NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **ST1201/ST6201 Statistiske metoder**

Faglig kontakt under eksamen: Bo Lindqvist

Tlf: 975 89 418

Eksamensdato: 12. desember 2018

Eksamenstid (fra–til): 09:00 – 13:00

Hjelpemiddelkode/Tillatte hjelpemidler: Hjelpemiddelkode C:

– Tabeller og formler i statistikk, Tapir forlag,
– K.Rottman: Matematisk formelsamling,
– Ett gult ark (A4 med stempel) med egne håndskrevne formler og notater,
– Bestemt, enkel kalkulator

Annen informasjon:

Alle svar må begrunnes.

Du må ha med nok mellomregninger til at tenkemåten din klart fremgår.

Oppgaven består av 10 delpunkter som har lik vekt ved sensur.

Målform/språk: bokmål

Antall sider: 6

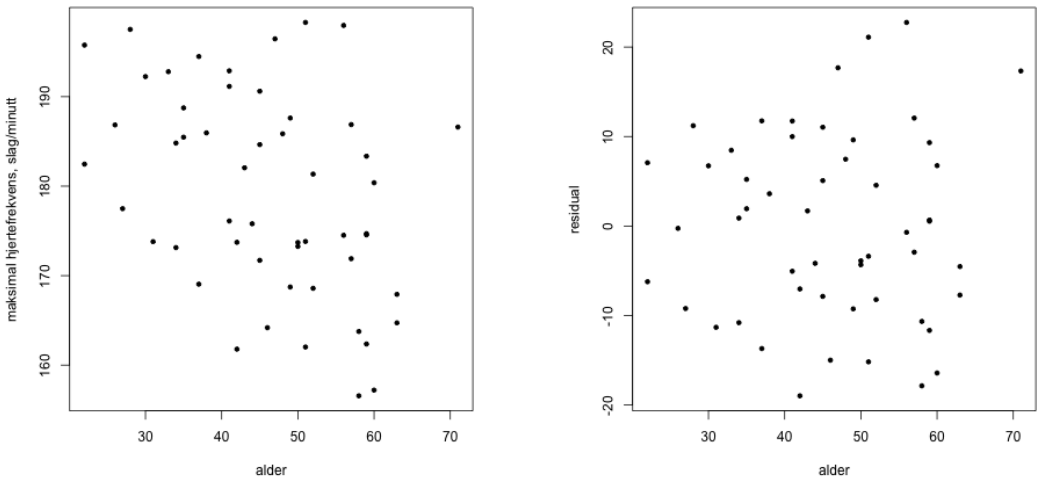
Antall sider vedlegg: 0

Kontrollert av:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
sort/hvit ☒	**farger** ☐
skal ha flervalgskjema ☐	

Dato Sign

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.



Figur 1: Plott av y mot x (venstre) og residualplott (høyre) for målingene av maksimal hjertefrekvens og alder for kvinner.

Oppgave 1 Enkel lineær regresjon

Vi ønsker å studere sammenhengen mellom maksimal hjertefrekvens i slag pr minutt (y) og alder i år (x) for kvinner, og innhenter et tilfeldig utvalg av størrelse $n = 52$. Se venstre panel av figur 1 for et plott av y mot x.

Deretter tilpasser vi en enkel lineær regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ for } i = 1, \ldots, n,$$

der vi antar at feilleddene ε_i er normalfordelte med forventning 0 og varians σ^2. Videre antas at $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er uavhengige.

Fra utvalget finner vi følgende numeriske verdier for minste kvadratsums estimatorene for skjæringspunktet, $\hat{\beta}_0 = 197$, og for stigningstallet, $\hat{\beta}_1 = -0.66$. I høyre panel av figur 1 finner du et residualplott.

a) Lag en skisse av den estimerte regresjonslinjen basert på de oppgitte estimatene for $\hat{\beta}_0$ og $\hat{\beta}_1$.

Hvordan ville du forklare til en person som ikke kjenner til lineær regresjon, hva det betyr at $\hat{\beta}_1 = -0.66$?

Hva er definisjonen av residualet e_i til observasjonsparet (x_i, y_i)?

Tegn inn de to observasjonene $x = 27, y = 178$ og $x = 52, y = 181$ i skissen. Marker residualet for hver av de to observasjonene på skissen.

Basert på residualplottet i figur 1, hvordan vil du evaluere om regresjonsmodellen passer for dette utvalget? Begrunn kort.

Vi ønsker også å studere sammenhengen mellom maksimal hjertefrekvens og alder for menn, og har trukket et utvalg av størrelse $n^* = 52$.

Vi har tilpasset en enkel lineær regresjonsmodell for maksimal hjertefrekvens y^* for menn, der parametrene er β_0^* og β_1^*, og der det antas at feilleddene ε_j^* er uavhengige og normalfordelte med forventning 0 og varians σ^2. Merk at vi antar samme varians σ^2 som for regresjonsmodellen for kvinner i punkt (a).

