

i **Forside**

Institutt for matematiske fag

Eksamensoppgave i TMA4245 Statistikk

Eksamensdato: 27.05.2024

Eksamenstid (fra-til): 9:00 - 13:00

Hjelpemiddelkode: C

Tillatte hjelpemidler: Tabeller og formler i statistikk, Akademika (markeringer med tekstmarkør i formelheftet er tillatt); Bestemt, enkel kalkulator; Gult stemplet A5-ark med håndskrevne notater.

Faglig kontakt under eksamen:

Ingeborg Hem Sørmoen tlf: 99 45 46 94

Thea Bjørnland, tlf: 41 12 38 49

Faglig kontakt møter i eksamenslokalet: NEI

ANNEN INFORMASJON:

Les oppgavene nøye, gjør dine egne antagelser og presiser i besvarelsen hvilke forutsetninger du har lagt til grunn i tolkning/avgrensing av oppgaven. Faglig kontaktperson kontaktes kun dersom det er direkte feil eller mangler i oppgavesettet. Henvend deg til en eksamensvakt hvis du mistenker feil og mangler. Noter spørsmålet ditt på forhånd.

Håndtegninger: I oppgavene **4, 5, 6, 7** er det lagt opp til å besvare på ark. Andre oppgaver skal besvares direkte i Inspira. Nederst i oppgaven finner du en sjusifret kode. Fyll inn denne koden øverst til venstre på arkene du ønsker å levere. Det anbefales å gjøre dette underveis i eksamen. Dersom du behøver tilgang til kodene etter at eksamenstiden har utløpt, må du klikke «Vis besvarelse».

Vekting av oppgavene: *I oppgave 1-3 teller hver deloppgave (f.eks. 1a) 5 poeng. I oppgave 4-7 teller hver deloppgave (f.eks. 4 eller 5a) 10 poeng.*

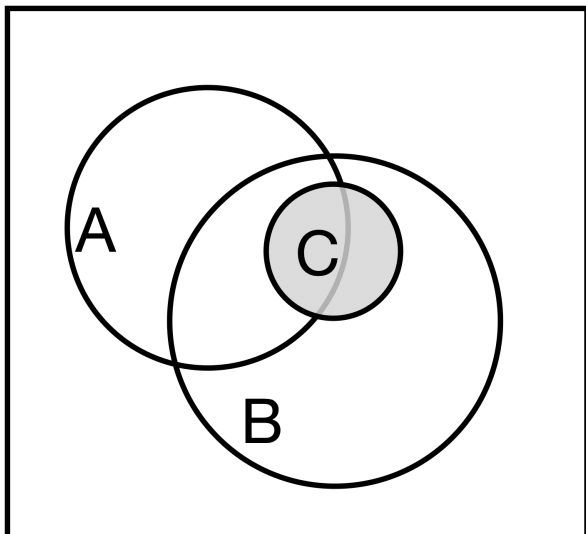
Varslinger: Hvis det oppstår behov for å gi beskjeder til kandidatene underveis i eksamen (f.eks. ved feil i oppgavesettet), vil dette bli gjort via varslinger i Inspira. Et varsel vil dukke opp som en dialogboks på skjermen. Du kan finne igjen varselet ved å klikke på bjella øverst til høyre.

Trekk fra/avbrutt eksamen: Blir du syk under eksamen, eller av andre grunner ønsker å levere blankt/avbryte eksamen, gå til "hamburgermenyen" i øvre høyre hjørne og velg «Lever blankt». Dette kan ikke angres selv om prøven fremdeles er åpen.

Tilgang til besvarelse: Etter eksamen finner du besvarelsen din i arkivet i Inspira. Merk at det kan ta én virkedag før eventuelle håndtegninger vil være tilgjengelige i arkivet.

1 Hendelser og sannsynlighet

a) Det grå området i venndiagrammet viser hendelsen C. Hvilket av uttrykkene under er også en riktig beskrivelse av det grå området?



Velg ett alternativ:

- ☐ $B \cup C$
- ☐ $A \cap C$
- ☐ C'
- ☐ $(A \cap B)'$
- ☐ $(B \cap C)'$
- ☐ $B \cap C$

b) Dersom $P(A) = 0.25$, $P(B) = 0.70$ og $P(A \cup B) = 0.88$, hva er $P(A'|B)$?

