UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Fasit til eksamen Tirsdag 3. juni 2008
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabeller over normal-, t-, F- og χ^2-fordelingene.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

(a) $T = \frac{\bar{X}}{s}\sqrt{n} \sim t_{n-1}$ fordelt under H_0, p-verdi $P(t_{n-1} \geq |T|)$.

Hvis X_i-ene ikke er normalfordelt holder testen holder likevel tilnærmet når n er stor, pga. sentralgrenseteoremet og konsistens av empirisk standardavvik s.

(b) Symmetri om $x = \mu$ når $f(\mu + \Delta|\mu, \tau) = f(\mu - \Delta|\mu, \tau)$.

Men $f(\mu + \Delta|\mu, \tau) = \frac{1}{\tau}\frac{\exp(\Delta/\tau)}{(1+\exp(\Delta/\tau))^2}$ og

$$f(\mu-\Delta|\mu,\tau) = \frac{1}{\tau}\frac{\exp(-\Delta/\tau)}{(1+\exp(-\Delta/\tau))^2}\frac{\exp(2\Delta/\tau))}{\exp(2\Delta/\tau)} = \cdots = f(\mu+\Delta|\mu,\tau)$$

(c) Likelihood

$$L(\mu,\tau) = \prod_{i=1}^{n} f(X_i|\mu,\tau) = \tau^{-n}\frac{\exp(\sum_{i=1}^{n}(X_i-\mu)/\tau)}{\prod_{i=1}^{n}[1+\exp((X_i-\mu)/\tau)]^2}$$

(Fortsettes side 2.)

som gir log-likelihood

$$l(\mu,\tau) = \log(L(\mu,\tau)) = -n\log(\tau) + \sum_{i=1}^{n} \frac{X_i - \mu}{\tau} - 2\sum_{i=1}^{n} \log[1+\exp((X_i-\mu)/\tau)]$$

Scorefunksjon

$$s(\mu,\tau) = \begin{pmatrix} \frac{\partial l(\mu,\tau)}{\partial \mu} \\ \frac{\partial l(\mu,\tau)}{\partial \tau} \end{pmatrix} = \begin{pmatrix} s_1(\mu,\tau) \\ s_2(\mu,\tau) \end{pmatrix}$$

der $s_1(\mu,\tau) = -\frac{n}{\tau} + \frac{2}{\tau}\sum_{i=1}^{n} \frac{\exp((X_i-\mu)/\tau)}{1+\exp((X_i-\mu)/\tau)}$

og $s_2(\mu,\tau) = -\frac{n}{\tau} - \frac{1}{\tau^2}\sum_{i=1}^{n}(X_i-\mu) + \frac{2}{\tau^2}\sum_{i=1}^{n} \frac{\exp((X_i-\mu)/\tau)}{1+\exp((X_i-\mu)/\tau)}(X_i-\mu)$.

Observert informasjonsmatrise er minus matrisen av annenderiverte av log-likelihood, her

$$\bar{J}(\mu,\tau) = -\begin{bmatrix} \frac{\partial^2 l(\mu,\tau)}{\partial \mu^2} & \frac{\partial^2 l(\mu,\tau)}{\partial \mu \partial \tau} \\ \frac{\partial^2 l(\mu,\tau)}{\partial \mu \partial \tau} & \frac{\partial^2 l(\mu,\tau)}{\partial \tau^2} \end{bmatrix}$$

(d) Newton-Raphson-algoritme

$$\begin{bmatrix} \mu_{(k+1)} \\ \tau_{(k+1)} \end{bmatrix} = \begin{bmatrix} \mu_{(k)} \\ \tau_{(k)} \end{bmatrix} + \bar{J}(\mu_{(k)},\tau_{(k)})^{-1} s(\mu_{(k)},\tau_{(k)})$$

Fisher-scoring består i å bytte ut observert informasjon med forventet informasjon $\bar{I}(\mu_{(k)},\tau_{(k)}) = \mathrm{E}[\bar{J}(\mu_{(k)},\tau_{(k)})]$. Forventet informasjon er vanskelig å regne ut analytisk og ikke så aktuell her.

Naturlig å starte fra $\mu_{(0)} = \bar{X}$ og $\tau_{(0)} = s\sqrt{3}/\pi$ der s^2 er empirisk varians (siden $\mathrm{Var}[X_i] = \frac{\pi^2}{3}\tau^2$).

(e) Tilnærmet fordeling for $\hat{\mu} \sim \mathrm{N}(\mu, 0.131)$ (ved MLE's asymptotiske egenskaper).

Test-basert på MLE $\hat{\mu}$: $Z = \frac{\hat{\mu}}{\sqrt{\hat{\sigma}_{11}}} = \frac{1.0413}{\sqrt{0.1310}} = 2.88$. Dette gir en p-verdi 0.004 sml. med $\mathrm{N}(0,1)$ og p-verdi = 0.01 sml. med t_{19}.

t-test: $T = \frac{\bar{X}}{s}\sqrt{n} = \frac{1.0495}{\sqrt{3.057}}\sqrt{20} = 2.68$ gir p-verdi $2\mathrm{P}(t_{19} \geq 2.68) < 2\mathrm{P}(t_{19} \geq 2.539) = 0.02$, dvs. forkaster på 2 prosent nivå.

