TMA4240/TMA4245
Statistikk
Eksamen august 2019

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a) For å være en gyldig sannsynlighetsfordeling må vi ha at $\sum_x p(x) = 1$, slik at $k = 0.1$.

Den kumulative fordelingsfunksjonen til X er gitt ved

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < -1 \\ 0.3 & -1 \leq x < 0 \\ 0.9 & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}.$$

Forventningsverdien er gitt som

$$\mathrm{E}[X] = \sum_x x \cdot p(x) = -1 \cdot 0.3 + 0 \cdot 0.6 + 1 \cdot 0.1 = -0.2.$$

For å finne variansen bruker vi at $\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$, der

$$\mathrm{E}[X^2] = \sum_x x^2 \cdot p(x) = (-1)^2 \cdot 0.3 + 0^2 \cdot 0.6 + 1^2 \cdot 0.1 = 0.4.$$

Variansen til X er dermed

$$\mathrm{Var}[X] = 0.4 - (-0.2)^2 = 0.36.$$

Oppgave 2

a) For $t \geq 0$ har vi at

$$\begin{aligned} P(T \leq t) &= \int_0^t \frac{1}{\beta} e^{-x/\beta} dx \\ &= \left[-e^{-x/30} \right]_0^t \\ &= 1 - e^{-t/30}, \end{aligned}$$

mens $P(T \leq t) = 0$ for $t < 0$.

Vi har at

$$P(T < 20) = P(T \leq 20) = F(20) = 1 - e^{-20/30} = 0.487,$$

og at

$$\begin{aligned} P(T < 20 \cup T > 40) &= P(T < 20) + P(T > 40) \\ &= P(T < 20) + 1 - P(T < 40) \\ &= 0.487 + 1 - (1 - e^{-40/30}) \\ &= 0.750. \end{aligned}$$

For at $P(T \leq k) = 0.5$ må vi ha at

$$1 - e^{-k/30} = 0.5 \iff k = -30 \cdot \ln(0.5) \approx 20.79.$$

b) Vi har at

$$\begin{aligned} P(T \geq t + s | T \geq t) &= \frac{P(T \geq t + s \cap T \geq t)}{P(T \geq t)} \\ &= \frac{P(T \geq t + s)}{P(T \geq t)} \\ &= \frac{e^{-(t+s)/\beta}}{e^{-t/\beta}} \\ &= e^{-s/\beta} \\ &= P(T \geq s). \end{aligned}$$

Vi ønsker å vise at sannsynlighetstettheten til $Y = \frac{2}{\beta}T$ er khikvadratfordelt med 2 frihetsgrader, det vil si at sannsynlighetstettheten til Y er gitt ved

$$f_Y(y) = \frac{1}{2}e^{-y/2},$$

for $y \geq 0$. Ettersom transformasjonen er en-til-en kan vi ta i bruk transformasjonsformelen, som gir

$$\begin{aligned} f_Y(y) &= f_T\left(\frac{y\beta}{2}\right)\left|\frac{d}{dy}\frac{y\beta}{2}\right| \\ &= \frac{1}{\beta}e^{-\frac{y\beta/2}{\beta}}\frac{\beta}{2} \\ &= \frac{1}{2}e^{-y/2}, \end{aligned}$$

for $y \geq 0$ og 0 ellers, som var det vi skulle vise.

c) Ettersom $T_1, T_2, \ldots, T_n$ er uavhengige av hverandre er også $\frac{2}{\beta}T_1, \frac{2}{\beta}T_2, \ldots, \frac{2}{\beta}T_n$ uavhengige av hverandre. Det er kjent at summen av uavhengige khikvadratfordelte variabler også er khikvadratfordelte, med frihetsgrader lik summen av de individuelle frihetsgradene. Ettersom $\frac{2}{\beta}T_i$ er khikvadratfordelt med 2 frihetsgrader er $\frac{2}{\beta}\sum_{i=1}^n T_i$ khikvadratfordelt med $2n$ frihetsgrader.

