UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST102 — Videregående kurs i statistikk.
Eksamensdag:	Tirsdag 29. mai 2001.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 6 sider.	
Vedlegg:	Utskrifter fra MINITAB til oppgave 2. Tabeller over den kumulative standard normalfordelingen, kumulative kji-kvadrat fordelinger og kumulative t-fordelinger.
Tillatte hjelpemidler:	Formelsamlinger for ST101 og ST102, lommeregner, Haugens "Formler og tabeller," Jahren og Knutsens "Formelsamling i matematikk," Rottmanns "Mathematische Formelsammlung."

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

Denne oppgaven dreier seg om en spesiell type minste kvadraters metode, nemlig vektet minste kvadraters metode.

I modellen

$$Y_i = \beta_0 + \beta_1 x_i + e_i\,; \qquad i = 1, 2, \ldots n\,; \tag{1}$$

antar vi at støyleddene e_i er uavhengige og normalfordelte med forventning null og

$$Var(e_i) = \sigma_i^2 = w_i^2 \sigma^2,$$

(Fortsettes side 2.)

hvor w_i-ene er kjente positive konstanter.

a) Forklar hvorfor situasjonen ovenfor oppstår når responsen Y_i er gjennomsnittet av n_i observasjoner i punktet x_i. Hva blir w_i-ene i dette tilfellet?

For å kunne bruke teori for den vanlige lineære regresjonsmodellen, kan vi transformere modellen (1). Vi innfører da $Z_i = Y_i/w_i$.

b) Vis at

$$Z_i = \beta_0 u_i + \beta_1 v_i + \delta_i\,; \qquad i = 1, 2, \ldots n\,; \tag{2}$$

der $u_i = 1/w_i$, $v_i = x_i/w_i$ og $\delta_i = e_i/w_i$. Vis at denne modellen oppfyller forutsetningene for den vanlige lineære regresjonsmodellen.

c) Vis at minste kvadraters metode benyttet på modellen (2), er ekvivalent med å minimere

$$\sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{w_i^2},$$

dvs. at observasjoner med stor varians får minst vekt i kvadratsummen.

d) En annen situasjon der modellen (1) er aktuell, er når standardavvikene til støyleddene vokser proporsjonalt med x_i, dvs. $\sigma_i = \sigma x_i$. Skriv ut den transformerte modellen (2) i dette tilfellet. Hva blir estimatorene for β_0 og β_1?

Oppgave 2.

Ved hjelp av intervjuer og spørreskjema kan en beregne en indeks – kalt IDS (Inventory of Drinking Situations) – som angir i hvilken grad ulike sosiale og emosjonelle situasjoner leder en person til å ønske å drikke alkohol. En psykolog har brukt denne indeksen til å studere alkoholvanene til amerikanske studenter. (Dataene nedenfor er konstruerte, men de er i samsvar med resultatene fra den aktuelle studien – publisert i "Journal of Counseling Psychology.")

I studien var det til sammen med 84 studenter. Av disse hadde 28 vanligvis et lite (L) alkoholkonsum, 28 hadde vanligvis et moderat (M) alkoholkonsum, mens 28 vanligvis hadde et stort (S) alkoholkonsum. De 28 studentene i hver komsumgruppe ble delt i to like store grupper:

(Fortsettes side 3.)

- Den ene gruppen ble utsatt for en situasjon der studentene kom i konflikt med andre personer. Denne situasjonen betegnes i det følgende som negativ (neg).
- Den andre gruppen deltok i hyggelig sosialt samvær. Denne situasjonen betegnes i det følgende som positiv (pos).

IDS-indeksen ble beregnet for hver av de 14 studentene i de $3 \cdot 2 = 6$ gruppene en får ved å kombinere typisk alkoholkonsum (L, M eller S) og sosial/emosjonell situasjon (neg, pos).

En utskrift fra MINITAB er vedlagt. Den gir blant annet deskriptiv statistikk for variabelen IDS for de seks gruppene av studenter vi skal studere. Lneg er gruppen av studenter med vanligvis lite alkoholkonsum i den negative situasjonen, Lpos er gruppen med vanligvis lite alkoholkonsum i den positive situasjonen, osv. Vi kan anta at IDS-observasjonene er uavhengige og følger normalfordelingen.

