NTNU
Kunnskap for en bedre verden

Institutt for matematiske fag

# Eksamensoppgave i **TMA4245 Statistikk**

**Faglig kontakt under eksamen:**
**Tlf:**

**Eksamensdato:**
**Eksamenstid (fra–til):**
**Hjelpemiddelkode/Tillatte hjelpemidler:** A: Alle hjelpemidler tillatt.

**Annen informasjon:**

**Målform/språk:** bokmål
**Antall sider:** 4
**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)** Medianen til datasett A er litt høyere enn medianen til datasett B. Siden A er symmetrisk så vil gjennomsnittet være ganske likt medianen. B er ikke symmetrisk. Det er ikke lett å si noe om varians men det vi ser er at A har en større variasjonsbredde ('range') enn B.

**b)** I plott A er de to stokastiske variablene X og Y positivt korrelert, som betyr at det er en lineær sammenheng mellom X og Y. I plott B er X og Y ukorrelerte og ser også ut til å være uavhengige. I plott C er X og Y ukorrelerte, men det er tydelig at X og Y ikke er uavhengige.

## Oppgave 2

**a)** Vi kan definere hendelsene:

- $M_0$ Ingen skårer mål (0-0)
- $M_1$ United skårer første mål
- $M_2$ motstander skårer første mål

Fra teksten vi vet at $P(M_1) = 0.3$, $P(M_0) = 0.05$, og fra dette kan vi regne ut

$$\begin{aligned} P(M_2) &= 1 - P(M_1) - P(M_0) \\ &= 1 - 0.3 - 0.05 = 0.65. \end{aligned}$$

Vi vet også at $P(\text{vinn}|M_1) = 0.6$ og $P(\text{vinn}|M_2) = 0.1$. I tillegg er det åpenbart at $P(\text{vinn}|M_0) = 0$. Ved å bruke setningen om total sannsynlighet kan vi regne ut

$$\begin{aligned} P(\text{vinn}) &= P(\text{vinn}|M_1)P(M_1) + P(\text{vinn}|M_2)P(M_2) + P(\text{vinn}|M_0)P(M_0) \\ &= 0.6 \cdot 0.3 + 0.1 \cdot 0.65 + 0 \cdot 0.05 = 0.245 \end{aligned}$$

**b)** Riktig svar er $(A \cap C) \cup (C \cap (A \cup B)')$

**c)** Riktig svar er 'Sann kun dersom A og B er like sannsynlige'

## Oppgave 3

**a)** Vi kan regne $P(X = 0)$ ved $P(X = 0) = \sum_{y=0}^{3} P(X = 0, Y = y) = 0.1 + 0.05 + 0.2 + 0.1 = 0.45$

**b)** Vi har at

$$P(Y \leq 1 | X = 0) = \frac{P(Y \leq 1, X = 0)}{P(X = 0)} = \frac{0.1 + 0.05}{0.45} = 0.333$$

**c)** Nei $X$ og $Y$ er ikke uavhengige. Vi kan for eksempel se at $P(X = 0|Y = 0) = \frac{0.1}{0.15} = 0.667 \neq P(X = 0)$.

**d)** Vi bruker formelen $E(g(X,Y)) = \sum_x \sum_y g(x,y)P(X = x, Y = y)$. Siden $g(x,y) = 0$ når $x = 0$ kan vi regne ut

$$\begin{aligned} E(g(X,Y)) &= \sum_{y=0}^{3}(1 + 2y)P(X = 1, Y = y) \\ &= 1 \cdot 0.05 + (1 + 2) \cdot 0.15 + (1 + 4) \cdot 0.3 + (1 + 6) \cdot 0.05 = 2.35 \end{aligned}$$

## Oppgave 4

**a)** Forventningsverdien til en eksponensialfordeling med parameter $\lambda = 0.8$ er $1/\lambda = 1.25$

**b)** Vi har at

$$P(X > 1) = 1 - F(x) = 1 - \int_0^x f(x)dx = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x} = e^{-0.8} = 0.449$$

**c)** Siden komponentene er koblet i serie så er levetiden til maskinen lik levetiden til den komponenten som først slutter å fungere, dvs $\min\{X_1, \ldots, X_5\}$. Videre er minimum av 5 eksponensialfordelte variabler med samme parameter $\lambda$ eksponensialfordelt med parameter $5 \cdot \lambda$. Dermed er $P(\min(X_1, \ldots, X_5) > 1) = e^{-5 \cdot 0.8 \cdot 1} = 0.018$.

