Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1, vejledende besvarelse

9. april 2015

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

Opgave 1

1. Vi regner først på middelværdi og varians for X_i:

$$\mathrm{E}_\theta X_i = \sum_{x=1}^{\infty} \theta x \Big(\frac{1}{2}\Big)^{x+1} = \frac{\theta}{2}\sum_{x=1}^{\infty} x\Big(\frac{1}{2}\Big)^{x} = \frac{\theta}{2}\frac{\frac{1}{2}}{(1-\frac{1}{2})^2} = \theta$$

$$\mathrm{E}_\theta X_i^2 = \sum_{x=1}^{\infty} \theta x^2 \Big(\frac{1}{2}\Big)^{x+1} = \frac{\theta}{2}\sum_{x=1}^{\infty} x^2\Big(\frac{1}{2}\Big)^{x} = \frac{\theta}{2}\frac{\frac{1}{2}\cdot\frac{3}{2}}{(1-\frac{1}{2})^3} = 3\theta$$

$$\mathrm{V}_\theta X_i = E_\theta X_i^2 - \big(\mathrm{E}_\theta X_i\big)^2 = 3\theta - \theta^2$$

   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

$$\mathrm{E}_\theta \tilde{\theta} = \theta, \quad \mathrm{V}_\theta \tilde{\theta} = \frac{1}{n}(3\theta - \theta^2)$$

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

2. 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 leddet $(\frac{1}{2})^{x+1}$ afhænger ikke af θ og er derfor blot en proportionalitetsfaktor. Der er $n-n_0$ af disse x_i der derfor i alt bidrager med θ^{n-n_0}. I alt får formlen fra opgaven.

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

$$\ell_x(\theta) = -n_0 \log\Big(1-\frac{\theta}{2}\Big) - (n-n_0)\log\theta$$
$$S_x(\theta) = \frac{n_0}{2-\theta} - \frac{n-n_0}{\theta}$$
$$I_x(\theta) = \frac{n_0}{(2-\theta)^2} + \frac{n-n_0}{\theta^2}.$$

3. 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 ℓ_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_θ 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

$$\mathrm{E}_\theta \hat{\theta} = \frac{2}{n} \cdot \frac{n\theta}{2} = \theta$$
$$\mathrm{V}_\theta \hat{\theta} = \frac{4}{n^2} \cdot n\frac{\theta}{2}\left(1 - \frac{\theta}{2}\right) = \frac{1}{n}(2\theta - \theta^2)$$

   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

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

   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

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

   Værdien skal vurderes i χ^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	θ	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

$$\mathrm{E}X = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathrm{V}X = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

   Da $\det(\mathrm{V}X) = 0$, er $\mathrm{V}X$ 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

$$Y \sim N(0, C\mathrm{V}XC^T),$$

   og når man regner på variansmatricen får man

$$\mathrm{V}X = \begin{pmatrix} 18 & 0 \\ 0 & 18 \end{pmatrix}$$

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

3. Betragt så

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

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

$$\mathrm{V}Z = \begin{pmatrix} 1 & c \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ c \end{pmatrix} = 1 + 4c^2 + 4c$$

   Kravet er at Z er konstant med sandsynlighed 1, altså at $\mathrm{V}Z = 0$. Men

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

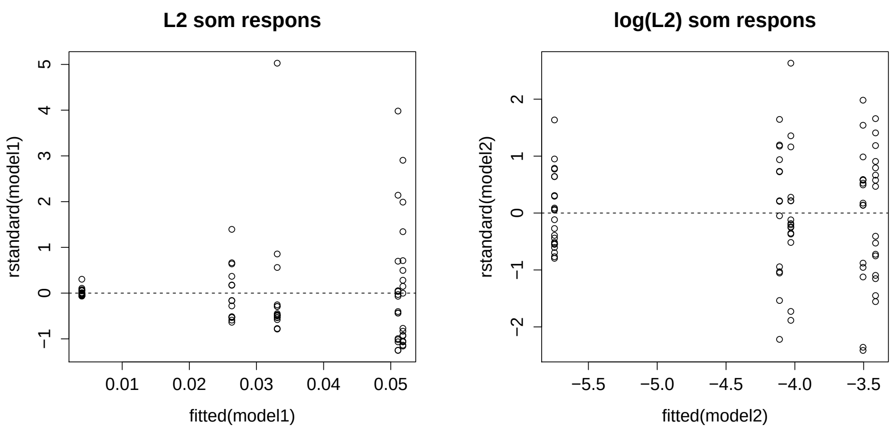
   Der gælder altså at $X_1 - X_2/2$ er 0 med sandsynlighed 1.

Opgave 3

1. Modellerne fittes 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 `L2` bruges som respons. Residualplottet ser meget bedre ud når `log(L2)` benyttes som respons.



2. 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.

3. 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(L2)` er forskellig i de fire halthedsgrupper.

4. Eftersom der ikke er forskel på de fire halthedgrupper, 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}).