

i Eksamensoppgave i TMA4245 Statistikk**Dato:** 28.05.2025**Tid:** 09.00-13.00**Faglig kontakt:** Håkon Tjelmeland, Geir-Arne Fuglstad**Møter i eksamenslokalet:** Ja, cirka 10.45-11.30**Hjelpemiddelkode/Tillatte hjelpemidler:** C

- Tabeller og formler i statistikk (Akademika).
- Bestemt enkel kalkulator i henhold til NTNUs generelle reglement.
- Ett A5-ark med egne formler og notater.

ANNEN INFORMASJON:

Les oppgavene nøye og gjør dine egne antagelser. Presiser i besvarelsen hvilke forutsetninger du har lagt til grunn i tolkning/avgrensing av oppgaven.

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

Dersom det er direkte feil eller mangler i oppgavesettet og du ikke kan gjøre dine egne antagelser, vises det til informasjon om klage på formelle feil på NTNU sin nettside om «Begrunnelse og klage»

FAGSPESIFIKK INFORMASJON

Papirhåndtegninger: I oppgavene 7, 8 og 9 er det lagt opp til å besvare på ark. Andre oppgaver skal kun 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».

Du er selv ansvarlig for å fylle inn riktige koder på eventuelle håndtegningsark. Les derfor informasjonen på omslagsarket nøye. Eksamenskontoret kan ikke garantere at feilaktig utfylte ark blir lagt til besvarelsen din.

Vekting av oppgavene: I oppgave 1 til 6 teller hver oppgave 5 poeng. I oppgave 7, 8 og 9 teller hver av deloppgavene (7a, 7b, 7c, 7d, 8, 9a og 9b) 10 poeng.

Trekk fra/avbrutt eksamen:

Dersom du 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 under tidligere prøver i Inspira. Merk at det kan ta én virkedag før eventuelle håndtegninger vil være tilgjengelige under «tidligere prøver».

- 1 **Innledning:** La X være en normalfordelt stokastisk variabel med forventningsverdi lik **2** og varians lik **9**.

Oppgave: Finn følgende sannsynligheter. Angi svarene med to desimaler etter komma. Eventuell mellomregning bør gjøres med fire desimaler etter komma.

$$P(X > 3) : \boxed{}$$

$$P(X > 2 | X < 3) : \boxed{}$$

Maks poeng: 5

- 2 **Innledning:** La X og Y være to uavhengige og normalfordelte stokastiske variabler. La X ha forventningsverdi lik **1** og varians lik **9**, og la Y ha forventningsverdi lik **2** og varians lik **4**.

Oppgave: Finn følgende størrelser. Angi svarene med to desimaler etter komma. Eventuell mellomregning bør gjøres med fire desimaler etter komma.

$$P(X + Y > 0) : \boxed{}$$

$$\text{Var}[2X - 4Y] : \boxed{}$$

Maks poeng: 5

- 3 Innledning:** La X og Y være to diskret fordelte stokastiske variabler, der $X, Y \in \{0, 1, 2\}$. Simultan punktsannsynlighet for X og Y , $f(x, y)$, er som angitt i følgende tabell.

$x \backslash y$	0	1	2
0	0.12	0.20	0.00
1	0.10	0.12	0.20
2	0.04	0.10	0.12

Oppgave: Finn følgende sannsynligheter. Angi svarene med to desimaler etter komma. Eventuell mellomregning bør gjøres med fire desimaler etter komma.

$$P(X = 0) : \boxed{}$$

$$P(X \leq 1 | Y \leq 1) : \boxed{}$$

Maks poeng: 5

- 4 Innledning:** La A , B og C være tre hendelser i et utfallsrom S . Det oppgis at hendelsene A og C er disjunkte, og at hendelsene B og C er disjunkte. Dessuten oppgis det at $P(A) = 0.30$, $P(B) = 0.25$, $P(C) = 0.40$ og $P(A \cap B) = 0.15$.

Oppgave: Finn følgende sannsynligheter. Angi svarene med to desimaler etter komma. Eventuell mellomregning bør gjøres med fire desimaler etter komma.

$$P(A \cup B) : \boxed{}$$

$$P(A|C') : \boxed{}$$

Maks poeng: 5

- 5 **Innledning:** Betrakt følgende skall til en pythonfunksjon som skal generere en realisasjon fra en binomisk fordeling med parametre n og p .