Velg ett alternativ

- ☐ 0.90
- ☐ 0.55
- ☐ 0.23
- ☐ 0.75
- ☐ 0.60
- ☐ 0.81

Maks poeng: 10

2 Simultanfordeling

Tabellen viser simultanfordelingen til to stokastiske variabler X og Y , $P(X = x, Y = y)$:

	$y = 2$	$y = 3$	$y = 4$
$x = -1$	0.10	0.10	0.05
$x = 0$	0.02	0.00	0.08
$x = 1$	0.09	0.15	0.06
$x = 2$	0.07	0.07	0.21

Fra tabellen kan vi f.eks. lese av at $P(X = 0, Y = 4) = 0.08$ og $P(X = 2, Y = 2) = 0.07$.

Angi svarene med to desimaler etter komma.

a) Hva blir $P(Y = 2)$?

b) Hva blir $E(X)$?

Maks poeng: 10

3 Poissonfordeling

a) La X og Y representere uavhengige Poissonfordelte stokastiske variabler med forventning $E(X) = 3$ og $E(Y) = 2$. Hva er $P(X+Y \leq 5)$?

Velg ett alternativ:

- ☐ 0.7290
- ☐ 0.4071
- ☐ 0.3174
- ☐ 0.5238
- ☐ 0.6160
- ☐ 0.2450

b) En Poissonprosess i tid med intensitet (rate) λ skal observeres i 3 ikke-overlappende tidsintervaller av lengde t_1 , t_2 og t_3 . La X_1 , X_2 og X_3 representere antall hendelser i hvert tidsintervall. Hvilket uttrykk representerer *ikke* en forventningsrett estimator for λ ?

Velg ett alternativ

- ☐ $\frac{X_1}{t_1}$
- ☐ $\frac{\sum_{i=1}^3 X_i}{\sum_{i=1}^3 t_i}$
- ☐ $0.5 \cdot \left(\frac{X_2}{t_2} + \frac{X_3}{t_3} \right)$
- ☐ $\sum_{i=1}^3 \frac{X_i}{t_i}$
- ☐ $\frac{X_1+X_2}{t_1+t_2}$
- ☐ $\frac{X_1}{t_1} + \frac{X_2}{t_2} - \frac{X_3}{t_3}$

Maks poeng: 10

4 Feilforplantning

Darcy-Weisbachs falltapslikning kan brukes til å beregne trykkfall som følge av friksjon langs en gitt lengde av et rør når det fører en væskestrøm. I et sylinderformet rør vil trykkfallet Δp langs en lengde l være gitt ved

$$\Delta p = l \cdot f_d \cdot \frac{\rho}{2} \cdot \frac{v^2}{d}$$

der l er lengden (i m), f_d er en kjent koeffisient kalt Darcy-Weisbachs friksjonsfaktor, ρ er tettheten av fluidet som strømmer i røret (kg/m^3), d er den indre diameteren av røret (målt i m) og v er den gjennomsnittlige hastigheten til fluidstrømmen (målt i m/s).

Anta at l , f_d og ρ er kjente størrelser. Hastigheten v er ukjent og målinger av denne kan representeres med den stokastiske variabelen V , der $V \sim N(v, \sigma_V^2)$. Diameteren d er også ukjent og målinger, D , av denne er normalfordelte med forventning d og varians σ_D^2 . Videre er de stokastiske variablene V og D uavhengige. Det tilhørende målte trykkfallet blir dermed en realisasjon av den stokastiske variabelen

$$X = l \cdot f_d \cdot \frac{\rho}{2} \cdot \frac{V^2}{D}.$$

(i) Forventningsverdien til målingen X kan approksimeres med $E(X) \approx l \cdot f_d \cdot \frac{\rho}{2} \cdot \frac{v^2}{d}$. Bruk feilforplantningsformelen til å finne tilnærmet verdi av $\text{Var}(X)$. Oppgi svaret som en funksjon av v , d , σ_V og σ_D .

