Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen 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

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

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

   Vi får således

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

   hvor vi til sidst har benyttet at $\mathrm{E}_\theta X_i = \theta$.

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 asympotiske fordeling af $\hat{\theta}$ er

$$\hat{\theta} \overset{as}{\sim} N\big(\theta, i(\theta)^{-1}\big), \text{ dvs. } \hat{\theta} \overset{as}{\sim} N\Big(\theta, \frac{\theta^2}{2n}\Big)$$

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

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

Alle X_i'er Y_i'er uafhængige, så $\frac{1}{\theta}S_X$ og θS_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 θ at

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

således at

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

er et eksakt 95% konfidensinterval for θ.

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

   - $\Sigma_{33} \geq 0$, dvs. $\varphi \geq 0$
   - Den øvre 2×2 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 Σ en lovlig variansmatrix hvis og kun hvis $0 \leq \varphi \leq 2$.

   Fordelingen er regulær hvis og kun hvis Σ er invertibel. Determinanten

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

   er ikke-negativ for alle de lovlige værdier af φ 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
$$C\Sigma C^T = \begin{pmatrix} 5+2\varphi & 0 \\ 0 & \varphi \end{pmatrix}.$$

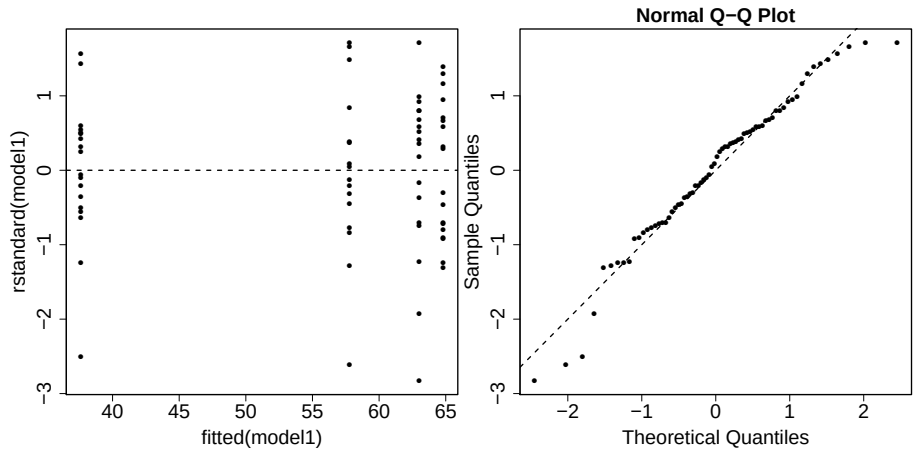
3. Fra spørgsmål 2 ses at $X_1 + X_2 \sim N(0, 5+2\varphi)$. Det følger at
$$\mathrm{E}\tilde{\varphi} = \frac{\mathrm{E}(X_1+X_2)^2 - 5}{2} = \frac{5+2\varphi-5}{2} = \varphi$$
$$\mathrm{Var}(\tilde{\varphi}) = \frac{1}{4}\mathrm{Var}(X_1+X_2)^2 = \frac{1}{4}2(5+2\varphi)^2 = \frac{1}{2}(5+2\varphi)^2$$
hvor vi har benyttet vinket om at hvis $Z \sim N(0,\sigma^2)$, så er $\mathrm{Var}(Z^2) = 2\sigma^2$.
Desuden er $X_3 \sim N(0,\varphi)$, så
$$\mathrm{E}\hat{\varphi} = \mathrm{E}X_3^2 = \varphi, \quad \mathrm{Var}(\hat{\varphi}) = \mathrm{Var}(X_3^2) = 2\varphi^2$$
Både $\tilde{\varphi}$ og $\hat{\varphi}$ er altså centrale estimatorer for φ.
Bemærk evt. at
$$\mathrm{Var}(\tilde{\varphi}) = \frac{1}{2}(5+2\varphi)^2 > \frac{1}{2}(2\varphi)^2 = 2\varphi^2 = \mathrm{Var}(\hat{\varphi}),$$
så $\hat{\varphi}$ har mindst varians. (Dette er ikke så overraskende: φ 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 $x = (x_1, \ldots, x_{70})$ 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 linien. (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 δ_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 δ_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 α_{1c}, α_{1f}, α_{2c} og α_{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 α-parameterene i rækkefølgen α_{1c}, α_{2c}, α_{1f}, α_{2f}, således at

$$\bar{\delta} = \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \alpha.$$

Vi kan derfor benytte eksempel 10.30 med $\psi^T = \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. Vi får

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

Den ønskede parametrisering opnås med kommandoen

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

og estimat $\hat{\alpha}$ og variansmatrix $\mathrm{Var}(\hat{\alpha})$ fås fx. med `coef` og `vcov`. Vi får $\hat{\bar{\delta}} = -16.213$ med estimeret spredning $\mathrm{SE}(\hat{\bar{\delta}}) = 2.870$. 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

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

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

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

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 $LR(X) \overset{as}{\sim} \chi^2_1$ 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. $x \in \mathscr{K}$.

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:

	Relativ hyppighed hvormed hypotesen forkastes	
Sand værdi af σ^2	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 stærkere/bedre end det andet.