end{pmatrix}, \quad \psi(\theta) = \frac{e^{\theta_1}}{\theta_2} - \log \theta_2.$$

Det kanoniske parameterrum bliver  $\Theta = \mathbb{R} \times \mathbb{R}_+$  som er en åben og konveks delmængde af  $\mathbb{R}^2$ . Familien er minimalt repræsenteret, idet  $\alpha_1 X - \alpha_2 Y$  kun er konstant næsten sikkert, hvis  $\alpha_1 = \alpha_2 = 0$ .

## Spørgsmål 1.2

Vi differentierer kumulantfunktionen og får

$$\mathbf{E}(X) = \frac{\partial}{\partial \theta_1} \psi(\theta) = \frac{e^{\theta_1}}{\theta_2} = \frac{\lambda/\beta}{1/\beta} = \lambda, \quad \mathbf{E}(Y) = -\frac{\partial}{\partial \theta_2} \psi(\theta) = \frac{e^{\theta_1}}{\theta_2^2} + \frac{1}{\theta_2} = \frac{\lambda+1}{\theta_2} = \beta(\lambda+1).$$

Bemærk, at vi selvfølgelig vidste at  $\mathbf{E}(X) = \lambda$  fra specifikationen af fordelingen, hvor  $X$  var Poisson med middelværdi  $\lambda$ .

For varianserne får vi tilsvarende

$$\mathbf{V}(X) = \frac{\partial^2}{\partial \theta_1^2} \psi(\theta) = \frac{e^{\theta_1}}{\theta_2^2} = \lambda, \quad \mathbf{V}(Y) = \frac{\partial^2}{\partial \theta_2^2} \psi(\theta) = 2\frac{e^{\theta_1}}{\theta_2^3} + \frac{1}{\theta_2^2} = \beta^2(2\lambda+1)$$

og for kovariansen

$$\mathbf{V}(X, Y) = -\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \psi(\theta) = \frac{e^{\theta_1}}{\theta_2^2} = \beta\lambda.$$

## Spørgsmål 1.3

I en regulær eksponentiel familie er MLE bestemt som momentestimator for den kanoniske stikprøvefunktion, dvs vi skal løse ligningen

$$\bar{X}_n = \lambda, \quad \bar{Y}_n = \beta(\lambda+1),$$

hvilket har den entydige løsning

$$\hat{\lambda}_n = \bar{X}_n, \quad \hat{\beta}_n = \frac{\bar{Y}_n}{\bar{X}_n + 1},$$

som er veldefineret med ssh 1.

Dog kan  $\hat{\lambda}_n = 0$  med positiv ssh., men  $\lim_{n \rightarrow \infty} P(\hat{\lambda}_n > 0) = 1$  og  $\hat{\beta}_n > 0$  altid så MLE er derfor asymptotisk veldefineret.

## Spørgsmål 1.4

Man kan bruge deltametoden på kovariansmatricen for  $X, Y$  eller finde informationsmatricen for  $(\lambda, \beta)$  direkte. Vi gør det sidste. Vi har (ignorerer konstanter i parameteren)

$$\ell_n(\lambda, \beta) = \log \lambda \sum_i X_i - \log \beta \sum_i Y_i - \sum_i Y_i / \beta - n\lambda - n \log \beta$$

og videre ved differentiation

$$S_n(\lambda, \beta) = \begin{pmatrix} \sum_i X_i / \lambda - n \\ -\sum_i X_i / \beta + \sum_i Y_i / \beta^2 - n / \beta \end{pmatrix}$$

og videre ved fortegnsskift og differentiation

$$I_n(\lambda, \beta) = \begin{pmatrix} \sum_i X_i / \lambda^2 & 0 \\ 0 & -\sum_i X_i / \beta^2 + 2 \sum_i Y_i / \beta^3 - n / \beta^2 \end{pmatrix}.$$

Fisherinformationen findes nu som middelværdien af informationsfunktionen:

$$i_n(\lambda, \beta) = \begin{pmatrix} n/\lambda & 0 \\ 0 & -n\lambda/\beta^2 + 2n\beta(\lambda+1)/\beta^3 - n/\beta^2 \end{pmatrix} = n \begin{pmatrix} 1/\lambda & 0 \\ 0 & (\lambda+1)/\beta^2 \end{pmatrix}.$$

Derfor er  $\hat{\lambda}_n$  og  $\hat{\beta}_n$  asymptotisk uafhængige og normalfordelte

$$\hat{\lambda}_n \xrightarrow{\text{as}} N(\lambda, \lambda/n), \quad \hat{\beta}_n \xrightarrow{\text{as}} N\left(\beta, \frac{\beta^2}{n(\lambda+1)}\right).$$

## Spørgsmål 1.5

Hvis man formulerer hypotesen i de kanoniske parametre får man:  $H_0 : \theta_1 = 0$  og dette er en lineær delfamilie af den oprindelige familie, derfor (Thm 2.14 i BMS) igen en regulær eksponentiel familie.

## Spørgsmål 1.6

Igen skal vi blot løse momentligningen

$$\bar{Y}_n = \mathbf{E}(Y) = \beta(\beta+1)$$

som er en andengrads ligning med præcis en positiv løsning

$$\hat{\beta}_n = \frac{\sqrt{4\bar{Y}_n + 1} - 1}{2}.$$

## Spørgsmål 1.7

Vi har  $n = 10$ ,  $\bar{x} = 2.3$  og  $\bar{y} = 2.653$  og derfor

$$\hat{\lambda}_n = 2.3, \quad \hat{\beta}_n = 2.653/3.3 = 0.8039, \quad \hat{\hat{\beta}}_n = 1.204.$$

Det giver kvotientteststørrelsen

$$\Lambda_n = 2n \left( \bar{X}_n \log \hat{\lambda}_n - \bar{X}_n \log \hat{\beta}_n - \frac{\bar{Y}_n}{\hat{\beta}_n} - \hat{\lambda}_n - \log \hat{\beta}_n + \frac{\bar{Y}_n}{\hat{\hat{\beta}}_n} + \hat{\beta}_n + \log \hat{\hat{\beta}}_n \right).$$

Vi indsætter værdierne og får  $\Lambda_n = 12.58$  hvilket skal sammenlignes med en  $\chi^2$ -fordeling med 1 frihedsgrad, så den er højsignifikant.

Man kan også lave et Waldbaseret test. Vi får den asymptotiske varians for  $\hat{\lambda}_n - \hat{\beta}_n$  til

$$\mathbf{V}(\hat{\lambda}_n - \hat{\beta}_n) \approx \frac{1}{n} \left( \lambda + \frac{\beta^2}{\lambda+1} \right) = \frac{1}{n} \beta \left( \frac{2\beta+1}{\beta+1} \right),$$

hvor det sidste lighedstegn kun gælder under hypotesen.

Vi får så de to Waldstørrelser

$$W_n = n \frac{\hat{\beta}_n + 1}{\hat{\beta}_n(2\hat{\beta}_n + 1)} (\hat{\lambda}_n - \hat{\beta}_n)^2 = 12.02$$

og

$$\tilde{W}_n = n \frac{\hat{\lambda}_n + 1}{\hat{\beta}_n^2 + \hat{\lambda}_n^2 + \hat{\lambda}_n} (\hat{\lambda}_n - \hat{\beta}_n)^2 = 8.97$$

hvilket fører til samme klare konklusion.

## Spørgsmål 2.1

Vi indlæser data til opgaven.