(ii) Python-funksjonen funX under kan brukes for å approksimere $E(X)$ og $\text{Var}(X)$ med stokastisk simulering. Fyll inn de manglende feltene `[*]` og `[**]`.

```
import numpy as np

def funX(n, l, f_d, rho, v, d, sigma_V, sigma_D):
    simV = np.random.normal(v, sigma_V, size = n)
    simD = [*]
    simX = l*f_d*(rho/2)*(simV**2)/simD
    mux = [**]
    varx = np.sum((simX-mux)**2)/(n-1)
    return([mux, varx])
```

Maks poeng: 10

5 Binomisk fordeling

La $X_1, X_2, \dots, X_n$ representere n uavhengige Bernoulli-fordelte stokastiske variabler med suksess-sannsynlighet p , det vil si $X_i \sim \text{Binom}(1, p)$ for $i = 1, \dots, n$.

a) Bruk regneregler for momentgenererende funksjoner til å vise at summen $Y = \sum_{i=1}^n X_i$ er binomisk fordelt med parametre n og p , det vil si $Y \sim \text{Binom}(n, p)$. Bruk videre den momentgenererende funksjonen til Y til å vise at $E(Y) = np$.

b) La $Y \sim \text{Binom}(n, p)$ og $W \sim \text{Binom}(m, 1 - p)$ være to uavhengige binomisk fordelte stokastiske variabler. Med utgangspunkt i realisasjonene y av Y og w av W , utled et uttrykk for sannsynlighetsmaksimeringsestimatoren for parameteren p .

Maks poeng: 10

6 Velferd i lakseoppdrett

Lakseprodusenten SalFisk AS ønsker å undersøke om en avlusingsbehandling påvirker velferd hos laks. SalFisk AS vil gjennomføre et forsøk der 15 laksemerder behandles med samme metode og under like omstendigheter. En egnet numerisk skadeindikator skal måles i hver av de 15 merdene før og etter behandling. En *økning* i denne indikatoren kan indikere dårligere velferd som følge av behandling, men målingene er forbundet med usikkerhet.

La de uavhengige stokastiske variablene $D_1, D_2, \dots, D_{15}$ representere differansen i skadeindikator fra før til etter behandling.

a) Vi antar først at differansene $D_1, D_2, \dots, D_{15}$ er uavhengige og normalfordelte med forventning μ og standardavvik σ . SalFisk AS vil med utgangspunkt i denne vurderingen teste

$$H_0 : \mu = 0 \text{ mot } H_1 : \mu > 0.$$

For å gjennomføre testen skal de bruke testobservatoren

$$T = \frac{\bar{D}}{\sqrt{S^2/15}}$$

$$\text{der } \bar{D} = \frac{1}{15} \sum_{i=1}^{15} D_i \text{ og } S^2 = \frac{1}{14} \sum_{i=1}^{15} (D_i - \bar{D})^2.$$

(i) Skisser fordelingen til testobservatoren under antagelsen om at H_0 er sann. Inkluder testens forkastningsområde, ved signifikansnivå $\alpha = 0.05$, i skissen.

(ii) Etter forsøket finner SalFisk AS at $\sum_{i=1}^{15} d_i = 8.55$ og $\sum_{i=1}^{15} (d_i - \bar{d})^2 = 17.64$. Hva blir konklusjonen av testen?

(iii) I en lignende eller samme skisse som over, inkluder resultatet av testen og skisser tilhørende p -verdi (du behøver ikke regne ut et tall for p -verdien).

b) SalFisk AS er usikre på normalantagelsen, og vil i stedet gjennomføre en fortegnstest for å undersøke hypotesene

$$H_0 : \tilde{\mu} = 0 \text{ mot } H_1 : \tilde{\mu} > 0,$$

der $\tilde{\mu}$ er medianen til hver av de 15 målingene $D_1, D_2, \dots, D_{15}$. De ønsker å gjennomføre testen ved signifikansnivå $\alpha = 0.05$.

(i) Skriv ned en egnet testobservator og beslutningsregel for fortegnstesten.

(ii) Hva blir det faktiske signifikansnivået til testen, og hvorfor kan vi ikke oppnå et signifikansnivå nøyaktig lik 0.05 i slike tester?

