NTNU
Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4240/TMA4245 Statistikk**

Faglig kontakt under eksamen: Geir-Arne Fuglstad, og Jarle Tufto

Tlf:

Eksamensdato: 4. august 2025

Eksamenstid (fra–til): 09:00–13:00

Hjelpemiddelkode/Tillatte hjelpemidler: C:

- Tabeller og formler 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: 8

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

a) Vi har at $Z = (X + Y - 1)/5$ er standardnormalfordelt gjennom vanlig standardisering. Dermed kan vi skrive

$$\mathrm{P}(X + Y > 1) = \mathrm{P}\left(\frac{X + Y - 1}{5} > \frac{1 - 1}{5}\right) = \mathrm{P}(Z > 0) = 0.5.$$

Dette følger enten direkte fra at standardnormalfordelingen er symmetrisk rundt 0 eller ved å slå opp i tabell.

Videre har vi

$$\begin{aligned}\mathrm{P}((X + Y)^2 > 1) = \mathrm{P}(|X + Y| > 1) &= \mathrm{P}(X + Y > 1) + \mathrm{P}(X + Y < -1) \\ &= 0.5 + \mathrm{P}(Z \leq \frac{-1 - 1}{5}) = 0.5 + \mathrm{P}(Z \leq -0.4) \\ &= 0.5 + 0.3446 \approx 0.84\end{aligned}$$

der vi har slått opp i tabell i formelsamlingen.

Oppgave 2

a) Standard regneregler gir

$$\mathrm{E}[2X + Y] = 2\mathrm{E}[X] + \mathrm{E}[Y] = 2 \cdot 1 + (-1) = 1$$

og

$$\begin{aligned}\mathrm{Var}[2X + Y] &= 2^2\mathrm{Var}[X] + \mathrm{Var}[Y] + 2 \cdot 2\mathrm{Cov}[X, Y] \\ &= 4 \cdot 4 + 3 + 4 \cdot (-2) \\ &= 11.\end{aligned}$$

Oppgave 3

a) Vi finner den første sannsynligheten direkte i tabellen i formelsamlingen,

$$\mathrm{P}(X \leq 7) = 0.7440 \approx 0.74.$$

Videre vet vi at $X + Y \sim \mathrm{Poisson}(4 + 6)$ siden $X \sim \mathrm{Poisson}(6)$ og $Y \sim \mathrm{Poisson}(4)$ er uavhengige. Dermed har vi fra tabell i formelsamlingen,

$$\mathrm{P}(X + Y \leq 7) = 0.2202 \approx 0.22.$$

Oppgave 4

a) Vi har

$$\begin{aligned}\mathrm{P}(X>5) &= \int_5^\infty f(x)\,\mathrm{d}x = \int_5^\infty \exp(-x/5)/5\,\mathrm{d}x \\ &= [-\exp(-x/5)]_{x=5}^{x=\infty} = -0+\exp(-1) \approx 0.37.\end{aligned}$$

Dette er en eksponentialfordeling som er minneløs. Vi har derfor

$$\mathrm{P}(X>10|X>5) = \mathrm{P}(X>10-5) = \exp(-1) \approx 0.37.$$

Oppgave 5

a) Fra $\mathrm{P}(A\cup B)\leq 1$ har vi

$$\begin{aligned}\mathrm{P}(A)+\mathrm{P}(B)-\mathrm{P}(A\cap B) &\leq 1 \\ 0.1 &\leq \mathrm{P}(A\cap B),\end{aligned}$$

og **hendelsene er ikke disjunkte**.

For $\mathrm{P}(A\cap B)=0.18$ er A og B uavhengige, men for $\mathrm{P}(A\cap B)=0.1$ er de avhengige. Begge deler er mulig under informasjonen gitt i oppgaven. Dermed **har vi ikke tilstrekkelig informasjon til å bestemme om hendelsene er uavhengige eller avhengige.**

Oppgave 6

a) Vi skal velge 1 av 3 målvakter og 10 av 17 utespillere. Vi får dermed

$$\binom{3}{1}\cdot\binom{17}{10} = 58344$$

ulike lag.

b) Vi skal velge 1 av 3 målvakter, 4 av 7 forsvarspillere, og 6 av 10 angrepsspillere. Vi får dermed

$$\binom{3}{1}\cdot\binom{7}{4}\cdot\binom{10}{6} = 22050$$

ulike lag.

