UNIVERSITETET I OSLO

Matematisk Institutt

EKSAMEN I:	**ST 202 – Statistiske slutninger for den eksponensielle fordelingsklasse**
TID FOR EKSAMEN:	**Fredag 16. desember 1994 kl. 9:00–15:00**
HJELPEMIDLER:	**Kalkulator, formelsamlinger for ST 101, 102, 103**

Det er ett oppgavesett for kurset ST 202 og ett for kurset ST 202A; kontroller at du har fått riktig sett utlevert. Nærværende ST 202-sett inneholder tre oppgaver og er på fire sider.

Oppgave 1

Vi sier at X er Gamma-fordelt med parametre (a, b), hvor a og b er positive, dersom tettheten er $\{b^a/\Gamma(a)\}\, x^{a-1}e^{-bx}$ for $x > 0$. Man kaller ofte a for fasongparameteren og b for skalaparameteren.

Et datamateriale består av observerte uavhengige levetider $Y_1, \ldots, Y_n$ for en viss type tekniske komponenter. Det er også registrert bestemte kovariater $x_i = (x_{i,1}, \ldots, x_{i,p})'$ for hver komponent, og man vil analysere disses betydning for levetiden. Fra tidligere erfaring antas Y_i-ene å være tilnærmet Gamma-fordelte.

(a) La $Y_i \sim \text{Gamma}(a, c_i)$ for $i = 1, \ldots, n$, der

$$c_i = \gamma' x_i = \gamma_1 x_{i,1} + \cdots + \gamma_p x_{i,p}.$$

Vis at dette definerer en generalisert lineær modell. Identifiser det som i vanlig GLM-notasjon heter θ_i, η_i, μ_i og $b(\theta_i)$, samt link-funksjonen $\eta_i = g(\mu_i)$. Identifiser også parameterområdet for $\gamma = (\gamma_1, \ldots, \gamma_p)'$ (mengden av slike der modellen er veldefinert).

(b) Bruk kjente resultater om forventning og varians i en generalisert lineær modell, uttrykt ved $b(\theta_i)$-funksjonen, til å finne forventning og varians for Y_i. Finn også variansfunksjonen i denne generaliserte lineære modellen.

(c) *Anta i punktene* (c)–(g) at parameteren a er et *kjent tall*. Sett opp log-likelihood-funksjonen, og sett opp likelihood-ligninger til bestemmelse av $\widehat{\gamma} = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p)'$, sannsynlighetsmaksimeringsestimatene (maximum likelihood-estimatene) for γ_j-ene. Nevn noen stikkord om en numerisk algoritme som klarer å beregne disse estimatene.

(d) Beskriv den tilnærmede fordelingen til $\widehat{\gamma}$. Gi spesielt en oppskrift for hvordan man kan estimere dennes kovariansmatrise.

(e) Finn et uttrykk for deviansen $D(y, \widehat{\mu})$, der $y = (y_1, \ldots, y_n)$ er vektoren av data og der $\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)$ er vektoren av estimater $\widehat{\mu}_i = \mu_i(\widehat{\gamma})$. Diskuter kort hvordan deviansen kan bli benyttet til å sjekke om den underliggende modellen er god nok.

(f) Forklar så hvordan man kan bruke observerte devianser til å teste hypotesen $\gamma_2 = 0$ mot $\gamma_2 \neq 0$. Kommenter (kort) forskjellen på denne type deviansanvendelse og den i punkt (e).

(g) Man kan tenke seg andre måter å modellere c_i-ene på enn den over. Forklar (kort) hvorfor

$$c_i = \exp(-\beta' x_i) = \exp\{-(\beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})\}$$

har visse modelleringsmessige fordeler fremfor det som er brukt over. Hva er linkfunksjonen i denne alternative generaliserte lineære modellen? Hva er den tilnærmede kovariansmatrisen for sannsynlighetsmaksimeringsestimatoren $\widehat{\beta}$?

(h) Den statistiske analyse over har bygget på at fasongparameteren a er kjent. Anta til slutt i denne oppgaven at også denne er ukjent, i den modellen vi startet med, altså den hvor $c_i = \gamma' x_i$. Foreslå en måte å estimere a på. For å illustrere på hvilken måte den statistiske analysen blir forandret, beskriv hvordan man kan lage et konfidensintervall for γ_2 i denne mer kompliserte modellen. Blir et slikt intervall lengre, kortere, eller omtrent like langt, som det tilsvarende konfidensintervall utledet under betingelsen om kjent a?

Oppgave 2

Vi skal se på en modell for parrede (matchede) binomiske forsøk. Anta at to medisinske behandlinger utføres på to pasientgrupper ved hvert av ti sentre (for eksempel i ti forskjellige land). Ved behandling A blir $Y_{i,1}$ av i alt $m_{i,1}$ pasienter friske mens $Y_{i,2}$ av i alt $m_{i,2}$ blir friske ved behandling B. Altså kan vi anta at

$$Y_{i,1} \sim \text{Bin}(m_{i,1}, \pi_{i,1}) \quad \text{og} \quad Y_{i,2} \sim \text{Bin}(m_{i,2}, \pi_{i,2}) \quad \text{for } i = 1, \ldots, 10,$$

og alle 20 binomiske variable er uavhengige.