Vi antar videre at utvalgene av menn og av kvinner er trukket uavhengig av hverandre.

Numeriske verdier for dette utvalget av menn og den tilpassede enkle lineære regresjonen er presentert i tabellen under, sammen med tilsvarende numeriske verdier for utvalget av kvinner fra punkt (a). Her er $\bar{x}$ gjennomsnittet av observasjonene x_i av alder for kvinner og $\bar{x}^*$ gjennomsnittet av observasjonene x_j^* av alder for menn.

Kvinner	Menn
$n = 52$	$n^* = 52$
$\hat{\beta}_0 = 197$	$\hat{\beta}_0^* = 224$
$\hat{\beta}_1 = -0.66$	$\hat{\beta}_1^* = -0.70$
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 6969$	$\sum_{j=1}^{n^*}(x_j^* - \bar{x}^*)^2 = 8409$
$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 5581$	$\sum_{j=1}^{n^*}(y_j^* - \hat{\beta}_0^* - \hat{\beta}_1^* x_j^*)^2 = 4669$

b) Nå er β_1 stigningstallet for den enkle lineære regresjonen for kvinner (fra (a)), og β_1^* stigningstallet for den enkle lineære regresjonen for menn. Vi ønsker å teste om disse to stigningstallene er ulike.

Formuler dette problemet ved å sette opp en nullhypotese og en alternativ hypotese.

Finn forventning og varians til differansen mellom minste kvadratsums estimatorene i de to utvalgene, $\hat{\beta}_1 - \hat{\beta}_1^*$.

Du trenger ikke utlede estimatorene $\hat{\beta}_1$ og $\hat{\beta}_1^*$, men kan isteden bruke en formel for disse gitt i "Tabeller og formler i statistikk".

c) Foreslå en estimator for den felles variansen σ^2 for de to utvalgene.

Foreslå en testobservator for å teste hypotesene fra punkt (b).

Hva er testobservatorens fordeling når nullhypotesen er sann?

Utfør testen. Bruk signifikansnivå 10% .

Oppgave 2 Forsøksplanlegging

En produsent av sportsutstyr utførte et forsøk for å undersøke slitasjeegenskaper for to typer materiale for skosåler, betegnet A og B. Ti forsøkspersoner ble valgt ut til å delta i forsøket, og hver person ble utstyrt med et par sko der den ene sålen var laget med materiale A og den andre sålen med materiale B. Det ble gjort loddtrekning om hvilken fot, høyre eller venstre, de to forskjellige skosålene skulle være på. Etter en viss tid ble slitasjen på skosålene målt. Observert slitasje er gitt i tabellen under. Høyre eller venstre fot er angitt i parentes.

Person	Slitasje med A	Slitasje med B	Forskjell i slitasje B - A
1	13.2 (V)	14.0 (H)	0.8
2	8.2 (V)	8.8 (H)	0.6
3	10.9 (H)	11.2 (V)	0.3
4	14.3 (V)	14.2 (H)	-0.1
5	10.7 (H)	11.8 (V)	1.1
6	6.6 (V)	6.4 (H)	-0.2
7	9.5 (V)	9.8 (H)	0.3
8	10.8 (V)	11.3 (H)	0.5
9	8.8 (H)	9.3 (V)	0.5
10	13.3 (V)	13.6 (H)	0.3

La Y_{ij} være slitasje på skosålen brukt av i-te forsøksperson og laget av j-te materiale ($j = 1$ for A, $j = 2$ for B). Følgende modell ble først antatt å kunne beskrive dataene:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \quad i = 1, \ldots, 10, \; j = 1, 2,$$

der ϵ_{ij} er uavhengige og normalfordelte med forventning 0 og varians σ^2. La $\overline{Y}_{\cdot j} = \sum_{i=1}^{10} Y_{ij}/10$ og $\overline{Y}_{\cdot\cdot} = \sum_{i=1}^{10}\sum_{j=1}^{2} Y_{ij}/20$. En måte å splitte opp variasjonen i dataene på er:

$$\sum_{i=1}^{10}\sum_{j=1}^{2}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{10}\sum_{j=1}^{2}(Y_{ij} - \overline{Y}_{\cdot j})^2 + \sum_{i=1}^{10}\sum_{j=1}^{2}(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2 \qquad (1)$$

a) Hva beskriver μ_j og σ?

Hva beskriver de tre kvadratsummene gitt i ligning (1)?

En delvis utfylt variansanalysetabell er gitt nedenfor:

Kilder	Frihetsgrader	Kvadratsummer	Gjennomsnittlig kvadratsum	F
Materiale	*	0.841	*	*
Feil	*	*	*	
Total	*	112.005		

b) Fyll ut variansanalysetabellen.

Formuler og utfør en test for å undersøke om det er ulik slitasje med de to materialene. Hva blir konklusjonen? Bruk signifikansnivå 5%.

