Løsningsforslag til eksamen i

STK1110 5. desember 2007

Oppgave 1

a) Vi antar at slaktevektene til kalkunene fra Virginia er observerte verdier av stokastiske variabler X_1, X_2, X_3, X_4 som er uavhengige og $N(\mu_1, \sigma^2)$-fordelte, og at slaktevektene til kalkunene fra Wisconsin er observerte verdier av stokastiske variabler Y_1, Y_2, Y_3, Y_4, Y_5 som er uavhengige og $N(\mu_2, \sigma^2)$-fordelte. Vi antar også at X_i-ene og Y_j-ene er uavhengige.

Vi lar $\bar{X}$ være gjennomsnittet av X_i-ene og $\bar{Y}$ være gjennomsnittet av Y_j-ene. Videre lar vi S_1^2 være den empiriske variansen for X_i-ene og S_2^2 den empiriske variansen for Y_j-ene. Endelig innfører vi

$$S_p^2 = \frac{4-1}{4+5-2}\,S_1^2 + \frac{5-1}{4+5-2}\,S_2^2 = \frac{3}{7}\,S_1^2 + \frac{4}{7}\,S_2^2.$$

Et 90% konfidensintervall for $\mu_2 - \mu_1$ er gitt ved (se avsnitt 11.2.1 i Rice):

$$(\bar{Y} - \bar{X}) \pm t_{0.95} \cdot S_p \sqrt{\frac{1}{4} + \frac{1}{5}},$$

der $t_{0.95} = 1.895$ er 95% fraktilen i t-fordelingen med $4+5-2 = 7$ frihetsgrader.

For tallene i oppgaven har vi $\bar{X} = 12.625$, $\bar{Y} = 13.600$, $S_1 = 0.655$, $S_2 = 1.440$ og

$$S_p = \sqrt{\frac{3}{7} \cdot 0.655^2 + \frac{4}{7} \cdot 1.440^2} = 1.170.$$

Et 90% konfidensintervall blir:

$$(13.600 - 12.625) \pm 1.895 \cdot 1.170 \cdot \sqrt{\frac{1}{4} + \frac{1}{5}},$$

dvs. 0.975 ± 1.487. Det betyr at vi med rimelig grad av sikkerhet kan si at $\mu_2 - \mu_1$ ligger mellom -0.51 kg og 2.46 kg. (Merk at det er galt å si det er 90% sannsynlig at $\mu_2 - \mu_1$ ligger mellom -0.51 og 2.46. For $\mu_2 - \mu_1$ er et (ukjent) reelt tall, som enten ligger mellom -0.51 og 2.46 eller ikke gjør det. Vi kan derimot si at vi har 90% tiltro til at $\mu_2 - \mu_1$ ligger mellom -0.51 og 2.46. Med det mener vi at konfidenintervall som er beregnet slik vi har gjort det her, vil inneholde den sanne differansen i 90% av tilfellene i "det lange løp".)

b) En test med signifikansnivå 5% for nullhypotesen $H_0 : \mu_1 = \mu_2$ mot den alternative hypotesen $H_1 : \mu_1 \neq \mu_2$, forkaster H_0 hvis (se avsnitt 11.2.1 i Rice)

$$T = \frac{\bar{Y} - \bar{X}}{S_p \sqrt{\frac{1}{4} + \frac{1}{5}}} > t_{0.975} \quad \text{eller hvis} \quad T < -t_{0.975},$$

der $t_{0.975} = 2.365$ er 97.5% fraktilen i t-fordelingen med $4+5-2 = 7$ frihetsgrader.

Med tallene i oppgaven finner vi

$$T = \frac{13.600 - 12.625}{1.170\sqrt{\frac{1}{4} + \frac{1}{5}}} = 1.242.$$

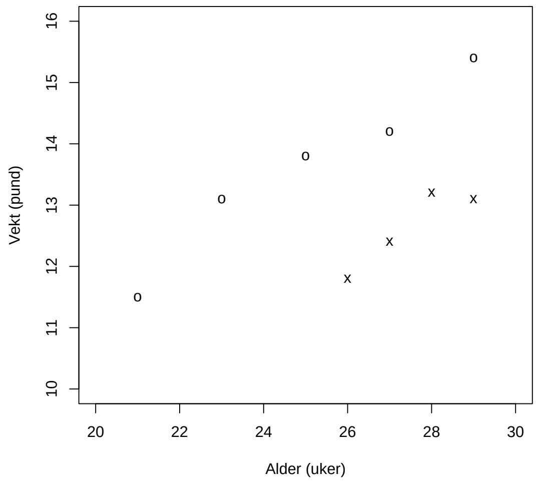
Vi ser at $T = 1.242 < 2.365$, så vi forkaster ikke H_0 på 5% nivå.

c) P-verdien til en test er det minste signifikansnivået som gir forkastning av nullhypotesen. Litt mer uformelt, kan vi si at P-verdien er sannsynligheten for å få et resultat som det vi har fått, eller et resultat som er enda mer i favør av den alternative hypotesen, bare på grunn av tilfeldigheter (dvs. når H_0 er sann).

