TMA4245 Statistics
Exam June 2019

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a) Siden X er binomisk fordelt har vi punktsannsynlighet

$$P(X=4) = \binom{17}{4} 0.2^4 (1-0.2)^{17-4} = 0.209.$$

Bruker komplementærsetningen

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - \sum_{x=0}^{3} \binom{17}{x} 0.2^x (1-0.2)^{17-x} \\ &= 1 - (0.023 + 0.096 + 0.191 + 0.239) \\ &= 0.451. \end{aligned}$$

Bruker definisjonen av betinget sannsynlighet

$$\begin{aligned} P(X \geq 6 | X \geq 4) &= \frac{P(X \geq 6 \cap X \geq 4)}{P(X \geq 4)} \\ &= \frac{P(X \geq 6)}{P(X \geq 4)} \\ &= \frac{1 - P(X \leq 5)}{P(X \geq 4)} \\ &= \frac{1 - (0.023 + 0.096 + 0.191 + 0.239 + 0.209 + 0.136)}{0.451} \\ &= 0.235. \end{aligned}$$

b) Sentralgrenseteoremet sier at dersom $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte stokastiske variabler med forventningsverdi $\mathrm{E}(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$ vil sannsynlighetsfordelingen til

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

går mot standard normalfordeling når $n \to \infty$.
La $X_1, X_2, \ldots, X_{215}$ være uavhengige stokastiske variabler der $X_i = 1$ dersom pasient nummer i hadde en bruddskade etter en sparkesykkelulykke og 0 ellers. Da er X_i

Bernoullifordelt med suksess-sannsynlighet p for $i = 1, 2, \ldots, 215$. Vi har videre at $\mu = \mathrm{E}(X_i) = p$ og $\sigma^2 = \mathrm{Var}(X_i) = p(1-p)$. Det er kjent at $\hat{p} = \frac{X}{n}$ er sannsynlighetsmaksimeringsestimatoren til p, der $X = \sum_{i=1}^{215} X_i$.
Fra sentralgrenseteoremet har vi at

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

går mot standard normalfordeling. Vi bytter p i nevneren med $\hat{p}$ og får

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx n(z; 0, 1).$$

Vi tar utgangspunkt i

$$P\left(-z_{0.025} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{0.025}\right) = 0.95$$

og løser ulikhetene med hensyn på p:

$$P\left(\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

Et 95% konfidensintervall for p er derfor

$$\left[\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Innsatt tallene får vi $[0.193, 0.309]$.

Oppgave 2

a) Vi ønsker å utføre følgende hypotesetest

$$H_0 : \mu = 3000 \text{ mot } H_1 : \mu > 3000.$$

Siden både forventningsverdien og variansen er ukjent har vi under nullhypotesen at

$$T = \frac{\bar{X} - 3000}{\sqrt{\frac{S^2}{17}}} \sim t_{16}$$

Vi forkaster nullhypotesen dersom $T^{obs} \geq t_{16,0.05} = 1.745$. Innsatt tallene våre har vi

$$T^{obs} = \frac{3200 - 3000}{\sqrt{\frac{300^2}{17}}} = 2.749.$$

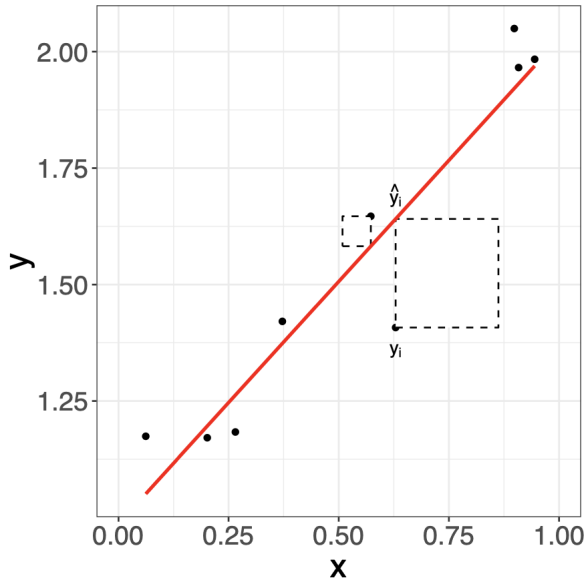
Vi forkaster altså nullhypotesen.
Ja, det er grunnlag til å tro at stømforbruket er høyere enn 3000 kilowattimer.

