ST1201 Statistiske metoder

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsforslag - Eksamen desember 2008

Oppgave 1

a)

$$L(\theta) = \prod_{i=1}^{n} f(x_i) \prod_{i=1}^{n} \left[\frac{1}{2\theta} e^{-\frac{|x_i|}{\theta}} \right]$$

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \left[-\ln 2 - \ln\theta - \frac{|x_i|}{\theta} \right] = -n\ln 2 - n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} |x_i|$$

$$l'(\theta) = \frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} |x_i|$$

$$l'(\theta) = 0 \Rightarrow \theta = \frac{1}{n}\sum_{i=1}^{n} |x_i|$$

SME blir dermed

$$\underline{\underline{\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} |X_i|}}$$

Vi har at

$$\mathrm{E}(|X_i|) = \int_{-\infty}^{\infty} |x| f(x) \mathrm{d}x = \int_{-\infty}^{\infty} \frac{|x|}{2\theta} e^{-\frac{|x|}{\theta}} \mathrm{d}x = \int_{0}^{\infty} \frac{x}{\theta} e^{-\frac{x}{\theta}} \mathrm{d}x = \left[-x e^{-\frac{x}{\theta}} \right]_0^{\infty} + \int_0^{\infty} 1 \cdot e^{-\frac{x}{\theta}} \mathrm{d}x$$
$$= \int_0^{\infty} e^{-\frac{x}{\theta}} \mathrm{d}x = \left[-\theta e^{-\frac{x}{\theta}} \right]_0^{\infty} = 0 + \theta e^0 = \theta.$$

slik at

$$\mathrm{E}\left(\widehat{\theta}\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} |X_i| \right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(|X_i|) = \frac{1}{n}\sum_{i=1}^{n} \theta = \theta$$

Dvs. $\underline{\underline{\widetilde{\theta} \text{ er forventningsrett}}}$

Oppgave 2

a)

$\widehat{\mu}_x - \widehat{\mu}_y$ er en lineærkombinasjon av X_i'ene og Y_i'ene som er uavhengige og normalfordelt. Dermed blir også $\widehat{\mu}_x - \widehat{\mu}_y$ normalfordelt.

$$\mathrm{E}\left(\widehat{\mu}_x - \widehat{\mu}_y\right) = \mathrm{E}\left(\widehat{\mu}_x\right) - \mathrm{E}\left(\widehat{\mu}_y\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) - \frac{1}{m}\mathrm{E}\left(\sum_{i=1}^{m} Y_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i\right) - \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu_x - \frac{1}{m}\sum_{i=1}^{m}\mu_y = \underline{\underline{\mu_x - \mu_y}}$$

$$\mathrm{Var}\left(\widehat{\mu}_x - \widehat{\mu}_y\right) = \mathrm{Var}\left(\widehat{\mu}_x\right) + (-1)^2\mathrm{Var}\left(\widehat{\mu}_y\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) + \frac{1}{m^2}\mathrm{Var}\left(\sum_{i=1}^{m} Y_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) + \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}\left(Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + \frac{1}{m^2}\sum_{i=1}^{m}k\sigma^2 = \frac{\sigma^2}{n} + \frac{k\sigma^2}{m} = \underline{\underline{\sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)}}$$

b)

Fra **a**) vet vi at

$$\widehat{\mu}_x - \widehat{\mu}_y \sim \mathrm{N}\left(\mu_x - \mu_y, \sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)\right).$$

Dermed får vi at

$$\frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)}} \sim \mathrm{N}(0, 1).$$

Dermed

$$\mathrm{P}\left(-z_{\frac{\alpha}{2}} \leq \frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Løser hver av de to ulikhetene med hensyn på $\mu_x - \mu_y$:

$$-z_{\frac{\alpha}{2}} \leq \frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)}} \Rightarrow \mu_x - \mu_y \leq \widehat{\mu}_x - \widehat{\mu}_y + z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)}$$

$$\frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)}} \leq z_{\frac{\alpha}{2}} \Rightarrow \mu_x - \mu_y \geq \widehat{\mu}_x - \widehat{\mu}_y - z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{k}{m}\right)}$$