For t-fordelingen med 7 frihetsgrader finner vi av tabellen at $t_{0.80} = 0.896$ og $t_{0.90} = 1.415$. Det betyr at en test med nivå 40% vil forkaste H_0, mens en test med nivå 20% ikke vil gjøre det. P-verdien er derfor mellom 20% og 40%. (Bedre klarer vi ikke å bestemme P-verdien av tabellen.)

Oppgave 2

a) Diagrammet gir et plott av vekten til kalkunene mot alderen ved slakting (Virginia=`x`, Wisconsin=`o`).



Vi ser at punktene ligger tilnærmet på to parallelle rette linjer, en linje for Virginia og en linje for Wisconsin.

b) Parameteren β_1 angir forventet vektøkning for kalkunene i løpet av én uke. Parameteren β_2 angir forventet forskjell i vekt mellom like gamle kalkuner fra Wisconsin og Virginia.

Modellen (1) sier at forventet vekt som funksjon av alder følger to parallelle rette linjer. Stigningstallet for linjene er β_1 og avstanden mellom dem er β_2. Det er rimelig ut fra plottet.

c) En test med signifikansnivå 5% for nullhypotesen $H_0 : \beta_2 = 0$ mot den alternative hypotesen $H_1 : \beta_2 \neq 0$, forkaster H_0 hvis (se avsnitt 14.4.5 i Rice)

$$T = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} > t_{0.975} \quad \text{eller hvis} \quad T < -t_{0.975},$$

der $t_{0.975} = 2.447$ er 97.5% fraktilen i t-fordelingen med $9 - 3 = 6$ frihetsgrader.

Med tallene i oppgaven finner vi $T = 2.094/0.235 = 8.91$. Vi ser at $T = 8.91 > 2.447$, så vi forkaster H_0 på 5% nivå. (Vi har $t_{0.995} = 3.707$, så P-verdien er mindre enn 1%. Mer nøyaktig klarer vi ikke å bestemme P-verdien ut fra tabellen.)

Når vi sammenligner like gamle kalkuner, slik vi gjør her, veier kalkunene fra Wisconsin signifikant mer enn kalkunene fra Virginia. I punkt b i oppgave 1, tar vi ikke hensyn til alderen til kalkunene, og da er det ikke signifikant forskjell i vekten for kalkuner fra de to statene.

d) Vi innfører matrisen $\mathbf{C} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ og skriver

$$\mathbf{C} = \begin{pmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{pmatrix}$$

Da er kovariansmatrisen til $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^T$ gitt ved $\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}} = \sigma^2\mathbf{C}$ (se avsnitt 14.4.2 i Rice). Spesielt gir det at $\mathrm{Var}\hat{\beta}_1 = \sigma^2\, c_{11}$.

Videre er $S^2 = RSS/(9-3)$ en forventningsrett estimator for σ^2 (se avsnitt 14.4.3 i Rice). Den estimerte standardfeilen til $\hat{\beta}_1$ er dermed $S_{\hat{\beta}_1} = \sqrt{S^2\, c_{11}}$.

Med tallene i oppgaven blir $S^2 = 0.5648/6 = 0.0941$ og $c_{11} = 0.0222$. Den estimerte standardfeilen til $\hat{\beta}_1$ er dermed $S_{\hat{\beta}_1} = \sqrt{0.0941 \cdot 0.0222} = 0.0457$.

Et 95% konfidensintervall for β_1 er gitt ved (se avsnitt 14.4.5 i Rice):

$$\hat{\beta}_1 \pm t_{0.975} \cdot S_{\hat{\beta}_1},$$

der $t_{0.975} = 2.447$ er 97.5% fraktilen i t-fordelingen med $9 - 3 = 6$ frihetsgrader. Med tall får vi $0.448 \pm 2.447 \cdot 0.0457$, dvs. 0.448 ± 0.112. Det betyr at vi med rimelig grad av sikkerhet kan si at forventet vektøking for kalkunene er mellom 0.34 kg og 0.56 kg per uke. (Se også punkt a i oppgave 1 når det gjelder fortolking av konfidensintervall.)

Oppgave 3

a) La Y ha sannsynlighetstettheten (2). For $y > 0$ er den kumulative fordelingen til Y gitt ved:

$$F(y \,|\, \theta) = \int_{-\infty}^{y} f(u \,|\, \theta)\, du = \int_{0}^{y} \frac{u}{\theta}\, e^{-u^2/2\theta}\, du$$

$$= \left[-e^{-u^2/2\theta}\right]_0^y = 1 - e^{-y^2/2\theta}$$

Medianen $y_{0.50}$ er gitt ved at $F(y_{0.50} \,|\, \theta) = 0.50$. Det gir:

$$\begin{aligned} 1 - \exp\left(-\frac{y_{0.50}^2}{2\theta}\right) &= \frac{1}{2} \\ -\frac{y_{0.50}^2}{2\theta} &= -\log 2 \\ y_{0.50} &= \sqrt{2\theta \log 2} \end{aligned}$$

b) $Y_1, Y_2, \ldots, Y_{20}$ er uavhengige og identisk fordelte med sannsynlighetstetthet gitt ved (2). Da er likelihooden:

$$\mathrm{lik}(\theta) = \prod_{i=1}^{20} f(Y_i \,|\, \theta) = \prod_{i=1}^{20} \frac{Y_i}{\theta}\, e^{-Y_i^2/2\theta}$$

