# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

| | |
|---|---|
| Eksamen i: | ST110 — Statistiske metoder og dataanalyse |
| Eksamensdag: | Fredag 23. mai 2003. |
| Tid for eksamen: | 09.00 – 15.00. |
| Oppgavesettet er på 4 sider. | |
| Vedlegg: | Tabell over kumulative normalfordelinger. |
| Tillatte hjelpemidler: | Lommeregner, Formelsamling for ST100 og ST110, 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.

Vi skal i denne oppgaven anta at de tilfeldige variablene $X_1, \ldots, X_n$ er uavhengige, identisk gammafordelte med tetthet

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x},\ 0 \le x < \infty$$

der $\alpha$ og $\lambda$ er parametre som er positive.

a) Finn forventning og varians i gammafordelingen, og utled momentestimatorene for $\alpha$ og $\lambda$.

b) Angi ligningene som bestemmer sannsynlighetsmaksimeringsestimatorene (SME) for $\alpha$ og $\lambda$. Hva er den tilnærmede fordelingen når antallet observasjoner, $n$, er stort?

c) Vis hvordan ligningssystemet kan løses ved Newton-Raphsons algoritme. Hva vil du bruke som startverdier? Forklar også hvordan skåringsalgoritmen blir i dette tilfellet.

*(Fortsettes side 2.)*

Algoritmen fra punkt c) kan forbedres på to måter, som vi skal se nærmere på.

d) For det første kan problemet reduseres til løsning av en ikke-lineær ligning i variabelen $\alpha$ . Forklar hvordan, og vis også hvordan en endimensjonal Newton-Raphson algoritme blir i dette tilfellet.

e) For det andre kan en fjerne kravet om at parameteren skal være positiv ved å innføre $a = \log(\alpha)$. Forklar hvorfor, og vis hvordan løsningsalgoritmen for det endimensjonale systemet blir i dette tilfellet. Hvordan velger du startverdien nå?

f) Forklar hvordan du kan konstruere et tilnærmet 95% konfidensintervall for $\alpha$ på grunnlag av den tilnærmede fordelingen fra punkt b), og hvordan dette også kan gjøres ved å bootstrappe sannsynlighetsmaksimeringsestimatoren til $\alpha$ og bruke kvantilene fra den empiriske fordelingen til de genererte estimatorene.

Resultatet av å tilpasse modellen ovenfor til et datasett med 48 observasjoner gav at sannsynlighetsmaksimeringsestimatene ble $\hat{\alpha} = 0.353$ og $\hat{\lambda} = 1.608$. Den tilsvarende inverse informasjonsmatrisen ble estimert til

$$\bar{I}^{-1}(\hat{\alpha}, \hat{\lambda}) = \begin{pmatrix} 0.003 & 0.015 \\ 0.015 & 0.221 \end{pmatrix}.$$

1000 bootstrap replikasjoner, med bruk av de oppgitte estimatene, gav at den 25'te ordnede bootstappede verdi for $\hat{\alpha}$, $\hat{\alpha}^*_{(25)}$, ble 0.265 og den 975'te ordnede, $\hat{\alpha}^*_{(975)}$, ble 0.509.

g) Bruk disse resultatene til å beregne 95% konfidensintervaller for $\alpha$ ved metodene du beskrev i punkt f).

# Oppgave 2.

a) Anta at $X_1, X_2, \ldots, X_n$ er uvhengige variable som er normalfordelte med ukjent forventning $\mu$ og med kjent varians $\sigma^2$. Finn på grunnlag av dette en test på 5 % nivå for hypotesen $\mathrm{H}_0 : \mu = 0$ mot $\mathrm{H}_1 : \mu > 0$. Finn styrkefunksjonen for denne testen.

En lege vil prøve ut et nytt allergimiddel A på 100 pasienter. Alle lider av sesongbestemt pollenallergi. Før de får det nye middelet, blir de samlete symptomene (snue, kløe osv.) vurdert mest mulig objektivt på en skala med verdiene 0, 1, 2 og 3. Her betyr 0 ingen symptomer, og 3 betyr svært alvorlige symptomer. Etter dette tar hver pasient den nye medisinen i 3 uker, og kommer så tilbake til kontroll. Igjen blir symptomene vurdert etter den samme skalaen. Legen ønsker å teste virkningen av allergimedisinen på dette grunnlaget.