Oppgave 3

a) Vi kan finne minste kvadraters metode estimatorer b_0 og b_1 til β_0 og β_1 ved å minimere summen av kvadratfeilene for den estimerte modellen:

$$\text{SSE} = \sum_{i=1}^{25}(y_i - \hat{y})^2 = \sum_{i=1}^{25}(y_i - b_0 - b_1 x_i)^2.$$

Dette gjøres ved å sette den deriverte $\partial SSE/\partial b_0$ og $\partial SSE/\partial b_1$ lik 0 og løse det lineære ligningssystemet for b_0 og b_1. Ved lineær regresjon har vi følgende modellantagelser



- lineær sammenheng for E$(Y|x)$: i Figur 1(a) er det en klar lineær sammenheng mellom x og y.
- varians uavhengig av x, de vil si Var$(Y|x) = \sigma^2$: det er ingen tydelig trend i spredning av predikert avrenning i Figur 1(b) mot residualene.
- støyleddene ϵ_i er normalfordelte og uavhengige: observasjonene virker å være jevnt fordelt omkring den tilpassede linja i Figur 1(a) uten noe tydeleg trend i spredning. Residualene virker og å være sentrert rundt 0. En kunne også ha sett på et normalsannsynlighetsplott av residualene for å avgjøre om de er normalfordelte.

b) Estimert forventet avrenning når $x = 2000$ er $-1364 + 1.08 \cdot 2000 = 796$.
Vi har at

$$\begin{aligned} E(\hat{Y}_0 - Y_0) &= E(\hat{\beta}_0 + \hat{\beta}_1 x_0 - \beta_0 - \beta_1 x_0 - \epsilon) \\ &= E(\hat{\beta}_0) + x_0 E(\hat{\beta}_1) - \beta_0 - \beta_1 x_0 - E(\epsilon) \\ &= \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 \\ &= 0 \end{aligned}$$

siden $\hat{\beta}_0$ og $\hat{\beta}_1$ er forventningsrette estimatorer. Videre er

$$\begin{aligned}\text{Var}(\hat{Y}_0 - Y_0) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 - \beta_0 - \beta_1 x_0 - \epsilon) \\ &= \text{Var}(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 - \epsilon) \\ &= \text{Var}(\bar{Y}) + (x_0 - \bar{x})^2 \text{Var}(\hat{\beta}_1) + \text{Var}(\epsilon)\end{aligned}$$

der vi har brukt at $\bar{Y}$ og $\hat{\beta}_1$ er uavhengige stokastiske variabler. Vi trenger $\text{Var}(\hat{\beta}_1)$:

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \frac{1}{(\sum_{i=1}^{25}(x_i - \bar{x})^2)^2} \text{Var}\left(\sum_{i=1}^{25}(x_i - \bar{x})Y_i\right) \\ &= \frac{1}{(\sum_{i=1}^{25}(x_i - \bar{x})^2)^2} \sum_{i=1}^{25}(x_i - \bar{x})^2 \text{Var}(Y_i) \\ &= \frac{\sigma^2 \sum_{i=1}^{25}(x_i - \bar{x})^2}{(\sum_{i=1}^{25}(x_i - \bar{x})^2)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}.\end{aligned}$$

Ved å sette dette inn får vi

$$\begin{aligned}\text{Var}(\hat{Y}_0 - Y_0) &= \text{Var}(\bar{Y}) + (x_0 - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2} + \text{Var}(\epsilon) \\ &= \frac{\sigma^2}{25} + \frac{\sigma^2 (x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2} + \sigma^2 \\ &= \sigma^2 \left(1 + \frac{1}{25} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}\right).\end{aligned}$$

Vi finner et 95% prediksjonsintervall for Y_0 ved å se på

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\sigma^2 \left(1 + \frac{1}{25} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}\right)}}$$

som er standard normalfordelt. Siden σ^2 er ukjent får vi

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{s^2 \left(1 + \frac{1}{25} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}\right)}} \sim t_{23}.$$

Ta utgangspunkt i

$$P\left(-t_{23,0.025} \leq \frac{\hat{Y}_0 - Y_0}{\sqrt{s^2 \left(1 + \frac{1}{25} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}\right)}} \leq t_{23,0.025}\right) = 0.95.$$

Et 95% prediksjonsintervall for Y_0 er gitt som

$$\left[\hat{Y}_0 - t_{23,0.025}\sqrt{s^2\left(1+\frac{1}{25}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{25}(x_i-\bar{x})^2}\right)}, \hat{Y}_0 + t_{23,0.025}\sqrt{s^2\left(1+\frac{1}{25}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{25}(x_i-\bar{x})^2}\right)}\right].$$

Oppgave 4

a) En god estimator μ^* for den ukjente parameteren μ er forventningsrett, det vil si $\mathrm{E}(\mu^*) = \mu$. Av flere forventningsrette estimatorer $\mu_1^*, \ldots, \mu_K^*$ velger vi den estimatoren med lavest varians.