Oppgave 7

a)

- Siden vi har et tilfeldig utvalg er rimelighetsfunksjonen

$$\begin{aligned} L(b; x_1, \dots, x_n) &= \prod_{i=1}^{n} g(x_i) = \prod_{i=1}^{n} \frac{1}{b^5 \Gamma(5)} x_i^4 \exp(-x_i/b) \\ &= \frac{1}{b^{5n}} \frac{1}{\Gamma(5)^n} \left(\prod_{i=1}^{n} x_i \right)^4 \exp\left(-\frac{1}{b} \sum_{i=1}^{n} x_i \right). \end{aligned}$$

  Dette gir en log-rimelighetsfunksjon

$$\begin{aligned} l(b; x_1, \dots, x_n) &= \log\left(L(b; x_1, \dots, x_n) \right) \\ &= -5n \log(b) - n \log(\Gamma(5)) + 4 \log\left(\prod_{i=1}^{n} x_i \right) - \frac{1}{b} \sum_{i=1}^{n} x_i. \end{aligned}$$

  Målet er nå å finne $\hat{b}$ som maksimerer log-rimelighetsfunksjonen. Vi deriverer

$$\frac{\mathrm{d}}{\mathrm{d}b} l(b; x_1, \dots, x_n) = -\frac{5n}{b} + \frac{1}{b^2} \sum_{i=1}^{n} x_i,$$

  og finner deretter kritisk punkt $\hat{b}$ der den deriverte er lik 0,

$$\begin{aligned} -\frac{5n}{\hat{b}} + \frac{1}{\hat{b}^2} \sum_{i=1}^{n} x_i &= 0 \\ \hat{b} &= \frac{1}{5n} \sum_{i=1}^{n} x_i. \end{aligned}$$

  Dette gir estimatoren

$$\hat{B} = \frac{1}{5n} \sum_{i=1}^{n} X_i = \bar{X}/5.$$

  NB: Man kunne sjekket at dette er et maksimum ved å se på den andrederiverte av log-rimelighetsfunksjonen, men det var ikke nødvendig å gjøre dette på eksamenen.

- Estimatet blir $\hat{b} = 13.049/5 = 2.6098$.
- Forventningsverdi til estimatoren er

$$\mathrm{E}[\hat{B}] = \frac{1}{5n} \sum_{i=1}^{n} \mathrm{E}[X_i] = \frac{1}{5n} \cdot n \cdot 5b = b$$

fordi $\mathrm{E}[X_i] = 5b$, $x = 1, \ldots, n$. Dette kan man slå opp i formelsamlingen under gammafordelingen. Tilsvarende blir variansen

$$\mathrm{Var}[\hat{B}] = \frac{1}{(5n)^2} \sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{1}{25n^2} \cdot n \cdot 5b^2 = \frac{b^2}{5n},$$

fordi $\mathrm{Var}[X_i] = 5b^2$, $x = 1, \ldots, n$. Dette kan man slå opp i formelsamlingen under gammafordelingen.

b)

- Basert på sentralgrenseteoremet vil $\bar{X}$ være omtrent normalfordelt. Det kan diskuteres om $n = 10$ er nok, men vi kan se på hver X_i som en sum av 5 uavhengige eksponentialfordelte stokastiske variabler. Basert på 8a får vi da at $\hat{B}$ er omtrent $\mathrm{N}(b, b^2/(5n))$.
- En mulig tilnærming er å bytte ut $b^2/(5n)$ med $\hat{b}^2/(5n)$ slik at $\hat{B}$ er omtrent $\mathrm{N}(b, \hat{b}^2/(5n))$. *Merk: Man kan også regne videre uten å tilnærme variansen videre, og vil da få et annet svar i neste kulepunkt.*
- Hvis vi baserer konfidensintervallet på pivotalen $Z = \frac{\hat{B}-b}{\hat{b}/\sqrt{5n}} \sim N(0,1)$ får vi da et omtrent 95% konfidensintervall for b gitt ved $\hat{b} \pm z_{0.025}\hat{b}/\sqrt{5n}$ der $z_{0.025} = 1.96$. Dette gir numerisk verdi 2.61 ± 0.72 eller alternativt $1.89 < b < 3.33$.

c)