*(Fortsettes side 3.)*

b) Foreslå en testmetode, og argumenter for at en eventuell test som bygger på at dataene er normalfordelte, likevel kan brukes som en tilnærmelse i en slik situasjon.

Legen er interessert i å vurdere følsomheten av testmetoden. Han vurderer først variasjonen i skalaverdiene før og etter behandling, og finner at gjennomsnittsverdien av de to variansene kan settes lik 1. Anta at det skal brukes en ensidig test med nivå 5% for forventet forbedring. Sett opp nullhypotese og alternativ.

c) Finn en tilnærmet formel for styrken av testen.Beregn styrken av testen for følgende verdier av forventet forbedring: 0.2, 0.5, 0.8. Hvordan vil du forklare med enkle ord for legen hva disse beregnete verdiene betyr?

d) Utfør testen på 5% nivå når de 100 målingene resulterte i følgende verdier for differansen mellom poeng før medikering og poeng etter medikering:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Diff. : | $-3$ | $-2$ | $-1$ | 0 | 1 | 2 | 3 |
| Antall : | 0 | 4 | 8 | 23 | 20 | 26 | 19 |

(Hint: Tenk over hvordan regneformlene for middeltall og eventuelt for empirisk standardavvik blir for slike grupperte data.)

Senere vil legen sammenligne allergimiddelet A med et noe rimeligere allergimiddel B. Han prøver ut middel B på 100 nye pasienter, som vi kan regne med er trukket fra den samme populasjonen. Undersøkelsen gjøres på samme måte. Tilsvarende resultater for differanse mellom poeng før og etter medikering for middel B ble:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Diff. : | $-3$ | $-2$ | $-1$ | 0 | 1 | 2 | 3 |
| Antall : | 0 | 6 | 14 | 25 | 27 | 15 | 13 |

e) Finn ut fra dette et 95% konfidensintervall for forskjellen i forventet poengforbedring mellom allergimidlene A og B.

# Oppgave 3.

Som et ledd i en større undersøkelse om kroppslig utvikling hos gutter og jenter ble høyde $x$ (i cm) og vekt $y$ (i kg) målt på $n = 10$ jenter, alle 18 år gamle. Resultatene ble:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x$ | 169.6 | 166.8 | 157.1 | 181.1 | 158.4 | 165.6 | 166.7 | 156.5 | 168.1 | 165.3 |
| $y$ | 71.2 | 58.2 | 56.0 | 64.5 | 53.0 | 52.4 | 56.8 | 49.2 | 55.6 | 77.8 |

For disse dataene er $\sum_i x_i = 1655.2$, $\sum_i y_i = 594.7$, $\sum_i x_i^2 = 274440.78$, $\sum_i y_i^2 = 36098.77$ og $\sum_i x_i y_i = 98709.53$.

*(Fortsettes side 4.)*

a) Sett opp en regresjonsmodell for dataene, og presiser hvilke forutsetninger som gjøres. Er alle disse forutsetningene nødvendige, for eksempel for estimeringsspørsmålet i neste punkt?

b) Estimer stigningsforholdet $\beta_1$ for regresjonslinjen og skjæringspunktet $\beta_0$ med $y$-aksen. Bruk minste kvadraters estimatorer, og vis at disse er forventningsrette.

c) Plott dataene sammen med den estimerte regresjonslinjen, og diskuter hvor godt modellen ser ut til å passe.

d) Finn en formel for standardavviket til estimatoren $\hat{\beta}_1$, og estimer denne når det oppgis at residualkvadratsummen er 572.06.

e) Sett opp en tilsvarende regresjonsmodell med $\beta_0 = 0$, og finn minste kvadraters estimator for $\beta_1$ i forhold til denne. Kommenter forholdet mellom de to $\beta_1$-estimatene.

SLUTT