Vi skal først se på IDS-indeksene i gruppene Lneg og Lpos.

a) Utled en test med nivå 5% for nullhypotesen om at gruppene har lik varians mot et tosidig alternativ. Aktuelle fraktiler for F-fordelingen med 13 og 13 frihetsgrader er $F_{0.025}(13, 13) = 0,32$ og $F_{0.975}(13, 13) = 3,12$. Hva blir konklusjonen av testen?

b) Utled et 95% konfidensintervall for forskjellen i forventet IDS i de to gruppene. Du kan anta at de to gruppene har lik varians, uansett hva du kom fram til i a). Kommenter resultatet.

Vi skal så se på effekten av å vanligvis ha et lite, moderat eller stort alkoholkonsum på lysten til å drikke alkohol i sosialt samvær. I vedlegget finner du resultatet av en en-veis variansanalyse for de tre gruppene Lpos, Mpos og Spos. Noen av tallene i variansanalysetabellen er fjernet.

c) La Y_{ij} være j-te observasjon (IDS-indeks) i i-te gruppe. Beskriv modellen som ligger til grunn for variansanalysen. Forklar kort hva som framkommer i variansanalysetabellen, og fyll inn de tallene som mangler i denne.

d) Hvilken nullhypotese er P-verdien i tabellen relatert til? Forklar hvordan P-verdien er regnet ut og gi din konklusjon på testen.

I vedlegget finner du også resultatet av en to-veis variansanalyse av IDS-observasjonene, der de to faktorene er typisk alkoholkonsum (tre nivåer: L, M og S) og situasjon (to nivåer: pos og neg).

e) La Y_{ijk} være k-te observasjon i celle ij. Skriv opp modellen for en to-veis variansanalyse med interaksjonsledd. Diskuter, i lys av de vedlagte utskrifter og plott, hvilken betydning sosial situasjon og studentenes vanlige alkoholkonsum har for IDS.

(Fortsettes side 4.)

Oppgave 3.

I medisinsk forskning er en interessert i å finne ut om en sjelden, ikke smittsom sykdom forekommer oftere i en bestemt gruppe av personer enn i befolkningen forøvrig. Vi skal i denne oppgaven se på metoder som kan brukes til å studere dette. Spesielt vil vi undersøke om forekomsten av kreft i åndedrettsorganene (lungekreft, m.m.) er høyere blant menn som har arbeidet i smelteverk enn den er for andre menn. Dataene i tabellen nedenfor stammer fra et smelteverk i Montana i USA, og de er fra perioden 1960-69.

Aldersgruppe (i)	40-49 år	50-59 år	60-69 år	70-79 år
Antall krefttilfeller (X_i)	7	28	44	15
Antall personår (n_i)	16 120	13 660	7 550	2 725

I tabellen er det, for hver av aldersgruppene 40-49 år, 50-59 år, 60-69 år og 70-79 år, gitt følgende:

- Antall tilfeller av kreft i åndedrettsorganene.
- Antall personår. I antall personår teller en arbeider ett år for hvert år han er observert i det aktuelle aldersintervallet i perioden 1960-69.

Vi vil først studere problemstillingen generelt. Dataene om de amerikanske smelteverksarbeiderene vender vi tilbake til i punktene d) og e).

Vi ser altså på en spesiell, ikke smittsom sykdom og en bestemt gruppe av personer. Vi lar X_i være antallet i gruppen som får sykdommen i aldersgruppe i; $i = 1, 2, \ldots, I$. Antall personår i aldersgruppe i er n_i. Vi vi i hele oppgaven betrakte n_i-ene som gitte, ikke-stokastiske størrelser.

a) Forklar hvorfor det er rimelig å anta at X_i-ene er uavhengige og Poisson fordelte og at forventningen til X_i kan gis på formen $\lambda_i n_i$. Parameteren λ_i kan tolkes som en sykdomsintensitet, dvs. som forventet antall sykdomstilfeller per personår i aldersgruppe i.

