NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4240 Statistikk**

Faglig kontakt under eksamen: Mette Langaas[a], Ingelin Steinsland[b], Geir-Arne Fuglstad[c]

Tlf: [a] 988 47 649 , [b] 926 63 096 , [c] 452 70 806

Eksamensdato: 21. desember 2016

Eksamenstid (fra–til): 09:00–13:00

Hjelpemiddelkode/Tillatte hjelpemidler: C: *Tabeller og formler i statistikk* (Tapir forlag, Fagbokforlaget), *Matematisk formelsamling* (K. Rottmann), ett stemplet gult A5-ark med egne håndskrevne notater, bestemt enkel kalkulator.

Annen informasjon:
Alle svar skal begrunnes, og besvarelsen skal inneholde naturlig mellomregning.

Målform/språk: bokmål

Antall sider: 7

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 Elektriske komponenter

En bedrift produserer elektriske komponenter. Komponentene kan ha to typer feil. Vi velger tilfeldig ut en komponent fra produksjonen og definerer to hendelser: A=komponenten har en feil av type A, og B=komponenten har en feil av type B. La A' og B' være de tilhørende komplementære hendelsene.

Det er kjent at $P(B) = 0.09$, $P(A \mid B) = 0.5$ og $P(A \mid B') = 0.01$.

a) Vi ser på en tilfeldig valgt komponent fra produksjonen.

Hva er sannsynligheten for at komponenten både har en type A og en type B feil, dvs. $P(A \cap B)$?

Hva er sannsynligheten for at komponenten har en feil av type A, dvs. $P(A)$?

Gitt at komponenten har en feil av type A, hva er sannsynligheten for at komponenten har en feil av type B, dvs. $P(B \mid A)$?

Vi er nå kun interessert i om en komponent er feilfri eller ikke. Ledelsen i bedriften har over mange år overvåket produksjonen, og er sikre på at sannsynligheten for at en tilfeldig valgt komponent er feilfri er 0.9. Vi velger tilfeldig ut 20 komponenter fra produksjonen, og undersøker om komponentene er feilfrie. La X være en stokastisk variabel som angir antall feilfrie komponenter.

b) Hvilken fordeling har X? Begrunn svaret.

Hva er sannsynligheten for at akkurat 19 komponenter er feilfrie?

Hva er sannsynligheten for at flere enn 15 komponenter er feilfrie?

Ledelsen i bedriften har innført noen endringer i produksjonsprosessen og håper at det har ført til en økt andel feilfrie komponenter. Kall denne ukjente andelen av feilfrie komponenter for p. Vi trekker et tilfeldig utvalg på n komponenter fra den nye produksjonsprosessen og lar X være antall feilfrie komponenter.

En intuitiv estimator for p er andelen feilfrie komponenter i utvalget, dvs. $\hat{P} = \frac{X}{n}$. Når vi har observert $X = x$ feilfrie komponenter kan vi regne ut et estimat for p som $\hat{p} = \frac{x}{n}$. Det tilfeldige utvalget av størrelse n er så stort at vi kan anta at $\frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}}$ er tilnærmet standard normalfordelt.

c) Utled et 90% konfidensintervall for p.

Regn ut konfidensintervallet når $n = 500$ og $x = 470$.

Gi en kort tolkning av intervallet.

Oppgave 2 Varians og kovarians

I denne oppgaven skal vi se på hvordan vi kan regne ut forventningsverdi og varians til et gjennomsnitt når observasjonene som inngår i gjennomsnittet er avhengige.

La X_1 og X_2 være stokastiske variabler med $\mathrm{E}(X_1) = \mathrm{E}(X_2) = 2$, $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1$ og $\mathrm{Cov}(X_1, X_2) = \frac{1}{2}$.

Finn $\mathrm{E}(\frac{1}{2}X_1 + \frac{1}{2}X_2)$ og $\mathrm{Var}(\frac{1}{2}X_1 + \frac{1}{2}X_2)$.

La videre $X_1, X_2, \ldots, X_{10}$ være stokastiske variabler med $\mathrm{E}(X_i) = 2$ og $\mathrm{Var}(X_i) = 1$ for $i = 1, 2, \ldots, 10$ og $\mathrm{Cov}(X_i, X_j) = \frac{1}{2}$ for alle $i = 1, 2, \ldots, 10$ og $j = 1, 2, \ldots, 10$ der $i \neq j$. La $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$.

Finn $\mathrm{E}(\bar{X})$ og $\mathrm{Var}(\bar{X})$.

Hint: du kan bruke følgende formel for variansen til en sum (som også finnes i *Tabeller og formler i statistikk*)

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i + b) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \,\mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} a_i a_j \,\mathrm{Cov}(X_i, X_j).$$

Oppgave 3 Høyde til mannlige studenter

Vi studerer en populasjon av mannlige studenter, og antar at høyden til en tilfeldig valgt mann fra populasjonen er normalfordelt med forventningsverdi μ og varians σ^2.

a) Anta (kun i dette punktet) at $\mu = 181\,\text{cm}$ og $\sigma = 6\,\text{cm}$. Vi trekker tilfeldig ut to mannlige studenter fra populasjonen og lar X_1 angi høyden til den første studenten og X_2 høyden til den andre. Vi antar at X_1 og X_2 er uavhengige stokastiske variabler.

