NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4240 Statistikk**

Fagleg kontakt under eksamen: Håkon Tjelmeland[a], Sara Martino[b]

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

Fagleg kontakt kjem til eksamenslokalet:

Eksamensdato: Fredag 22. desember 2023

Eksamenstid (frå–til): 09.00-13.00

Hjelpemiddelkode/Tillatne hjelpemiddel: C: *Tabeller og formler i statistikk*, Akademika; Bestemt, enkel kalkulator; Gult stempla A5-ark med eigne handskrevne notater.

Annan informasjon:

Fagleg kontakt kjem til eksamenslokalet: Ja, ein gong i tidsrommet 10.30-12.00

Skaff deg eit overblikk over oppgåvesettet før du byrjar å svare på oppgåvene.

Alle svar skal grunngjevast og utrekninga skal innehalde naturleg mellomrekning slik at det er heilt klart kva som er gjort.

Les oppgåvene nøye, gjer deg opp dine eigne meiningar og presiser i svara dine kva for føresetnadar du har lagt til grunn i tolking/avgrensing av oppgåva. Fagleg kontaktperson kontaktast berre dersom du meiner det er direkte feil eller manglar i oppgåvesettet. Vend deg til ei eksamensvakt om du meiner det er feil eller manglar. Noter spørsmålet ditt på førehand.

Handteikningar: Alle oppgåvene skal du svara på ark. Du får utdelt ein sjusifra kode. Fyll inn denne koden øvst til venstre på dei arka du ynskjer å levere. Det er tilrådd å gjere dette undervegs i eksamen.

Vekting av oppgåvene: Oppgåvesettet inneheld ti deloppgåver, oppgåve 1, 2, 3a, 3b, 3c, 3d, 4a, 4b, 4c og 4d. Alle dei ti deloppgåvene gis lik vekt ved sensur.

Varslingar: Dersom det oppstår behov for å gje beskjedar til kandidatane medan eksamen er i gang vil dette bli gjort munnleg.

Trekk frå/avbroten eksamen: Blir du sjuk under eksamen, eller av andre grunnar ynskjer å levere blankt/avbryte eksamen, ta kontakt med eksamensvaktene.

Tilgang til svara dine: Etter eksamen finn du svara dine i arkivet i Inspera. Merk at det kan ta ein vyrkedag før eventuelle handteikningar vert tilgjengelege i arkivet.

Oppgåve 1

Anta at X og Y er to uavhengige og normalfordelte variablar. Anta at X har forventingsverdi lik 0 og standardavvik lik 1, medan Y har forventingsverdi lik 1 og standardavvik lik 3.

Finn sannsyna

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

Oppgåve 2

I figur 1 er det gjeve ein python-funksjon med to input parametrar, eit positivt heiltall n og ein parameter phi (φ) som må være større enn null. Funksjonen genererer n uavhengige realisasjonar av ein kontinuerleg stokastisk variabel Y. Ut frå denne koden, bestem sannsynstettleiken til Y, dvs $f_Y(y)$ for $-\infty < y < \infty$.

La ein annan kontinuerleg stokastisk variabel X ha kumulativ fordelingsfunksjon gjeve ved

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

der $\theta > 0$ er ein parameter. Utlei korleis man ved å ta utgongspunkt i ein uniformfordelt variabel på intervallet $[0, 1]$ kan generere ein realisasjon av X. Skriv så ein python-funksjon, simX, som genererer n uavhengige realisasjonar 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: Phyton-funksjon som genererer n uavhengige realisasjonar av ein kontinuerleg stokastisk variabel Y, der phi (φ) er ein parameter i sannsynsfordelinga til Y.

Oppgåve 3

Levetida (målt i månadar) til nokon typar mekaniske komponentar har vist seg å følgje ei fordeling med kumulativ fordelingsfunksjon

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

der $\theta > 0$ er ein parameter som beskriv kvaliteta til komponentane.

a) Rekn ut følgjande to sannsyn når $\theta = 100$.

- Kva er sannsynet for at ein slik komponent framleis fungerer etter 12 månadar?
- Om me veit at ein slik komponent framleis fungerer etter 12 månadar, kva er sannsynet for at den også fungerer etter 15 månadar?