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

```
def simBinom(n,p):
    # her skal det legges til noen linjer som genererer en verdi x

    return x
```

Nedenfor er det fem forslag til kodelinjer som kan benyttes til å generere x .

Oppgave: Marker hvilke (et eller flere) av alternativene som korrekt genererer en realisasjon fra en binomisk fordeling med parametre n og p .

Velg ett eller flere alternativer

☐ $u = np.random.normal(0,p,n)$
 $x = np.sum(u * u)$

☐ $u = np.random.uniform(0,1,n)$
 $v = -np.log(1 - u)$
 $success = (v \leq p)$
 $x = np.sum(success)$

☐ $u = np.random.uniform(0,1,n)$
 $success = (u \leq p)$
 $x = np.sum(success)$

☐ $u = np.random.uniform(0,1,2*n)$
 $success = (u \leq p)$
 $x = np.sum(success)/2$

☐ $u = np.random.uniform(0,1)$
 $values = np.arange(n+1)$ # lager vektoren $(0,1,2,\dots,n)$
 $F = binom.cdf(values, n, p)$
 $x = np.sum(F \leq u)$

Maks poeng: 5

- 6 Innledning:** Anta at vi har en urne med 25 kuler. Av disse kulene er 10 kuler røde, mens resten er blå. Fra urna trekker vi 5 kuler uten tilbakelegging. Anta at vi trekker kulene på en slik måte at alle kulene har like stor sannsynlighet for å bli trukket ut. Vi er så interessert i hvilke kuler som er trukket ut.

Oppgave: Finn følgende antall. Oppgi svarene som heltall.

- På hvor mange mulige måter kan man trekke ut de 5 kulene?

- På hvor mange gunstige måter kan man trekke ut nøyaktig 3 røde kuler?

Maks poeng: 5

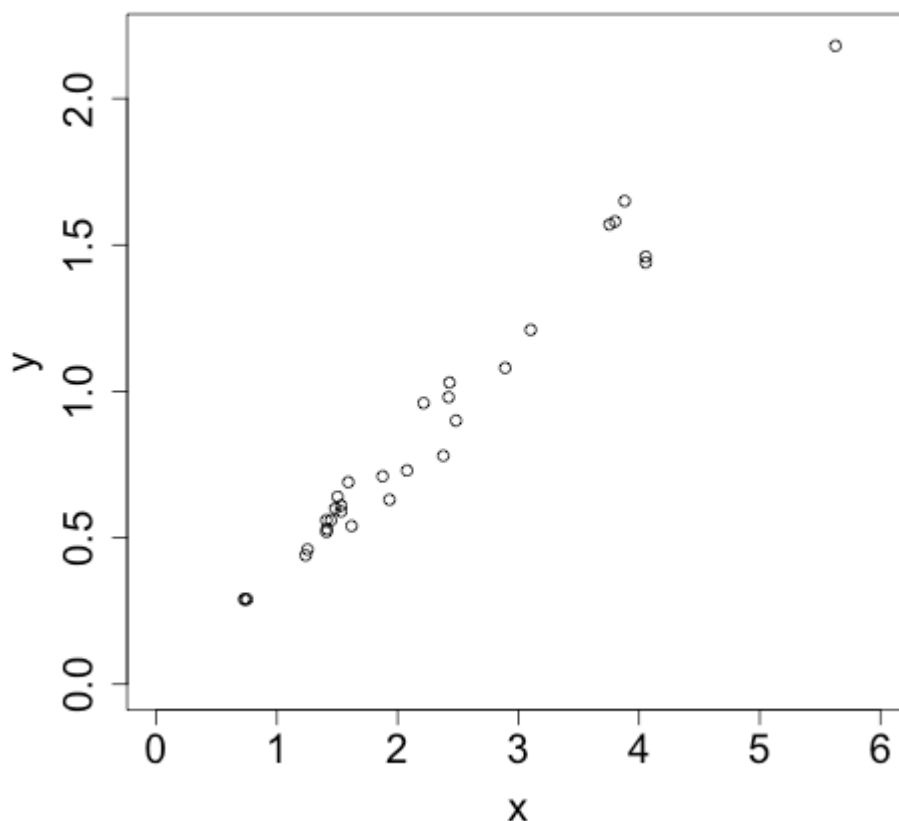
- 7 I denne oppgaven skal vi analysere et datasett hvor man for $n = 31$ trær har målt høyden (i meter) av treet, diameteren (i meter) til stammen 1.37 meter over bakken, og volumet (i kubikkmeter) av treet som kan brukes til plankeproduksjon. For $i = 1, 2, \dots, n$ lar vi h_i , d_i og y_i betegne henholdsvis målt høyde, diameter og volum som kan brukes til plankeproduksjon for tre nummer i . Vi lar videre $x_i = \pi d_i^2 h_i / 4$ være volumet av en sylinder med samme diameter og høyde som tre nummer i . I vedlagte spredningsplott (se under) er x_i plottet langs x-aksen og y_i langs y-aksen.

