UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Løsningsforslag
Eksamensdag:	Torsdag 14. desember 2017
Tid for eksamen:	09.00 – 13.00
Oppgavesettet er på 4 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Formelsamling for STK1100/STK1110. Godkjent kalkulator.

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

Oppgave 1

a

Under H_0 er $t = (\bar{X} - \mu_0)\sqrt{n}/S \sim t_{24}$-fordelt siden antall frihetsgrader er $n-1 = 24$. Med 5% nivå forkaster man da hvis observert $|t| > t_{0.025} = 2.064$ 97.5% persentilen i t_{24} fordelingen.

Her finner vi $t = (2.72 - 2) * 5/\sqrt{3.14} = 2.032$ og altså mindre enn 2.064. Vi forkaster altså ikke nullhypotesen på 5% nivå.

Men $t = 2.032 > 1.711$ som 95-persentilen i t_{24} fordelingen. Dermed ligger P-verdien i intervallet $(0.05, 0.10)$.

b

At $\bar{X} \pm t_{\alpha/2}S/\sqrt{n}$ er et $(1-\alpha)100\%$ konfidensintervall betyr at

$$\mathrm{P}(\bar{X} - t_{\alpha/2}S/\sqrt{n} < \mu_0 < \bar{X} + t_{\alpha/2}S/\sqrt{n}) = 1 - \alpha$$

Men denne begivenheten er ekvivalent med

$$-t_{\alpha/2} < t = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} < t_{\alpha/2}$$

og $t = \frac{(\bar{X}-\mu_0)\sqrt{n}}{S}$ er den t_{n-1}-fordelte testobservatoren. Nullhypotesen $H_0 : \mu = \mu_0$ forkastes altså nettopp *ikke* for slike verdier μ_0

Her blir konfidensintervallet lik $2.27 \pm 2.064 * \sqrt{3.14}/5 = (1.99, 3.45)$ og $\mu_0 = 2$ ligger (akkurat) innenfor dette intervallet, hvilket altså også innebærer at nullhypotesen $H_0 : \mu = \mu_0 = 2$ ikke forkastes.

(Fortsettes på side 2.)

c

Type II feilen er å unnlate å forkaste når den alternative hypotesen er sann, så for en θ i samsvar med H_a er $\gamma(\theta) = 1 - P(\text{Type II feil} \mid \theta)$.

Man forkaster altså hvis $\bar{X} > \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}$ eller $\bar{X} < \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$. Men dette er det samme som at $Z = (\bar{X} - \mu)\sqrt{n}/\sigma > (\mu_0 - \mu)\sqrt{n}/\sigma + z_{\alpha/2}$ eller $Z = (\bar{X} - \mu)\sqrt{n}/\sigma < (\mu_0 - \mu)\sqrt{n}/\sigma - z_{\alpha/2}$.

Men siden $Z \sim N(0,1)$ har den første begivenheten sannsynlighet $1-\Phi(z_{\alpha/2} - (\mu - \mu_0)\sqrt{n}/\sigma)$ og den andre $\Phi(-z_{\alpha/2} - (\mu - \mu_0)\sqrt{n}/\sigma)$.

Oppgave 2

a

Likelihooden blir $L(\lambda) = \frac{\lambda^X}{X!}\exp(-\lambda)$ hvilket gir log-likelihood $l(\lambda) = -\ln(X!) + X\ln(\lambda) - \lambda$ som har derivert $l'(\lambda) = X/\lambda - 1$. Ved å løse $l'(\hat{\lambda}) = 0$ fås MLE $\hat{\lambda} = X$.

Momentestimatoren er løsningen av ligningen $E[X] = h(\lambda) = X$. Men her er $E[X] = h(\lambda) = \lambda$.

b

Vi har $0.95 \approx P(-1.96 < Z = (\hat{\lambda} - \lambda)/\sqrt{\hat{\lambda}} < 1.96) = P(\hat{\lambda} - 1.96\sqrt{\hat{\lambda}} < \lambda < \hat{\lambda} + 1.96\sqrt{\hat{\lambda}})$ og $\hat{\lambda} \pm 1.96\sqrt{\hat{\lambda}}$ er et tilnærmet 95% konfidensintervall for λ

Intervallet blir $135 \pm 1.96 * \sqrt{1}35 = (112.2, 157.8)$

Normaltilnærmingen regnes som bra for $\lambda > 10$ og med $X = \hat{\lambda} = 135$ er det god grunn til å anta at λ er vensentlig større enn 10 og dermed at tilnærmingen er god.

c

Dette intervallet gis ved $-1.96 < Z_0 = (\hat{\lambda} - \lambda)/\sqrt{\lambda} < 1.96$ som genererer en 2. gradsulikhet i λ gitt ved $(\hat{\lambda} - \lambda)^2 < 1.96^2\lambda$ som igjen er ekvivalent $\lambda^2 - (2\hat{\lambda} + 1.96^2)\lambda + \hat{\lambda}^2 < 0$. Løsningen av denne er gitt i formelen.