Finn eit uttrykk for sannsynstettleiken $f_X(x)$ og bruk dette til å vise at

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

For å undersøkje om kvaliteten til ein ny type av slike komponentar er betre enn den typen som har vore nytta til no prøver man ut $n = 50$ av dei nye komponentane. Vi lar $x_1, x_2, \ldots, x_n$ beteikne dei observerte levetidene for dei nye komponentane og betrakter desse som realisasjonar av tilhøyrande stokastiske variablar $X_1, X_2, \ldots, X_n$. Vi antek at $X_1, X_2, \ldots, X_n$ er eit tilfeldeg utval frå fordelinga gjeve over.

Vidare i oppgåva kan du utan bevis nytta at sannsynsmaksimeringsestimatoren for θ basert på $X_1, X_2, \ldots, X_n$ er

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

og at $\text{Var}[X_i^2] = \theta^2$. Hugs dessutan at vi frå deloppgåve **a**) har at $E[X_i^2] = \theta$.

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

Forklar så kvifor sentralgrenseteoremet i situasjonen me betraktar i denne oppgåva gjev at

$$\widehat{\theta}$$

er tilnærma normalfordelt med forventingsverdi lik θ og varians lik θ^2/n.

Det er kjent at den type komponentar som har vore i bruk til no har kvalitetsparameter $\theta = \theta_0 = 100$. Me ynsker altså no å nytte observerte verdiar $x_1, x_2, \ldots, x_n$ til å avgjere om det er grunnlag for å påstå at den nye typen komponentar har betre kvalitet enn den typen som har vore i bruk til no. Dei 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 gjev $\sum_{i=1}^{n} x_i^2 = 6084.853$.

c) Formuler situasjonen som ein hypotesetest ved å spesifisere null- og alternativ hypotese. Ved å nytte resultatet frå deloppgåve **b**), angi ein testobservator du kan nytte og kva (tilnærma) sannsynsfordeling denne testobservatoren har når H_0 er sann. Finn (tilnærma) p-verdi når observerte verdiar er som angjeve over.

Diskuter kort om den utrekna p-verdien gjev grunnlag for å forkasta H_0.

Merk at du i oppgåve 2 vart beden om å skrive ein python-funksjon for å simulera realisasjonar av X. Uavhengig om du fekk til å skrive ein slik funksjon eller ikkje, kan du i neste deloppgåve forutsettje at du har ein slik python-funksjon tilgjengleg.

I oppgåve **c**) rekna du ut ein p-verdi basert på ein approksimasjon av sannsynsfordelinga til testobservatoren når H_0 er sann. Den berekna verdien er altså berre ein approksimasjon til den eksakte p-verdien for den angjevne testen.

d) Forklar korleis du ved hjelp av stokastisk simulering kan formulere ein forventingsrett estimator for den eksakte p-verdien, og korleis du kan rekne ut eit tilhøyrande estimat. Du bør innføre notasjon slik at du kan gi ei presis formulering av prosedyra.

Skriv python-kode som utfører prosedyra 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
Høgde	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
Høgde	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		
Høgde	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: Målingar på 31 tre av arten romhegg. Verdiane er henta frå *Ryan, T.A., Joiner, B.L. and Ryan, B.L. (1976). Minitab student handbook, Duxbury Press.*

Oppgåve 4

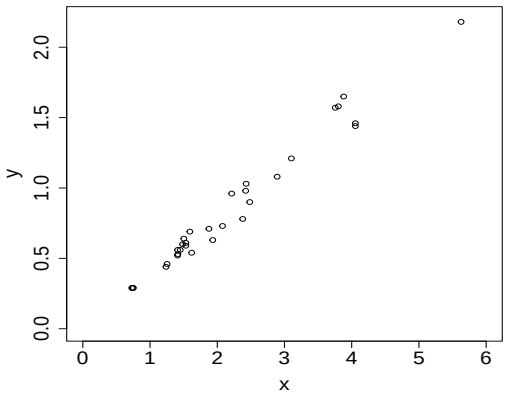
For å undersøkje om man ut ifrå målt diameter og høgde på eit tre kunne berekne kor stort volum av treet som kunne utnyttast til å lage plankar, har man for 31 tre av arten romhegg målt høvevis volum (m^3) som kunne bli nytta til plankeproduksjon, diameter (m) på treet målt 1.37 meter over bakken og høgde (m). Dei målte verdiane er gjeve i Tabell 1. For $i = 1, 2, \ldots, 31$ lar vi d_i, h_i og y_i være høvevis målt diameter, høgde og volum for tre nummer i. Vi lar videre $x_i = \pi d_i^2 h/4$ være volumet av ein sylinder med diameter lik d_i og høgde h_i. Figur 2 viser eit spreiingsplott der x_i er plotta langs x-aksa og y_i langs y-aksa. Det oppgis at dei observerte verdiene for diameter og høgde gjev $\sum_{i=1}^n x_i^2 = 193.3769$ og $\sum_{i=1}^n 1/x_i^2 = 13.5248$.