Vi modellerer dataene ved en regresjonsmodell, der vi betrakter observert verdi y_i som en realisasjon av en stokastisk variabel Y_i og antar at

$$Y_i = \beta x_i + x_i \varepsilon_i \text{ for } i = 1, 2, \dots, n,$$

og antar at $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ er uavhengige og normalfordelte stokastiske variabler med forventningsverdi lik 0 og varians lik σ^2 . Merk at dette betyr at $Y_1, Y_2, \dots, Y_n$ er uavhengige stokastiske variabler, og at Y_i er normalfordelt med forventningsverdi lik βx_i og varians lik $x_i^2 \sigma^2$. Modellen har altså to parametre, β og σ^2 , og verdien til begge disse er ukjente.

Det oppgis at de observerte verdiene for høyde, diameter og volum gir $\sum_{i=1}^n x_i = 68.6439$, $\sum_{i=1}^n x_i^2 = 193.3769$, $\sum_{i=1}^n x_i^4 = 2605.785$, $\sum_{i=1}^n \frac{y_i}{x_i} = 11.93827$ og $\sum_{i=1}^n \left(\frac{y_i}{x_i} \right)^2 = 4.624633$.



- 7(a) Oppgave:** Diskuter hvorfor det ut fra situasjonen vi betrakter her og ut fra hva vi ser i spredningsplottet kan være mer naturlig å benytte regresjonsmodellen beskrevet over, heller enn å benytte en vanlig regresjonsmodell som også inneholder et konstantledd og hvor variansen til Y_i er lik for alle $i = 1, 2, \dots, n$.

Maks poeng: 10

- 7(b) Innledning:** For å estimere parameteren β er følgende tre estimatorene foreslått.

$$\tilde{\beta} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \hat{\beta} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i} \quad \text{og} \quad \beta^* = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Det oppgis at $E[\hat{\beta}] = \beta$ og $\text{Var}[\hat{\beta}] = \frac{\sigma^2}{n}$.

Oppgave: Hvilken av de tre estimatorene er best? Husk at svaret skal begrunnes.

Maks poeng: 10

- 7(c) Innledning:** Uansett hvilken av de tre estimatorene for β du fant å være den beste, skal du videre i oppgaven ta utgangspunkt i estimatoren $\hat{\beta}$. Videre i oppgaven kan du uten bevis benytte at en forventningsrett estimator for σ^2 er

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{Y_i - \hat{\beta} x_i}{x_i} \right)^2.$$

Det oppgis at $\frac{(n-1)S^2}{\sigma^2}$ er kjikvadratfordelt med $n - 1$ frihetsgrader, og dessuten at $\frac{(n-1)S^2}{\sigma^2}$ og $\hat{\beta}$ er uavhengige stokastiske variabler.

Oppgave: Utled et $(1 - \alpha) \cdot 100\%$ konfidensintervall for β . Finn først en formel for konfidensintervallet, og regn deretter ut intervallet numerisk når $\alpha = 0.05$.

Maks poeng: 10

- 7(d) Innledning:** Betrakt nå et nytt tre, der man har målt høyde og diameter og regnet ut tilhørende verdi x_0 , men har foreløpig ikke målt volumet, Y_0 . Det er nå av interesse å predikere Y_0 basert på de observerte dataene.

