

Institutt for matematiske fag

Eksamensoppgave i TMA4245 Statistikk

Faglig kontakt under eksamen: Håkon Tjelmeland^a, Geir-Arne Fuglstad^b

Tlf: ^a48 22 18 96, ^b45 27 08 06

Faglig kontakt møter i eksamenslokalet:

Eksamensdato: Onsdag 28. mai 2025

Eksamenstid (fra–til): 09.00–13.00

Hjelpemiddelkode/Tillatte hjelpemidler: C

Tabeller og formles i statistikk, Akademika,
Bestemt, enkel kalkulator,
Ett A5-ark med egne håndskrevne notater.

Annen informasjon:

Alle svar skal begrunnes og besvarelsen skal inneholde naturlig mellomregning slik at det er helt klart hvilke regneregler som benyttes.

Målform/språk: bokmål

Antall sider (uten forside): 12

Antall sider vedlegg: 0

Kontrollert av:

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig ☐ 2-sidig ☒

sort/hvit ☒ farger ☐

skal ha flervalgskjema ☐

Dato

Sign

Oppgave 1

Ved å standardisere får vi at

$$P(X > 3) = 1 - P(X \leq 3) = 1 - \Phi\left(\frac{3-2}{\sqrt{9}}\right) = 1 - \Phi(0.33) = 1 - 0.6293 = \underline{\underline{0.3707}}.$$

Ved å starte med definisjonen av betinget sannsynlighet og deretter standardisere får vi at

$$\begin{aligned} P(X > 2|X < 3) &= \frac{P(X > 2, X < 3)}{P(X < 3)} = \frac{P(2 < X < 3)}{P(X < 3)} = \frac{P(X < 3) - P(X \leq 2)}{P(X < 3)} \\ &= \frac{P(X \leq 3) - P(X \leq 2)}{P(X \leq 3)} = \frac{\Phi\left(\frac{3-2}{\sqrt{9}}\right) - \Phi\left(\frac{2-2}{\sqrt{9}}\right)}{\Phi\left(\frac{3-2}{\sqrt{9}}\right)} = \frac{\Phi(0.33) - \Phi(0)}{\Phi(0.33)} \\ &= \frac{0.6293 - 0.5}{0.6293} = \underline{\underline{0.2055}}. \end{aligned}$$

Oppgave 2

Siden vi har at $X + Y \sim N(1 + 2, 9 + 4)$ får vi, ved å standardisere, at

$$\begin{aligned} P(X + Y > 0) &= 1 - P(X + Y \leq 0) = 1 - \Phi\left(\frac{0 - 3}{\sqrt{13}}\right) = 1 - \Phi(-0.83) \\ &= 1 - 0.2033 = \underline{\underline{0.7967}}. \end{aligned}$$

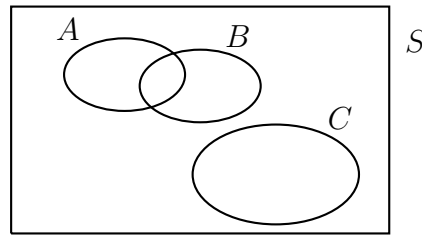
Ved å benytte regneregler for varians får vi

$$\begin{aligned} \text{Var}[2X - 4Y] &= \text{Var}[2X] + \text{Var}[-4Y] = 2^2\text{Var}[X] + (-4)^2\text{Var}[Y] \\ &= 4 \cdot 9 + 16 \cdot 4 = \underline{\underline{100}}. \end{aligned}$$

Oppgave 3

Ved å benytte de oppgitte simultane punktsannsynlighetene får man at

$$\begin{aligned} P(X = 0) &= P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 0, Y = 2) \\ &= 0.12 + 0.20 + 0.00 = \underline{\underline{0.32}}, \\ P(X \leq 1|Y \leq 1) &= \frac{P(X \leq 1, Y \leq 1)}{P(Y \leq 1)} = \frac{0.12 + 0.20 + 0.10 + 0.12}{0.12 + 0.20 + 0.10 + 0.12 + 0.04 + 0.10} \\ &= \underline{\underline{0.7941}}. \end{aligned}$$



