NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4240 Statistikk**

Fagleg 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 (frå–til): 09:00–13:00

Hjelpemiddelkode/Tillatne hjelpemiddel: C: *Tabeller og formler i statistikk* (Tapir forlag, Fagbokforlaget), *Matematisk formelsamling* (K. Rottmann), eit stempla gult A5-ark med eigne handskrivne notat, bestemd enkel kalkulator.

Annan informasjon:
Alle svara skal grunngis og besvarelsen skal innehalde naturleg mellomrekning.

Målform/språk: nynorsk

Sidetal: 7

Sidetal vedlegg: 0

Kontrollert av:

Informasjon om trykking av eksamensoppgåve	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
svart/kvit ☒	**fargar** ☐
skal ha fleirvalskjema ☐	

Dato Sign

Oppgåve 1 Elektriske komponentar

Ei bedrift produserer elektriske komponentar. Komponentane kan ha to typar feil. Vi vel tilfeldig ut ein komponent frå produksjonen og definerer to hendingar: A=komponenten har ein feil av type A, og B=komponenten har ein feil av type B. La A' og B' vere dei tilhøyrande komplementære hendingane.

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

a) Vi ser på ein tilfeldig valt komponent frå produksjonen.

Kva er sannsynet for at komponenten både har ein feil av type A og ein feil av type B, dvs. $P(A \cap B)$?

Kva er sannsynet for at komponenten har ein feil av type A, dvs. $P(A)$?

Gitt at komponenten har ein feil av type A, kva er sannsynet for at komponenten har ein feil av type B, dvs. $P(B \mid A)$?

Vi er no kun interessert i om ein komponent er feilfri eller ikkje. Leiinga i bedrifta har over mange år overvaka produksjonen, og er sikre på at sannsynet for at ein tilfeldig valt komponent er feilfri er 0.9. Vi vel tilfeldig ut 20 komponentar frå produksjonen, og undersøkjer om komponentane er feilfrie. La X vere ein stokastisk variabel som angjev talet på feilfrie komponentar.

b) Kva for fordeling har X? Grunngi svaret.

Kva er sannsynet for at akkurat 19 komponentar er feilfrie?

Kva er sannsynet for at fleire enn 15 komponentar er feilfrie?

Leiinga i bedrifta har innført nokre endringar i produksjonsprosessen og håper at det har ført til auka andel feilfrie komponentar. Kall den ukjende andelen av feilfrie komponentar for p. Vi trekkjer eit tilfeldig utval på n komponentar frå den nye produksjonsprosessen og lar X vere talet på feilfrie komponentar.

Ein intuitiv estimator for p er andelen feilfrie komponentar i utvalet, dvs. $\hat{P} = \frac{X}{n}$. Når vi har observert $X = x$ feilfrie komponentar kan vi rekne ut eit estimat for p som $\hat{p} = \frac{x}{n}$. Det tilfeldige utvalet av storleik n er så stort at vi kan anta at $\frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}}$ er tilnærma standard normalfordelt.

c) Utlei eit 90% konfidensintervall for p.

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

Gjev ei kort tolking av intervallet.

Oppgåve 2 Varians og kovarians

I denne oppgåva skal vi sjå på korleis vi kan rekne ut forventningsverdi og varians til eit gjennomsnitt når observasjonane som inngår i gjennomsnittet er avhengige.

La X_1 og X_2 vere stokastiske variablar 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 vidare $X_1, X_2, \ldots, X_{10}$ vere stokastiske variablar 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 følgjande formel for variansen til ein sum (som du og finn i *Tabeller og formler i statistikk*)

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

Oppgåve 3 Høgde til mannlege studentar

Vi studerer ein populasjon av mannlege studentar, og antar at høgda til ein tilfeldig valt mann frå 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 trekkjer tilfeldig ut to mannlege studentar frå populasjonen og lar X_1 angi høgda til den første studenten og X_2 høgda til den andre. Vi antar at X_1 og X_2 er uavhengige stokastiske variablar.

Rekn ut følgjende sannsyn:

$\mathrm{P}(X_1 > 190)$

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

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

To forskingsgrupper har uavhengig av kvarandre estimert forventa høgde til mannlege studentar, μ. Forskingsgruppe 1 henta eit tilfeldig utval av storleik n og observerte høgdene $x_1, x_2, \ldots, x_n$, og forskingsgruppe 2 henta eit tilfeldig utval av storleik m og observerte høgdene $y_1, y_2, \ldots, y_m$. Dei to utvala blei trekte uavhengig av kvarandre frå den gitte populasjonen.

Begge forskingsgruppene 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 forskingsgruppe 1 fann $\bar{x} = 180\,\text{cm}$ og forskingsgruppe 2 fann $\bar{y} = 183\,\text{cm}$.

Du har studert statistikk og veit at du kan kombinere estimatar frå uavhengige studium til å lage eit estimat for μ som har lågare usikkerheit enn kvart av estimata separat. Du har bestemt deg for å bruke estimatoren

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

der a og b er reelle tal.

b) Forklar kva for to eigenskapar som kjenneteiknar ein god estimator.

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

Kva blir ditt estimat for μ dersom $n = 64$ og $m = 192$?

Ved nærare ettertanke synes du at skilnaden mellom estimata til dei to forskingsgruppene er urimelig stor i forhold til deira utvalsstorleikar $n = 64$ og $m = 192$. Din påstand er at du trur at dei to forskingsgruppene ikkje har henta utvala sine frå den same populasjonen.