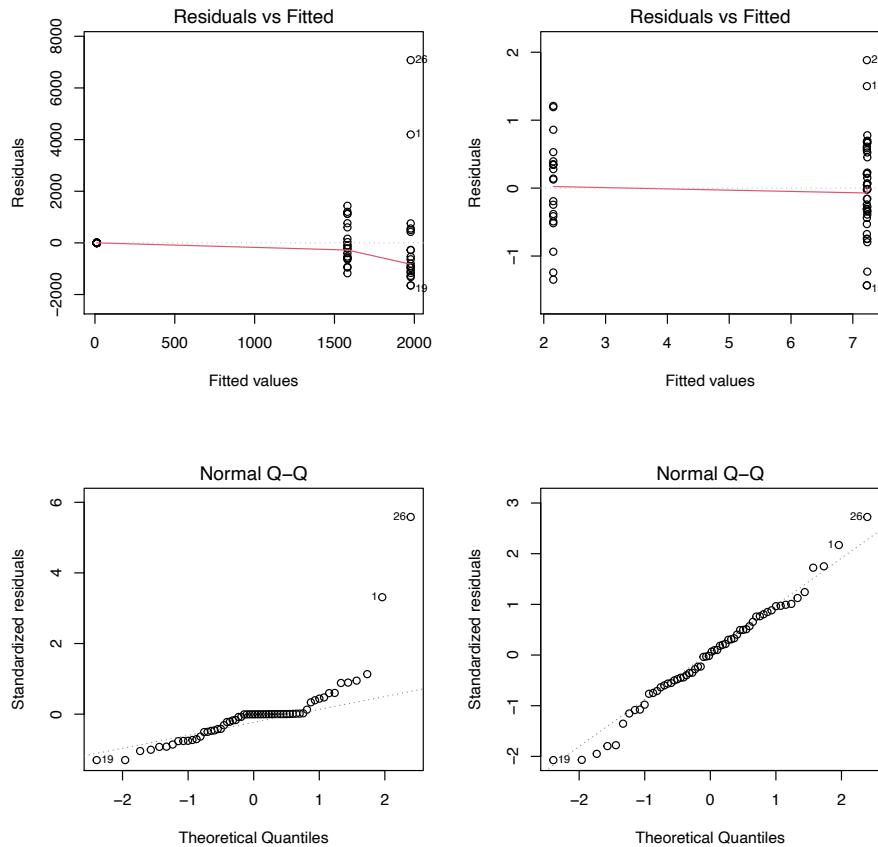
```
library(tidyverse)
vaccine <- read_csv("covid19vaccine.txt", col_types = cols(U = col_factor()))
vaccine_uge5 <- filter(vaccine, U == 5)
```

Vi fitter de to modeller som angivet i opgaven med `lm`.

```
vaccine_lm_1 <- lm(X ~ T, data = vaccine_uge5)
vaccine_lm_1_log <- lm(log(X) ~ T, data = vaccine_uge5)
```

Standard residual- og QQ-plot for begge modeller laves.

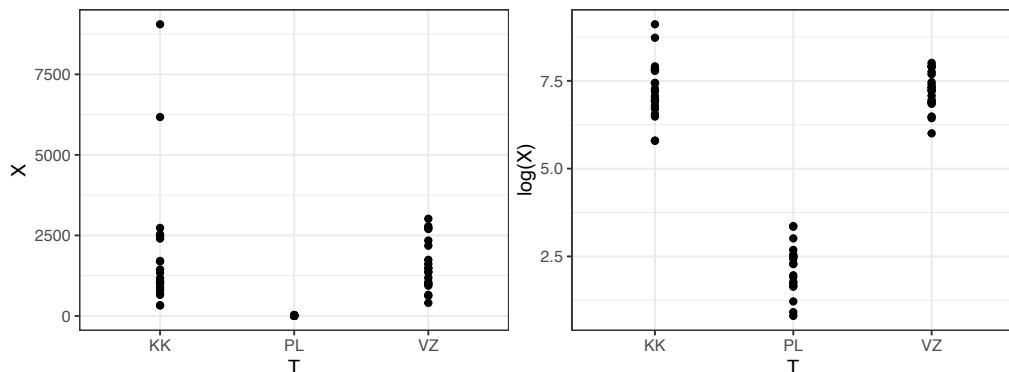
```
plot(vaccine_lm_1, c(1, 2))
plot(vaccine_lm_1_log, c(1, 2))
```



Vi ser (venstre figurer) at den utransformerede model ikke fitter særligt godt. Residualplottet viser klar variansheterogenitet, og QQ-plottet viser at residualerne har en asymmetrisk fordeling. Den log-transformerede model (højre figurer) fitter derimod data godt.

Bemærk at simple plot af datapunkterne for de tre grupper giver en tilsvarende konklusion.

```
ggplot(vaccine_uge5, aes(T, X)) + geom_point()
ggplot(vaccine_uge5, aes(T, log(X))) + geom_point()
```



En normalfordelingsmodel med ens varianser fitter ikke for de utransformerede data, men for de log-transformerede data ser det fornuftigt ud. Det er under alle omstændigheder nyttigt at visualisere data på den måde.

## Spørgsmål 2.2

På baggrund af delopgave 2.1 vælger vi at fortsætte analysen med log-transformerede antistofniveauer. Vi benytter anova til at teste hvorvidt alle tre grupper har samme middelværdi i uge 5 på log-skalaen.

```
anova(vaccine_lm_1_log)
```

Analysis of Variance Table

```
Response: log(X)
          Df Sum Sq Mean Sq F value    Pr(>F)
T          2  343.62 171.808  341.53 < 2.2e-16 ***
Residuals 57   28.67    0.503
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Som det fremgår at tabellen er  $F$ -testet meget stort og  $p$ -værdier meget lille. Vi kan derfor afvise hypotesen om at middelværdierne er ens i de tre grupper. Det fremgår endvidere også ret klart af figuren ovenfor, at antistofniveauet i placebogruppen er markant lavere end i de vaccinerede grupper. Vi har derfor dokumenteret en forskel, og figuren ovenfor viser, at begge vacciner giver et betydeligt højere antistofniveau end placebo.

Men antistofniveauerne i de to grupper, der er vaccineret, ser ikke markant forskellige ud på figuren. Testet afslører ikke, om der er forskel på de to grupper. Vi kan imidlertid se på de estimerede parametre.

```
summary(vaccine_lm_1_log) %>% coef()
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.227355037	0.1585963	45.57077784	1.548578e-46
TPL	-5.073055425	0.2242890	-22.61838860	3.757015e-30
TVZ	0.006945167	0.2242890	0.03096526	9.754054e-01

Her er parameteren angivet som TVZ ovenfor et estimat for netop forskellen i middelværdierne i de to vaccinerede grupper. (Interceptet er middelværdien i KK gruppen). Estimatet for forskellen er tæt på 0, og med en lille  $t$ -værdi og

en  $p$ -værdi på 0.98 kan vi bestemt ikke afvise at antistofniveauet har samme middelværdi (på log-skalaen) for de to vacciner. Vi kan derfor ikke dokumentere nogen forskel på de to vacciner.