Figur 1: Et venndiagram for situasjonen gitt i oppgave 4

Oppgave 4

Situasjonen gitt i oppgaven kan illustreres med venndiagrammet i figur 1. Ut fra dette venndiagrammet, samt definisjonen av betinget sannsynlighet, får vi at

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.30 + 0.25 - 0.15 = \underline{0.40},$$

$$P(A|C') = \frac{P(A \cap C')}{P(C')} = \frac{P(A)}{1 - P(C)} = \frac{0.30}{1 - 0.40} = \underline{0.50}.$$

Oppgave 5

Av de fem forslagene til kodelinjer er det to som vil generere en realisasjon fra en binomisk fordeling.

- `u = np.random.uniform(0,1,n)`
`success = (u <= p)`
`x = np.sum(success)`

Her blir hvert element i vektoren `success` lik 1 med sannsynlighet `p` og null med sannsynlighet `1-p`. Elementene i `success` er dessuten uavhengige av hverandre slik at den genererte verdien `x` blir binomisk fordelt med parametre `n` og `p`.

- `u = np.random.uniform(0,1)`
`values = np.arange(n+1) # lager vektoren (0,1,2,...,n)`
`F = binom.cdf(values, n, p)`
`x = np.sum(F <= u)`

Her genereres `u` uniformfordelt mellom 0 og 1, og vektoren `F` inneholder verdiene til $F(x)$, $x = 0, 1, \dots, n$ for en binomisk fordeling med parametre `n` og `p`. Den siste linja vil dermed beregne $x = F^{-1}(u)$, som betyr at `x` blir binomisk fordelt med parametre `n` og `p`.

- `u = np.random.normal(0,p,n)`
`x = np.sum(u * u)`

Her er elementene i u uavhengige og normalfordelt med forventningsverdi lik null og standardavvik lik p . Verdien til x regnes ut ved å kvadrere disse verdiene og deretter summere. Verdien til x vil da ikke ha en fordeling som vi har diskutert. Og spesielt er x ikke binomisk fordelt.

- `u = np.random.uniform(0,1,n)`
`v = - np.log(1 - u)`
`success = (v <= p)`
`x = np.sum(success)`

Her blir elementene i v uavhengige og eksponensialfordelt med parameter 1. Den utregnede verdien x blir binomisk fordelt med n forsøk og med sannslighet for suksess lik $P(Y \leq p)$ når Y er eksponensialfordelt med parameter 1. Siden $P(Y \leq p) \neq p$ vil dermed den genererte verdien x ikke være fra korrekt binomisk fordeling.

- `u = np.random.uniform(0,1,2*n)`
`success = (u <= p)`
`x = np.sum(success)/2`

Her vil $2x$ være binomisk fordelt med parametre n og $2p$, med x selv vil ikke være binomisk fordelt.

Oppgave 6

Siden vi her har trekning uten tilbakelegging vil antall mulige måter å trekke ut fem kuler fra 25 være

$$\binom{25}{5} = \frac{25 \cdot 24 \cdot 23 \cdot 22 \cdot 21}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \underline{\underline{53\,130}}.$$

For å trekke ut nøyaktig 3 røde kuler må man trekke 3 kuler fra de 10 røde og $5-3=2$ kuler fra de 15 blå. Antall måter dette kan gjøres på er

$$\binom{10}{3} \cdot \binom{15}{2} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} \cdot \frac{15 \cdot 14}{2 \cdot 1} = \underline{\underline{12\,600}}.$$

Oppgave 7

- a) Det er naturlig å sette konstantleddet lik null siden $x_i = 0$ tilsvarer at enten $d_i = 0$ eller $h_i = 0$, og dersom $d_i = 0$ eller $h_i = 0$ vil nødvendigvis også y_i være lik null. Dessuten ser man fra spredningsplottet at en linje som passer med de observerte verdiene vil gå gjennom eller svært nær punktet $(0, 0)$.

Fra spredningsplottet ser vi at variabiliteten til y_i vokser med økende x_i . Det kan se ut som om standardavviket til Y_i er proporsjonalt med verdien til x_i , noe man får ved å anta at $\text{Var}[Y_i]$ er proporsjonalt med x_i^2 .