Regn ut følgende sannsynligheter:
$\text{P}(X_1 > 190)$

$\text{P}(X_1 > 190 | X_1 > 185)$

$\text{P}(X_1 > 190 | X_2 > 185)$

To forskningsgrupper har uavhengig av hverandre estimert forventet høyde til mannlige studenter, μ. Forskningsgruppe 1 innhentet et tilfeldig utvalg av størrelse n og observerte høydene $x_1, x_2, \ldots, x_n$, og forskningsgruppe 2 innhentet et tilfeldig utvalg av størrelse m og observerte høydene $y_1, y_2, \ldots, y_m$. De to utvalgene ble trukket uavhengig av hverandre fra den gitte populasjonen.

Begge forskningsgruppene brukte empirisk middelverdi (gjennomsnitt) som estimator for μ, $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$ og $\bar{Y} = (Y_1 + Y_2 + \ldots + Y_m)/m$, og forskningsgruppe 1 fant $\bar{x} = 180\,\text{cm}$ og forskningsgruppe 2 fant $\bar{y} = 183\,\text{cm}$.

Du har studert statistikk og vet at du kan kombinere estimater fra uavhengige studier til å lage et estimat for μ som har lavere usikkerhet enn hvert av estimatene separat. Du har bestemt deg for å bruke estimatoren

$$\hat{\mu} = a\bar{X} + b\bar{Y},$$

der a og b er reelle tall.

b) Forklar hvilke to egenskaper som kjennetegner en god estimator.

Finn uttrykk for a og b (som funksjoner av n og m) slik at $\hat{\mu}$ er en estimator for μ som oppfyller egenskapene over.

Hva blir ditt estimat for μ hvis $n = 64$ og $m = 192$?

Etter nærmere ettertanke synes du at forskjellen mellom estimatene til de to forskningsgruppene er urimelig stor i forhold til deres utvalgsstørrelser $n = 64$ og $m = 192$. Din påstand er at du tror at de to forskningsgruppene ikke har tatt utvalgene sine fra den samme populasjonen.

Anta at forskningsgruppe 1 tok et tilfeldige utvalg fra en normalfordelt populasjon med forventningsverdi μ_1 og standardavvik σ_1, og at forskningsgruppe 2 tok et tilfeldige utvalg fra en normalfordelt populasjon med forventningsverdi μ_2 og standardavvik σ_2. Du har tidligere fått oppgitt at $\bar{x} = 180\,\text{cm}$ og $\bar{y} = 183\,\text{cm}$. Du kontakter forskningsgruppene og de sender deg de empiriske standardavvikene for sine observasjoner, $s_1 = 6.0$ for forskningsgruppe 1 og $s_2 = 5.5$ for forskningsgruppe 2.

c) Bruk påstanden din (gitt tidligere i teksten) til å formulere en null- og en alternativ hypotese.

Det er oppgitt at formelen for antall frihetsgrader i en test av forskjell i forventningsverdier når σ_1 kan være ulik σ_2, er

$$\nu = \frac{(s_1^2/n + s_2^2/m)^2}{(s_1^2/n)^2/(n-1) + (s_2^2/m)^2/(m-1)} = 100.6$$

innsatt numeriske verdier for s_1, s_2, n og m som oppgitt i oppgaven.

Argumenter for hvorfor denne testen kan brukes og finn forkastningsområdet til testen når signifikansnivået er $\alpha = 0.05$.

Hva blir konklusjonen av hypotesetesten når du bruker dataene som er oppgitt?

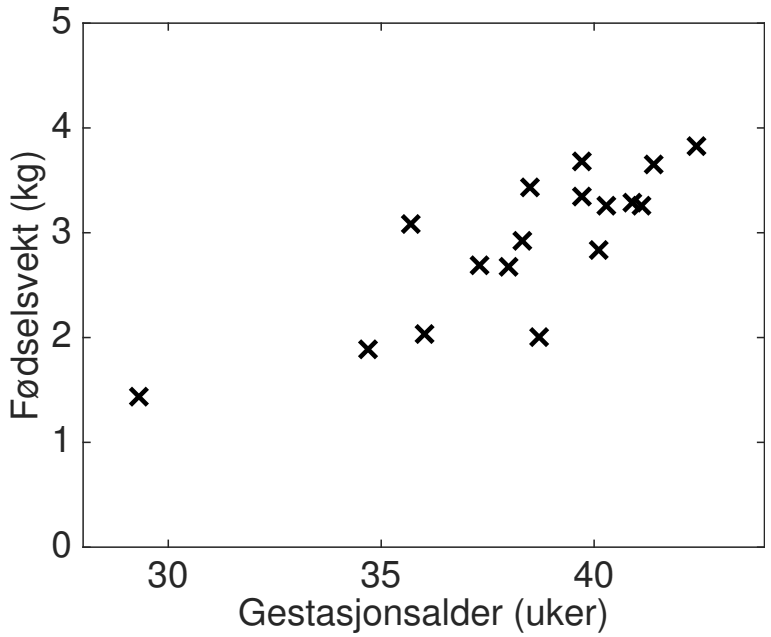
Oppgave 4 Fødselsvekt og gestasjonsalder

