UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Fasit
Eksamensdag:	Tirsdag 2. juni 2009.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabeller over normal-, t-, F- og χ^2-fordelingene.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

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

Oppgave 1.

(a) Y_i = antall jenter i familie i er bin$(12, \theta)$ blir $\pi_y = P(Y_i = y; \theta) = \binom{12}{y}\theta^y(1-\theta)^{12-y}$ og forventet antall familier med y jenter blir $6125\pi_y$.

Vi har at totalt antall jenter i de 6125 familiene er lik $(\sum_{y=0}^{12} yO_y)$ og det er ialt 12*6125 barn. Under antagelsen er $\sum_{i=1}^{6125} Y_i \sim \text{bin}(12 * 6125, \theta)$ og MLE for θ er andelen jenter.

(b) Pearson-kjivadrat er tilnærmet kji-kvadrat fordelt med (13-1)-1 = 11 frihetsgrader under H_0. 13-1 er antall parametre under den fulle modellen, mens det bare er en parameter (θ) under H_0.

Med $X^2 = 107.9$ har vi soleklar forkastning av nullhypotesen, 99.5% persentilen i χ^2_{11} er lik 26.76.

(c) Vi ser at E-ene under den beta-binomiske modellen generelt er nærmere de observerte verdiene (O-ene) enn E-ene under den binomiske modellen. Tilsvarende blir $(O-E)^2/E$-ene mindre.

Her får Pearson X^2 10 frihetsgrader under nullhypotesen siden nullhypotesemodellen har to parametre.

(Fortsettes side 2.)

Med $X^2 = 13.32$ kan vi ikke forkaste, siden $X^2 < 15.99 = 90$ persentilen i χ^2_{10}, dvs. p-verdien > 0.10. Dataene er altså i samsvar med denne modellen.

Oppgave 2.

(a) Vi har at $(O_0, O_1, , \dots, O_{12})$ er multinomisk fordelt, dermed blir punktsannsynlighetene proporsjonale (avhenger ikke av parametrene) med $\prod_{y=0}^{12} \pi_y^{o_y}$ der o_y er mulige observerte verdier av O_y-ene, Under modellen er $\pi_y = \pi_y(\alpha, \beta)$

Scorefunksjonen er definert som

$$s(\alpha, \beta) = \begin{bmatrix} \frac{\partial \log(L(\alpha,\beta))}{\partial \alpha} \\ \\ \frac{\partial \log(L(\alpha,\beta))}{\partial \beta} \end{bmatrix}$$

der $\log(L(\alpha, \beta)) = \sum_{y=0}^{12} O_y[\log(\binom{n}{y}) + \log(\Gamma(\alpha+\beta)) + \log(\Gamma(\alpha+y)) + \log(\Gamma(\beta+12-y)) - \log(\Gamma(\alpha+\beta+n) - \log(\Gamma(\alpha)) - \log(\Gamma(\beta))]$. Kun ledd 3 og 4 i denne summen avhenger av både y og parametrene. De øvrige summerer seg $6125 = \sum_{y=0}^{12} O_y$ ganger dette opprinnelige leddet. Dessuten har vi $\frac{\partial \log(\Gamma(\alpha))}{\partial \alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. Derav følger score-funksjonen.

Den observerte informasjonsmatrisen blir

$$\bar{J}(\alpha, \beta) = -\begin{bmatrix} \frac{\partial^2 \log(L(\alpha,\beta))}{\partial \alpha^2} & \frac{\partial^2 \log(L(\alpha,\beta))}{\partial \alpha \partial \beta} \\ \\ \frac{\partial^2 \log(L(\alpha,\beta))}{\partial \alpha \partial \beta} & \frac{\partial^2 \log(L(\alpha,\beta))}{\partial \beta^2} \end{bmatrix} = \begin{bmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{bmatrix}$$

der med $\Psi(\alpha) = \frac{\partial^2 \log(\Gamma(\alpha))}{\partial \alpha^2}$:

$J_{11} = 6125[\Psi(\alpha) + \Psi(\alpha+\beta+12) - \Psi(\alpha+\beta)] - \sum_{y=0}^{12} O_y \Psi(\alpha+y)$

$J_{12} = 6125[\Psi(\alpha+\beta+12) - \Psi(\alpha+\beta)]$ og

$J_{22} = 6125[\Psi(\beta) + \Psi(\alpha+\beta+12) - \Psi(\alpha+\beta)] - \sum_{y=0}^{12} O_y \Psi(\beta+12+y)$

(b) MLE gis ved $s(\hat{\alpha}, \hat{\beta}) = 0$. Dette er to ikke-lineære ligninger som må løses numerisk. Her er det mest aktuelt å benytte Newton-Raphson algoritmen:

$$(\alpha^{ny}, \beta^{ny})' = (\alpha^g, \beta^g)' + \bar{J}(\alpha^g, \beta^g))^{-1} s(\alpha^g, \beta^g)$$

siden den forventede informasjonsmatrisen neppe kan uttrykkes enkelt.

(Fortsettes side 3.)