Alternativt kan vi fitte modellen med en fælles middelværdi for de to grupper og bruge anova til at udføre testet.

```
vaccine_lm_2_log <- lm(log(X) ~ S, data = vaccine_uge5)
anova(vaccine_lm_2_log, vaccine_lm_1_log)
```

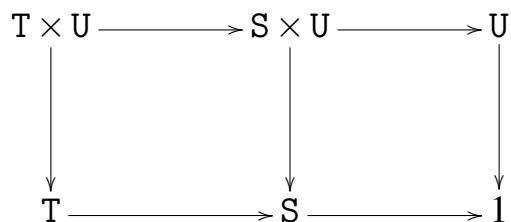
#### Analysis of Variance Table

	Model 1: $\log(X) \sim S$	Model 2: $\log(X) \sim T$				
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	58	28.675				
2	57	28.674	1	0.00048235	0.001	0.9754

Konklusionen er den samme.  $F$ -teststørrelsen bliver meget lille, og  $p$ -værdien bliver den samme (husk,  $F$ -testet og  $t$ -testet er her ækvivalente).

### Spørgsmål 2.3

Vi ved fra konstruktionen af  $S$  at  $S \leq T$ . Forsøgsdesignet er sådan, at  $U$  ikke er sammenlignelig med  $S$  eller  $T$  (hvilket bekræftes af tabellerne nedenfor), og vi får derfor følgende faktorstrukturdiagram.



For at undersøge om designet er  $\wedge$ -stabilt, skal vi finde minimum for alle par af ikke-sammenlignelige faktorer, dvs.  $(T, S \times U)$ ,  $(T, U)$  og  $(S, U)$ .

Vi ser først på  $(T, S \times U)$ , og finder ved krydstabulering

```
table(vaccine$T, interaction(vaccine$U, vaccine$S))
```

	1.PL	2.PL	3.PL	4.PL	5.PL	1.VA	2.VA	3.VA	4.VA	5.VA
KK	0	0	0	0	0	20	20	20	20	20
PL	20	20	20	20	20	0	0	0	0	0
VZ	0	0	0	0	0	20	20	20	20	20

Tabellen viser, at designgrafen har to sammenhængskomponenter. En svarende til placebovaccine og en svarende til vaccine, dvs.  $T \wedge S \times U = S$ .

Vi ser dernæst på  $(T, U)$ , og finder ved krydstabulering

```
table(vaccine$T, vaccine$U)
```

	1	2	3	4	5
KK	20	20	20	20	20
PL	20	20	20	20	20
VZ	20	20	20	20	20

Alle indgange er  $20 > 0$ , og dermed er designgrafen fuldstændig og specielt sammenhængende. Dvs.  $T \wedge U = 1$ .

Endelig ser vi (kunne også afgøres med endnu en tabel), at

$$S \wedge U = (S \wedge T) \wedge U = S \wedge (T \wedge U) = S \wedge 1 = 1.$$

Dermed er  $T \wedge S \times U = S \in \mathbb{G}$  og  $T \wedge U = S \wedge U = 1 \in \mathbb{G}$ , og designet  $\mathbb{G}$  er  $\wedge$ -stabil, jf. også EH eksempel 14.20 for et tilsvarende design.

## Spørgsmål 2.4

EH lemma 14.11 giver at sammenlignelige faktorer er geometrisk ortogonale, så det er igen nok at tjekke balancealigningen for de tre par  $(T, S \times U)$ ,  $(T, U)$  og  $(S, U)$ .

Tabellen for  $(T, S \times U)$  ovenfor ses at opfylder balancealigningen indenfor hver af de to sammenhængskomponenter idet alle indgange i hver komponent er 20. Ifølge EH sætning 14.8 er  $T$  og  $S \times U$  geometrisk ortogonale.

Tabellen for  $(T, U)$  ovenfor opfylder balancealigningen, da alle kombinationer forekommer lige mange gange (designet er balanceret), og  $T$  og  $U$  er derfor geometrisk ortogonale ifølge EH sætning 14.8 (her kan EH lemma 13.11 også bruges).

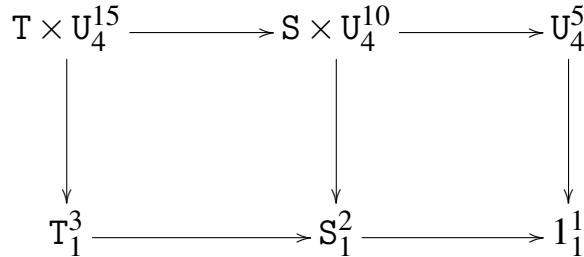
Da  $T \wedge U = 1 \leq S \leq T$  er  $S$  og  $U$  også geometrisk ortogonale ifølge EH lemma 14.12. Det kunne også afgøres ved at bemærke, at tabellen

```
table(vaccine$S, vaccine$U)
```

	1	2	3	4	5
PL	20	20	20	20	20
VA	40	40	40	40	40

ligeledes opfylder balancealigningen.

Vi annoterer faktorstrukturdiagrammet med dimensioner i henhold til den ortogonale dekomposition i EH sætning 14.21, idet vi bemærker at alle faktorer inklusiv de to produktfaktorer er surjektive.



Eftersom  $T \leq T \times U$  og  $S \times U \leq T \times U$  er  $L_T \subseteq L_{T \times U}$  og  $L_{S \times U} \subseteq L_{T \times U}$ , og dermed er

$$L_T + L_{S \times U} \subseteq L_{T \times U}.$$

Vi har  $\dim(L_{T \times U}) = 15$  og

$$\dim(L_T + L_{S \times U}) = 1 + 1 + 1 + 4 + 4 = 11$$

svarende til den ortogonale dekomposition

$$L_T + L_{S \times U} = V_1 + V_S + V_T + V_U + V_{S \times U}.$$

Vi kunne også have fundet dimensionen på to andre måder. Dels

$$\dim(L_T + L_{S \times U}) = \dim(L_T) + \dim(L_{S \times U}) - \dim(L_S) = 3 + 10 - 2 = 11$$

ved brug af EH (13.1), eller ved at observere at

$$V_{T \times U} \perp L_T + L_{S \times U}$$

og

$$L_{T \times U} = L_T + L_{S \times U} + V_{T \times U}$$

så

$$\dim(L_T + L_{S \times U}) = \dim(L_{T \times U}) - \dim(V_{T \times U}) = 15 - 4 = 11.$$

## Spørgsmål 2.5

Det væsentlige i denne delopgave er det forståelsesmæssige indhold i de modeller, som testes. Den tekniske løsning er at teste modellen  $L_{S \times U}$  mod  $L_{T \times U}$ , men med denne løsning skal der følge en forklaring. Alle modeller vil være af  $\log(X)$ , som i delopgave 1 og 2.

Modellen  $L_U$  beskriver at antistofniveauet afhænger af uge, men ikke af vaccine. Vi må forvente at placebovaccinen ikke giver noget antistofrespons, men at vaccinerne gør. De additive modeller  $L_S + L_U$  og  $L_T + L_U$  beskriver antistofresponset som konstant henover alle ugerne. I den første model med den samme konstante forskel for begge vacciner i forhold til placebo, og i den anden model med forskellige konstanter for de to vacciner.