(a) Som en liten startutredning, før $\pi_{i,j}$-ene skal modelleres, la Y være binomisk fordelt (m, π), og parametriser π som $e^\theta/(1+e^\theta)$. Finn sannsynlighetsmaksimeringsestimatoren θ^* for θ og vis at

$$\theta^* \text{ er tilnærmet } \mathcal{N}\{\theta, v^2\}, \quad \text{der } v^2 = \frac{1}{m\pi} + \frac{1}{m(1-\pi)},$$

når m ikke er for liten. — Man får en noe bedre approksimasjon ved å anvende

$$\widehat{\theta} = \log \frac{Y + \frac{1}{2}}{m - Y + \frac{1}{2}} \approx \mathcal{N}\{\theta, v^2\}.$$

(At denne approksimasjonen faktisk er bedre behøver du ikke vise idag.)

(b) Når 'halvkorreksjonen' brukes som over er det også vanlig å bruke variansestimatoren

$$\widehat{v}^2 = \frac{1}{Y + \frac{1}{2}} + \frac{1}{m - Y + \frac{1}{2}}.$$

Vis at denne er konsistent i den forstand at $\widehat{v}^2/v^2$ konvergerer mot 1 (i sannsynlighet, eller nesten sikkert) når m vokser.

(c) En modell for sammenhengen mellom de 20 $\pi_{i,j}$-ene er som følger:

$$\pi_{i,1} = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} \quad \text{og} \quad \pi_{i,2} = \frac{\exp(\theta_i + \delta)}{1 + \exp(\theta_i + \delta)}.$$

Forklar hvordan man kan tolke parametrene $\theta_1, \ldots, \theta_{10}, \delta$.

(d) Innfør variablene

$$Z_{i,j} = \log \frac{Y_{i,j} + \frac{1}{2}}{m_{i,j} - Y_{i,j} + \frac{1}{2}} \quad \text{for } i = 1, \ldots, 10 \text{ og } j = 1, 2.$$

Foreslå estimatorer for de elleve modellparametrene basert på disse størrelsene. Gi et konfidensintervall for δ med konfidensgrad tilnærmet lik 90%.

(e) Diskuter (kort) hvordan man kan undersøke om modellantagelsene holder.

(f) Diskuter eventuelle andre metoder som kan anvendes for å estimere δ.

Oppgave 3

At X er eksponensielt fordelt med parameter θ betyr at tettheten er $\theta e^{-\theta x}$ for positive x. I denne oppgaven kan du videre bruke at tettheten til en Fisher-fordelt variabel med m og m frihetsgrader, altså en variabel som kan representeres som en brøk mellom to uavhengige χ^2_m-størrelser, er

$$g(x) = \frac{\Gamma(m)}{\Gamma(\frac{1}{2}m)^2} \frac{x^{m/2-1}}{(1+x)^m} \quad \text{for } x > 0.$$

Observasjonene $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ er uavhengige, hvor X_i-ene er eksponensielt fordelte med parameter θ_1 mens Y_i-ene er eksponensielt fordelte med parameter θ_2.

(a) Hva er fordelingen til $2\theta_1 X_i$ og til $2\theta_2 Y_i$? Lag en fornuftig test for hypotesen $\theta_1 = \theta_2$ mot alternativet $\theta_2 > \theta_1$, med gitt nivå 0.05.

(b) Vis at enhver suffisiensbestemt test som også er invariant må avhenge av observasjonsmaterialet kun gjennom $F = \sum_{i=1}^n Y_i / \sum_{i=1}^n X_i$.

(c) Finn den overalt sterkeste test for $\theta_1 = \theta_2$ mot $\theta_2 > \theta_1$, blant alle 0.05-nivå-suffisiensbestemte tester som også er invariante.

(d) Nå forlater vi suffisienskravet og ser hvor langt invarianskravet alene kan bringe oss. Innfør

$$U_i = X_i/Z \text{ for } i = 1, \ldots, n \quad \text{og} \quad V_i = Y_i/Z \text{ for } i = 1, \ldots, n,$$

der $Z = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$. Vis at enhver invariant test må være en funksjon av $(U_1, \ldots, U_n, V_1, \ldots, V_{n-1})$.

(e) Ved hjelp av metoder fra kurset ST 103 kan man regne seg frem til at simultantettheten til $(U_1, \ldots, U_n, V_1, \ldots, V_{n-1})$ kan skrives

$$\Gamma(2n) \frac{\theta_1^n \theta_2^n}{\{\theta_1 \sum_{i=1}^n u_i + \theta_2 (1 - \sum_{i=1}^n u_i)\}^{2n}}$$

i det området der u_i-ene og v_i-ene er positive med samlet sum $\sum_{i=1}^n u_i + \sum_{i=1}^{n-1} v_i < 1$ (dette skal altså ikke vises på eksamen). Vis at det til enhver invariant test finnes en annen test (eventuelt randomisert) som er basert på data bare gjennom F, og som har nøyaktig samme styrkefunksjon som den opprinnelige. Forklar hvorfor dette leder til et sterkere resultat enn det i (c).

(f) Til slutt i denne oppgaven skal du se på det relaterte problemet der alternativet til hypotesen $\theta_1 = \theta_2$ er det tosidige $\theta \neq \theta_2$. Vis at hver suffisiensbestemt og invariant test må være en funksjon av data bare gjennom størrelsen $W = \max(F, 1/F)$. Finn deretter en overalt sterkest test med nivå 0.05, blant alle slike suffisiensbestemte og invariante tester.

SLUTT