

Institutt for matematiske fag

Eksamensoppgave i TMA4240 Statistikk

Faglig kontakt under eksamen: Håkon Tjelmeland^a, Sara Martino^b

Tlf: ^a 48 22 18 96 , ^b 99 40 33 30

Faglig kontakt møter i eksamenslokalet:

Eksamensdato: Fredag 22. desember 2023

Eksamenstid (fra–til): 09.00–13.00

Hjelpemiddelkode/Tillatte hjelpemidler: C: *Tabeller og formler i statistikk*, Akademika; Bestemt, enkel kalkulator; Gult stemplet A5-ark med egne håndskrevne notater.

Annen informasjon:

Faglig kontakt møter i eksamenslokalet: Ja, en gang i tidsrommet 10.30–12.00

Skaff deg overblikk over oppgavesettet før du begynner på besvarelsen din.

Alle svar skal begrunnes og besvarelsen skal inneholde naturlig mellomregning slik at det er helt klart hva som er gjort.

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: Alle oppgavene skal besvares på ark. Du får utdelt en sjusifret kode. Fyll inn denne koden øverst til venstre på arkene du ønsker å levere. Det anbefales å gjøre dette underveis i eksamen.

Vekting av oppgavene: Oppgavesettet består av ti deloppgaver, oppgave 1, 2, 3a, 3b, 3c, 3d, 4a, 4b, 4c og 4d. Alle de ti deloppgavene gis lik vekt ved sensur.

Varslinger: Hvis det oppstår behov for å gi beskjeder til kandidatene underveis i eksamen vil dette bli gitt muntlig.

Trekk fra/avbrutt eksamen: Blir du syk under eksamen, eller av andre grunner ønsker å levere blankt/avbryte eksamen, ta kontakt med eksamensvaktene.

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.

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

Oppgave 1

Anta at X og Y er to uavhengige og normalfordelte stokastiske variabler. Anta at X har forventningsverdi lik 0 og standardavvik lik 1, mens Y har forventningsverdi lik 1 og standardavvik lik 3

Finn sannsynlighetene

$$P(X \geq 2), \quad P(13X - 2Y \leq 12) \quad \text{og} \quad P(Y \geq 4 | X \geq 2).$$

Oppgave 2

I figur 1 er det gitt en python-funksjon med to input-parametre, et positivt heltall n og en parameter phi (φ) som må være større enn null. Funksjonen genererer n uavhengige realisasjoner av en kontinuerlig stokastisk variabel Y . Ut fra denne koden, bestem sannsynlighetstettheten til Y , dvs $f_Y(y)$ for $-\infty < y < \infty$.

La en annen kontinuerlig stokastisk variabel X ha kumulativ fordelingsfunksjon gitt ved

$$F_X(x) = \begin{cases} 1 - \exp\left\{-\frac{x^2}{\theta}\right\} & \text{dersom } x \geq 0, \\ 0 & \text{ellers,} \end{cases}$$

der $\theta > 0$ er en parameter. Utled hvordan man ved å ta utgangspunkt i en uniformfordelt variabel på intervallet $[0, 1]$ kan generere en realisasjon av X . Skriv så en python-funksjon, `simX`, som genererer n uavhenige realisasjoner av X .

```
import numpy as np

def simY(n, phi):
    u = np.random.uniform(size=n)
    y = phi * np.sqrt(u)

    return y
```

Figur 1: Python-funksjon som genererer n uavhengige realisasjoner av en kontinuerlig stokastisk variabel Y , der phi (φ) er en parameter i sannsynlighetsfordelingen til Y .

Oppgave 3

Levetiden (målt i måneder) til noen typer mekaniske komponenter har vist seg å følge en fordeling med kumulativ fordelingsfunksjon

$$F_X(x) = \begin{cases} 1 - \exp\left\{-\frac{x^2}{\theta}\right\} & \text{dersom } x \geq 0, \\ 0 & \text{ellers,} \end{cases}$$

der $\theta > 0$ er en parameter som beskriver kvaliteten til komponentene.

a) Regn ut følgende to sannsynligheter når $\theta = 100$.

- Hva er sannsynligheten for at en slik komponent fremdeles fungerer etter 12 måneder?
- Hvis vi vet at en slik komponent fremdeles fungerer etter 12 måneder, hva er sannsynligheten for at den også fungerer etter 15 måneder?

Finn et uttrykk for sannsynlighetstettheten $f_X(x)$ og bruk denne til å vise at

$$E[X^2] = \theta.$$

For å undersøke om kvaliteten til en ny type av slike komponenter er bedre enn den typen som hittil har vært brukt prøver man ut $n = 50$ av de nye komponentene. Vi lar $x_1, x_2, \dots, x_n$ betegne de observerte levetidene for de nye komponentene og betrakter disse som realisasjoner av tilhørende stokastiske variabler $X_1, X_2, \dots, X_n$. Vi antar at $X_1, X_2, \dots, X_n$ er et tilfeldig utvalg fra fordelingen gitt over.