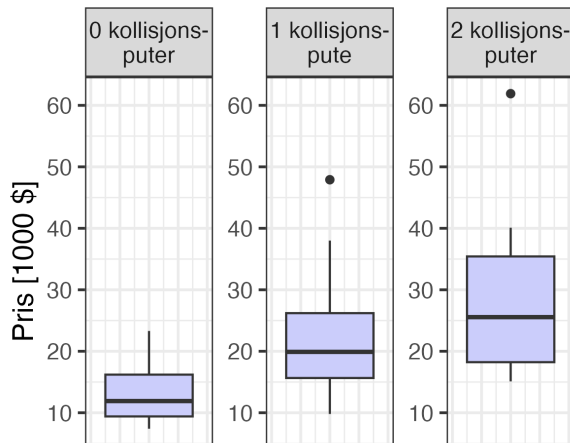
(iii) SalFisk AS observerte at skadeindikatoren økte i 10 av 15 merder. Hva blir resultatet av testen?

Maks poeng: 10

7 93CARS

Det åpent tilgjengelige datasettet 93CARS inneholder informasjon om pris, bilmerke, modell, etc. for 93 tilfeldige valgte biler som ble produsert og solgt i USA i år 1993. Bilmodellene i datasettet har enten 0, 1 (bare til fører), eller 2 (fører og passasjer i framsete) kollisjonsputer. I denne oppgaven skal vi studere sammenhengen mellom pris (i 1000 US dollar, 1000 \$) og antall kollisjonsputer.

a)



Figuren over viser boksplott av salgspris for henholdsvis 34 biler med 0 kollisjonsputer, 43 biler med 1 kollisjonspute, og 16 biler med 2 kollisjonsputer.

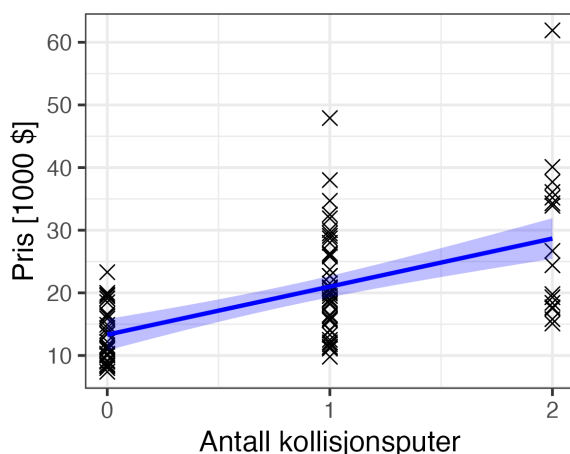
(i) Ta utgangspunkt i boksplottet for salgspris for biler med 0 kollisjonsputer. Forklar hva plottet viser. Bruk begrepene median, 0.25-kvantil og 0.75-kvantil i forklaringen din. Forklaringen skal knyttes opp mot de aktuelle dataene.

(ii) Ved å sammenligne de tre boksplottene, ser det ut til å være forskjell i pris mellom de tre ulike kategoriene av biler? [Drøft observasjonene med maksimalt 5 setninger.]

b) Vi skal videre studere en lineær regresjonsmodell på formen

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

der $\varepsilon_i, i = 1, \dots, 93$, er uavhengige og normalfordelte støy-ledd med forventning 0 og varians σ^2 . Kovariaten x_i tar verdiene 0, 1 eller 2 og representerer antall kollisjonsputer. Responsvariabelen Y_i representerer pris (i 1000 \$).



Figuren over viser observasjoner av pris for ulikt antall kollisjonsputer, samt en estimert enkel lineær regresjonslinje med 95%-konfidensintervall. Det oppgis at gjennomsnittlig pris i 1000\$ for de 34 bilene uten kollisjonsputer er 13.17, gjennomsnittlig pris i 1000\$ for de 43 bilene med kun én kollisjonspute er 21.22 og gjennomsnittlig pris i 1000\$ for de 16 bilene med to kollisjonsputer er 28.37.

- (i) Regn ut estimer for skjæringspunktet og stigningstallet til regresjonslinja.
- (ii) Hva blir predikert pris for en bil med 1 kollisjonspute (i 1993)? Bruk svaret ditt fra (i) eller figuren til å finne prediksjonen. Hvorfor, og hvordan, vil et 95% prediksjonsintervall være annerledes enn intervallet som er skissert i figuren over?
- (iii) Kommenter om modellantagelsene i enkel lineær regresjon ser ut til å stemme for dataene gitt i figuren.

Kilde: Lock, R. H. (1993) 1993 New Car Data. *Journal of Statistics Education* 1(1).
doi:10.1080/10691898.1993.11910459

Maks poeng: 10