- b) For å bestemme hvilken estimator som er best starter vi med å undersøke hvilke estimatorer som er forventningsrette. Det er oppgitt i oppgaveteksten at $E[\hat{\beta}] = \beta$, så vi trenger bare å regne ut $E[\tilde{\beta}]$ og $E[\beta^*]$,

$$\begin{aligned} E[\tilde{\beta}] &= E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n} E\left[\sum_{i=1}^n Y_i\right] = \frac{1}{n} \sum_{i=1}^n E[Y_i] = \frac{1}{n} \sum_{i=1}^n \beta x_i = \beta \bar{x}, \\ E[\beta^*] &= E\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{1}{\sum_{i=1}^n x_i^2} \cdot E\left[\sum_{i=1}^n x_i Y_i\right] = \frac{\sum_{i=1}^n E[x_i Y_i]}{\sum_{i=1}^n x_i^2} \\ &= \frac{\sum_{i=1}^n x_i E[Y_i]}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i \cdot \beta x_i}{\sum_{i=1}^n x_i^2} = \frac{\beta \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} = \beta. \end{aligned}$$

Vi ser at $\hat{\beta}$ og β^* er forventningsrette, mens $\tilde{\beta}$ er forventningsskjev. Vi foretrekker en forventningsrett estimator. For å bestemme hvilken av $\hat{\beta}$ og β^* som er best må vi sammenligne variansen til disse. Det er oppgitt at $\text{Var}[\hat{\beta}] = \frac{\sigma^2}{n}$, så vi trenger bare å regne ut $\text{Var}[\beta^*]$. I denne utregningen benytter vi at $Y_1, Y_2, \dots, Y_n$ er antatt å være uavhengige stokastiske variabler,

$$\begin{aligned} \text{Var}[\beta^*] &= \text{Var}\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] = \left(\frac{1}{\sum_{i=1}^n x_i^2}\right)^2 \text{Var}\left[\sum_{i=1}^n x_i Y_i\right] \\ &= \frac{\sum_{i=1}^n \text{Var}[x_i Y_i]}{(\sum_{i=1}^n x_i^2)^2} = \frac{\sum_{i=1}^n x_i^2 \text{Var}[Y_i]}{(\sum_{i=1}^n x_i^2)^2} \\ &= \frac{\sum_{i=1}^n x_i^2 \cdot x_i^2 \sigma^2}{(\sum_{i=1}^n x_i^2)^2} = \frac{\sigma^2 \sum_{i=1}^n x_i^4}{(\sum_{i=1}^n x_i^2)^2}. \end{aligned}$$

Vi må så avgjøre hvilken varians som er minst, $\text{Var}[\hat{\beta}]$ eller $\text{Var}[\beta^*]$. Den enkleste måten å gjøre dette på er å sette inn verdien for n de observerte verdiene for x_i i uttrykkene for de to variansene,

$$\begin{aligned} \text{Var}[\hat{\beta}] &= \frac{\sigma^2}{31} = 0.0323 \cdot \sigma^2, \\ \text{Var}[\beta^*] &= \frac{\sigma^2 \cdot 2605.785}{193.3769^2} = 0.0697 \sigma^2. \end{aligned}$$

Vi ser at variansen til $\hat{\beta}$ er minst, og siden vi foretrekker den (forventningsrette) estimatoren som har minst varians betyr dette at $\hat{\beta}$ er den beste estimatoren.

- c) Vi ser at $\hat{\beta}$ er en lineær funksjon av Y_i -ene, som er uavhengige og normalfordelt. Dermed vil også $\hat{\beta}$ være normalfordelt. Ved å benytte de oppgitte uttrykkene for $E[\hat{\beta}]$ og $\text{Var}[\hat{\beta}]$ får vi ved å standardisere at

$$Z = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1).$$

Siden verdien til σ^2 er ukjent erstatter vi denne i uttrykket for Z med estimatoren S^2 og kaller uttrykket vi da får, for T ,