Grafisk kan løsningen finnes å plotte f.eks. $(\hat{\lambda} - \lambda)^2/\lambda$ mot λ og se hvor denne kurven skjærer verdien $1.96^2 \approx 3.84$.

Oppgave 3

a

β_0 estimeres til 311.4, dette er anslaget for antall dødsulykker i 2000. β_1 estimeres til -11.6, så dataene tilsier at vi har hatt en nedgang i dødsulykker på 11.6 per år. Dessuten fås et estimat for standardavviket til Y_i, altså σ, lik 20.88.

Vi har $t = (\hat{\beta}_1 - \beta_1)/se(\beta_1) \sim t_{18-2}$ og dermed blir, med 97.5-persentilen i t_{16} fordelingen lik 2.12, 95% konfidensintervallet for β_1 gitt som $-11.6 \pm$

(Fortsettes på side 3.)

$2.12 * 0.95 = (-13.6, -9.6)$. Siden -10 er med i dette intervallet kan vi ikke forkaste $\mathrm{H}_0 : \beta_1 = -10$ på 5% nivå.

Vi har en forklart andel av variasjon $R^2 = 0.90$, altå er det meget sterk sammenheng mellom år og antall dødsulykker over denne perioden.

b

Minste kvadraters estimatorene for (β_0, β_1) er de verdiene som minimerer $\sum_{i=1}^{18}(Y_i - \beta_0 - \beta_1 x_i)^2 = f(\beta_0, \beta_1)$.

Hvis ε_i er normalfordelte (og uavhengige) med forventning 0 og varians σ^2 blir $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ der $\mu_i = \beta_0 + \beta_1 x_i$ og også uavhengige. Dermed fås likelihood

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{18} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(Y_i - \mu_i)^2) = (2\pi\sigma^2)^{-9} \exp(-\frac{1}{2\sigma^2} f(\beta_0, \beta_1))$$

Minimering av $f(\beta_0, \beta_1)$ er dermed ekvivalent med maksimering av $L(\beta_0, \beta_1, \sigma^2)$ for enhver gitt σ^2.

c

Siden $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$ blir $\mathrm{E}(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$ og dermed fås

$$\begin{aligned} \mathrm{E}[\hat{\beta}_1] &= \frac{\sum_i (x_i - \bar{x})\mathrm{E}[(Y_i - \bar{Y})]}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ &= \beta_1 \frac{\sum_i (x_i - \bar{x})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \beta_1 \end{aligned}$$

så vi har forventningsretthet av $\hat{\beta}_1$.

Siden Y_i-ene er uavhengige blir

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})^2 \mathrm{Var}(Y_i)}{[\sum_i (x_i - \bar{x})^2]^2} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

d

I det første plottet ser vi $(\hat{Y}_i, e_i)$ der $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ er de predikerte verdiene og $e_i = Y_i - \hat{Y}_i$ residualene fra modellen. Hvis man ser mønstre i plottet indikerer dette avvik fra linearitetsantagelsen. For det aktuelle plottet er det ikke slike avvik.

Man kan også se om variansen varierer med $\mu_i = \beta_0 + \beta_1 x_i$ hvis variasjonen i e_i øker eller avtar med $\hat{Y}_i$. Heller ikke dette synes å være tilfelle.

Det andre plottet er et qq-plott over de standardiserte residualene e_i^*, som skalerer e_i med S (og influensverdiene slik at alle e_i^* har lik varians). Hvis punktene ligger klart utenfor den rette linja tyder dette på at normalfordelingsantagelsen ikke er tilfredstilt. Dette er heller ikke noe problem her.

Det 3. plottet viser $(\hat{Y}_i, \sqrt{|e_i^*|})$ og brukes til å vurdere mer spesifikt om variansen avhenger av μ_i. Kurven antyder en tendens til litt større varians med store μ_i, men sammenhengen er ikke monoton og ikke spesielt klar.

(Fortsettes på side 4.)

I henhold til Poissonantagelsen bør $\text{Var}(Y_i) = \mu_i$. Plottene viste ikke klart en slik sammenheng. Vi har dessuten at estimatet for σ er $S = 20.88$ og altså $S^2 = 20.88 = 436$. Dette er større enn alle $\hat{Y}_i$ som vil variere fra 125 til 323. Dette indikerer at variasjonen er større enn det Poissonfordelingen tilsier.

SLUTT