Modellen  $L_{S \times U}$  beskriver at antistofresponset for de to vacciner er det samme, men kan være forskelligt fra placebo. I modsætning til  $L_S + L_U$  behøves forskellen til placebo ikke at være konstant henover ugerne. Modellen  $L_{T \times U}$  beskriver at antistofresponset for de to vacciner kan være forskellige og ligeledes forskellige fra placebo. Derfor tester vi  $L_{S \times U}$  mod  $L_{T \times U}$  for at undersøge, om antistofresponset er det samme for de to vacciner.

```
vaccine_lm <- lm(log(X) ~ T * U, data = vaccine)
vaccine_lm_null <- lm(log(X) ~ S * U, data = vaccine)
anova(vaccine_lm_null, vaccine_lm)
```

Analysis of Variance Table

	Model 1: $\log(X) \sim S * U$	Model 2: $\log(X) \sim T * U$			
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	290	165.00			
2	285	142.22	5	22.784	9.1316 4.517e-08 ***
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

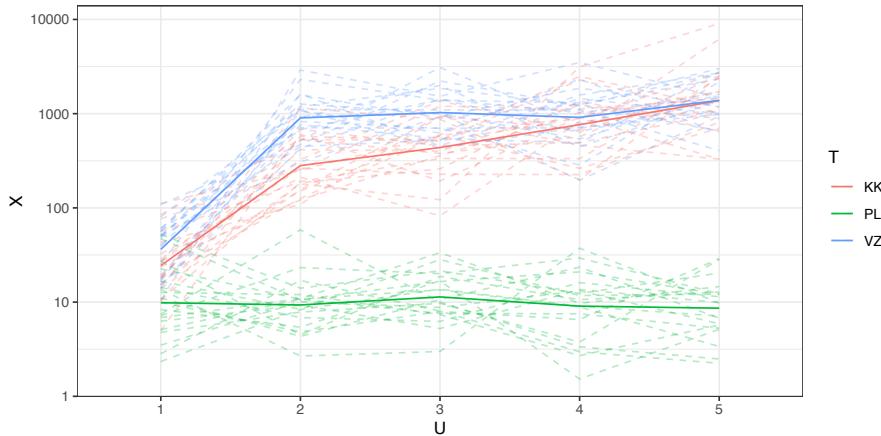
Vi aflæser af ovenstående ANOVA-tabel at  $p$ -værdien for  $F$ -testet for hypotesen

$$H_0 : \xi \in L_{S \times U}$$

er meget lille, og vi forkaster hypotesen. Vi konkluderer, at antistofresponset er forskelligt for de to vacciner.

Data og modellen  $L_{T \times U}$  bør visualiseres, f.eks. på følgende måde:

```
ggplot(vaccine, aes(U, X, color = T)) +
  geom_line(linetype = 2, aes(group = Id), alpha = 0.3) +
  geom_line(stat = "summary", fun = "mean", aes(group = T)) +
  scale_y_log10()
```

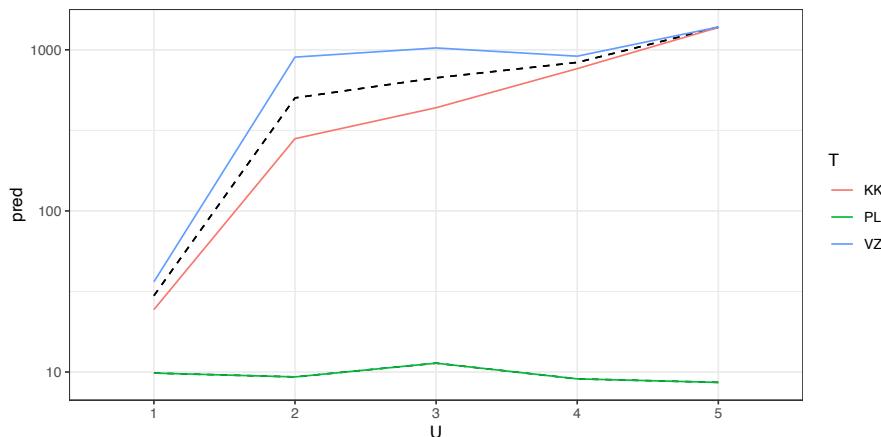


Her viser de fuldtoptrukne linjer middelværdierne svarende til  $L_{T \times U}$  og de stiplede linjer de individuelle antistofresponser. Heraf fremgår det klart at antistofresponset for begge vacciner er forskelligt fra placebo, og at forskellen ikke kan beskrives ved en additiv model. Vi kan også se en mindre forskel mellem vaccinerne, og det er denne forskel som testet ovenfor detekterer.

Vi kunne også visualisere modellen ved at beregne fittede værdier, og her visuelt sammenligne med modellen under hypotesen.

```
vaccine_pred <- mutate(
  vaccine,
  pred = exp(fitted(vaccine_lm)),
  pred_null = exp(fitted(vaccine_lm_null)))
)

ggplot(vaccine_pred, aes(U, pred, color = T, group = T)) +
  geom_line(aes(y = pred_null), linetype = 2, color = "black") +
  geom_line() +
  scale_y_log10()
```



Her er den sorte stiplede linje det fittede antistofrespons for begge vacciner under hypotesen, og vi kan her se afvigelserne til den blå og røde kurve fra modellen  $L_{T \times U}$ .

Inspireret af delopgave 4 er det også muligt at lave en successiv analyse ved at skyde hypotesen  $L_T + L_{S \times U}$  ind mellem  $L_{S \times U}$  og  $L_{T \times U}$ .

```
vaccine_lm_add <- lm(log(X) ~ T + S * U, data = vaccine)
anova(vaccine_lm_null, vaccine_lm_add, vaccine_lm)
```

### Analysis of Variance Table

```

Model 1: log(X) ~ S * U
Model 2: log(X) ~ T + S * U
Model 3: log(X) ~ T * U
Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     290 165.00
2     289 151.46  1   13.5431 27.1400 3.635e-07 ***
3     285 142.22  4    9.2407  4.6295  0.001232 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Vi ser, at  $F$ -testet for

$$H_1 : \xi \in L_T + L_{S \times U}$$

har en lille  $p$ -værdi på ca. 0.001, så selv  $H_1$  forkastes.

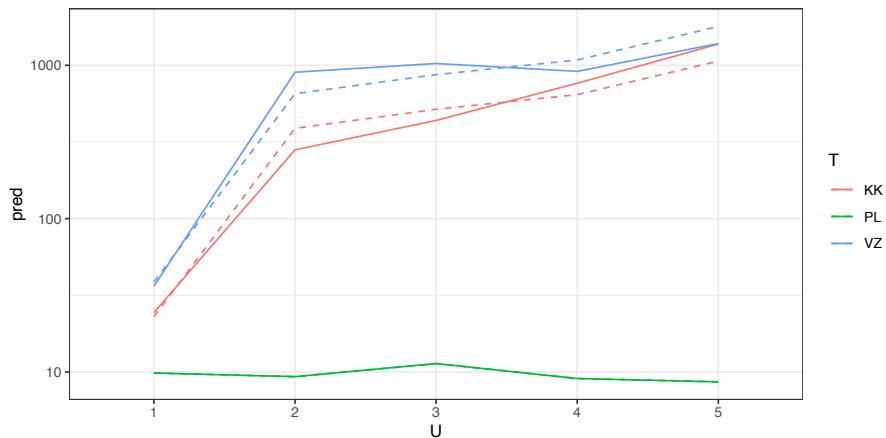
Vi kan også visuelt sammenligne  $L_T + L_{S \times U}$  med  $L_{T \times U}$

```

vaccine_pred$pred_add <- exp(fitted(vaccine_lm_add))

ggplot(vaccine_pred, aes(U, pred, color = T, group = T)) +
  geom_line(aes(y = pred_add), linetype = 2) +
  geom_line() +
  scale_y_log10()

```



Modellen  $L_T + L_{S \times U}$  giver en konstant forskel mellem de to vacciner. Vi kan se, at de to modeller ikke er voldsomt forskellige, men den additive model underestimerer forskellene i uge 2 og 3 og overestimerer forskellene i uge 4 og uge 5.

Den samlede analyse viser, at antistofresponset bedst beskrives af modellen  $L_{T \times U}$ , dvs. at det udvikler sig forskelligt for de to vacciner. I uge 5 når responset omtrædt samme niveau for begge vacciner, men særligt i uge 2 og uge 3 er antistofniveauet mindre for KK-vaccinen end for VZ-vaccinen.

### Spørgsmål 3.1

Estimatet for odds-ratio udregnes for tabellen som i BMS, side 140,

$$\hat{\phi}_{OR} = \frac{223 \times 16}{234 \times 27} = 0.56$$

```

OR_hat <- (223 * 16) / (234 * 27)
OR_hat

```

```
[1] 0.5647357
```

Vi tester hypotesen om samme hyppigheder ved Pearson's  $\chi^2$ , jf. BMS side 139.