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\frac{S^2}{n}}}$$

Vi observerer så at vi kan omskrive uttrykket for T på formen

$$T = \frac{\frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2} \cdot \frac{1}{n-1}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}}, \quad (1)$$

der vi har innført notasjonen

$$V = \frac{(n-1)S^2}{\sigma^2}.$$

Vi har fått oppgitt at $V \sim \chi_{n-1}^2$. Vi vet allerede at $Z \sim N(0, 1)$. Og dessuten har vi fått oppgitt at $\hat{\beta}$ og V er uavhengige stokastiske variabler, som medfører at også Z og V er uavhengige stokastiske variabler. Utrykket i (1) gir dermed at T er student t-fordelt med $n-1$ frihetsgrader. Ved vanlig notasjon for kvantiler i en t-fordeling har vi dermed at

$$P(-t_{\frac{\alpha}{2}, n-1} \leq T \leq t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha,$$

$$P\left(-t_{\frac{\alpha}{2}, n-1} \leq \frac{\hat{\beta} - \beta}{\sqrt{\frac{S^2}{n}}} \leq t_{\frac{\alpha}{2}, n-1}\right) = 1 - \alpha.$$

Løser så hver av ulikhetene inne i dette sannsynlighetsuttrykket med hensyn

på β . For den venstre ulikheten får vi

$$\begin{aligned} -t_{\frac{\alpha}{2}, n-1} &\leq \frac{\hat{\beta} - \beta}{\sqrt{\frac{S^2}{n}}} \\ -t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} &\leq \hat{\beta} - \beta \\ -\hat{\beta} + t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} &\leq -\beta \\ \hat{\beta} + t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} &\geq \beta \\ \beta &\leq \hat{\beta} + t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}}. \end{aligned}$$

For den høyre uliketen får vi tilsvarende

$$\begin{aligned} \frac{\hat{\beta} - \beta}{\sqrt{\frac{S^2}{n}}} &\leq t_{\frac{\alpha}{2}, n-1} \\ \hat{\beta} - \beta &\leq t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} \\ -\beta &\leq -\hat{\beta} + t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} \\ \beta &\geq \hat{\beta} - t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} \\ \hat{\beta} - t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} &\leq \beta. \end{aligned}$$

Dermed har vi at

$$P \left(\hat{\beta} - t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} \leq \beta \leq \hat{\beta} + t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{S^2}{n}} \right) = 1 - \alpha,$$

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

$$\left[\hat{\beta} - t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{s^2}{n}}, \hat{\beta} + t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{s^2}{n}} \right].$$

For å regne ut numerisk verdi for intervallet trenger vi å regne ut observert verdi for $\hat{\beta}$ og S^2 . Ved å benytte oppgitte summeverdier i oppgaveteksten får

vi at

$$\begin{aligned}\hat{\beta} &= \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i} = \frac{1}{31} \cdot 11.93827 = 0.3851055, \\ s^2 &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i - 2\hat{\beta}x_i}{x_i} \right)^2 = \frac{1}{n-1} \sum_{i=1}^n \frac{y_i^2 - 2\hat{\beta}x_i y_i + \hat{\beta}^2 x_i^2}{x_i^2} \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i^2}{x_i^2} - 2\hat{\beta} \cdot \frac{y_i}{x_i} + \hat{\beta}^2 \right) = \frac{1}{n-1} \left[\sum_{i=1}^n \left(\frac{y_i}{x_i} \right)^2 - 2\hat{\beta} \sum_{i=1}^n \frac{y_i}{x_i} + n\hat{\beta}^2 \right] \\ &= \frac{1}{31-1} [4.624633 - 2 \cdot 0.3851055 \cdot 11.93827 + 31 \cdot 0.3851055^2] \\ &= 0.0009046585.\end{aligned}$$

Når $\alpha = 0.05$ blir den aktuelle kvantilen i t-fordelinga $t_{\frac{\alpha}{2}, n-1} = t_{0.025, 30} = 2.042$. Konfidensintervallet blir dermed

$$\begin{aligned}&\left[0.3851055 - 2.042 \cdot \sqrt{\frac{0.0009046585}{31}}, 0.3851055 + 2.042 \cdot \sqrt{\frac{0.0009046585}{31}} \right] \\ &= \underline{\underline{[0.3740744, 0.3961366]}}.\end{aligned}$$