En innvending mot analysen ovenfor er at forsøkspersonene kan bruke skoene veldig forskjellig, noe som kan føre til stor variasjon i slitasje mellom personer. Det ble derfor foreslått at analysen isteden skulle utføres som ved et randomisert blokkdesign.

c) Beskriv forsøksopplegget som et randomisert blokkdesign.

Skriv opp modellen det nå er naturlig å bruke.

Du får opplyst at $\sum_{i=1}^{10}\sum_{j=1}^{2}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2 = 110.491$. Hva blir nå konklusjonen på testproblemet i (b)?

I siste kolonne i tabellen gitt i innledningen, er det oppgitt forskjellen i slitasje mellom skosåle B og skosåle A for de 10 forsøkspersonene. Gjennomsnittet av disse tallene er 0.41.

d) Siden materiale B er billigere enn materiale A, ønsker produsenten å undersøke om slitasjen med materiale B er større enn slitasjen med materiale A.

Formuler en test for dette tilfellet.

Utfør testen. Hva blir konklusjonen med signifikansnivået satt til 1%?

Oppgave 3 Wilcoxons to-utvalgstest

La X og Y være uavhengige stokastiske variabler der X har tetthet $f_X(x)$ og Y har tetthet $f_Y(y)$.

Vi skal teste nullhypotesen

$$H_0: \ f_X(x) = f_Y(x) \ \text{ for alle } x \in (-\infty, +\infty)$$

mot den alternative hypotesen

$$H_1: \ f_Y(x) = f_X(x - c) \ \text{ for alle } x \in (-\infty, +\infty)$$

der $c \neq 0$ er en vilkårlig konstant

La $X_1, X_2, \ldots, X_{n_1}$ og $Y_1, Y_2, \ldots, Y_{n_2}$ være uavhengige tilfeldige utvalg fra henholdsvis f_X og f_Y.

a) Tegn en skisse og forklar med ord hva de to hypotesene H_0 og H_1 betyr.

Forklar kort hvordan man beregner testobservatoren i Wilcoxons to-utvalgstest (*The Wilcoxon Rank Sum Test*).

Regn ut testobservatoren og finn p-verdien for Wilcoxons to-utvalgstest for H_0 mot H_1 når $n_1 = 4,\ n_2 = 5$ og dataene er

X_i :	13.1	16.6	8.8	14.1	
Y_j :	15.7	19.1	16.9	18.9	8.2

For å finne p-verdien kan du bruke tabellen på side 24 i "Tabeller og formler i statistikk". Merk at tabellen gir fordelingen for

$$U_1 = W_1 - n_1(n_1 + 1)/2,$$

der W_1 er summen av rangene for de n_1 X-ene i utvalget.

Observatoren U_1 kalles *Mann-Whitney observatoren.* Den ble opprinnelig definert ved

$$U_1 = \text{ antall par } (i, j) \text{ med } Y_j < X_i, \tag{2}$$

der $i = 1, 2, \ldots, n_1$ og $j = 1, 2, \ldots, n_2$.

b) Finn U_1 for datasettet i punkt (a) ved å bruke definisjonen (2). Skriv ned de par som teller med i U_1 for dette datasettet.

Vis at med U_1 definert ved (2), vil vi ha at

$$U_1 = W_1 - n_1(n_1 + 1)/2,$$

der W_1 er summen av rangene for X-observasjonene.

(*Vink:* La $X'_1 < X'_2 < \cdots < X'_{n_1}$ være ordningsobservatoren for de n_1 X-observasjonene og la $R_1, \ldots, R_{n_1}$ være de tilhørende rangene (som blir brukt ved beregning av W_1). Forklar hvorfor

det er $R_1 - 1$ Y-observasjoner som er mindre enn X'_1;

det er $R_2 - 2$ Y-observasjoner som er mindre enn X'_2;

$\vdots$

det er $R_{n_1} - n_1$ Y-observasjoner som er mindre enn X'_{n_1}.

Bruk også at $1 + 2 + \ldots + n_1 = n_1(n_1 + 1)/2$.)

c) La X og Y være som definert i begynnelsen av oppgaven, og la

$$p = P(Y < X)$$

Vis at $p = 1/2$ hvis H_0 gjelder.

Vis at med definisjonen av U_1 gitt ved (2), er

$$\hat{p} = \frac{U_1}{n_1 n_2}$$

en forventningsrett estimator for p.

Bruk dette til å vise at under H_0 er

$$E(W_1) = \frac{n_1(n_1 + n_2 + 1)}{2}$$

(*Vink:* For å vise at $\hat{p}$ er forventningsrett, kan det lønne seg å innføre variablene

$$Z_{ij} = \begin{cases} 1 & \text{hvis } Y_j < X_i \\ 0 & \text{ellers} \end{cases}$$

og begrunne at $U_1 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} Z_{ij}$.)