Anta at forskingsgruppe 1 henta eit tilfeldige utval frå ein normalfordelt populasjon med forventningsverdi μ_1 og standardavvik σ_1, og at forskingsgruppe 2 henta et tilfeldige utval frå ein normalfordelt populasjon med forventningsverdi μ_2 og standardavvik σ_2. Du har tidlegare fått oppgitt at $\bar{x} = 180\,\text{cm}$ og $\bar{y} = 183\,\text{cm}$. Du kontaktar forskingsgruppene og dei sender deg dei empiriske standardavvika for sine observasjonar, $s_1 = 6.0$ for forskingsgruppe 1 og $s_2 = 5.5$ for forskingsgruppe 2.

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

Det er oppgitt at formelen for talet på fridomsgrader i ein test av forskjell i forventningsverdier når σ_1 kan vere ulik σ_2, er

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

innsett numeriske verdiar for s_1, s_2, n og m som oppgitt i oppgåva.

Argumenter for kvifor denne testen kan brukast og finn forkastingsområde for testen når signifikansnivået er $\alpha = 0.05$.

Kva blir konklusjonen av hypotesetesten når du bruker data som er oppgitt?

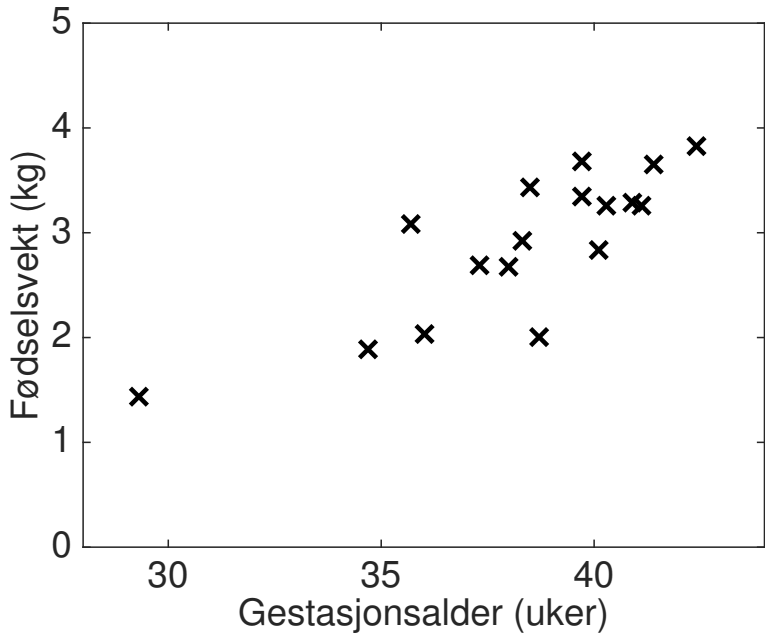
Oppgåve 4 Fødselsvekt og gestasjonsalder

I figur 1 finn du eit kryssplott (scatter plot) av fødselsvekt (målt i kg) og gestasjonsalder (tid frå første dag i siste menstruasjonsperiode til mor, målt i veker) for $n = 17$ fødsler.

Vi ønsker å tilpasse ein enkel lineær regresjonsmodell

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

der kvar ϵ_i er ein normalfordelt stokastisk variabel med forventningsverdi 0 og varians σ^2. Vidare 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.

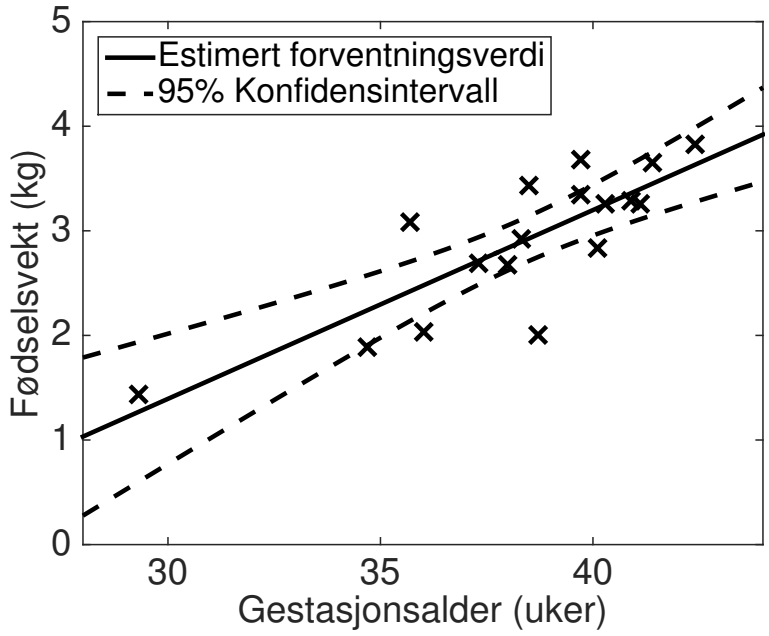
a) Diskuter om det er rimelig å nytte ein lineær regresjonsmodell for observasjonane vist i figur 1.

Forklar korleis minste kvadraters (least squares) metode kan brukast til å finne estimatorar B_0 for β_0 og B_1 for β_1, og illustrer med ein figur. Du skal ikkje utleie uttrykka for estimatorane.

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

b) Finn eit uttrykk for variansen til $\hat{Y}_0 = B_0 + B_1 x_0$, der B_0 og B_1 er minste kvadraters estimatorane (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 variablar.

Studer figur 2 og bruk uttrykket for variansen til $\hat{Y}_0$ til å forklare kvifor estimatet for forventningsverdien $\mathrm{E}(\hat{Y}_0)$ er meir 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.

Oppgåve 5 Generere data

Anta at Y er uniformt fordelt med sannsynstettleik

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

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

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

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

Kva blir sannsynstettleiken $f_X(x)$ til X, og kva for kjent statistisk fordeling er dette?