- d) Den nye observasjonen Y_0 er uavhengig av de gamle observasjonene $Y_1, Y_2, \dots, Y_n$. Dermed blir Y_0 uavhengig av $\hat{\beta}$, som er en funksjon av de gamle observasjonene. Vi har dermed at $Y_0 - \hat{\beta}x_0$ er en lineær funksjon av uavhengige og normalfordelte variabler, og er dermed selv normalfordelt. Dessuten har vi

$$\begin{aligned}E[Y_0 - \hat{\beta}x_0] &= E[Y_0] - E[\hat{\beta}]x_0 = \beta x_0 - \beta x_0 = 0, \\ \text{Var}[Y_0 - \hat{\beta}x_0] &= \text{Var}[Y_0] + \text{Var}[\hat{\beta}](-x_0)^2 = x_0^2 \sigma^2 + \frac{\sigma^2}{n} \cdot x_0^2 \\ &= x_0^2 \sigma^2 \left(1 + \frac{1}{n} \right),\end{aligned}$$

der vi i utregningen av variansen har beyttet at Y_0 og $\hat{\beta}$ er uavhengige stokastiske variabler. Ved å standardisere får vi dermed at

$$Z = \frac{Y_0 - \hat{\beta}x_0}{\sqrt{x_0^2 \sigma^2 \left(1 + \frac{1}{n} \right)}} \sim N(0, 1).$$

Siden verdien til σ^2 er ukjent erstatter vi σ^2 i dette uttrykket med estimatoren S^2 . Tilsvarende som i forrige deloppgave kan vi da vise at uttrykket vi får er

student t-fordelt med $n - 1$ frihetsgrader,

$$T = \frac{Y_0 - \hat{\beta}x_0}{\sqrt{x_0^2 \sigma^2 \left(1 + \frac{1}{n}\right)}} = \frac{\frac{Y_0 - \hat{\beta}x_0}{\sqrt{x_0^2 \sigma^2 \left(1 + \frac{1}{n}\right)}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}} \sim t_{n-1},$$

siden $Z \sim N(0, 1)$, $V \sim \chi_{n-1}^2$, og Z og V er uavhengige. Dermed har vi at

$$P(-t_{\frac{\alpha}{2}, n-1} \leq T \leq t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha$$

$$P\left(-t_{\frac{\alpha}{2}, n-1} \leq \frac{Y_0 - \hat{\beta}x_0}{\sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)}} \leq t_{\frac{\alpha}{2}, n-1}\right) = 1 - \alpha.$$

Løser så hver av ulikhetene inne i dette sannsynlighetsuttrykket med hensyn på Y_0 . For den venstre ulikheten får vi

$$-t_{\frac{\alpha}{2}, n-1} \leq \frac{Y_0 - \hat{\beta}x_0}{\sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)}}$$

$$-t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)} \leq Y_0 - \hat{\beta}x_0$$

$$\hat{\beta}x_0 - t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)} \leq Y_0.$$

For den venstre ulikheten får vi tilsvarende,

$$\frac{Y_0 - \hat{\beta}x_0}{\sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)}} \leq t_{\frac{\alpha}{2}, n-1}$$

$$Y_0 - \hat{\beta}x_0 \leq t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)}$$

$$Y_0 \leq \hat{\beta}x_0 + t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)}.$$

Dermed får vi at

$$P\left(\hat{\beta}x_0 - t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)} \leq Y_0 \leq \hat{\beta}x_0 + t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)}\right) = 1 - \alpha,$$

som betyr at et $(1 - \alpha) \cdot 100\%$ prediksjonsintervall for Y_0 er

$$\left[\hat{\beta}x_0 - t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)}, \hat{\beta}x_0 + t_{\frac{\alpha}{2}, n-1} \sqrt{x_0^2 S^2 \left(1 + \frac{1}{n}\right)} \right].$$

Oppgave 8

Punktsannsynligheten til X er

$$f_X(x; p) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \dots, n,$$

mens punktsannsynligheten til Y er

$$f_Y(y; p) = \binom{m}{y} \left(\frac{p}{2}\right)^y \left(1 - \frac{p}{2}\right)^{m-y}, y = 0, 1, \dots, m.$$

