NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **ST0103 Brukerkurs i statistikk**

Faglig kontakt under eksamen: Jarle Tufto

Tlf: 99 70 55 19

Eksamensdato: 3. desember 2016

Eksamenstid (fra–til): 09:00-13:00

Hjelpemiddelkode/Tillatte hjelpemidler: C: Bestemt enkel kalkulator. Tabeller og formler i statistikk (Tapir akademisk forlag). Ett gult A4-ark med egne håndskrevne notater.

Annen informasjon:
Noen formler for bruk i regresjonsanalyse er gitt i vedlegg.

Alle svar skal begrunnes (f.eks. ved at mellomregning tas med eller ved henvisning til teori eller eksempler fra pensum).

I vurderingen teller hvert av de ti bokstavpunktene likt.

Målform/språk: bokmål

Antall sider: 5

Antall sider vedlegg: 1

Kontrollert av:

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

Dato Sign

Oppgave 1

Ved et botanisk institutt gjorde man en undersøkelse av forekomsten av en bestemt type sopp i myrjord. Dette ble gjort ved at man på et bestemt myrområde tok opp sylindre av jord med en jordskrue, vasket røttene, og deretter bestemte andelen av rotceller med sopp. Andelen ble oppgitt med enhet prosent.

La X være resultatet av en slik prøve. Fra lang erfaring regnes som kjent at X er normalfordelt med forventning 42.0 og standardavvik 4.0. Resultater fra forskjellige prøver antas å være stokastisk uavhengige.

a) Hva er sannsynligheten for at en måling av X er mindre enn 35?

Finn sannsynligheten for at en måling er mellom 35 og 45.

Tegn inn de to sannsynlighetene som arealer på en enkel skisse av sannsynlighetstettheten til X.

b) Anta i dette punktet at det ble gjort fire prøver.

Hva er sannsynligheten for at gjennomsnittet av de fire prøvene er under 35?

Hva er sannsynligheten for at minst tre av de fire prøvene er mellom 35 og 45?

Oppgave 2

La situasjonen være som i Oppgave 1. På en annen del av myrområdet består jordsmonnet hovedsakelig av bleket sand. Man ønsket å finne ut om andelen av sopp er en annen i denne jordtypen enn i myrjorda der prøvene beskrevet i Oppgave 1 ble tatt.

Det ble tatt 20 prøver i det nye området på samme måte som beskrevet i begynnelsen av Oppgave 1. Dette ga målingene $X_1, X_2, \ldots, X_{20}$ gitt nedenfor. Målingene antas å være realisasjoner av uavhengige og identisk normalfordelte variabler med ukjent forventning μ og samme standardavvik som for de første prøvene, $\sigma = 4.0$. *(Analysene nedenfor skal altså gjøres med kjent σ.)*

Målinger:

37.99	46.49	36.92	48.10	42.70	49.91	39.10	46.43	39.52	40.40
47.11	41.32	42.61	47.99	44.19	47.40	49.42	45.52	44.84	41.45

Du kan bruke at $\sum_{i=1}^{20} X_i = 879.41$.

a) Finn et punktestimat for μ basert på målingene, og beregn dets standardfeil, dvs. standardavviket for estimatoren.

Finn også et 95% konfidensintervall for μ.

b) Forklar kort hvorfor botanikernes problemstilling medfører testing av

$$H_0 : \mu = 42.0 \text{ mot } H_1 : \mu \neq 42.0.$$

Gjennomfør testingen med de gitte dataene og angi konklusjonen når signifikansnivået settes til 0.05. Beregn også den tilhørende p-verdi. Hvilken fortolkning har den?

c) Hva menes med type I-feil og type II-feil i hypotesetesting?

Hva er sannsynligheten for type I-feil i testen i forrige punkt?

Hva er sannsynligheten for type II-feil ved denne testen dersom μ i virkeligheten er 45.0?

Oppgave 3

Som en del av en større geologisk undersøkelse har man studert kjerneprøver av sand med formål å beskrive sammenhengen mellom effektiv diffusjon og graden av sementering.

Den effektive diffusjon beskriver hvor lett gasser diffunderer gjennom sandkjernen og representeres ved en måling d. Graden av sementering for en sandkjerne er et tall mellom 0 og 1 og betegnes med c.

Man ønsker å beskrive d som funksjon av c og forventer en sammenheng

$$d = \theta(1 - c)^{\beta}, \tag{1}$$

der θ og β er ukjente parametre.