Vi har dermed at

$$P\left(\chi^2_{1-\alpha/2,2n} \leq \frac{2\sum_{i=1}^n T_i}{\beta} \leq \chi^2_{\alpha/2,2n}\right) = 1 - \alpha.$$

Ved å løse for β får vi at

$$P\left(\frac{2\sum_{i=1}^{n} T_i}{\chi^2_{\alpha/2,2n}} \leq \beta \leq \frac{2\sum_{i=1}^{n} T_i}{\chi^2_{1-\alpha/2,2n}}\right) = 1-\alpha,$$

som vil si at et $(1-\alpha)\cdot 100\%$ konfidensintervall for β er

$$\left[\frac{2\sum_{i=1}^{n} T_i}{\chi^2_{\alpha/2,2n}}, \frac{2\sum i=1^{n} T_i}{\chi^2_{1-\alpha/2,2n}}\right],$$

hvor $P(V \geq \chi^2_{\alpha/2,\nu}) = \alpha/2$, der V er khikvadratfordelt med ν frihetsgrader.

Innsatt verdiene fra oppgaven får vi intervallet

$$\left[\frac{2\cdot 30}{59.342}, \frac{2\cdot 30}{24.433}\right] = [1.011, 2.456].$$

Oppgave 3

a) Sannsynligheten for å sovne innen 20 minutter er

$$P(X \leq 20) = P\left(Z \leq \frac{20-35}{10}\right) = P(Z \leq -1.5) = 0.067.$$

Sannsynligheten for å ikke ha sovnet innen 40 minutter, gitt at man fortsatt var våken etter 20 minutter er

$$\begin{aligned} P(X \geq 40 | X \geq 20) &= \frac{P(X \geq 40 \cap X \geq 20)}{P(X \geq 20)} \\ &= \frac{P(X \geq 40)}{P(X \geq 20)} \\ &= \frac{1 - P(Z \leq \frac{40-35}{10})}{1 - P(X \leq 20)} \\ &= \frac{1-0.691}{0.933} \\ &= 0.331. \end{aligned}$$

La X_1 være tiden det tar før den første pasienten sovner og X_2 være tiden før den andre pasienten sovner, begge normalfordelte med forventningsverdi 35 minutter og standardavvik 10 minutter. Ettersom de to pasientene tar sovemedisin uavhengig av hverandre, og en lineærkombinasjon av uavhengige og normalfordelte stokastiske variable også er normalfordelt, har vi at $X_1 + X_2$ er normalfordelt med forventningsverdi $35 + 35 = 70$ minutter og standardavvik $\sqrt{10^2 + 10^2}$ minutter. Dermed får vi at

$$\begin{aligned} P(X_1 + X_2 \geq 90) &= 1 - P(X_1 + X_2 \leq 90) \\ &= 1 - P\left(Z \leq \frac{90-70}{\sqrt{10^2+10^2}}\right) \\ &= 1 - P(Z \leq 1.41) \\ &= 1 - 0.9207 \\ &= 0.0793. \end{aligned}$$

b) Vi bruker $\bar{X} - \bar{Y}$ som estimator for $\mu - \theta$. Ettersom $\bar{X}$ og $\bar{Y}$ er uavhengige og normalfordelte stokastiske variable har vi at $\bar{X} - \bar{Y}$ også er normalfordelt, med forventningsverdi lik 0 og standardavvik $\sqrt{\frac{10^2}{n} + \frac{12^2}{n}}$, under nullhypotesen. Ved signifikansnivå $\alpha = 0.1$ forkaster vi nullhypotesen dersom

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{10^2}{n} + \frac{12^2}{n}}} \geq z_{0.1} = 1.282.$$

Det kreves at testens styrke skal være minst 95% når $\mathrm{H}_1 : \mu - \theta = 5$ er sann, altsåhar vi at

$$\begin{aligned}
0.95 &\leq P(\text{forkast } \mathrm{H}_0 \text{ når } \mathrm{H}_1 : \mu - \theta = 5) \\
&= P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{10^2}{n} + \frac{12^2}{n}}} \geq 1.282 \text{ når } \mathrm{H}_1 : \mu - \theta = 5 \right) \\
&= P\left(\bar{X} - \bar{Y} \geq 1.282 \sqrt{\frac{10^2}{n} + \frac{12^2}{n}} \text{ når } \mathrm{H}_1 : \mu - \theta = 5 \right) \\
&= P\left(Z \geq \frac{1.282\sqrt{\frac{10^2}{n} + \frac{12^2}{n}} - 5}{\sqrt{\frac{10^2}{n} + \frac{12^2}{n}}} \text{ når } \mathrm{H}_1 : \mu - \theta = 5 \right) \\
&= P\left(Z \geq 1.282 - \frac{5}{\sqrt{\frac{10^2}{n} + \frac{12^2}{n}}} \text{ når } \mathrm{H}_1 : \mu - \theta = 5 \right).
\end{aligned}$$