Dermed får man at

$$\mathrm{P}\left(\widehat{\mu}_x-\widehat{\mu}_y-z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{k}{m}\right)}\leq \mu_x-\mu_y\leq \widehat{\mu}_x-\widehat{\mu}_y+z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{k}{m}\right)}\right)=1-\alpha.$$

Et $(1-\alpha)\cdot 100\%$ konfidensintervall for $\mu_x-\mu_y$ er dermed

$$\underline{\underline{\left[\widehat{\mu}_x-\widehat{\mu}_y-z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{k}{m}\right)},\widehat{\mu}_x-\widehat{\mu}_y+z_{\frac{\alpha}{2}}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{k}{m}\right)}\right]}}.$$

c)

$$\begin{aligned}\mathrm{E}\left(S_p^2\right)&=\mathrm{E}\left(\frac{n-1}{n+m-2}S_x^2+\frac{m-1}{k(n+m-2)}S_y^2\right)\\ &=\frac{\sigma^2}{n+m-2}\mathrm{E}\left(\frac{(n-1)S_x^2}{\sigma^2}\right)+\frac{\sigma^2}{n+m-2}\mathrm{E}\left(\frac{(m-1)S_y^2}{k\sigma^2}\right)\\ &=\frac{\sigma^2}{n+m-2}\cdot(n-1)+\frac{\sigma^2}{n+m-2}\cdot(m-1)=\frac{\sigma^2}{n+m-2}(n-1+m-1)=\underline{\underline{\sigma^2}}\end{aligned}$$

Siden X_i'ene er uavhengig av Y_i'ene blir S_x^2 uavhengig av S_y^2. Dessuten har vi at

$$\frac{(n+m-2)S_p^2}{\sigma^2}=\frac{(n-1)S_x^2}{\sigma^2}+\frac{(m-1)S_y^2}{k\sigma^2}.$$

Siden $\frac{(n+m-2)S_p^2}{\sigma^2}$ er en sum av to uavhengige χ^2-fordelte variable blir også $\frac{(n+m-2)S_p^2}{\sigma^2}$ kji-kvadrat fordelt. Antall frihetsgrader blir summen av antall frihetsgrader for $\frac{(n-1)S_x^2}{\sigma^2}$ og for $\frac{(n-1)S_y^2}{k\sigma^2}$, dvs. $(n-1)+(m-1)=n+m-2$. Vi har dermed at

$$\underline{\underline{\frac{(n+m-2)S_p^2}{\sigma^2}\sim\chi^2_{n+m-2}}}.$$

Vi har at T kan skrives på formen

$$T=\frac{\frac{(\widehat{\mu}_x-\widehat{\mu}_y)-(\mu_x-\mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{k}{m}\right)}}}{\sqrt{\frac{\frac{(n+m-2)S_p^2}{\sigma^2}}{n+m-2}}}=\frac{Z}{\sqrt{\frac{V}{n+m-2}}}$$

der

$$Z=\frac{(\widehat{\mu}_x-\widehat{\mu}_y)-(\mu_x-\mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{k}{m}\right)}}\sim \mathrm{N}(0,1)\quad\text{og}\quad V=\frac{(n+m-2)S_p^2}{\sigma^2}\sim\chi^2_{n+m-2}.$$

Siden Z er kun en funksjon av $\bar{X}$ og $\bar{Y}$, mens S_p^2 kun er en funksjon av S_x^2 og S_y^2 og $(\bar{X}, \bar{Y})$ er uavhengig av (S_x^2, S_y^2) blir Z og V uavhengige. Per definsjon av t-fordeling blir dermed T Student t-fordelt og antall frihetsgrader blir lik antall frihetsgrader for V, dvs $n+m-2$.

d)
Siden T er Student t-fordelt med $n+m-2$ frihetsgrader har vi at

$$\mathrm{P}\left(-t_{\frac{\alpha}{2},n+m-2} \leq \frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}} \leq t_{\frac{\alpha}{2},n+m-2}\right) = 1 - \alpha.$$