For å estimere θ og β gjøres et eksperiment der man starter med å måle den effektive diffusjon D for en sandkjerne med $c = 0$. Deretter økes sementeringen gradvis ved tilsetting av koppersulfat, og man måler de korresponderende verdier av c og D. Dette gir tilsammen 10 sammenhørende verdier av D og c, $(D_1, c_1), (D_2, c_2), \ldots, (D_{10}, c_{10})$.

a) Vis, ved å ta den naturlige logaritmen på hver side av likheten i (1), at vi kan beskrive den forventede sammenhengen (1) ved

$$y = \alpha + \beta x$$

der $y = \ln d$, $\alpha = \ln \theta$ og $x = \ln(1 - c)$.

Basert på dette bestemmer man seg for å analysere dataene med en enkel lineære regresjonsmodell

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \ldots, 10 \tag{2}$$

der $Y_i = \ln D_i$ og $x_i = \ln(1 - c_i)$.

Hvilke antagelser ligger generelt til grunn for bruk av modellen (2)?
Diskuter kort i hvilken grad de kan antas oppfylt i den gitte situasjonen.

Tabellen nedenfor gir verdiene av c_i og de tilhørende responser $D_i = d_i$, samt de transformerte x_i og y_i som brukes i regresjonsanalysen.

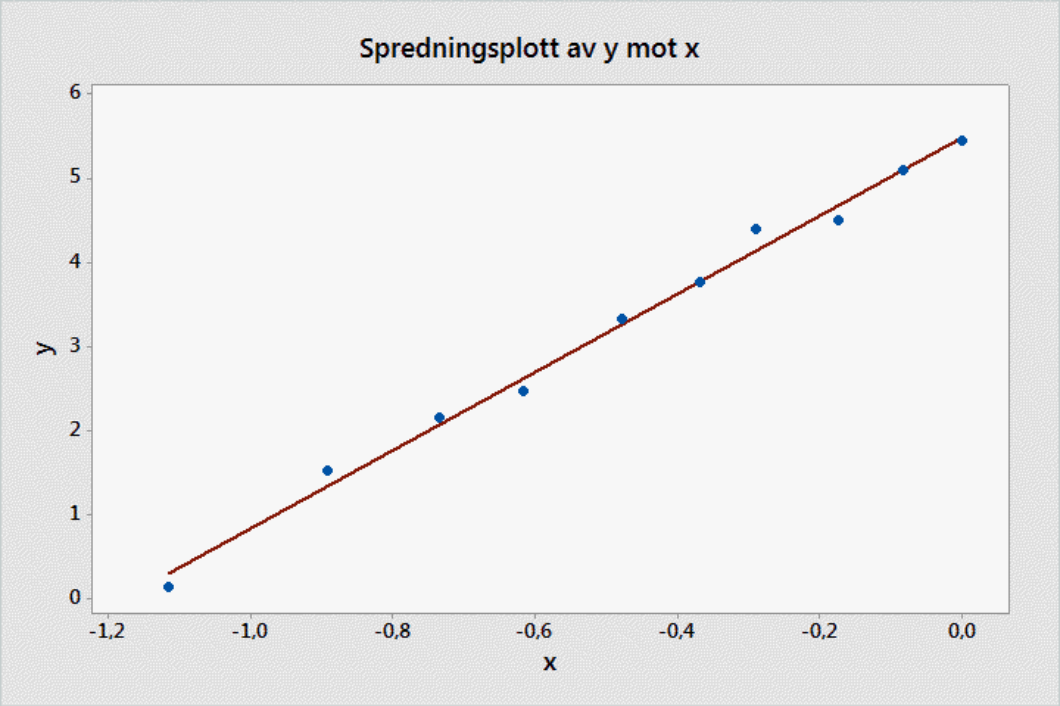
i	c_i	d_i	$x_i = \ln(1 - c_i)$	$y_i = \ln d_i$
1	0,000	234.102	0.000	5.456
2	0.080	165.961	-0.083	5.112
3	0.160	90.990	-0.174	4.511
4	0.250	82.496	-0.288	4.413
5	0.307	43.413	-0.367	3.771
6	0.380	28.156	-0.478	3.338
7	0.460	11.903	-0.616	2.477
8	0.520	8.747	-0.734	2.169
9	0.590	4.589	-0.892	1.524
10	0.672	1.165	-1.115	0.153

Det oppgis at $\sum x_i = -4.747$, $\sum x_i^2 = 3.440$, $\sum y_i = 32.924$, $\sum y_i^2 = 134.273$, $\sum x_i y_i = -10.108$.