Vi har

$$E(\hat{\mu}) = E(Y) = \mu$$

og

$$E(\tilde{\mu}) = E\left(\frac{1}{2}(X+Y)\right) = \frac{1}{2}E(X) + \frac{1}{2}E(Y) = \frac{\mu}{2} + \frac{\mu}{2} = \mu.$$

Siden begge estimatorene er forventningsrette må vi avgjøre hvem av dem som har minst varians:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(Y) = 0.5^2 = 0.25$$

og

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}) &= \mathrm{Var}\left(\frac{1}{2}(X+Y)\right)\\ &= \left(\frac{1}{2}\right)^2\mathrm{Var}(X) + \left(\frac{1}{2}\right)^2\mathrm{Var}(Y) + 2\cdot\frac{1}{2}\cdot\frac{1}{2}\mathrm{Cov}(X,Y)\\ &= \frac{1}{4}\cdot 1^2 + \frac{1}{4}\cdot 0.5^2 + \frac{1}{2}\cdot(-0.2)\\ &= 0.2125.\end{aligned}$$

Siden $\mathrm{Var}(\tilde{\mu}) < \mathrm{Var}(\hat{\mu})$ vil vi foretrekke $\tilde{\mu}$ som estimator for μ.

Oppgave 5

a) To stokastiske variabler V og W har same sannsynlighetsfordeling dersom de momentgenererende funksjonene deres er like for alle t. Vi må altså vise at den momentgenererende funksjonen til Y er $M_Y(t) = \frac{1}{(1-\beta t)^{n\alpha}}$.
Merk: det opprinnelige eksamenssettet inneholdt en typografisk feil i definisjonen for den momentgenererende funksjonen da det stod $M_X(t) = (1-t/\beta)^{-\alpha}$. Utregninga (og konklusjonen) ville ha vært på samme måte som vist under. Denne feilen blir tatt hensyn til ved sensur.

Siden $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variabler har vi at den momentgenererende funksjonen til Y er gitt ved

$$\begin{aligned}M_Y(t) &= M_{X_1+X_2+\cdots+X_n}(t)\\ &= M_{X_1}(t)M_{X_2}(t)\ldots M_{X_n}(t)\\ &= \frac{1}{(1-\beta t)^{\alpha}}\frac{1}{(1-\beta t)^{\alpha}}\cdots\frac{1}{(1-\beta t)^{\alpha}} \quad = \frac{1}{(1-\beta t)^{n\alpha}}\end{aligned}$$

som er den momentgenererende funksjonen til en gammafordelt stokastisk variabel med parametre $n\alpha$ og β.

Rimelighetsfunksjonen basert på det tilfeldige utvalget $Y = y$ er

$$L(\beta; y) = \frac{1}{\beta^{n\alpha}\Gamma(n\alpha)} y^{n\alpha-1} e^{-y/\beta}$$

for $y > 0$. Log-rimelighetsfunksjonen er

$$l(\beta; y) = -n\alpha \log \beta - \log \Gamma(n\alpha) + (n\alpha - 1)\log(y) - \frac{y}{\beta}.$$

Vi deriverer log-rimelighetsfunksjonen, setter uttrykket lik 0 og løser likningen for β:

$$\frac{dl(\beta; y)}{\beta} = -\frac{n\alpha}{\beta} + \frac{y}{\beta^2} = 0$$

Vi får da

$$\beta n\alpha = y \implies \beta = \frac{y}{n\alpha}.$$

Det vil si at sannsynlighetsmaksimeringsestimatoren for β basert på y er $\hat{\beta} = \frac{Y}{n\alpha}$.