Løser hver av de to ulikhetene med hensyn på $\mu_x - \mu_y$:

$$-t_{\frac{\alpha}{2},n+m-2} \leq \frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}} \Rightarrow \mu_x - \mu_y \leq \widehat{\mu}_x - \widehat{\mu}_y + t_{\frac{\alpha}{2},n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}$$

$$\frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}} \leq t_{\frac{\alpha}{2},n+m-2} \Rightarrow \mu_x - \mu_y \geq \widehat{\mu}_x - \widehat{\mu}_y - t_{\frac{\alpha}{2},n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}$$

Dermed får man at

$$\mathrm{P}\left(\widehat{\mu}_x - \widehat{\mu}_y - t_{\frac{\alpha}{2},n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)} \leq \mu_x - \mu_y \leq \widehat{\mu}_x - \widehat{\mu}_y + t_{\frac{\alpha}{2},n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}\right) = 1-\alpha.$$

Et $(1-\alpha)\cdot 100\%$ konfidensintervall for $\mu_x - \mu_y$ er dermed

$$\underline{\underline{\left[\widehat{\mu}_x - \widehat{\mu}_y - t_{\frac{\alpha}{2},n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}, \widehat{\mu}_x - \widehat{\mu}_y + t_{\frac{\alpha}{2},n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}\right]}}.$$

Lengden av konfidensintervaller blir

$$L = 2t_{\frac{\alpha}{2},n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)} = 2t_{\frac{\alpha}{2},n+m-2}\sqrt{\frac{(n+m-2)S_p^2}{\sigma^2}} \cdot \sqrt{\frac{\sigma^2}{n+m-2}\left(\frac{1}{n} + \frac{k}{m}\right)}$$

Vi får dermed at

$$\mathrm{E}(L) = 2t_{\frac{\alpha}{2},n+m-2}\mathrm{E}\left(\sqrt{\frac{(n+m-2)S_p^2}{\sigma^2}}\right)\sqrt{\frac{\sigma^2}{n+m-2}\left(\frac{1}{n} + \frac{k}{m}\right)}.$$

Siden $(n+m-2)S_p^2/\sigma^2 \sim \chi^2_{n+m-2}$ må $E((n+m-2)S_p^2/\sigma^2)$ kun være en funksjon av summen $n+m$ og ikke n og m hver for seg. For å minimere lengden av konfidensintervallet gitt at $n+m=N$ er fiksert må vi dermed minimere

$$g(n) = \frac{1}{n} + \frac{k}{N-n}$$

med hensyn på n. Deriverer og setter lik null:

$$g'(n) = -\frac{1}{n^2} + \frac{k}{(N-n)^2} = 0 \Rightarrow \frac{1}{n^2} = \frac{k}{(N-n)^2} \Rightarrow \frac{1}{n} = \frac{\sqrt{k}}{N-n} \Rightarrow n = \frac{N}{1+\sqrt{k}}$$

Konfidensintervallet blir dermed kortest ved å velge

$$\underline{\underline{n = \frac{N}{1+\sqrt{k}} \quad \text{og} \quad m = N-n = \frac{N\sqrt{k}}{1+\sqrt{k}}}}.$$

Hvis N og k er slik at disse svarene ikke blir heltall må man selvfølgelig forsøke seg frem med avrunding nedover og oppover for å bestemme hvilke av disse to som gir det korteste intervallet.