```
Y <- matrix(c(223, 234, 27, 16), 2, 2)
EXP <- c(250, 250) %% c(457 / 500, 43 / 500)
Xsq <- det(Y)^2 / 500^2 * (sum(1/EXP))
Xsq
```

```
[1] 3.078724
```

Det giver en  $\chi^2$ -teststørrelse på ca. 3.08. Under hypotesen er  $X^2$  asymptotisk  $\chi^2$ -fordelt med 1 frihedsgrad (BMS s. 138,  $k = 1$ ). Vi finder  $p$ -værdien

```
pval <- pchisq(Xsq, 1, lower.tail = FALSE)
pval
```

```
[1] 0.07932275
```

som er ca. 0.08, og altså ikke specielt lille, så vi kan ikke forkaste hypotesen om at hyppighederne er ens.

En odds-ratio på omtrent 0.5 indikerer, at migræne er mindre hyppigt forekommende ved hjemmearbejde end ved arbejde på kontoret, men vi kan ikke afvise via testet at hyppigheden er den samme (hvilket er ækvivalent med en odds-ratio på 1.)

Testet kan også udføres i R via `chisq.test` – uden Yates' korrektion for at få det samme som i BMS.

```
chisq.test(Y, correct = FALSE)
```

Pearson's Chi-squared test

```
data: Y
X-squared = 3.0787, df = 1, p-value = 0.07932
```

## Spørgsmål 3.2

Bemærk først, at for MLE af log-odds-ratio er den asymptotiske varians,  $se^2 = \frac{1}{V(\hat{\theta}_{10})}$ , angivet i BMS på side 141, hvor altså

$$\hat{\theta}_{10} = \log \hat{\phi}_{OR} \stackrel{as}{\sim} \mathcal{N}(\theta_{10}, se^2),$$

og den estimerede standard error bliver:

```
se_hat <- sqrt(1/234 + 1/223 + 1/16 + 1/27)
se_hat
```

```
[1] 0.3290818
```

Idet  $\phi_{OR} = e^{\theta_{10}}$  følger det af deltametoden (med  $f = \exp$  er  $f' = \exp$ ) at

$$\hat{\phi}_{OR} = e^{\hat{\theta}_{10}} \stackrel{as}{\sim} \mathcal{N}(\phi_{OR}, \underbrace{e^{2\hat{\theta}_{10}} se^2}_{=\phi_{OR}^2}).$$

Vi kunne have startet med den asymptotiske bivariate fordeling af  $(\hat{\pi}_0, \hat{\pi}_1)$  og brugt deltametoden. Det giver (selvfølgelig) det samme, men er noget mere bøvlet.

Dvs. standard error for  $\hat{\phi}_{OR}$  kan estimeres som

$$\hat{\phi}_{OR} \hat{se},$$

og vi får følgende estimat og 95% konfidensinterval.

```

se_OR_hat <- OR_hat * se_hat
se_OR_hat

[1] 0.1858442

OR_hat + c(-1, 1) * 1.96 * se_OR_hat

[1] 0.2004810 0.9289904

```

Vi kan imidlertid også transformere intervallet for log-odds-ratio, som angivet i BMS (7.9), hvilket giver

```

exp(log(OR_hat) + c(-1, 1) * 1.96 * se_hat)

[1] 0.2962955 1.0763793

```

Vi bemærker, at det sidste interval er lidt bredere og skubbet til højre i forhold til det første. Mest bemærkelsesværdigt er det, at det første interval (ret klart) ikke indeholder 1, hvilket modsiger konklusionen fra delopgave 3.1. Det andet interval er ikke i modstrid med konklusionen fra delopgave 3.1. Vi kunne sammenholde med Fishers eksakte test