I figur 1 finner du et kryssplott (scatter plot) av fødselsvekt (målt i kg) og gestasjonsalder (tid fra første dag i siste menstruasjonsperiode til mor, målt i uker) for $n = 17$ fødsler.

Vi ønsker å tilpasse en enkel lineær regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

der hver ϵ_i er en normalfordelt stokastisk variabel med forventningsverdi 0 og varians σ^2. Videre er $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ uavhengige, og Y_i er fødselsvekt og x_i er gestasjonsalder.



Figur 1: Kryssplott av fødselsvekt, y_i, og gestasjonsalder, x_i for $i = 1, \ldots, 17$ barn.

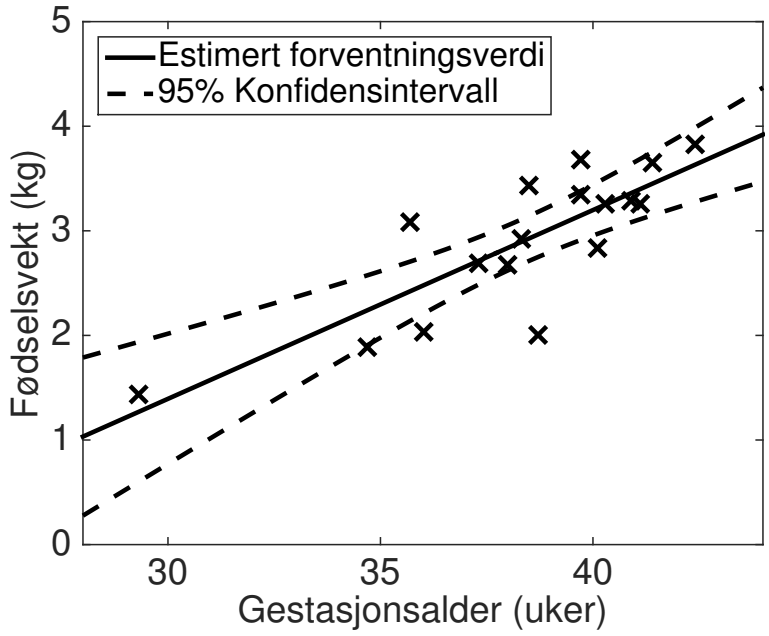
a) Diskuter om det er rimelig å bruke en lineær regresjonsmodell for observasjonene vist i figur 1.

Forklar hvordan minste kvadraters (least squares) metode kan brukes til å finne estimatorer B_0 for β_0 og B_1 for β_1, og illustrer med en figur. Du skal ikke utlede uttrykkene for estimatorene.

Det er oppgitt at estimatet for β_0 blir -4.02 og for β_1 blir 0.18. Finn predikert fødselsvekt for barn ved gestasjonsalder 40 uker.

b) Finn et uttrykk for variansen til $\hat{Y}_0 = B_0 + B_1 x_0$, der B_0 og B_1 er minste kvadraters estimatorene (least squares estimators) for β_0 og β_1. Du kan bruke at $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ og B_1 er uavhengige stokastiske variabler.

Studer figur 2 og bruk uttrykket for variansen til $\hat{Y}_0$ til å forklare hvorfor estimatet for forventningsverdien $\mathrm{E}(\hat{Y}_0)$ er mer usikkert ved $x_0 = 29$ uker enn ved $x_0 = 39$ uker.



Figur 2: Kryssplott av fødselsvekt og gestasjonsalder for 17 barn med estimert forventningsverdi for fødselsvekt (regresjonslinje) og grenser for 95% konfidensintervall for forventet fødselsvekt som funksjon av gestasjonsalder.

Oppgave 5 Generere data

Anta at Y er uniformt fordelt med sannsynlighetstetthet

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1, \\ 0, & \text{ellers.} \end{cases}$$

Finn kumulativ fordeling $F_Y(y)$ til Y.

På en datamaskin generer vi ofte observasjoner fra en fordeling ved først å generere en observasjon fra en uniform fordeling og så transformere observasjonen. Vi skal se på transformasjonen $X = -\ln(Y)/\lambda$, der $\lambda > 0$.

Bruk $F_Y(y)$ til å finne kumulativ fordeling $F_X(x)$ til X.

Hva blir sannsynlighetsfordelingen $f_X(x)$ til X, og hvilken kjent statistisk fordeling er dette?