Siden X og Y er uavhengige stokastiske variabler blir rimelighetsfunksjonen dermed

$$L(p) = f_X(x; p) \cdot f_Y(y; p) = \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{y} \left(\frac{p}{2}\right)^y \left(1 - \frac{p}{2}\right)^{m-y}.$$

Log-rimelighetsfunksjonen blir dermed

$$\begin{aligned} \ell(p) &= \ln L(p) = \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p) + \ln \binom{m}{y} + y \ln \left(\frac{p}{2}\right) + (m-y) \ln \left(1 - \frac{p}{2}\right) \\ &= \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p) + \ln \binom{m}{y} + y \ln p - y \ln 2 + (m-y) \ln \left(1 - \frac{p}{2}\right). \end{aligned}$$

Den deriverte av $\ell(p)$ blir

$$\begin{aligned} \ell'(p) &= 0 + x \cdot \frac{1}{p} + (n-x) \cdot \frac{1}{1-p} \cdot (-1) + 0 + y \cdot \frac{1}{p} - 0 + (m-y) \cdot \frac{1}{1-\frac{p}{2}} \cdot \left(-\frac{1}{2}\right) \\ &= \frac{x}{p} - \frac{n-x}{1-p} + \frac{y}{p} - \frac{m-y}{2-p} = \frac{x+y}{p} - \frac{n-x}{1-p} - \frac{m-y}{2-p}. \end{aligned}$$

For å finne verdien til p som maksimerer $\ell(p)$ setter vi den deriverte lik null,

$$\begin{aligned} \ell'(p) = 0 &\Leftrightarrow \frac{x+y}{p} = \frac{n-x}{1-p} + \frac{m-y}{2-p} \\ (x+y)(1-p)(2-p) &= (n-x)p(2-p) + (m-y)p(1-p) \\ (x+y)(2-p-2p+p^2) &= (n-x)(2p-p^2) + (m-y)(p-p^2) \\ (x+y+n-x+m-y)p^2 &+ (-3x-3y-2n+2x-m+y)p + 2(x+y) = 0 \\ (n+m)p^2 - (2n+m+x+2y)p &+ 2(x+y) = 0. \end{aligned}$$

Setter vi inn de oppgitte tallene for n , m , x og y får vi følgende andregradsligning som må løses,

$$(30 + 40)p^2 - (2 \cdot 30 + 40 + 26 + 2 \cdot 19)p + 2(26 + 19) = 0$$

$$70p^2 - 164p + 90 = 0.$$

Dette gir at

$$p = \frac{-(-164) \pm \sqrt{(-164)^2 - 4 \cdot 70 \cdot 90}}{2 \cdot 70} = \frac{164 \pm 41.18252}{140},$$

som gir løsningene $p = 1.465589$ og $p = 0.8772677$. Siden vi må ha at $p \in [0, 1]$, er estimatet vi er interessert i

$$\underline{\underline{p = 0.8772677.}}$$

Eventuelt kan man sjekke andrederiverte for å se at $p = 0.8772677$ er et lokalt maksimumspunkt for $\ell(p)$, mens $p = 1.465589$ er et lokalt minimumspunkt. Strengt tatt bør man også overbevise seg om at $\ell(p)$, $p \in [0, 1]$ ikke tar sitt maksimum i et av randpunktene $p = 0$ eller $p = 1$.

Oppgave 9

a) Sentralgrenseteoremet gir at

$$\frac{\bar{X} - E[X_i]}{\sqrt{\frac{\text{Var}[X_i]}{n}}} = \frac{\bar{X} - \frac{2}{\lambda}}{\sqrt{\frac{2}{n\lambda^2}}}$$

er tilnærmet standard normalfordelt når n er stor. En mye brukt tommelfingerregel sier at approksimasjonen er god når $n \geq 30$. Siden Per og Pål bare har $n = 20$ kan det diskuteres hvor god approksimasjonen er, men man skal huske at X_i kan skrives som en sum av to uavhengige eksponensialfordelte variabler slik at $\bar{X}$ kan alternativt uttrykkes som gjennomsnitt over $2n = 40$ uavhengige eksponensialfordelte variabler.