```

fisher.test(Y)

```

```

Fisher's Exact Test for Count Data

data: Y
p-value = 0.1098
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.2765924 1.1215914
sample estimates:
odds ratio
0.565378

```

som giver et konfidensinterval (og en konklusion), der stemmer overens med det asymptotiske interval baseret på log-odds-ratio.

**Forklaring:** Asymptotikken virker dårligere på odd-ratio skalaen end på log-odds-ratio skalaen, så den direkte brug af den asymptotiske normalfordeling for  $\hat{\phi}_{OR}$  kan lede til konfidensintervaller med forkert / for lille dækningsgrad. De transformerede intervaller vil være mere korrekte. Denne forklaring kunne belyses gennem et lille simulationsekspерiment, men det er ikke forventet af besvarelsen.

# Matematisk Statistik: Vejledende besvarelse af reeksamen

Steffen Lauritzen og Niels Richard Hansen

26. august, 2021

## Spørgsmål 1.1

Vi omskriver tæthedens idet vi har  $\lambda = (\theta_1 - e^{\theta_2})^{-1}$  og  $\beta = e^{\theta_2}$  så:

$$f_{\theta}(x, y) = (\theta_1 - e^{\theta_2}) \exp\{-\theta_1 x + y\theta_2\} \frac{x^y}{y!}$$

hvorfaf vi får den kanoniske stikprøvefunktion og kumulantfunktionen til

$$t(x, y) = \binom{-x}{y}, \quad \psi(\theta) = -\log(\theta_1 - e^{\theta_2}).$$

Det kanoniske parameterrum bliver

$$\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}_+ \times \mathbb{R} : \theta_1 > e^{\theta_2}\}$$

som er en åben og konveks delmængde af  $\mathbb{R}^2$  idet

$$\theta_1 > e^{\theta_2} \text{ og } \eta_1 > e^{\eta_2}$$

medfører at

$$(\theta_1 + \eta_1)/2 > (e^{\theta_2} + e^{\eta_2})/2 > e^{(\theta_2 + \eta_2)/2}.$$

Familien er minimalt repræsenteret, idet  $\alpha_1 X - \alpha_2 Y$  kun er konstant næsten sikkert, hvis  $\alpha_1 = \alpha_2 = 0$ .

## Spørgsmål 1.2

Vi differentierer kumulantfunktionen og får

$$\mathbf{E}(X) = -\frac{\partial}{\partial \theta_1} \psi(\theta) = \frac{1}{\theta_1 - e^{\theta_2}} = \lambda, \quad \mathbf{E}(Y) = \frac{\partial}{\partial \theta_2} \psi(\theta) = \frac{e^{\theta_2}}{\theta_1 - e^{\theta_2}} = \beta \lambda.$$

Bemærk, at vi selvfølgelig vidste at  $\mathbf{E}(X) = \lambda$  fra specifikationen af fordelingen, hvor  $X$  var eksponentialfordelt med middelværdi  $\lambda$ .

For varianserne får vi tilsvarende

$$\mathbf{V}(X) = \frac{\partial^2}{\partial \theta_1^2} \psi(\theta) = \frac{1}{(\theta_1 - e^{\theta_2})^2} = \lambda^2$$

og

$$\mathbf{V}(Y) = \frac{\partial^2}{\partial \theta_2^2} \psi(\theta) = \frac{e^{\theta_2}(\theta_1 - e^{\theta_2}) + e^{2\theta_2}}{(\theta_1 - e^{\theta_2})^2} = \beta \lambda + \beta^2 \lambda^2 = \beta \lambda (1 + \beta \lambda).$$

og for kovariansen

$$\mathbf{V}(X, Y) = -\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \psi(\theta) = \frac{e^{\theta_2}}{(\theta_1 - e^{\theta_2})^2} = \beta \lambda^2.$$

### Spørgsmål 1.3

I en regulær eksponentiel familie er MLE bestemt som momentestimator for den kanoniske stikprøvefunktion, dvs vi skal løse ligningen

$$\bar{X}_n = \lambda, \quad \bar{Y}_n = \beta\lambda,$$

hvilket har den entydige løsning

$$\hat{\lambda}_n = \bar{X}_n, \quad \hat{\beta}_n = \bar{Y}_n / \bar{X}_n.$$

Her er  $\hat{\lambda}_n > 0$  altid og  $\lim_{n \rightarrow \infty} P(\hat{\beta}_n > 0) = 1$  og MLE er derfor asymptotisk veldefineret.

### Spørgsmål 1.4

Man kan bruge deltametoden på kovariansmatricen for  $X, Y$  eller finde informationsmatricen for  $(\lambda, \beta)$  direkte. Vi gør det sidste. Vi har (ignorerer konstanter i parameteren)

$$\ell_n(\lambda, \beta) = -n \log \lambda - \frac{\sum_i X_i}{\lambda} + \log \beta \sum_i Y_i - \beta \sum_i X_i$$

og videre ved differentiation

$$S_n(\lambda, \beta) = \left( \sum_i X_i / \lambda^2 - n / \lambda, \sum_i Y_i / \beta - \sum_i X_i \right)$$

og videre ved fortegnsskift og differentiation

$$I_n(\lambda, \beta) = \begin{pmatrix} 2 \sum_i X_i / \lambda^3 - n / \lambda^2 & 0 \\ 0 & \sum_i Y_i / \beta^2 \end{pmatrix}.$$

Fisherinformationen findes nu som middelværdien af informationsfunktionen:

$$i_n(\lambda, \beta) = \begin{pmatrix} 2n / \lambda^2 - n / \lambda^2 & 0 \\ 0 & n\beta\lambda / \beta^2 \end{pmatrix} = n \begin{pmatrix} 1 / \lambda^2 & 0 \\ 0 & \lambda / \beta \end{pmatrix}.$$

Derfor er  $\hat{\lambda}_n$  og  $\hat{\beta}_n$  asymptotisk uafhængige og normalfordelte

$$\hat{\lambda}_n \xrightarrow{\text{as}} N(\lambda, \lambda^2/n), \quad \hat{\beta}_n \xrightarrow{\text{as}} N\left(\beta, \frac{\beta}{n\lambda}\right).$$

### Spørgsmål 1.5

Momentfunktionen  $m(\lambda, \beta)$  er givet som

$$m_1(\lambda, \beta) = \mathbf{E}_{(\lambda, \beta)}(X) = \lambda$$

og

$$m_2(\lambda, \beta) = \mathbf{E}_{(\lambda, \beta)}(XY) = \mathbf{V}_{(\lambda, \beta)}(X, Y) + \mathbf{E}_{(\lambda, \beta)}(X)\mathbf{E}_{(\lambda, \beta)}(Y) = 2\beta\lambda^2$$

hvoraf vi får følgende momentestimatorer

$$\tilde{\lambda}_n = \hat{\lambda}_n = \sum_i X_i / n, \quad \tilde{\beta}_n = \frac{\sum_i X_i Y_i / n}{2(\sum_i X_i / n)^2}.$$

### Spørgsmål 1.6

Vi anvender deltametoden på funktionen

$$f(u, v) = \frac{v}{2u^2}$$

og får

$$Df(u, v) = \left( \frac{-v}{u^3}, \frac{1}{2u^2} \right)$$

som taget i det relevante punkt bliver

$$Df(\lambda, 2\beta\lambda^2) = \left( \frac{-2\beta}{\lambda}, \frac{1}{2\lambda^2} \right).$$

Den asymptotiske varians for  $\tilde{\beta}_n$  bliver derfor

$$\frac{1}{n} \left( \frac{-2\beta}{\lambda}, \frac{1}{2\lambda^2} \right) \begin{pmatrix} \lambda^2 & 4\beta\lambda^3 \\ 4\beta\lambda^3 & 20\beta^2\lambda^4 + 6\beta\lambda^3 \end{pmatrix} \begin{pmatrix} \frac{-2\beta}{\lambda} \\ \frac{1}{2\lambda^2} \end{pmatrix} = \frac{1}{n} \left( \beta^2 + 3\frac{\beta}{2\lambda} \right)$$

og dermed har vi at  $\tilde{\beta}_n$  er asymptotisk normalfordelt

$$\tilde{\beta}_n \stackrel{\text{as}}{\sim} N \left\{ \beta, \frac{1}{n} \left( \beta^2 + 3\frac{\beta}{2\lambda} \right) \right\}.$$

Vi bemærker, at

$$\beta^2 + 3\frac{\beta}{2\lambda} > \frac{\beta}{\lambda}$$

så denne ‘regressionslignende’ momentestimator har betydelig større varians end maksimaliseringsestimatoren.

## Spørgsmål 1.7

Vi får  $\hat{\lambda} = 1.28$  og  $\hat{\beta} = 2.2656$  og dermed den approximative standardafvigelse på  $\hat{\beta}$

$$\sqrt{\hat{\beta}/(10\hat{\lambda})} = 0.421$$

så konfidensintervallet baseret på momentestimatoren bliver

$$\hat{\beta} \pm 1.96 \times 0.421 = (1.44, 3.09).$$

Tilsvarende har vi  $\tilde{\beta} = 1.4364$  med approximativ standardafvigelse

$$\sqrt{\frac{1}{10} \left( \tilde{\beta}^2 + \frac{3\tilde{\beta}}{2\hat{\lambda}} \right)} = 0.612$$

og det bredere konfidensinterval

$$\tilde{\beta} \pm 1.96 \times 0.612 = (0.24, 2.64).$$

## Spørgsmål 2.1

Vi indlæser først data.

```
hjemlos <- read_csv("hjemlos.txt")
```

Derefter laver vi krydstabulering for  $(B, S)$ ,  $(B, K)$  og  $(S, K)$  for at afgøre, om designet er  $\wedge$ -stabilt.