Oppgave 6

a) Fra komplementærsetningen har vi

$$\begin{aligned} P(2X > 3) &= 1 - P(X \leq 3/2) \\ &= 1 - \Phi\left(>\leq \frac{3/2 - 1}{1} \right) \\ &= 1 - \Phi(0.5) \\ &= 1 - 0.691 \\ &= 0.309 \end{aligned}$$

Siden X og Y er uavhengige har vi

$$P(2X > 3 | Y > 0) = P(2X > 3) = 0.309.$$

Siden $X - Y$ er en lineærkombinasjon av uavhengige normalfordelte variabler er den og normalfordelt med forventningsverdi

$$E(X - Y) = E(X) - E(Y) = 1 - 0 = 1$$

og varians

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1 + 1 = 2.$$

Vi har da

$$\begin{aligned} P(-1 \leq X - Y \leq 1) &= P\left(\frac{-1 - 1}{\sqrt{2}} \leq Z \leq \frac{1 - 1}{\sqrt{2}} \right) \\ &= \Phi(0) - \Phi(-\sqrt{2}) \\ &= 0.5 - 0.079 \\ &= 0.421 \end{aligned}$$

b) Hver X_i kan enten være mindre eller lik x, eller større enn x. Sannsynligheten for at en tilfeldig valgt X_i er mindre enn eller lik x er $p = P(X_i \leq x) = P(X \leq x)$ for $i = 1, 2, \ldots, 10$. Siden hver av hendelsene $\{X_1 \leq x\}, \ldots, \{X_{10} \leq x\}$ er uavhengige av hverandre har vi et binomisk forsøk med konstant suksess-sannsynlighet p og $n = 10$ forsøk. Vi har derfor

$$\begin{aligned} P(\text{høgst 5 av} X_i\text{-ene er større enn } x) &= \sum_{k=0}^{5} \binom{10}{k} p^k (1-p)^{10-k} \\ &= \sum_{k=0}^{5} \binom{10}{k} [P(X \leq x)]^k (1 - P(X \leq x))^{10-k} \\ &= \sum_{k=0}^{5} \binom{10}{k} [\Phi(x-1)]^k (1 - \Phi(x-1))^{10-k} \end{aligned}$$

der $\Phi(x)$ er den kumulative fordelingsfunksjonen til standard normalfordeling.
Siden X_i og Y_j er parvis uavhengige for alle par av i og j har vi

$$\begin{aligned} P(\max\{X_1, \ldots, X_{10}, Y_1, \ldots, Y_{15}\} \leq z) &= P(X_1 \leq z, \ldots, X_{10} \leq z, Y_1 \leq z, \ldots, Y_{15} \leq z) \\ &= \prod_{i=1}^{10} P(X_i \leq z) \prod_{j=1}^{15} P(Y_j \leq z) \\ &= [P(X \leq z)]^{10} [P(Y \leq z)]^{15} \\ &= [\Phi(z-1)]^{10} [\Phi(z)]^{15} \end{aligned}$$

der $\Phi(z)$ er den kumulative fordelingsfunksjonen til standard normalfordeling.

Oppgave 7

a) Under nullhypotesen er sannsynlighetstettheten til X_1 gitt ved

$$f(x; 0) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{elles.} \end{cases} .$$

Sannsynligheten for type I-feil for forkastningsregel 1

$$P(\text{forkast } H_0 \text{ når } H_0 \text{ er sann}) = P(X_1 \geq 0.95) = 0.05.$$

Vi ser på to tilfeller: for $0 < \tau < 0.95$

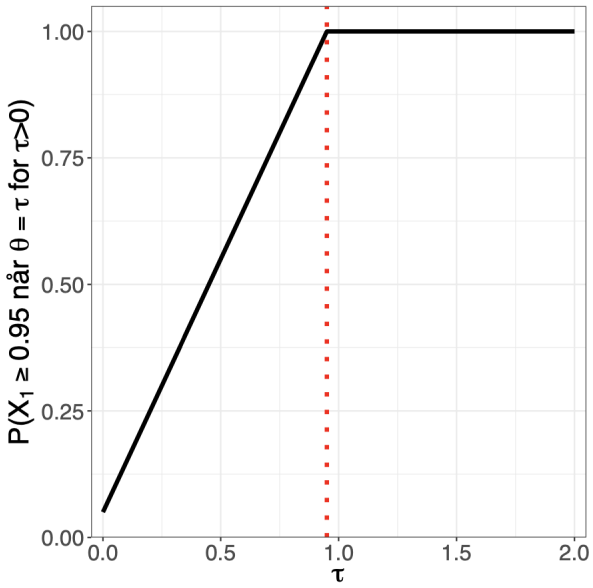
$$\begin{aligned} P(X_1 \geq 0.95 \text{ når } \theta = \tau \text{ for } \tau > 0) &= \int_{0.95}^{1+\tau} 1 dx \\ &= 1 + \tau - 0.95 \\ &= \tau + 0.05 \end{aligned}$$