Videre i oppgaven kan du uten bevis benytte at sannsynlighetsmaksimerings-estimatoren for θ basert på $X_1, X_2, \dots, X_n$ er

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i^2$$

og at $\text{Var}[X_i^2] = \theta^2$. Husk dessuten at vi fra deloppgave **a)** har at $E[X_i^2] = \theta$.

b) Angi antagelsene og resultatet i sentralgrenseteoremet og beskriv den generelle praktiske konsekvensen av dette resultatet.

Forklar så hvorfor sentralgrenseteoremet i situasjonen vi betrakter i denne oppgaven gir at

$$\hat{\theta}$$

er tilnærmet normalfordelt med forventningsverdi lik θ og varians lik θ^2/n .

Det er kjent at den type komponenter som har vært i bruk til nå har kvalitetsparameter $\theta = \theta_0 = 100$. Vi ønsker altså nå å benytte observerte verdier $x_1, x_2, \dots, x_n$ til å avgjøre om det er grunnlag for å påstå at den nye typen komponenter har bedre kvalitet enn den typen som har vært i bruk til nå. De observerte levetidene er:

7.17	16.95	7.57	3.06	8.67	9.02	8.62	13.53	13.88	5.00
11.44	17.88	11.79	9.06	14.26	11.39	12.88	9.28	9.26	6.68
11.26	3.65	8.24	24.79	2.78	13.61	8.53	20.63	9.14	7.60
8.27	8.93	1.67	1.76	12.85	7.71	9.30	12.01	9.13	19.07
7.81	3.95	8.41	15.75	14.21	10.66	4.33	8.18	5.53	11.05

Det oppgis at dette gir $\sum_{i=1}^n x_i^2 = 6084.853$.

- c) Formuler situasjonen som en hypotesetest ved å spesifisere null- og alternativ hypotese. Ved å benytte resultatet i deloppgave **b**), angi en testobservator du kan benytte og hvilken (tilnærmet) sannsynlighetsfordeling denne testobservatoren har når H_0 er sann. Finn (tilnærmet) p-verdi når observerte verdier er som angitt over.

Diskuter kort om den utregnede p-verdien gir grunnlag for å forkaste H_0 .

Merk at du i oppgave 2 ble bedt om å skrive en python-funksjon for å simulere realisasjoner av X . Uansett om du fikk til å skrive en slik funksjon eller ikke, kan du i neste deloppgave forutsette at du har en slik python-funksjon tilgjengelig.

I oppgave **c**) regnet du ut en p-verdi basert på en approksimasjon av sannsynlighetsfordelingen til testobservatoren når H_0 er sann. Det utregnede verdien er dermed bare en approksimasjon til den eksakte p-verdien for den angitte testen.

- d) Forklar hvordan du ved hjelp av stokastisk simulering kan formulere en forventningsrett estimator for den eksakte p-verdien, og hvordan du kan beregne et tilhørende estimat. Du bør innføre notasjon slik at du kan gi en presis formulering av prosedyren.

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

Nummer	1	2	3	4	5	6	7	8	9	10	11
Diameter	0.21	0.22	0.22	0.27	0.27	0.27	0.28	0.28	0.28	0.28	0.29
Høyde	21.3	19.8	19.2	21.9	24.7	25.3	20.1	22.9	24.4	22.9	24.1
Volum	0.29	0.29	0.29	0.46	0.53	0.56	0.44	0.52	0.64	0.56	0.69
Nummer	12	13	14	15	16	17	18	19	20	21	22
Diameter	0.29	0.29	0.30	0.30	0.33	0.33	0.34	0.35	0.35	0.36	0.36
Høyde	23.2	23.2	21.0	22.9	22.6	25.9	26.2	21.6	19.5	23.8	24.4
Volum	0.59	0.61	0.60	0.54	0.63	0.96	0.78	0.73	0.71	0.98	0.90
Nummer	23	24	25	26	27	28	29	30	31		
Diameter	0.37	0.41	0.41	0.44	0.44	0.45	0.46	0.46	0.52		
Høyde	22.6	21.9	23.5	24.7	25.0	24.4	24.4	24.4	26.5		
Volum	1.03	1.08	1.21	1.57	1.58	1.65	1.46	1.44	2.18		

Tabell 1: Målinger på 31 trær av arten romhegg. Verdiene er hentet fra *Ryan, T.A., Joiner, B.L. and Ryan, B.L. (1976). Minitab student handbook, Duxbury Press.*