Du kan også bruke at $SS_E = 0.1898$ og $\sum(x_i - \bar{x})^2 = 1.187$.

I regningen kan du bruke formler som er gitt i vedlegget og i tabellene.

Et spredningsplott for de beregnede (x_i, y_i) med inntegnet regresjonslinje er til orientering gitt på neste side.



b) Beregn punktestimatene $\hat{\alpha}$ og $\hat{\beta}$ for α og β basert på minste kvadraters metode ved å bruke de oppgitte resultatene.

Finn også et punktestimat for parameteren θ fra ligning (1).

Beregn punktestimatet S for σ og bruk dette til å finne standardfeilen for estimatet $\hat{\beta}$. Finn også et 95% konfidensintervall for β.

Hvor stor andel av variasjonen i responsene forklares av regresjonsmodellen? Gi en kommentar i lys av spredningsplottet.

Oppgave 4

En biolog er interessert i å undersøke forekomsten av en sjelden insektsart, i det følgende kalt art A, i et bestemt område. Det settes opp en såkalt malaisefelle (en stor, telt-lignende struktur) og det antas at insekter av art A fanges i fellen som en Poisson-prosess med intensitet (rate) λ pr. time. La X være antallet insekter av art A som fanges i løpet av en uke (= 168 timer). Biologen anslår før undersøkelsen at λ er 0.01. Denne verdi for λ skal brukes i punktene a) og b).

a) Gjør rede for at X er Poisson-fordelt med forventning $\mu = 1.68$.

Hva er sannsynligheten for at biologen ikke vil finne noen insekter av art A i fellen etter en uke?

Hva er sannsynligheten for at det er minst tre insekter av art A i fellen?

Anta nå at fellen ble stående i to uker. Gitt at det ble fanget minst tre insekter av art A i løpet av de to ukene, hva er sannsynligheten for at ingen ble fanget den første uken?

b) La T være tiden, målt i timer, til det første insektet av art A blir fanget i fellen. *(Se i dette punktet bort fra at fellen er montert bare en begrenset tid.)*

Hvilken fordeling har T?

Hva er sannsynligheten for at ingen insekter av art A fanges i løpet av de første 48 timene, dvs. $T > 48$?

Gitt at ingen insekter av art A fanges i løpet av de første 48 timene, hva er sannsynligheten for at det heller ikke fanges noen i løpet av de neste 48 timene? Kommenter resultatet.

Biologen er ikke lenger sikker på sitt tidligere anslag av λ og vil estimere λ. Han lar fellen stå i fem uker og finner Y insekter av art A.

c) Hvilken sannsynlighetsfordeling har Y?

Sett opp en forventningsrett estimator $\hat{\lambda}$ for λ basert på Y.

Hva blir estimatet hvis det observeres at $Y = 15$? Hva blir standardfeilen (dvs. estimert standardavvik) for dette estimatet?

Supplement til "Noen resultater fra regresjonsanalysen"
i *Tabeller og formler i statistikk, Tapir akademisk forlag.*

Formlene bygger på summene:

$$\sum_{i=1}^{n} x_i, \quad \sum_{i=1}^{n} x_i^2, \quad \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} y_i^2, \quad \sum_{i=1}^{n} x_i y_i$$

Minste kvadraters metode gir da:

$$\begin{aligned} \hat{\beta} &= \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2} \\ \hat{\alpha} &= \frac{(\sum_{i=1}^n y_i) - \hat{\beta}(\sum_{i=1}^n x_i)}{n} \end{aligned}$$

Fra læreboka:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i - \bar{y})^2}_{SS_R} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}_{SS_E}$$

Her er:

$$\begin{aligned} SS_T &= (\sum_{i=1}^{n} y_i^2) - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 \\ SS_E &= (\sum_{i=1}^{n} y_i^2) - \hat{\alpha}(\sum_{i=1}^{n} y_i) - \hat{\beta}(\sum_{i=1}^{n} x_i y_i) \end{aligned}$$

Forventningsrett estimator for σ^2:

$$S^2 = \frac{SS_E}{n-2}$$

Statistiske egenskaper ved estimatorene:

$$E(\hat{\alpha}) = \alpha, \qquad E(\hat{\beta}) = \beta$$

$$Var(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$