```
table(hjemlos[, c("B", "S")])
```

		S
B		gade herberg venner
Aalborg	3	7 1
Aarhus	0	5 5
København	3	9 3
Odense	0	4 0
Øvrige	2	10 5

```
table(hjemlos[, c("B", "K")])
```

		K	
		kvinde	mand
B			
Aalborg		3	8
Aarhus		2	8
København		1	14
Odense		0	4
Øvrige		4	13

```
table(hjemlos[, c("S", "K")])
```

		K	
		kvinde	mand
S			
gade		0	8
herberg		8	27
venner		2	12

Vi ser af ovenstående tre tabeller at designet er sammenhængende for alle tre par af faktorer – der er f.eks. i alle tabellerne en søjle og en række uden nuller. Derfor er  $B \wedge S = 1$ ,  $B \wedge K = 1$  og  $S \wedge K = 1$ , og det viser  $\wedge$ -stabilitet.

Da  $S$  og  $K$  udgør et sammenhængende design, er de geometrisk ortogonale netop hvis balanceligningen er opfyldt, jf. EH lemma 13.11. Men tabellen ovenfor for  $(S, K)$  indeholder et 0, og da både  $S$  og  $K$  er surjektive kan balanceligningen ikke være opfyldt. Designet er altså **ikke** ortogonal.

Faktorerne  $B$  og  $K$  såvel som  $B$  og  $S$  er med samme argument som ovenfor heller ikke geometrisk ortogonale.

## Spørgsmål 2.2

Iflg. EH (13.1) er

$$\dim(L_B + L_S) = \dim(L_B) + \dim(L_S) - \dim(L_B \cap L_S).$$

Faktorerne  $B$  og  $S$  er surjektive (ses f.eks. af tabellerne ovenfor) med 5 hhv. 3 forskellige værdier, og  $L_B \cap L_S = L_1$  iflg. EH lemma 13.10, da designet er sammenhængende. Derfor er

$$\dim(L_B + L_S) = 5 + 3 - 1 = 7.$$

At  $a \in L_K$  følger direkte af definition 12.4: Hvis to individer,  $i$  og  $j$ , har samme køn er  $a_i = a_j$  (og lig 1, hvis de er kvinder, og lig 0, hvis de er mænd).

Vi beregner nu residualen  $a - P_{B+S}(a)$  som i vinket til opgaven.

```
lm(K == "kvinde" ~ B + S, data = hjemlos) %>%
  residuals() %>%
  round(2) %>% # This line ...
  unname()    # ... and this line only to simplify output!
```

```
[1] -0.14  0.86 -0.26 -0.30 -0.14 -0.26 -0.03  0.70 -0.26  0.00 -0.19 -0.26
[13] -0.14  0.12 -0.14 -0.30  0.86 -0.35  0.00 -0.03 -0.04  0.00  0.70 -0.04
[25] -0.14  0.65 -0.35 -0.09 -0.24  0.74  0.70 -0.35 -0.09 -0.09 -0.14  0.81
[37] -0.35 -0.30  0.65 -0.03 -0.30 -0.14  0.12  0.00 -0.19 -0.14  0.65 -0.30
[49] -0.30 -0.14 -0.14  0.12 -0.19 -0.14 -0.19 -0.30 -0.14
```

Da  $a - P_{B+S}(a) \neq 0$  er  $P_{B+S}(a) \neq a$  og  $a \notin L_B + L_S$ . Nu er  $L_B + L_S \subset L_B + L_S + L_K$  (der gælder ikke lighed) så

$$\begin{aligned} 7 &= \dim(L_B + L_S) \\ &< \dim(L_B + L_S + L_K) \\ &= \dim(L_B + L_S) + \dim(L_K) - \dim((L_B + L_S) \cap L_K) \\ &\leq 7 + 2 - 1 = 8. \end{aligned}$$

Her har vi udnyttet at  $\dim((L_B + L_S) \cap L_K) \geq \dim(L_1) = 1$  og at  $K$  er surjektiv med 2 forskellige værdier. Pga. den skarpe ulighed ovenfor må

$$\dim(L_B + L_S + L_K) = 8.$$

### Spørgsmål 2.3

Tabellen for  $(B \times K, S \times K)$  er

```
table(interaction(hjemlos$B, hjemlos$K), interaction(hjemlos$S, hjemlos$K))
```

	gade.kvinde	herberg.kvinde	venner.kvinde	gade.mand	herberg.mand	venner.mand
Aalborg.kvinde	0	3	0	0	0	0
Aarhus.kvinde	0	1	1	0	0	0
København.kvinde	0	1	0	0	0	0
Odense.kvinde	0	0	0	0	0	0
Øvrige.kvinde	0	3	1	0	0	0
Aalborg.mand	0	0	0	3	4	1
Aarhus.mand	0	0	0	0	4	4
København.mand	0	0	0	3	8	3
Odense.mand	0	0	0	0	4	0
Øvrige.mand	0	0	0	2	7	4

Der er en nul-række og en nul-søjle svarende til, at ingen kvinder bor i Odense, og at ingen kvinder sover på gaden. Der er altså ikke nogen observationer (kanter) knyttet til de tilsvarende punkter, så hvis vi tog (Odense, kvinde) og (gade, kvinde) med som to sammenhængskomponenter, ville minimum ikke blive surjektiv. Vi ser derfor bort fra nul-rækken og nul-søjlen.

Tilbage står to sammenhængskomponenter, som vi af tabellen kan aflæse svarer til køn (K). Dvs.

$$(B \times K) \wedge (S \times K) = K.$$

Heraf fås fra EH (13.1) at

$$\begin{aligned} \dim(L_{B \times K} + L_{S \times K}) &= \dim(L_{B \times K}) + \dim(L_{S \times K}) - \dim(L_K) \\ &= 9 + 5 - 2 = 12 \end{aligned}$$

F-testet for  $H : L_B + L_S + L_K \vdash L_{B \times K} + L_{S \times K}$  har derfor en  $F$ -fordeling med  $(12 - 8, 57 - 12) = (4, 45)$  frihedsgrader.

### Spørgsmål 2.4

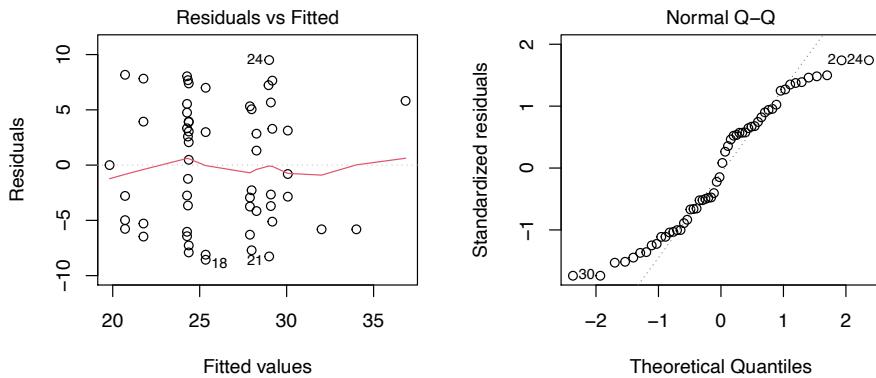
Vi fitter interaktionsmodellen:

```
BMI_int <- lm(BMI ~ B*K + S*K, data = hjemlos)
```

Dernæst ser vi på residual- og QQ-plot (der er kun en kvinde i København, observation 17, så den observation får leverage 1 og udelades af `plot.lm`)