Vi lar μ_i være sykdomsintensiteten i aldersgruppe i i den allmenne befolkningen. Sykdomsintensitetene μ_i er kjent fra offentlig statistikk. Vi innfører også de relative sykdomsintensitetene

$$\theta_i = \frac{\lambda_i}{\mu_i}; \qquad i = 1, 2, \ldots, I \tag{3}$$

I punktene b) – d) vil vi anta at de relative sykdomsintensitetene ikke avhenger av aldersgruppe, slik at $\theta_i = \theta$ for alle i. Da er X_i Poisson fordelt med parameter $\theta \mu_i n_i$.

(Fortsettes side 5.)

b) Sett opp likelihooden. Vis at maksimum likelihood estimatoren for θ er $\widehat{\theta} = O/E$, hvor $O = \sum_{i=1}^{I} X_i$ og $E = \sum_{i=1}^{I} n_i \mu_i$. Gi en fortolkning av O og E.

c) Bruk egenskapene til maksimum likelihood estimatoren til å vise at $\widehat{\theta}$ er tilnærmet $N(\theta, \theta/E)$-fordelt. Angi et tilnærmet 95% konfidensintervall for θ.

(*Vink:* Husk at egenskapene til maksimum likelihood estimatoren og sannsynlighetskvotetesten gjelder under svake betingelser. Spesielt trenger ikke observasjonene å være uavhengige og identisk fordelte. Husk også at maksimum likelihood estimatoren $\widehat{\theta}$ er tilnærmet $N(\theta, \sigma^2)$-fordelt, hvor $\sigma^2 = -1/E[l''(\theta)]$. Her er $l(\theta)$ log-likelihood funksjonen.)

Vi ser nå på dataene om kreft i åndedrettsorganene for amerikanske smelteverksarbeidere. I tillegg til tallene i tabellen over, trenger vi da sykdomsintensitetene μ_i for kreft i åndedrettsorganene blant amerikanske menn i sin alminnelighet. Disse er gitt i tabellen nedenfor.

Aldersgruppe (i)	40-49 år	50-59 år	60-69 år	70-79 år
μ_i	0,0003	0,0011	0,00240	0,00280

d) Hva blir estimatet for den relative sykdomsintensiteten θ for smelteverksarbeiderne? Beregn et tilnærmet 95% konfidensintervall for θ. Har smelteverksarbeiderne en høyere risiko for å få kreft i åndedrettsorganene enn andre amerikanske menn?

Vi har over antatt at de relative sykdomsintensitetene (3) er like i alle aldersgrupper.

e) Utled en test for nullhypotesen om at de relative sykdomsintensitetene ikke avhenger av aldersgruppe. Hva gir denne testen for de amerikanske smelteverksarbeiderne?

Oppgave 4.

La $X_1, X_2, \ldots, X_I$ være uavhengige og Poisson fordelte stokastiske variable. Vi antar at $E(X_i) = \theta w_i$, hvor $w_1, w_2, \ldots, w_I$ er gitte positive tall, og $\theta > 0$ er en ukjent parameter. Vi er interessert i å teste nullhypotesen $H_0 : \theta = 1$ mot den alternative hypotesen $H_1 : \theta \neq 1$.

(Fortsettes side 6.)

a) En test forkaster nullhypotesen så sant

$$\frac{|\sum_{i=1}^{I} X_i - \sum_{i=1}^{I} w_i|}{\sqrt{\sum_{i=1}^{I} w_i}} > 1,96$$

Forklar hvorfor denne testen har nivå tilnærmet 5% under passende betingelser. Diskuter disse betingelsene.

b) Utled et tilnærmet uttrykk for styrkefunksjonen til testen. Hvor stor må $\sum_{i=1}^{I} w_i$ være for at teststyrken skal blir omtrent 80% når $\theta = 2$?

c) Kan du foreslå en annen test for H_0 enn den som ble gitt i punkt a)?

SLUTT