**d)** Eksponensialfordelingen har en kjent momentgenererende funksjon, og det har også gammafordelingen. Videre vet vi at MGF-en til en sum av uavhengige identisk fordelte variabler er gitt ved produktet av MGF-ene. I dette tilfellet blir MGF til summen av de 5 eksponensialfordelte variablene $1/(1 - t/0.8)^5$. Nå kan vi kjenne igjen MGF-en til en gammafordeling hvis vi setter parameteren $\alpha = 5$ og parameteren $\beta = 1/0.8 = 1.25$.

**Oppgave 5**

**a)** Det er ikke rimelig å anta at $X$ og $Y$ er uavhengige stokastiske variabler. Blodtrykksmålingene $X$ og $Y$ er gjort på samme person (før og etter treningsopplegg) og må antas avhengige. Det vil for eksempel være rimelig å tro at blodtrykksnivået før treningsopplegget ($X$) gir en viss indikasjon på hva blodtrykket etter treningsopplegget ($Y$) vil være.

**b)** $E(D) = E(Y - X) = E(Y) - E(Y) = \mu_y - \mu_x$ uansett om $X$ og $Y$ er uavhengige eller ikke. Basert på observasjonene i tabellen er $\bar{d} = -6.88$ estimatet vårt for $E(D)$ og $s_d = 10.757$ er estimatet for standardavviket.

**c)** Vi velger testobservatoren $T = (\bar{D} - 0)/(S/\sqrt{10})$, der $\bar{D}$ er gjennomsnittet av utvalget og $S$ er estimatoren for standardavviket til D. Når nullhypotesen er sann er testobservatoren $t$-fordelt med 9 frihetsgrader. Ved signifikansnivå 0.01 skal vi forkaste nullhypotesen dersom vi observerer $t < -1.383$. Observert verdi av testobservatoren finner vi ved å sette inn estimat på forventningsverdi og standardavvik fra forrige del-oppgave, slik at $t = -2.022$. Dermed forkaster vi nullhypotesen. Dr. Hansen har grunnlag for å påstå at treningsopplegget virker.

**d)** En type-1 feil går ut på å forkaste nullhypotesen, og dermed påstå at treningsopplegget 'virker', til tross for at treningsopplegget egentlig ikke hadde noen blodtrykksdempende effekt. En type-2 feil går ut på å ikke forkaste nullhypotesen selv om treningsopplegget faktisk har en blodtrykksdempende effekt.

**e)** Riktig svar er 'Sannsynligheten for det vi har observert, eller noe mer ekstremt i retning av den alternative hypotesen, når vi antar at nullhypotesen er sann.'

**Oppgave 6**

**a)** Ved å plotte residualene man kan sjekke antagelsene om at feil-leddet i regresjonsmodellen (altså $\epsilon$) har forventning 0 og samme varians ($\sigma^2$) for alle verdier av $x$. Residualene fra modell A ser ut til å tilfredstille disse antagelsene. Residualene i modell B er er også sentrert rundt 0 men vi ser at variansen øker for høyere verdier av $x$.

**b)** Vi vet at i en enkel regresjonsmodell er

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}.$$

Dermed er et 95% konfidensintervall for $\beta$ gitt ved

$$\hat{\beta} \pm t_{13,0.025} \sqrt{\frac{s^2}{\sum_{i=1}^{15}(x_i - \bar{x})^2}}$$

hvor $t_{13,0.025} = -2.16$.

Når vi setter inn tall finner vi at konfidensintervallet er:

$$1.109 \pm 2.16 \cdot \sqrt{\frac{0.762}{18.028}} = [0.665, 1.553]$$