```
plot(BMI_int, 1:2)
```

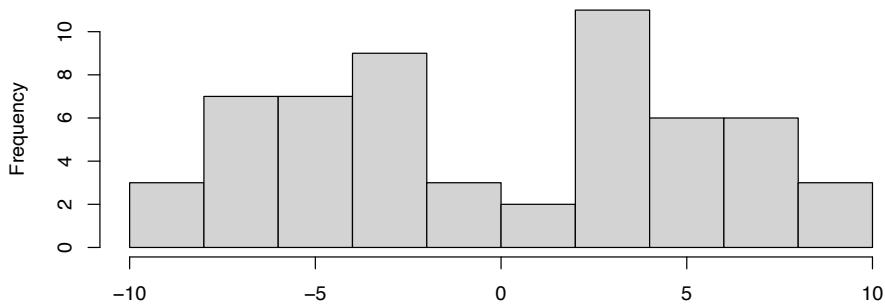
Warning: not plotting observations with leverage one:



Residualplottet ser fornuftigt ud. Der er ingen åbenlys tragtform eller andre indikationer af variansheterogenitet, og der er heller ingen indikationer af at middelværdien er misspecifieret. QQ-plottet viser at residualerne ikke helt følger en normalfordeling. Halerne ser for lette ud sammenlignet med en normalfordeling. Hvis vi i stedet ser på en histogram af residualerne

```
BMI_int %>% residuals() %>% hist()
```

**Histogram of .**



så ser residualfordelingen mere bimodal ud end normalfordelt. Konklusionen er, at modellens lineære middelværdistruktur og konstante varians ser ud til at fitte data godt, mens en normalfordelingsantagelse er mere problematisk.

## Spørgsmål 2.5

Vi fitter nu også den additive model.

```
BMI_add <- lm(BMI ~ B + S + K, data = hjemlos)
```

Dernæst kan vi teste hypotesen

$$H : L_B + L_S + L_K$$

ved et F-test. Vi bruger anova:

```
anova(BMI_add, BMI_int)
```

Analysis of Variance Table

Model 1: BMI ~ B + S + K					
Model 2: BMI ~ B * K + S * K					
	Res.Df	RSS	Df	Sum of Sq	F Pr(>F)
1	49	1840.4			
2	45	1672.0	4	168.39	1.133 0.3531

Med en relativt stor p-værdi på ca. 0.35 kan vi ikke afvise hypotesen. Bemærk at antallet af frihedsgrader, (4, 45), stemmer med de teoretiske udregninger ovenfor.

Man kunne problematisere brugen af F-testets fordeling, da residualerne ikke er normalfordelte, men det er ikke nødvendigt for en fuldstændig besvarelse af denne delopgave.

Vi ser på interceptet (svarende til at sove på gaden) og de to øvrige parameterestimater for  $S$ .

```
summary(BMI_add) %>% coef() %>% .[c(1, 6, 7), ]
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	34.892053	3.284236	10.6240988	2.596725e-14
Sherberg	-5.382758	2.546766	-2.1135658	3.966778e-02
Svenner	-1.243330	2.916430	-0.4263193	6.717422e-01

Modellen angiver at BMI i middel er ca. 5 mindre for hjemløse, der sover på herberg, og 1 mindre for hjemløse, der sover hos venner, sammenlignet med hjemløse, der sover på gaden.

Konfidensintervallerne er dog ret brede:

```
confint(BMI_add, c(1, 6, 7))
```

	2.5 %	97.5 %
(Intercept)	28.292132	41.4919728
Sherberg	-10.500677	-0.2648396
Svenner	-7.104116	4.6174546

Det viser sig, at hypotesen  $H' : L_S + L_K$  (by har ikke betydning) heller ikke kan afvises:

```
BMI_add_simple <- lm(BMI ~ S + K, data = hjemlos)
anova(BMI_add_simple, BMI_add)
```

#### Analysis of Variance Table

Model 1: BMI ~ S + K						
Model 2: BMI ~ B + S + K						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	53	1893.5				
2	49	1840.4	4	53.079	0.3533	0.8405

I den simplere model uden en effekt af by fås estimater for effekten af sovested svarende til estimatorne ovenfor, mens konfidensintervallerne er marginalt kortere.

```
confint(BMI_add_simple, 1:3)
```

	2.5 %	97.5 %
(Intercept)	28.778531	40.810418
Sherberg	-10.876559	-1.279843
Svenner	-7.502803	3.193667

## Spørgsmål 3.1

Vi har følgende estimator for ssh. for en blodprop for de to vacciner:

$$\hat{\pi}_A = 26/721798 = 3.60 \times 10^{-5}$$

og

$$\hat{\pi}_B = 5/221433 = 2.26 \times 10^{-5}.$$

Estimatet af  $\phi_{RR}$  er

$$\hat{\phi}_{RR} = \frac{\hat{\pi}_B}{\hat{\pi}_A} = \frac{5 \times 721798}{26 \times 221433} = 0.6269.$$

Hypotesen  $H : \phi_{RR} = 1$  svarer til hypotesen om homogenitet, dvs.  $\pi_A = \pi_B$ . Vi kan teste hypotesen ved hjælp af Pearsons teststørrelse, som under hypotesen er  $\chi^2$ -fordelt med 1 frihedsgrad. Her udregnes teststørrelse og p-værdi med chisq.test:

```
chisq.test(matrix(c(721772, 221428, 26, 5), 2, 2), correct = FALSE)
```

```
Pearson's Chi-squared test
```

```
data: matrix(c(721772, 221428, 26, 5), 2, 2)
X-squared = 0.93148, df = 1, p-value = 0.3345
```

Da p-værdien på ca. 0.33 er relativt stor kan vi ikke afvise hypotesen om, at  $\phi_{RR} = 1$ , dvs. at risikoen for en blodprop ikke afhænger af vaccinen.

### Spørgsmål 3.2

Estimatet af standard error for  $\log(\phi_{RR})$ ,  $\tilde{s}$ , findes som i BMS, side 140,

$$\begin{aligned}\tilde{s} &= \sqrt{\frac{1 - \hat{\pi}_B}{\hat{\pi}_B \times 221433} + \frac{1 - \hat{\pi}_A}{\hat{\pi}_A \times 721798}} \\ &= \sqrt{\frac{221428}{5 \times 221433} + \frac{721772}{26 \times 721798}} \\ &= 0.4883\end{aligned}$$

Da  $\phi_{RR} = e^{\log(\phi_{RR})}$  giver deltametoden (med  $f = \exp$  og  $f' = \exp$ ) at

$$\phi_{RR} \stackrel{\text{as}}{\sim} \mathcal{N}(\phi_{RR}, \underbrace{e^{2\log(\phi_{RR})}}_{=\phi_{RR}^2} \tilde{s}^2).$$

Dvs standard error af  $\hat{\phi}_{RR}$  estimeres til

$$\hat{\phi}_{RR} \tilde{s} = 0.6269 \times 0.4883 = 0.3061.$$

Et konfidensinterval kan laves på log-skalaen og transformeres, som i BMS side 140, hvilket giver:

```
s_tilde <- sqrt(221428 / (5 * 221433) + 721772 / (26 * 721798))
phi_hat <- (5 * 721798) / (26 * 221433)
phi_hat * exp(c(-1, 1) * 1.96 * s_tilde)
```

```
[1] 0.2407156 1.6324342
```

Alternativt kan intervallet laves direkte ved brug af den estimerede standard error, hvilket giver

```
phi_hat + c(-1, 1) * 1.96 * phi_hat * s_tilde
```

```
[1] 0.02688873 1.22682962
```

Begge intervaller dækker 1, og er dermed også i overensstemmelse med konklusionen fra hypotesetestet ovenfor, men normalt foretrækkes det første interval. Det skyldes at fordelingen af  $\log(\hat{\phi}_{RR})$  approksimeres bedre af en normalfordeling end fordelingen af  $\hat{\phi}_{RR}$ .