Oppgave 4

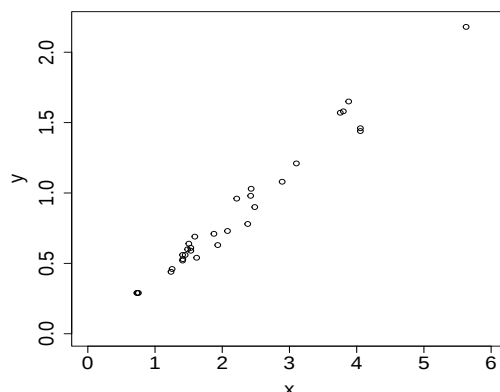
For å undersøke om man ut fra målt diameter og høyde på et tre kunne beregne hvor stort volum av treet som kunne utnyttes til å lage planker, har man for 31 trær av arten romhegg målt henholdsvis volum (m^3) som kunne bli brukt til plankeproduksjon, diameter (m) på treet målt 1.37 meter over bakken og høyde (m). De målte verdiene er gitt i Tabell 1. For $i = 1, 2, \dots, 31$ lar vi d_i , h_i og y_i være henholdsvis målt diameter, høyde og volum for tre nummer i . Vi lar videre $x_i = \pi d_i^2 h / 4$ være volumet av en sylinder med diameter lik d_i og høyde h_i . Figur 2 viser et spredningsplott hvor x_i er plottet langs x-aksen og y_i langs y-aksen. Det oppgis at de observerte verdiene for diameter og høyde gir $\sum_{i=1}^n x_i^2 = 193.3769$ og $\sum_{i=1}^n 1/x_i^2 = 13.5248$.

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

$$Y_i = \beta x_i + \varepsilon_i,$$

der residualene $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ er uavhengige og normalfordelte stokastiske variabler med forventningsverdi lik null og varians lik σ^2 . Legg merke til at vi velger ikke å inkludere noe konstantledd i regresjonsmodellen og at vi som vanlig i regresjonsmodeller velger å betrakte x_i 'ene som gitte tall.

Den antatte modellen har to parametre, β og σ^2 , og vi skal anta at verdien til begge disse er ukjent og skal estimeres. For å estimere parameteren β er det foreslått



Figur 2: Spredningsplott for observerte verdier $(x_i, y_i); i = 1, 2, \dots, 31$.

følgende tre estimatorer,

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

a) Hvilken av de tre estimatorene er best? Husk at svaret skal begrunnes!

I neste deloppgave skal du utlede formler for sannsynlighetsmaksimeringsestimatorene for β og σ^2 , og disse betegner vi med henholdsvis $\hat{\beta}$ og $\hat{\sigma}^2$.

b) Utled sannsynlighetsmaksimeringsestimatorene $\hat{\beta}$ og $\hat{\sigma}^2$.

Vis spesielt at $\hat{\beta}$ er identisk med en av de tre estimatorene foreslått over og at $\hat{\sigma}^2$ kan skrives på formen

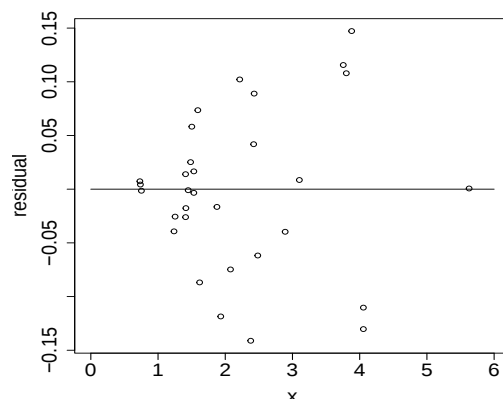
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta} x_i)^2.$$

c) Begrunn at

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

er kjikvadratfordelt med n frihetsgrader. Angi spesielt hvilke kjente egenskaper du benytter og hvordan du anvender disse.

Merk at deloppgaven fortsetter på neste side.



Figur 3: Residualplott for den antatte regresjonsmodellen i oppgave 4 når man benytter $\hat{\beta}$ som estimator for β .

Gi et kort argument for at man fra dette får at

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

er kjikvadratfordelt med $n - 1$ frihetsgrader. Merk at det her er tilstrekkelig at du gir et argument for resultatet, du skal ikke gi et bevis.

Benytt så dette siste resultatet til å bevise at $\hat{\sigma}^2$ er en forventingsskjev estimator for σ^2 . Foreslå en forventningsrett estimator for σ^2 og vis at estimatoren du foreslår faktisk er forventningsrett.

Som kjent kan man bruke et residualplott til å vurdere om en antatt regresjonsmodell ser ut til å passe med et observert datasett. Figur 3 viser residualplottet for datasettet og regresjonsmodellen gitt over når man bruker $\hat{\beta}$ som estimator for β .

- d) Forklar hvordan et residualplott konstrueres for regresjonsmodellen gitt over. Angi spesielt hva som plottes langs x-aksen og hva som plottes langs y-aksen. Diskuter hva du tolker ut fra residualplottet i figur 3. Hvis din konklusjon er at den antatte regresjonsmodellen ikke passer med det observerte datasettet, foreslå en alternativ regresjonsmodell som du mener vil passe bedre. Forklar spesielt hvorfor du mener at modellen du foreslår vil passe bedre.