Når H_0 er sann er $\lambda = \lambda_0$, og dermed får vi at

$$Z = \frac{\bar{X} - \frac{2}{\lambda_0}}{\sqrt{\frac{2}{n\lambda_0^2}}}$$

er tilnærmet standard normalfordelt.

For å finne approksimativ p-verdi må vi først finne observert verdi for test-observatoren,

$$z_{\text{obs}} = \frac{\bar{x} - \frac{2}{\lambda_0}}{\sqrt{\frac{2}{n\lambda_0^2}}} = \frac{\frac{109.34}{20} - \frac{2}{0.23}}{\sqrt{\frac{2}{n \cdot 0.23^2}}} = -2.35.$$

For å bestemme p-verdien må vi også avgjøre om vi skal forkaste H_0 når Z er stor eller når Z er liten. Siden $E[X_i] = \frac{2}{\lambda}$ vil en liten verdi for $\bar{X}$ indikere at H_1 er sann. Når $\bar{X}$ er liten blir også Z liten. Man skal derfor forkaste H_0 dersom $Z \leq k$, for en kritisk verdi k .

P-verdien blir dermed gitt ved

$$p = P(Z \leq z_{\text{obs}} | H_0 \text{ sann}) = \Phi(-2.3482) = \underline{\underline{0.0094}}.$$

Siden p-verdien er veldig liten er det rimelig å forkaste H_0 . Dette fordi dersom H_0 er sann er det svært lite sannsynlig å observere det vi har observert eller noe mer ekstremt, så dermed tror vi at H_0 er feil.

b) Som diskutert over har vi at p-verdien er gitt som

$$p = P(Z \leq z_{\text{obs}} | H_0 \text{ sann}).$$

Siden vi nå er interessert i den eksakte p-verdien kan vi ikke regne som om Z er standard normalfordelt. Vi må i stedet simulere mange realisasjoner av Z (under forutsetning av at H_0 er sann) og så bruke disse realisasjonene til å estimere sannsynligheten $P(Z \leq z_{\text{obs}} | H_0 \text{ sann})$.

For å simulere uavhengige realisasjoner $z_1, z_2, \dots, z_m$ av Z når H_0 er sann kan vi benytte oss av hvordan Z er definert. For å simulere en realisasjon z_k av Z kan vi først simulere uavhengige $x_1, x_2, \dots, x_n$ fra en gammafordeling med parametre $\alpha = 2$ og $\beta = \frac{1}{\lambda_0}$. Utfra disse kan vi så beregne tilhørende gjennomsnitt $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, og ut fra denne tilhørende verdi for z_k .

Vi har da verdier $z_1, z_2, \dots, z_m$, som er uavhengige realisasjoner fra (den eksakte) sannsynlighetsfordelingen til Z når H_0 er sann. Et estimat for p-verdien blir da

$$\hat{p} = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(z_i \leq z_{\text{obs}}),$$

der $\mathbb{I}(A)$ er indikatorfunksjonen som er lik 1 dersom A er sann og lik 0 dersom A ikke er sann.

I figur 2 er det en python-funksjon som returnerer estimert p-verdi. Funksjonen må kalles med $n=20$, $\lambda_0=0.23$ og $z_{\text{Obs}}=-2.3482$. Input-parameteren m må være et stort positivt heltall.

```
import numpy as np
from scipy.stats import gamma

def estimerPverdi(n, lambda0, zObs, m):

    alpha = 2
    beta = lambda0

    antallMindreEnnZobs = 0
    for i in np.arange(m):
        x = gamma.rvs(alpha, scale=1/beta, size=n)
        xbar = np.mean(x)
        zSim = (xbar - (2/lambda0))/np.sqrt(2/(n*(lambda0**2)))
        if zSim <= zObs:
            antallMindreEnnZobs = antallMindreEnnZobs + 1

    pHat = antallMindreEnnZobs / m

    return pHat
```

Figur 2: Python-funksjon som estimerer eksakt p-verdi.