- Fra formelsamling, har vi momementgenererende funksjoner

$$M_{X_i}(t) = (1 - bt)^{-5}, \quad i = 1, \ldots, n.$$

  La $S = \sum_{i=1}^{n} X_i$. Siden $X_1, \ldots, X_n$ er uavhengige har vi at momentgenererende funksjon for S er gitt ved

$$M_S(t) = \prod_{i=1}^{n} M_{X_i}(t) = ((1 - bt)^{-5})^n = (1 - bt)^{-5n}.$$

  Det vil is at S er gammafordelt med parametere $a = 5n$ og b.
- Den momentgenerende funksjonen til $10n\hat{B}/b = (2/b)S$ er gitt ved

$$M_{(2/b)S}(t) = M_S(2t/b) = (1 - 2t)^{-5n}.$$

  Denne momentgenererende funksjonen kjenner vi igjen (fra formelsamling) som en kjikvadratfordeling med $2 \cdot 5n = 10n$ frihetsgrader.

d)

- Vi har $\frac{10n}{b}\hat{B} \sim \chi^2_{10n}$. Med utgangspunkt i dette kan vi skrive

$$\mathrm{P}\left(\chi^2_{0.975,10n} < \frac{10n}{b}\hat{B} < \chi^2_{0.025,10n}\right) = 0.95,$$

der $\chi_{0.025,10n}$ og $\chi_{0.975,10n}$ angir kritiske verdier i en kjikvadratfordeling med $10n$ frihetsgrader. Vi kan manipulere uttrykket slik at

$$\mathrm{P}\left(\frac{\chi^2_{0.975,10n}}{10n\hat{B}} < \frac{1}{b} < \frac{\chi^2_{0.025,10n}}{10n\hat{B}}\right) = 0.95.$$

Vi inverterer så hvert ledd i ulikheten samtidig som vi passer på å snu ulikhetene

$$\mathrm{P}\left(\frac{10n\hat{B}}{\chi^2_{0.025,10n}} < b < \frac{10n\hat{B}}{\chi^2_{0.975,10n}}\right) = 0.95.$$

Vi har dermed at

$$\frac{10n\hat{b}}{\chi^2_{0.025,10n}} < b < \frac{10n\hat{b}}{\chi^2_{0.975,10n}}$$

er et eksakt 95% konfidensintervall.

- Med tallene i oppgaven får vi

$$\frac{10n\hat{b}}{\chi^2_{0.025,10n}} = \frac{10 \cdot 10 \cdot 2.6098}{129.561} \approx 2.014$$

og

$$\frac{10n\hat{b}}{\chi^2_{0.025,10n}} = \frac{10 \cdot 10 \cdot 2.6098}{74.222} \approx 3.516.$$

Dette gir 95% konfidensintervallet $2.01 < b < 3.52$.

- Vi ser at dette konfidensintervallet er noe forskjøvet i forhold til det tilnærmede intervallet $1.89 < b < 3.33$. Dette kan tyde på at n var for liten til å bruke sentralgrenseteoremet og/eller at n var for liten til å sette inn et estimat for variansen. Men vi kan ikke med sikkerhet vurdere om intervallet i 8b har for lav eller for høy konfidens uten videre utregninger eller simuleringer.

Oppgave 8

a)
- Vi bruker definisjonen av den kumulative fordelingsfunksjonen og ser på en $x \geq 1$,

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{1}^{x} 3t^{-4}\,\mathrm{d}t$$
$$= 3\left[\frac{t^{-3}}{-3}\right]_{t=1}^{t=x} = -(x^{-3} - 1).$$

Dermed har vi

$$F_X(x) = \begin{cases} 0, & x < 1, \\ 1 - x^{-3}, & x \geq 1. \end{cases}$$

- For å vise at Y har samme fordeling som X kan vi vise at de har samme kumulativ fordeling. Vi finner (for en $y \geq 1$)

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) = \mathrm{P}((1-U)^{-1/3} \leq y) \\ &= \mathrm{P}((1-U)^{-1} \leq y^3) = \mathrm{P}(1-U \geq y^{-3}) = \mathrm{P}(1-y^{-3} \geq U) \\ &= \mathrm{P}(U \leq 1-y^{-3}) = 1-y^{-3}. \end{aligned}$$

  Det er klart at $F_Y(y) = 0$ for $y < 1$ siden det er umulig å oppnå. Vi har dermed at F_Y og F_X er like som funksjoner, og at Y har samme fordeling som X.