Omtrent samme konklusjon, ikke særlig mye gevinst ved å bruke MLE i dette datamaterialet. Testen basert på MLE kan dessuten være følsom for antagelsen om logistisk fordeling.

(f) GLRT: $2(l(\hat{\theta}) - l(\theta^\star)) \sim \chi^2_{\mathrm{df}}$ under modell (nullhypotese) $\theta \in \omega$ der i utgangspunktet $\theta \in \Omega$. Her er $\hat{\theta}$ MLE over Ω og $\theta^\star$ MLE over ω og antall frihetsgrader $\mathrm{df} = \dim(\Omega) - \dim(\omega)$.

(Fortsettes side 3.)

I det aktuelle problemet blir GLRT $= 2(l(\hat{\mu}, \hat{\tau}) - l(0, \tau^\star)) \sim \chi^2_1$ under nullhypotesen. Har dessuten

$$l(\mu, \tau) = -n \log(\tau) + \frac{n(\bar{X} - \mu)}{\tau} - 2 \sum_{i=1}^{n} \log[1 + \exp((X_i - \mu)/\tau)]$$

slik at $l(\hat{\mu}, \hat{\tau}) = -20 \log(0.9453) + 20(1.0495 - 1.0413)/0.9453 - 2 * 20.153 = -39.01$ og $l(0, \tau^\star) = -20 \log(1.1425) + 20 * 1.0495/1.1425 - 2 * 29.074 = -42.440$ hvilket gir GLRT-observatoren lik 2 * (42.44-39.01)= 6.86 og p-verdi $P(\chi^2_1 > 6.94) < P(\chi^2_1 > 6.63) = 0.01$, dvs. samme konklusjon som i punkt (f).

Oppgave 2.

(a) R-utskriften tilsier en klart signifikant sammenheng mellom høyde og lungekapasitet (FEV1) slik at lungekapasiteten øker med ca. $\hat{\beta}_1 = 0.0585$ liter per cm.

Korrelasjonen mellom høyde og FEV1 blir + kvadratroten til R-squared. dvs. $+\sqrt{0.4618} = 0.68$ (+ siden $\hat{\beta}_1 > 0$.)

95% KI for β_1 blir tilnærmet $\hat{\beta}_1 \pm 1.96\text{se}(\hat{\beta}_1) = 0.058514 \pm 1.96 * 0.005124 = (0.04847, 0.06855)$. Dette intervallet inneholder ikke verdien 0.04, dermed forkastes nullhypotesen om at $\beta_1 = 0.04$ på 5% nivå.

(b) t-testen for ingen forskjell i FEV1 mellom menn og kvinner gir klar forkastning, p-verdi mindre enn 0.001.

Testen etter justering for høyde gir en p-verdi på 0.33, altså ingen forkastning.

Den observerte forskjellen i FEV1 mellom kvinner og menn kan forklares med at menn er høyere enn kvinner. Det er korrelasjon mellom kjønn og høyde, og høyde har en klar sammenheng med FEV1. Når vi utelater høyde fra analysen vil vi da få en tilsynelatende, "spuriøs", sammenheng mellom lungefunksjon og kjønn.

(c) Estimert sammenheng mellom FEV1 og bmi (justert for høyde) beskrives ved annengradskurven konstant+0.203648 bmi-0.003373 bmi^2. Denne kurven maksimeres for $0.203648/(2 * 0.003373) = 30.18$. En bmi såpass høy som 30 synes altså å være optimal for lungefunksjon.

Det er så mye som 25% av denne befolkningen som har bmi > 30. Hvis vi derfor bare tar med førstegradsleddet vil vi få en forholdsvis liten regresjonskoeffisient.

(Fortsettes side 4.)

(d) Vi har $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\beta$ Dermed blir (i)

$$\mathrm{E}[\hat{\beta}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}[\mathbf{Y}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\beta = \beta$$

der første likhet følger siden $\hat{\beta}$ er en lineærkombinasjon av Y_i-ene slik at forventningsoperatoren kan flyttes innenfor matriseproduktene og der annen likehet skyldes at $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}$ blir en identitetsmatrise.

For (ii) $\Sigma_{\hat{\beta}}$ = kovariansmatrisen til $\hat{\beta}$: Får først kovariansmatrisen til $\mathbf{X}^\top\mathbf{Y}$ lik

$$\Sigma_{\mathbf{X}^\top\mathbf{Y}} = \mathbf{X}^\top\sigma^2 I_{nxn}\mathbf{X} = \sigma^2\mathbf{X}^\top\mathbf{X}$$

der I_{nxn} er en n-dimensjonal identitetsmatrise. Dette gir

$$\Sigma_{\hat{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2$$

SLUTT