og for $\tau \geq 0.95$ vil alltid $X_1 \geq 0.95$. Det vil si at vi alltid forkaster H_0 når $\tau \geq 0.95$ ved forkastningsregel 1.

Til sammen har vi

$$P(\text{forkast } H_0 \text{ når } \theta = \tau \text{ for } \tau > 0) = P(X_1 \geq 0.95 \text{ når } \theta = \tau \text{ for } \tau > 0)$$
$$= \begin{cases} \tau + 0.05 & 0 < \tau < 0.95 \\ 1 & \tau \geq 0.95. \end{cases}.$$

En skisse av testen sin styrke er gitt i Figur 1. Oppgaven kan løses på flere måter. Merk



Figur 1: Testen sin styrke for forkastningsregel 1.

at under nullhypotesen er

$$f(x_1, x_2; \theta = 0) = f(x_1; 0) f(x_2; 0) = \begin{cases} 1 & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 \\ 0 & \text{ellers.} \end{cases}.$$

Ved å se på sannsynligheten $X_1 + X_2 > k$, gitt nullhypotesen, får vi

$$\begin{aligned} P(X_1 + X_2 > k \text{ når } \theta = 0) &= \int_{k-1}^{1} \int_{k-x_1}^{1} 1 dx_2 dx_1 \\ &= \int_{k-1}^{1} 1 - k + x_1 dx_1 \\ &= \frac{(2-k)^2}{2} \end{aligned}$$

Dersom vi krever sannsynlighet for type I-feil for forkastningsregel 1 må vi ha $P(X_1 + X_2 > k \text{ når } \theta = 0) = 0.05$. Det vil si

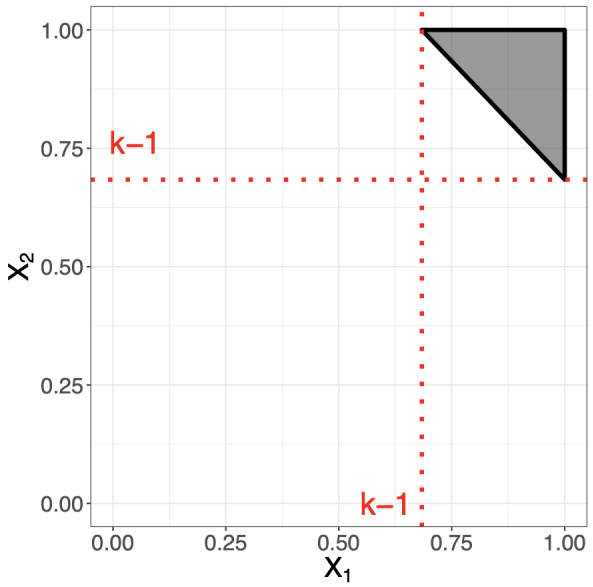
$$\frac{(2-k)^2}{2} = 0.05 \implies k = 2 - \sqrt{0.1}.$$

Husk at simultantettheten konstant for $(x_1, x_2) \in [0, 1] \times [0, 1]$. Et geometrisk argument kan sees fra Figur 2. Merk at den største verdien vi kan ha er $k = 2$ når $X_1 = X_2 = 1$.

Vi krever at arealet av den skraverte trekanten må være 0.05. Dersom vi fikserer $X_1 = 1$ er den minste verdien X_2 kan ha $X_2 = k - 1$ for at kravet om $X_1 + X_2 \geq k$ skal være oppfylt. På grunn av symmetri er den minste verdien X_1 kan ta dersom $X_2 = 1$ lik $X_1 = k - 1$. Arealet av trekanten er derfor

$$\frac{1}{2}((1-(k-1))(1-(k-1)) = \frac{(2-k)^2}{2}$$

som gir den samme verdien for k som over.



Figur 2: Skravert område viser der $X_1 + X_2 \geq k$ er slik at arealet er lik 0.05.