Me modellerer dataa ved ein regresjonsmodell der me betraktar observert verdi y_i som realisasjon av ein stokastisk variabel Y_i, og antek at

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

der residualane $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er uavhengige og normalfordelte stokastiske variablar med forventingsverdi lik null og varians lik σ^2. Legg merke til at me velger ikkje å inkludere noe konstantledd i regresjonsmodellen, og at vi som vanleg i regresjonsmodellar veljer å betrakte x_i'ane som gjevne tall.

Den antekne modellen har to parametrar, β og σ^2, og me skal anta at verdien til begge desse er ukjend og skal estimerast. For å estimere parameteren β er det



Figur 2: Spreiingsplott for observerte verdiar $(x_i, y_i); i = 1, 2, \ldots, 31$.

foreslått følgjande tre estimatorar,

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

a) Kva for ein av dei tre estimatorene er best? Hugs at svaret skal grunngjevast!

I neste deloppgåve skal du utleiie formlar for sannsynsmaksimeringsestimatorane for β og σ^2, og desse beteikner vi med høvevis $\widehat{\beta}$ og $\widehat{\sigma}^2$.

b) Utled sannsynsmaksimeringsestimatorane $\widehat{\beta}$ og $\widehat{\sigma}^2$.

Vis spesielt at $\widehat{\beta}$ er identisk med ein av dei tre estimatorane føreslege over og at $\widehat{\sigma}^2$ kan skrivast på formen

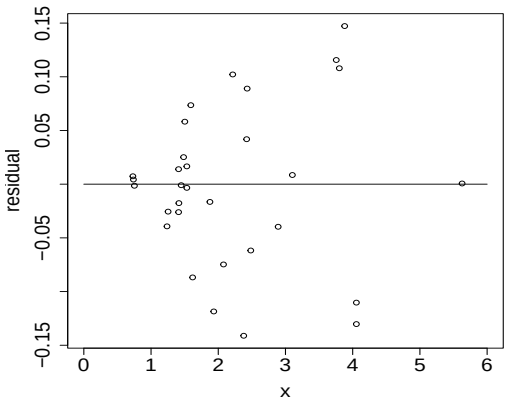
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \widehat{\beta} x_i\right)^2.$$

c) Grunngi at

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

er kjikvadratfordelt med n fridomsgrader. Angi spesielt kva kjende eigenskapar du nyttar og korleis du anvender desse.

Merk at deloppgåva held fram på neste side.



Figur 3: Residualplott for den antatte regresjonsmodellen i oppgåve 4 når man nytter $\widehat{\beta}$ som estimator for β.

Gje eit kort argument for at man frå dette får at

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

er kjikvadratfordelt med $n-1$ fridomsgrader. Merk at det her er tilstrekkeleg at du gjev eit argument for resultatet, du skal ikkje gje eit bevis.

Nytt så dette siste resultatet til å bevise at $\widehat{\sigma}^2$ er ein forventingsskeiv estimator for σ^2. Føreslå ein forventingsrett estimator for σ^2 og vis at estimatoren du føreslår faktisk er forventingsrett.

Som kjend kan man nytte eit residualplott til å vurdere om ein anteke regresjonsmodell ser ut til å passe med eit observert datasett. Figur 3 viser residualplottet for datasettet og regresjonsmodellen gjeve over når man nytter $\widehat{\beta}$ som estimator for β.

d) Forklar korleis eit residualplott konstrueres for regresjonsmodellen gjeve over. Angi spesielt kva som plottes langs x-aksa og kva som plottes langs y-aksa.

Diskuter kva du tolkar ut frå residualplottet i figur 3. Om din konklusjon er at den antekne regresjonsmodellen ikkje passar med det observerte datasettet, føreslå ein alternativ regresjonsmodell som du meiner vil passe betre. Forklar spesielt kvifor du meiner at modellen du føreslår vil passe betre.