Ettersom sannsynligheten for denne hendelsen er større enn eller lik 0.95, må sannsynligheten for komplementet av hendelsen være mindre enn eller lik 0.05, altså har vi

$$0.05 \geq P\left(Z \leq 1.282 - \frac{5}{\sqrt{\frac{10^2}{n} + \frac{12^2}{n}}} \text{ når } \mathrm{H}_1 : \mu - \theta = 5 \right).$$

Følgende får vi at

$$\begin{aligned}
-z_{0.05} = -1.645 &\geq 1.282 - \frac{5}{\sqrt{\frac{10^2}{n} + \frac{12^2}{n}}} \\
2.927\sqrt{\frac{10^2}{n} + \frac{12^2}{n}} &\leq 5 \\
\frac{10^2}{n} + \frac{12^2}{n} &\leq \left(\frac{5}{2.927} \right)^2 \\
n &\geq \frac{2.927^2 \cdot (10^2 + 12^2)}{5^2} \approx 83.56.
\end{aligned}$$

Hvert selskap må altså gi sovemedisin til minst 84 personer.

Oppgave 4

a) Under nullhypotesen har vi at $Z = \frac{\bar{X}-40}{\sqrt{0.05/10}}$ er standard normalfordelt. Vi har observert

$$Z_{\text{obs}} = \frac{40.2 - 40}{\sqrt{0.5^2/10}} = 1.27.$$

Fra definisjonen av $p-$verdi har vi at

$$\begin{aligned} p-\text{verdi} &= P(|Z| \geq 1.27) \\ &= 2P(Z \geq 1.27) \\ &= 2(1 - P(Z \leq 1.27)) \\ &= 2(1 - 0.898) \\ &= 0.204. \end{aligned}$$

Altså vil vi ikke forkaste nullhypotesen ved signifi kansnivå $\alpha = 0.1$.

Ta utgangsPunkt i $\bar{X} - X_0$, som er en lineÃ¦rkombinasjon av uavhengige og normalfordelte stokastiske variabler, og dermed også normalfordelt med forventningsverdi like 0 og med varians lik $\sigma^2(\frac{1}{n}+1)$. Vi har dermed at

$$P\left(-z_{0.025} \leq \frac{\bar{X} - X_0}{\sqrt{\sigma^2(\frac{1}{n}+1)}} \leq z_{0.025}\right) = 0.95.$$

Ved å isolere X_0 får vi

$$P\left(\bar{X} - z_{0.025}\sqrt{\sigma^2(\frac{1}{n}+1)} \leq X_0 \leq \bar{X} + z_{0.025}\sqrt{\sigma^2(\frac{1}{n}+1)}\right) = 0.95,$$

som betyr at et 95% prediksjonsintervall for X_0 er

$$\left[\bar{x} - z_{0.025}\sqrt{\sigma^2\left(\frac{1}{n}+1\right)}, \bar{x} + z_{0.025}\sqrt{\sigma^2\left(\frac{1}{n}+1\right)}\right],$$

hvor $\bar{x}$ er den observerte verdien av $\bar{X}$.

Oppgave 5

a) Rimelighetsfunksjonen er gitt ved

$$\begin{aligned} L(\beta_1) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_1 x_i)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 x_i)^2}. \end{aligned}$$

Log-rimelighetsfunksjonen er dermed lik

$$l(\beta_1) = -n\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 x_i)^2.$$

Vi deriverer med hensyn på β og setter lik 0.

$$\frac{\partial l(\beta_1)}{\partial \beta_1} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \beta_1 x_i)x_i = 0.$$

Løser vi dette uttrykket får vi at $\beta_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$, noe som vil si at sannsynlighetsmaksimeringsestimatoren for β_1 er lik