Log-likelihooden blir dermed:

$$l(\theta) = \sum_{i=1}^{20} \left\{ \log Y_i - \log\theta - \frac{Y_i^2}{2\theta} \right\}$$
$$= \sum_{i=1}^{20} \log Y_i - 20\log\theta - \frac{1}{2\theta}\sum_{i=1}^{20} Y_i^2$$

Den deriverte (m.h.p. θ) blir

$$l'(\theta) = -\frac{20}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{20} Y_i^2$$

Maksimum likelihood estimatoren $\hat{\theta}$ er løsningen av ligningen $l'(\hat{\theta}) = 0$. Det gir

$$\frac{20}{\hat{\theta}} = \frac{1}{2\hat{\theta}^2}\sum_{i=1}^{20} Y_i^2$$
$$\hat{\theta} = \frac{1}{40}\sum_{i=1}^{20} Y_i^2$$

c) Bedriften ønsker å teste

$$H_0 : y_{0.50} \leq 1000 \quad \text{mot} \quad H_1 : y_{0.50} > 1000$$

Nå har vi at

$$y_{0.50} \leq 1000$$
$$\sqrt{2\theta \log 2} \leq 10^3$$
$$2\theta\log 2 \leq 10^6$$
$$\theta \leq 10^6/(2\log 2)$$

Bedriften ønske derfor å teste

$$H_0 : \theta \leq \theta_0 \quad \text{mot} \quad H_1 : \theta > \theta_0$$

der $\theta_0 = 10^6/(2\log 2)$.

d) Siden den alternative hypotesen er $H_1 : \theta > \theta_0$, er det rimelig å forkaste H_0 hvis $\hat{\theta} > k$. Vi vil bestemme k slik at tesen får nivå 5%. Vi har at

$$P(\text{forkast } H_0 \,|\, H_0 \text{ sann}) = P(\hat{\theta} > k \,|\, \theta \leq \theta_0)$$
$$\leq P(\hat{\theta} > k \,|\, \theta = \theta_0)$$
$$= P\left(\frac{1}{40}\sum_{i=1}^{20} Y_i^2 > k \,|\, \theta = \theta_0\right)$$
$$= P\left(\frac{1}{\theta_0}\sum_{i=1}^{20} Y_i^2 > \frac{40k}{\theta_0} \,|\, \theta = \theta_0\right)$$
$$= P\left(\chi^2 > \frac{40k}{\theta_0}\right)$$

der χ^2 er kji-kvadrat-fordelt med 40 frihetsgrader.

Nå har vi av tabellen at

$$P(\chi^2 > 55.76) = 1 - P(\chi^2 \leq 55.76) = 1 - 0.95 = 0.05$$

Vi får derfor en test med nivå 5% hvis bestmmer k ved ligningen $40k/\theta_0 = 55.76$, dvs.

$$k = \frac{55.76\,\theta_0}{40} = \frac{55.76 \cdot 10^6}{40 \cdot 2\log 2} = 1.006 \cdot 10^6$$

e) Hvis $y_{0.50} = 1250$, har vi at $\sqrt{2\theta \log 2} = 1250$, dvs.

$$\theta = \frac{1250^2}{2\log 2} = 1.127 \cdot 10^6$$

For denne verdien av θ blir teststyrken:

$$\begin{aligned}
\gamma(\theta) &= P(\text{forkast } H_0 \,|\, \theta) \\
&= P(\hat{\theta} > k \,|\theta) \\
&= P\left(\frac{1}{40}\sum_{i=1}^{20} Y_i^2 > k \,|\, \theta\right) \\
&= P\left(\frac{1}{\theta}\sum_{i=1}^{20} Y_i^2 > \frac{40k}{\theta} \,|\, \theta\right) \\
&= P\left(\chi^2 > \frac{40k}{\theta}\right) \\
&= P\left(\chi^2 > \frac{40 \cdot 1.006 \cdot 10^6}{1.127 \cdot 10^6}\right) \\
&= P\left(\chi^2 > 35.70\right)
\end{aligned}$$

Av tabellen finner vi:

$$P(\chi^2 > 34.87) = 1 - P(\chi^2 \leq 34.87) = 1 - 0.30 = 0.70$$
$$P(\chi^2 > 37.13) = 1 - P(\chi^2 \leq 37.13) = 1 - 0.40 = 0.60$$

Det betyr at teststyrken er mellom 60% og 70%.

Vi gjør en feil av type II hvis vi ikke forkaster H_0 når den er gal. Sannsynligheten for feil av type II er derfor lik én minus teststyrken. Her blir sannsynligheten for feil av type II mellom 30% og 40%.