Vi har ved large sample egenskapene til maximum likelihood-estimatorer at tilnærmet er $\hat{\alpha} \sim \mathrm{N}(\alpha, 15.43)$. Dermed blir et tilnærmet 95% konfidensintervall for α gitt ved $\hat{\alpha} \pm 1.96\sqrt{15.43} = 31.88 \pm 1.96 * 3.92 = (24.18, 39.58)$

(c) Vi finner

(a) Avvik mellom gj.snittelig bootstrap-estimat og opprinnelig estimat < 0.4 som er ca. ett tiendels standardfeil, dvs. ikke betydelig skjevhet

(b) Standardfeil ved asymptotikk $\sqrt{15.43} = 3.92$ avviker ikke mye fra standardavvik for bootstrapestimatene (4.14), men er likevel litt mindre

(c) Persentil-intervallet blir (25.20,41.18) som er nokså likt det asymptotisk begrunnede intervallet, riktignok litt lenger. Standard bootstrap intervallet (som tar hensyn til evt. skjevestimering) blir tilsvarende

2 * 31.88 - (41.18,25.20) = (22.58, 38.56)

(d) Histogrammene er ikke perfekt normalfordelt, men det ser ut som den tunge halen i bootstrap-fordelingen bare utgjør en drøy prosent av alle bootstrap-verdiene.

(e) $\hat{\alpha}$ og $\hat{\beta}$ er sterkt korrelerte og resultatene for $\hat{\beta}$ blir tilsvarende

Inferens basert på MLE's asymptotiske egenskaper ser altså ut til å fungere akseptabelt. Dette er egentlig ikke så overraskende siden datamaterialet består av over 6000 observasjoner.

Oppgave 3.

(a) $\Omega = \{(\mu_1, \ldots, \mu_I, \sigma^2) : \mu_i \in \Re, \sigma^2 \in \Re^+\}$ har dimensjon $I + 1$

$\omega_0 = \{(\mu, \ldots, \mu, \sigma^2) : \mu \in \Re, \sigma^2 \in \Re^+\}$ har dimensjon 2

Dermed sier den generelle teorien for GLRT at $-2\log(\Lambda)$ under nullhypotesen er tilnærmet χ^2 med antall frihetsgrader lik $(I+1)-2 = I-1$.

(b) Under H_0 er eksakt $F \sim F_{I-1,I(J-1)}$, dvs. Fisherfordelt med $I-1$ og $I(J-1)$ frihetsgrader.

Vi kan skrive $F_{I-1,I(J-1)} = \frac{\chi^2_{I-1}/(I-1)}{\chi^2_{I(J-1)}/(I(J-1))}$ og $\mathrm{Var}(\chi^2_{I(J-1)}/(I(J-1))) = 2/(I(J-1))$ som går mot null når J går mot uendelig. Dermed vil $\chi^2_{(I(J-1))}/(I(J-1)) \approx 1$ som er dens forventning. Dermed får vi

$(I-1)F_{I-1,I(J-1)} \approx \chi^2_{I-1}$ når J er stor.

Vi har altså at $(I-1)F$ og $-2\log(\Lambda)$ har samme tilnærmede fordeling med mange replikasjoner J.

(Fortsettes side 4.)

(c) Maksimal log-likelihood i den fulle modellen er lik

$$\begin{aligned} l(\bar{Y}_{1\bullet},\ldots,\bar{Y}_{I\bullet},SS_W/n) &= -\tfrac{n}{2}\log(2\pi SS_W/n)) - -\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-Y_{i\bullet})^2}{2SS_W/n} \\ &= -\tfrac{n}{2}\log(2\pi\tfrac{SS_W}{n}) - \tfrac{n}{2} \end{aligned}$$

Tilsvarende under H_0 blir maksimal loglikelihood

$$\begin{aligned} l(\bar{Y}_{\bullet\bullet},\ldots,\bar{Y}_{\bullet\bullet},SS_{TOT}/n) &= -\tfrac{n}{2}\log(2\pi SS_{TOT}/n) - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-Y_{\bullet\bullet})^2}{2SS_{TOT}/n} \\ &= -\tfrac{n}{2}\log(2\pi\tfrac{SS_{TOT}}{n}) - \tfrac{n}{2} \end{aligned}$$

Dermed, siden $SS_{TOT} = SS_B + SS_W$, blir

$$\begin{aligned} -2\log(\Lambda) &= -2[l(\bar{Y}_{\bullet\bullet},\ldots,\bar{Y}_{\bullet\bullet},SS_{TOT}/n) - l(\bar{Y}_{1\bullet},\ldots,\bar{Y}_{I\bullet},SS_W/n)] \\ &= n\log(SS_{TOT}/SS_W) = n\log(1+\tfrac{(I-1)}{I(J-1)}F). \end{aligned}$$

Det er en 1-1 sammenheng mellom Λ og F slik at en stor F tilsvarer en liten Λ. Vi forkaster nullhypotesen når F er stor og ekvivalent når Λ er liten, testene er altså ekvivalente.

(d) Ved 1. orden Taylor er $\log(1+\epsilon) \approx \epsilon$ når ϵ er liten. Dermed fås for stor J at $-2\log(\Lambda) \approx n\frac{(I-1)}{I(J-1)}F = \frac{IJ}{I(J-1)}(I-1)F \approx (I-1)F$ siden $n = IJ$ og $\frac{IJ}{I(J-1)} \approx 1$. Men dette samsvarer med punkt (b) hvor det ble argumentert for at $-2\log(\Lambda)$ og $(I-1)F$ har samme tilnærmede fordeling.

SLUTT