Oppgave: Utled, ved å ta utgangspunkt i $Y_0 - \hat{\beta} x_0$, en formel for et $(1 - \alpha) \cdot 100\%$ prediksjonsintervall for Y_0 .

Maks poeng: 10

- 8 Innledning:** La X og Y være to uavhengige stokastiske variabler. Anta at X er binomisk fordelt med parametre n og p , mens Y er binomisk fordelt med parametre m og $\frac{p}{2}$.

Anta at vi kjenner verdiene til n og m , men verdien til parameteren $p \in [0, 1]$ er ukjent.

Oppgave: Benytt sannsynlighetsmaksimeringsprinsippet til å finne et estimat for p når $n = 30$, $m = 40$, og man har observert $x = 26$ og $y = 19$. (Merk at du ikke trenger å finne en generell formel for sannsynlighetsmaksimeringsestimatoren, det er tilstrekkelig at du finner estimatet for p for de oppgitte verdiene for n , m , x og y).

Maks poeng: 10

- 9 En landevei går forbi huset der brødrepåret Per og Pål bor. Da brødrepåret gikk på barneskolen brukte de mye tid på å telle biler som kjørte forbi huset, og fant da at i tidsperioden 14-15 på hverdager passerte det i gjennomsnitt 0.23 biler per minutt.

I dag studerer Per og Pål på NTNU og har nylig lært om blant annet poissonprosesser og hypotesetesting. De bestemmer seg for å gjøre en liten undersøkelse for å finne ut om det er grunnlag for å påstå at trafikken forbi huset i tidsperioden 14-15 på hverdager er større i dag enn da de gikk på barneskolen. De antar at i den aktuelle timen på hverdager passerer biler forbi huset der de bor ifølge en poissonprosess med parameter λ . De er ikke like tålmodige med å telle biler i dag som da de var mindre, men bestemmer seg for at de på $n = 20$ hverdager skal ta tiden fra de setter seg ned og begynner å observere landeveien til den andre bilen passerer. De lar X_i betegne den observerte tiden (i minutter) på dag nummer i . Fra egenskapene til en poissonprosess vet de da at X_i er gammafordelt med parametre $\alpha = 2$ og $\beta = \frac{1}{\lambda}$, slik at sannsynlighetstettheten til X_i er gitt som

$$f(x_i; \lambda) = \lambda^2 x_i e^{-\lambda x_i}, \quad x_i \geq 0,$$

og at $E[X_i] = \frac{2}{\lambda}$ og $\text{Var}[X_i] = \frac{2}{\lambda^2}$. Per og Pål formulerer så problemet sitt som en hypotesetest,

$$H_0 : \lambda = \lambda_0 \quad \text{mot} \quad H_1 : \lambda > \lambda_0,$$

der $\lambda_0 = 0.23$. Som testobservator velger de å benytte

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

som de konkluderer med er tilnærmet standard normalfordelt når H_0 er sann.

Ut fra sine observerte verdier $x_1, x_2, \dots, x_n$, regner Per og Pål ut at $\sum_{i=1}^n x_i = 109.34$ minutter.

- 9(a) **Oppgave:** Diskuter en grunn for at Per og Pål konkluderer med at Z er tilnærmet standard normalfordelt når H_0 er sann. Diskuter kort hvor god du tenker denne approksimasjonen er i situasjonen Per og Pål har.

Finn approksimativ p-verdi for hypotesetesten. Diskuter kort om den utregnede p-verdien gir grunnlag for å forkaste H_0 .

Maks poeng: 10

9(b) Innledning: I oppgave (a) regnet du ut en p-verdi basert på en approksimasjon av fordelingen til testobservatoren Z når H_0 er sann. Verdien du regnet ut er dermed bare en approksimasjon til den eksakte p-verdien for testen.

Oppgave: Forklar hvordan du ved hjelp av stokastisk simulering kan formulere en forventningsrett estimator for den eksakte p-verdien, og hvordan du kan beregne det tilhørende estimatet. Du bør innføre notasjon slik at du kan gi en presis beskrivelse.

Skriv pythonkode som utfører prosedyren med å estimere den eksakte p-verdien.

Maks poeng: 10