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Videre blir forventningsverdi og varians lik

$$\begin{aligned}
\mathrm{E}[\tilde{\beta}_1] &= \mathrm{E}\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] \\
&= \frac{\sum_{i=1}^n x_i \mathrm{E}[Y_i]}{\sum_{i=1}^n x_i^2} \\
&= \frac{\sum_{i=1}^n x_i \beta_1 x_i}{\sum_{i=1}^n x_i^2} \\
&= \beta_1,
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[\tilde{\beta}_1] &= \mathrm{Var}\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] \\
&= \frac{\sum_{i=1}^n x_i^2 \mathrm{Var}[Y_i]}{(\sum_{i=1}^n x_i^2)^2} \\
&= \frac{\sum_{i=1}^n x_i^2 \sigma^2}{(\sum_{i=1}^n x_i^2)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.
\end{aligned}$$

b) Estimert forventet verdi blir $0.2 + 1.32 \cdot 0.25 = 0.55$.

Følgende antakelser må være oppfylt for å kunne ta i bruk en enkel lineær regresjonsmodell:

- En lineær sammenheng mellom x og y. Ut i fra regresjonplottet kan det se ut som det er en svak positiv lineær sammenheng mellom x og y.
- Konstant varians σ^2. Fra residualplottet kan vi obserere at variansen øker ettersom verdiene til $\hat{y}_i$ øker. Man kan også se den sammen trenden fra regresjonsplottet, med større spredning av y-verdier ettersom x øker. Altså er ikke variansen konstant, men trolig heller en funksjon av x.
- Residualene er normalfordelte. Q-Q-plottet indikerer at kvantilene for de estimerte residualene avviker fra de teoretiske kvantilene. Det er ikke uvanlig at det er noe avvik ved randen, men i dette tilfellet er det avvik også i intervallet $[-1, 1]$, samt at avviket ved randen er relativt stort, noe som gir en indikasjon på at residualene ikke er normalfordelte.

- Y_i-ene er uavhengige. Dette er ikke like enkelt å avgjøre, men ingen av plottene gir noen antydning til at de skulle være avhengige.

Oppgave 6

a) Hver Z_i kan ta en av to verdier (enten 0 eller 1), og vi har kun ett forsøk med suksess (1) eller ikke suksess (0). Derfor må Z_i være bernoullifordelt, med suksessannsynlighet lik

$$\begin{aligned} p_i &= P(Z_i = 1) \\ &= P(X_i = 0 \cap X_3 = 0) \\ &= P(X_i = 0)P(X_3 = 0) \\ &= e^{-\mu_i}e^{-\mu_3} \\ &= e^{-(\mu_i+\mu_3)}. \end{aligned}$$

For en bernoullifordelt variabel er det kjent at $\mathrm{E}[Z_i] = p_i$ og $\mathrm{Var}[Z_i] = p_i(1-p_i)$.

Fra definisjonen av kovarians har vi at $\mathrm{Cov}[Z_1, Z_2] = \mathrm{E}[Z_1Z_2] - \mathrm{E}[Z_1]\mathrm{E}[Z_2]$, der

$$\begin{aligned} \mathrm{E}[Z_1Z_2] &= \sum_{z_1}\sum_{z_2} z_1z_2P(Z_1 = z_1, Z_2 = z_2) \\ &= 0 \cdot 0 \cdot P(Z_1 = 0, Z_2 = 0) + 1 \cdot 0 \cdot P(Z_1 = 1, Z_2 = 0) \\ &+ 0 \cdot 1 \cdot P(Z_1 = 0, Z_2 = 1) + 1 \cdot 1 \cdot P(Z_1 = 1, Z_2 = 1) \\ &= P(Z_1 = 1, Z_2 = 1) \\ &= P(X_1 = 0, X_2 = 0, X_3 = 0) \\ &= P(X_1 = 0)P(X_2 = 0)P(X_3 = 0) \\ &= e^{-\mu_1}e^{-\mu_2}e^{-\mu_3} \\ &= e^{-(\mu_1+\mu_2+\mu_3)}. \end{aligned}$$

Dermed får vi at

$$\begin{aligned} \mathrm{Cov}[Z_1, Z_2] &= \mathrm{E}[Z_1Z_2] - \mathrm{E}[Z_1]\mathrm{E}[Z_2] \\ &= e^{-(\mu_1+\mu_2+\mu_3)} - e^{-(\mu_1+\mu_3)}e^{-(\mu_2+\mu_3)} \\ &= e^{-(\mu_1+\mu_2+\mu_3)}(1 - e^{-\mu_3}). \end{aligned}$$