- Vi simulerer først n uniformt fordelte variabler og transformerer de:

```python
import numpy as np
def simX(n):
    u = np.random.uniform(size = n)
    y = (1-u)**(-1/3)
    return(y)
```

Oppgave 9

a)

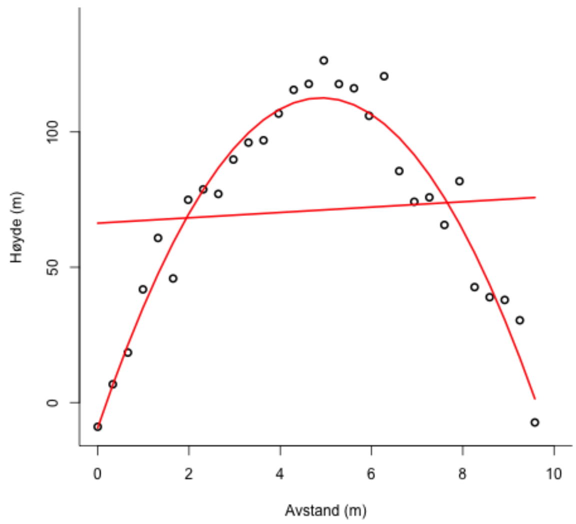
- Vi bruker formeler fra formelsamlingen til å regne ut

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{242.104}{245.210} \approx 0.98733 \approx 0.99, \\ \hat{\alpha}_1 &= \bar{y} - \hat{\beta}_1 \bar{x} = 71.0258 - 0.98733 \cdot 4.7895 \approx 66.30, \\ \hat{\beta}_2 &= \frac{\sum_{i=1}^n (z_i^2 - \bar{z^2}) y_i}{\sum_{i=1}^n (z_i^2 - \bar{z^2})^2} = \frac{-8178.659}{1610.039} \approx -5.07979 \approx -5.08, \\ \hat{\alpha}_2 &= \bar{y} - \bar{z^2} \hat{\beta}_1 \bar{x} = 71.0258 - (-5.07979) \cdot 8.1859 \approx 112.61. \end{aligned}$$

  Dette gir regresjonslinjene

$$\begin{aligned} &\text{Modell 1:} \quad \mathrm{E}[Y|x] = 66.30 + 0.99x \\ &\text{Modell 2:} \quad \mathrm{E}[Y|x] = 112.61 - 5.08(x-4.9)^2 \end{aligned}$$

- Basert på uttrykkene for regresjonslinjene kan vi lage en skisse lignende den vist i figur 1.



Figur 1: Skisse for oppgave 9a.

- Fra figur 1 virker spredningenen rundt regresjonslinjen fra modell 2 å være omtrent uavhengige med samme varians. Det er vanskelig å vurdere normalfordeling, men det er ingen åpenbare tegn på at det ikke er en normalfordeling (slik som skjevhet eller enkelte ekstreme verdier). Dermed virker modell 2 å oppfylle betingelsene i modelformuleringen.

  Siden modell 2 virker å oppfylle betingelsene virker det som om betingelsen som er brutt i modell 1 er at residualene ikke har forventningsverdi lik 0.

b)

- Variansparameteren i modell 2, σ_2^2, beskriver usikkerheten i målingene. Siden produsenten mener at $\sigma_2^2 = 5^2 = 25$, og vi ønsker å vise at $\sigma_2^2 > 25$ er rimelige nullhypotese og alternativ hypotese

$$H_0 : \sigma_2^2 = 25,$$
$$H_1 : \sigma_2^2 > 25.$$

- Vi tar utgangspunkt i testobservatoren $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha}_2 - \hat{\beta}_2 z_i^2)^2$. Denne testobservatoren tilfredstiller under H_0, $\frac{n-2}{25}S^2 \sim \chi^2_{n-2}$. Store verdier av S^2 er konsistent med vår alternative hypotese, og vi må dermed bestemme en k slik at

$$\mathrm{P}\left(\frac{n-2}{25}S^2 \geq k | H_0\right) = 0.05.$$

Vi finner fra tabell den kritiske verdien $k = \chi^2_{0.05,n-2} = \chi^2_{0.05,28} = 41.337$, og forkastningsregelen blir: “Forkast hvis $s^2 \geq 25\frac{41.337}{28} \approx 36.91$”.

- Vi regner ut

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha}_1 - \hat{\beta}_1(x_i - 4.9)^2)^2 = \frac{1}{28} \cdot 2615.707 \approx 93.41.$$

Vi ser at $s^2 \geq 36.91$, forkaster H_0, og konkluderer at usikkerheten i utstyret har større standardavvik enn 5 meter.