Oppgave 3

a)
Sannsynlighetstettheten for en χ^2_2-fordeling er

$$f(v) = \frac{1}{2^{\frac{2}{2}}\Gamma\left(\frac{2}{2}\right)} v^{\frac{2}{2}-1} e^{-\frac{v}{2}} = \frac{1}{2} e^{-\frac{v}{2}}.$$

For Y i oppgaveteksten får vi at kumulativ fordeling blir

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(2\theta(\ln X - \ln c) \le y) = \mathrm{P}\left(\ln X \le \frac{y}{2\theta} + \ln c\right)$$
$$= \mathrm{P}\left(X \le e^{\frac{y}{2\theta}+\ln c}\right) = F_X\left(e^{\frac{y}{2\theta}+\ln c}\right).$$

Finner sannsynlighetsfordelingen ved å derivere

$$f_Y(y) = F_Y'(y) = f_X\left(e^{\frac{y}{2\theta}+\ln c}\right) \cdot e^{\frac{y}{2\theta}+\ln c} \cdot \frac{1}{2\theta} = \frac{\theta c^\theta}{\left(e^{\frac{y}{2\theta}+\ln c}\right)^{\theta+1}} e^{\frac{y}{2\theta}+\ln c} \frac{1}{2\theta} = \frac{1}{2}\frac{1}{e^{\frac{y}{2}}} = \frac{1}{2} e^{-\frac{y}{2}}.$$

Ser dermed at $\underline{\underline{Y \sim \chi^2_2}}$.

b)
Ved å sette inn uttrykket for $\widehat{\theta}$ får vi at

$$\frac{2n\theta}{\widehat{\theta}} = \frac{2n\theta}{n}\left(\sum_{i=1}^n \ln X_i - n \ln c\right) = \sum_{i=1}^n (2\theta \ln X_i - \ln c) = \sum_{i=1}^n Y_i,$$

der $Y_i = 2\theta(\ln X_i - \ln c)$. Fra **a**) vet vi at $Y_i \sim \chi^2_2$. Siden X_i'ene er uavhengige blir også Y_i'ene uavhengige. Dermed er $2n\theta/\widehat{\theta}$ en sum av uavhengige χ^2-fordelte variable og er derfor selv χ^2-fordelt. Antall frihetsgrader blir $\sum_{i=1}^n 2 = 2n$. Har dermed vist at

$$\underline{\underline{\frac{2n\theta}{\widehat{\theta}} \sim \chi^2_{2n}}}.$$

Skal bestemme k fra kravet

$$\mathrm{P}\left(\widehat{\theta} \geq k \,\middle|\, H_0\right) = \alpha$$
$$\mathrm{P}\left(\frac{2n\theta}{\widehat{\theta}} \leq \frac{2n\theta}{k} \,\middle|\, \theta = \theta_0\right) = \alpha$$
$$\mathrm{P}\left(\frac{2n\theta}{\widehat{\theta}} \leq \frac{2n\theta_0}{k} \,\middle|\, \theta = \theta_0\right) = \alpha$$

Får dermed at man må ha

$$\frac{2n\theta_0}{k} = \chi^2_{1-\alpha,2n} \Rightarrow \underline{\underline{k = \frac{2n\theta_0}{\chi^2_{1-\alpha,2n}}}}$$

Innsatt $n = 20$, $\theta_0 = 2$ og $\chi^2_{0.99,40} = 22.164$ får vi at

$$k = \frac{2 \cdot 20 \cdot 2}{22.164} = \underline{\underline{3.61}}.$$

c)

Skal bestemme θ slik at

$$\mathrm{P}\left(\text{Ikke forkast } H_0 | \theta\right) = 0.05.$$

Kravet blir

$$\mathrm{P}\left(\widehat{\theta} < k \,\middle|\, \theta\right) = 0.05$$
$$\mathrm{P}\left(\frac{2n\theta}{\widehat{\theta}} > \frac{2n\theta}{k} \,\middle|\, \theta\right) = 0.05$$

Dette gir at vi må ha

$$\frac{2n\theta}{k} = \chi^2_{0.05,2n} \Rightarrow \theta = \frac{k}{2n}\chi^2_{0.05,2n} = \frac{\theta_0}{\chi^2_{1-\alpha,2n}} \cdot \chi^2_{0.05,2n}$$

Insatt $n = 20$, $\theta_0 = 2$, $\chi^2_{0.99,40} = 22.164$ og $\chi^2_{0.05,40} = 55.758$ får vi

$$\theta = \frac{2}{22.164} \cdot 55.758 = \underline{\underline{5.03}}.$$